Verbalized Confidence Calibration in Vision-Language Models with Semantic Perturbation

Anonymous ACL submission

Abstract

Vision-language models (VLMs) excel in various multimodal tasks but frequently suffer from poor calibration, resulting in misalignment between their verbalized confidence and response correctness. This miscalibration undermines user trust, especially when models confidently provide incorrect or fabricated information. In this work, we propose a novel Confidence Calibration through Semantic Perturbation (CSP) framework to improve the calibration of verbalized confidence for VLMs. We first introduce a perturbed dataset where Gaussian noise is applied to the key object regions to simulate visual uncertainty at different confidence levels, establishing an explicit mapping between visual ambiguity and confidence levels. We further enhance calibration through a two-stage training process combining supervised fine-tuning on the perturbed dataset with subsequent preference optimization. Extensive experiments on popular benchmarks demonstrate that our method significantly improves the alignment between verbalized confidence and response correctness while maintaining or enhancing overall task performance.

1 Introduction

Modern vision-language models (VLMs) have demonstrated remarkable success on tasks ranging from image captioning to visual question answering (Achiam et al., 2023; Bubeck et al., 2023). Beyond delivering correct results, real-world deployment of these models also requires trustworthy outputs. One key aspect of trustworthiness is a model's ability to verbalize its own confidence, typically by starting how certain it is about its answer, e.g. "I'm about 80% sure that this is a cat". This verbalized confidence help users appropriately weight its responses in downstream decisions. However, model often exhibits overconfidence, giving high certainty even when they provide incorrect answers, as shown in Figure 1 top. Therefore, even if the model produces incorrect answers, it is still essential to calibrate its confidence appropriately so that users can accurately gauge the model's uncertainty and avoid being misled by high-confidence yet erroneous responses, as shown in Figure 1 bottom. Thus, verbalized confidence calibration which ensure that the expressed confidence reliably reflects model's response correctness becomes crucial.

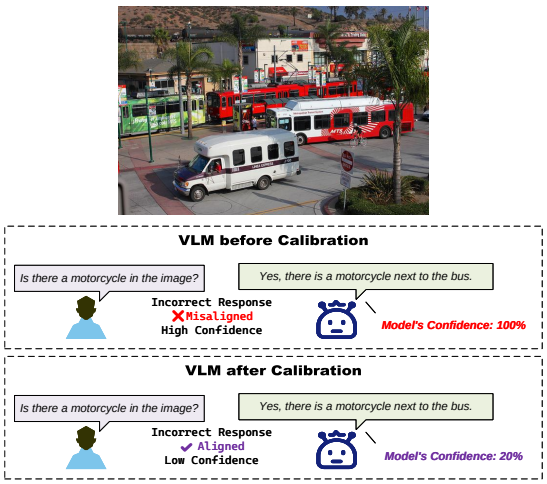

Figure 1: (Top): Most current VLMs tend to generate high verbalized confidence on incorrect response. (Bottom): After calibration, model's verbalized confidence will be aligned with response correctness.

While extensive research has addressed verbalized confidence calibration for text-only large language models (LLMs) (Kumar et al., 2023; Yin et al., 2023), these techniques are often inadequate when generalized to multimodal settings. Compared to text-based methods, VLMs face challenges in accurately verbalizing confidence due to difficulty of semantic understanding in visual features from two aspects. First, images may suffer from occlusion or poor lighting, which can obscure key objects, leading to incomplete semantic extraction and introducing uncertainty in visual understanding (Khan and Fu, 2024). Second, current VLMs relies heavily on textual cues such as language priors while neglecting critical visual content that causes multimodal imbalance (Zhao et al., 2024a). Consequently, due to the combined effects of introduced visual uncertainty and multimodal imbalance, the verbalized confidence calibration for VLMs still remains an unsolved issue.

To address these challenges, we introduce Confidence Calibration through Semantic Perturbation (CSP), a novel framework designed to improve the calibration of verbalized confidence for VLMs. Our key insights is to

introduce a perturbed dataset that modifies key visual elements based on different confidence levels, allowing the model to learn explicit mappings between visual uncertainties and verbalized confidence. To construct the perturbed dataset, we first extract key object regions referenced in a multimodal query using GroundingDINO (Liu et al., 2025) and Segment Anything (SAM) (Kirillov et al., 2023). We then apply varying levels of Gaussian noise to these regions, effectively creating progressively perturbed images that mimic different degrees of occlusion or distortion. Each perturbed image is associated with a ground-truth confidence label, teaching the model to explicitly modulate its verbalized confidence based on the severity of visual uncertainties. After that, we apply supervised fine-tuning and preference optimization to reinforce verbalized confidence calibration. In this way, our approach accounts for visual uncertainties in semantic extraction and encourages more visually-grounded confidence judgments, leading to more accurate and well-calibrated verbalized confidence in VLMs.

We conduct extensive experiments on widely-used VLM benchmarks across multiple state-of-the-art VLMs. Comparing their performance before and after calibration using our CSP framework, our results demonstrate substantial improvements in verbalized confidence calibration across a diverse set of evaluation metrics. We achieve consistent gains in accuracy, F1 score, and AUC, while simultaneously reducing Expected Calibration Error (ECE) and Brier Score (BS) between verbalized confidence and correct labels. These improvements show that models trained with CSP correlate their expressed confidence more faithfully with actual correctness, reducing the likelihood of misleadingly high-confidence errors. Moreover, these calibration improvements do not come at the cost of task performance. CSP preserves or enhances the VLMs' task accuracy, confirming that our method improves trustworthiness and interpretability without sacrificing predictive capability.

In summary, our primary contributions are:

1. **Novel Calibration Framework.** We introduce CSP, a new framework for training VLMs for better calibration with verbalized confidence and response correctness.

2. **Semantic Perturbation Data Construction.** We present a systematic approach to local image perturbation that simulates diverse levels of visual ambiguity, enabling more fine-grained calibration during training.

3. **Extensive Validation.** Empirical results on multiple benchmarks and model architectures show that CSP significantly reduces calibration error and improves trustworthiness, without sacrificing task accuracy.

2 Related Work

LLM Confidence Calibration. Confidence calibration has emerged as a critical challenge in LLMs to ensure reliable and trustworthy outputs. Early methods predominantly focused on calibrating internal confidence derived from model logits. These logit-based approaches often employ statistical techniques like temperature scaling or model-based re-calibration (Duan et al., 2024; Kuhn et al., 2023) to adjust probability distributions and mitigate overconfidence. More recent research extends beyond simple scaling by exploiting semantic features and contextual cues to better align token-level probabilities with actual prediction correctness (Burns et al., 2023). In particular, linguistic uncertainty modeling and sequence likelihood calibration have shown promise for managing generation tasks susceptible to compounding errors (Lin et al., 2022). Although effective in text-only settings, these calibration techniques largely overlook the additional complexity introduced by visual or other multimodal inputs.

Verbalized Confidence. While most traditional calibration work centers on the gap between predicted probabilities and true correctness, a growing line of research emphasizes the verbalized confidence explicitly stated by the model (Yin et al., 2023; Kumar et al., 2024). Such verbalized confidence can help users better gauge the reliability of model outputs. Recent studies have explored prompting strategies and self-reflection pipelines that encourage LLMs to articulate their confidence levels, thereby improving transparency (Xu et al., 2024; Xiong et al., 2023).

Towards Multimodal Calibration. Despite the progress in unimodal confidence calibration and verbalized confidence expression, few frameworks systematically address these issues in VLMs. Recent work highlights the need for calibration techniques tailored to multimodal data, which present unique challenges such as object occlusion and semantic ambiguity (Geng et al., 2024; Huang et al., 2023; Groot and Valdenegro-Toro, 2024). In contrast to existing methods that primarily focus on textual uncertainty or straightforward logit manipulation, our approach introduces semantic mask perturbation to simulate varying degrees of visual uncertainty, laying the groundwork for more trustworthy multimodal systems.

3 Methodology

In this section, we begin by formally defining the problem of verbalized confidence calibration. Then we introduce the CSP framework to improve the calibration of verbalized confidence. As illustrated in Figure 2, the proposed CSP framework consists of two stages: dataset construction and training. In dataset construction stage, we produce a dataset by generating images with different levels of visual uncertainties by systematically applying varying degrees of noise to the key visual regions

of the given visual input, and explicitly associates the visual uncertainties with corresponding confidence labels. In training stage, based on the constructed dataset, we perform supervised fine-tuning and preference optimization to achieve better verbalized confidence calibration. Together, these components form a cohesive framework that systematically enhances verbalized confidence calibration of VLMs, ensuring that VLMs produce more trustworthy and interpretable outputs.

3.1 Problem Definition of Verbalized Confidence Calibration

The goal of verbalized confidence calibration is to enable a VLM to assign a verbalized confidence score to a candidate answer to reflects the probability of correctness. In other words, the correct answer should be assigned with the highest confidence score, while the other incorrect answers should be assigned with lower scores. Specifically, given a visual input v_0 and a textual query q, considering there exists a set of candidate answers $\{a_1, a_2, \ldots, a_k\}$, the VLM will assign a verbalized confidence score $c(a_i)$ to each candidate response a_i. For the correct answer a_*, verbalized confidence calibration aims to enable the VLM to increase the verbalized confidence score $c(a_*)$ to be the highest, while decrease the verbalized confidence score of the other incorrect answers to be low. In this way, the objective of generating verbalized confidence score that accurately reflects the correctness of the VLM's answer can be achieve.

To evaluate the effectiveness of verbalized confidence calibration, a convenience way is to assess whether the answer $\hat{a}$ with highest verbalized confidence score matches the correct answer a_* with a metric M, namely calculate the similarity $M(a_*, \hat{a})$. Here, M can be any commonly used similar metrics, such as Accuracy, F1 Score, Area Under the Curve (AUC), Brier Score, Expected Calibration Error (ECE). In this way, we can evaluate both the correctness of the model's chosen responses and the fidelity of its verbalized confidence in reflecting true likelihoods of correctness.

3.2 Dataset Construction with semantic perturbation

To address the challenges of verbalized confidence calibration in VLMs, we construct a specialized dataset incorporating semantic perturbations. This design is motivated by two key challenges that hinder accurate confidence expression in VLMs. First, visual uncertainty arises from factors such as occlusion or poor lighting, which obscure key objects and lead to incomplete semantic extraction. Existing training data often lack explicit examples reflecting such conditions, making it difficult for VLMs to modulate their confidence in response to uncertain visual input. Second, VLMs exhibit a strong reliance on textual priors while often overlooking critical visual details, creating a multimodal imbalance. Without explicit guidance, VLMs tend to verbalize confidence based primarily on linguistic patterns rather than the actual uncertainty in visual perception. To mitigate this issue, we construct a dataset where varying levels of perturbations are applied according to different confidence levels. The perturbations are applied selectively to semantically relevant objects or regions rather than indiscriminately across the entire image, ensuring that verbalized confidence is directly grounded in affected visual features. Thus, this dataset provides a more structured learning framework for addressing verbalized confidence calibration in VLMs.

As illustrated in Figure 2, given a text query and a corresponding image, we begin by identifying the most relevant object mentions and localize those regions in the image using GroundingDINO and SAM. We then inject varying intensities of Gaussian noise into the segmented relevant object region, creating a series of perturbed images that mimic different levels of visual uncertainty according to a confidence label. The confidence label is sampled from 0% to 100%. Finally, we convert each sample into new confidence query which consist of original query and answer, and new confidence response which is the confidence label. By covering diverse noise levels and confidence targets, the expanded dataset forms the foundation for both supervised fine-tuning and preference optimization, ultimately improving the model's ability to align its verbalized confidence with true uncertainty.

Key Object Region Extraction: To construct the specialized dataset, we begin with extract the key object region that are relevant to the input query. This process consists of three main steps: key object descriptions extraction, object localization and semantic segmentation. Given a textual query q and a response r, first we extract key object descriptions $M_{\text{desc}}(q, r)$ by using a small LLM, for example, the word "steak" shown in the image. Then based on the extracted key object descriptions, we use GroundingDINO to localize the most relevant objects in the image that correspond to terms mentioned in the multimodal query-response pair. Given the corresponding image v_0, which has the size of $H \times W$, GroundingDINO outputs object bounding boxes that approximate the regions of interest. Next, we apply SAM to refine these localized object regions into precise semantic masks. Given the bounding boxes provided by GroundingDINO, SAM generates pixel-wise segmentation masks that more accurately delineate the object's shape and boundaries. The final binary mask m is computed as:

$$m = \text{SAM}(v_0, \text{GroundingDINO}(v_0, M_{\text{desc}}(q, r))),$$

where $m \in \{0, 1\}^{H \times W}$ denotes the binary mask of size $H \times W$ which is same as the visual input v_0. By combining object detection with fine-grained segmentation, we ensure that only the most relevant semantic regions are perturbed in the next step, allowing for precise control over visual uncertainty in the dataset.

Confidence Labeling and Semantic Perturbation Mechanism: To simulate visual uncertainty in a con-

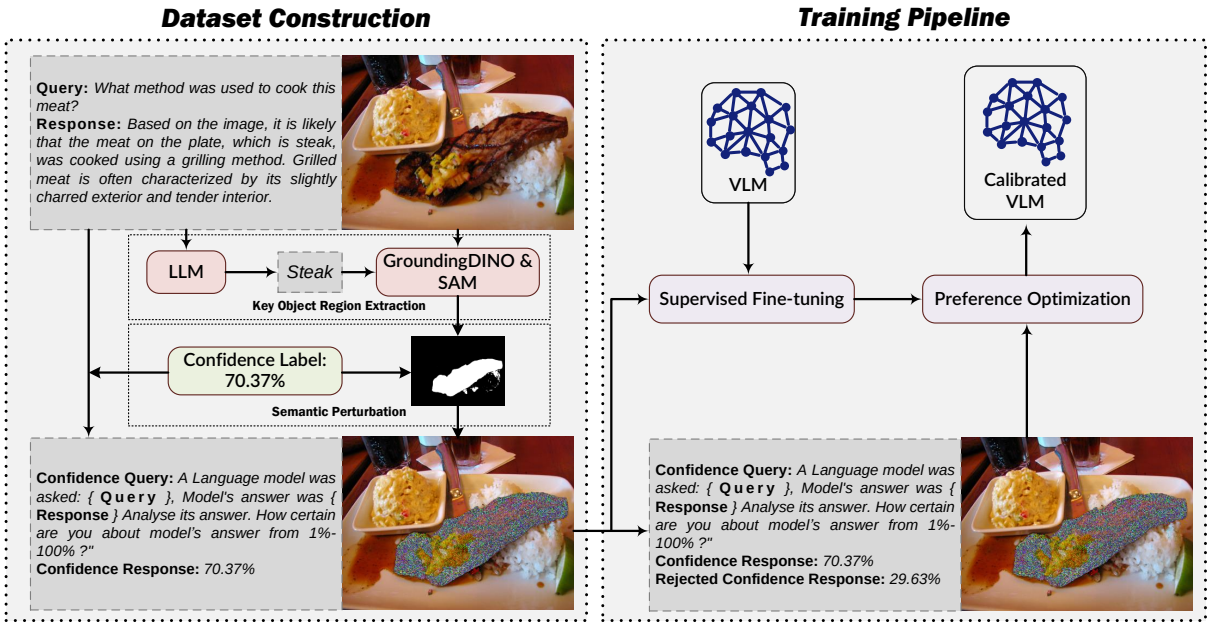

Figure 2: The image illustrates the dataset construction and training pipeline for improving confidence calibration in VLM. It highlights the two-stage process: **Dataset Construction**: Extracting key object regions using GroundingDINO and SAM, applying semantic perturbations, and assigning confidence labels based on noise levels. **Training Pipeline**: Fine-tuning the VLM with supervised learning, followed by preference optimization, to improve probability-confidence alignment and response calibration.

trolled way, we apply Gaussian noise only to the key object regions identified by the mask m. Let $c \in [0, 100]\%$ be the desired confidence label, where $c = 100\%$ indicates no perturbation and $c = 0\%$ indicates maximal noise. In this context, a higher confidence corresponds to lower uncertainty, so we inject less (or no) noise. Conversely, a lower confidence corresponds to higher uncertainty, so more noise is applied to the key object regions. Therefore, by decreasing confidence from 100% to 0%, we increase the noise to simulate a progressive increase in visual uncertainty.

Following the forward diffusion process in image generation (Ho et al., 2020), we inject Gaussian noise to the key object region by mapping c to a diffusion step T_c by a linear schedule:

$$T_c \;=\; \lfloor T_{\max} \times \big(1 - \frac{c}{100}\big) \rfloor,$$

where $T_{\max}$ is a chosen upper bound on the number of diffusion steps. Starting from the original image v_0, we iteratively sample:

$$v_t \;\sim\; \mathcal{N}\Big(\sqrt{1-\gamma}\, v_{t-1},\; \gamma\, \mathbf{I}\Big) \quad \text{for} \quad t = 1 \ldots T_c,$$

with $\mathcal{N}$ denotes a Gaussian distribution. γ is a predefined parameter that controls the noise intensity introduced in each step. A larger γ results in stronger noise injection per step, leading to a more rapid degradation of the image. This diffusion-based approach enables a gradual and controlled degradation of visual features, allowing for a continuous mapping between the confidence label and the level of perturbation. After T_c iterations, we combine the noised image v_{T_c} with the unperturbed background of v_0 using the binary mask m:

$$v_{\text{perturbed}} \;=\; m \odot v_{T_c} \;+\; \big(1 - m\big) \odot v_0,$$

where $\odot$ denotes element-wise multiplication.

This procedure distorts only the object region relevant to the given query-response pair, as illustrated in Figure 2 (where the region of the object "steak" is partially occluded, while the plate and background remain clear). Each confidence label c thus produces a distinct visually perturbed image, ranging from minimal to severe noise. By pairing these images with the appropriate textual instructions (e.g., "How certain are you about the model's answer from 1% to 100%?"), we give the model explicit supervision on how to verbalize confidence in accordance with visible uncertainty.

Dataset Integration: The resulting dataset is constructed through dataset modification and augmentation of the RLAIF dataset (Yu et al., 2024), which is a large-scale multimodal AI feedback dataset collected from a diverse sources. We enhance the original dataset by applying the proposed semantic perturbation technique, generating diverse image-query-response samples. Each sample in the dataset consists of a transformed query q_c and a corresponding confidence-labeled answer r_c. The transformed q_c is generated from the original query q and response r. To ensure objective and generalized confidence evaluation, we transform q_c in a Third-Person Perspective (TPP) format, framing the query as an external assessment rather than a direct model introspection,

as shown in Figure 2. The corresponding confidence-labeled answer r_c is assigned as: $r_c = c$ where c is the confidence label derived from the semantic perturbation process. Finally we construct modified semantic perturbed dataset $D = \{(v_{\text{perturbed}}, q_c, r_c)\}$. By framing confidence estimation in TPP format, we reduce self-referential bias, ensuring that the model assesses confidence based on visual and textual evidence rather than internal heuristics (Kumar et al., 2024). By simulating real-world visual uncertainties, this dataset is capable of supporting a robust framework for addressing verbalized confidence calibration challenges in vision-language models.

3.3 Training

Supervised Fine-tuning (SFT) Using the constructed dataset $D = \{(v_{\text{perturbed}}, q_c, r_c)\}$, we perform SFT to establish the model's capability to associate visual uncertainty with verbalized confidence. The objective of SFT minimizes the cross-entropy loss:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(v_{\text{perturbed}}, q_c, r_c)\sim D}\big[\log P_\theta(r_c \mid v_{\text{perturbed}}, q_c)\big],$$

where $P_\theta(r_c \mid v_{\text{perturbed}}, q_c)$ is the model's probability of generating the response given the visual and textual inputs. This step enables the model to learn the relationship between visual uncertainty (as affected by diffusion noise) and verbalized confidence.

Preference Optimization To further refine the model's verbalized confidence calibration, we adopt SimPO (Simple Preference Optimization) (Meng et al., 2024) on top of the SFT model. Specifically, for each training example in the perturbed dataset $\{(v_{\text{perturbed}}, q_c, r_c)\}$, we take r_c as the winning response and define the rejected response $r_{\text{rej}} = 100\% - c$. This pairwise preference setting encourages the model to produce confidence estimates that more closely reflect the visual uncertainty. Formally, let π_θ denote the policy model and (q_c, y_w, y_l) be a preference sample in which $y_w \equiv r_c$ and $y_l \equiv r_{\text{rej}}$. SimPO optimizes the following margin-based objective:

$$\mathcal{L}_{\text{SimPO}}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim D}$$
$$\left[\log \sigma\left(\frac{\beta}{|y_w|}\log \pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|}\log \pi_\theta(y_l \mid x) - \lambda\right)\right],$$

where σ is the sigmoid function, β is a scaling factor for the reward, λ is a target margin ensuring the policy assigns sufficiently higher probability to the winning response compared to the losing response. In conjunction with the SFT step, this preference-based fine-tuning further ensures that the final model's verbalized confidence provides a faithful reflection of the true visual uncertainty in the input.

4 Experiments

4.1 Experimental Settings

Dataset We conduct experiments on two popular datasets POPE (Li et al., 2023) and AMBER (Wang et al., 2023) to verify the effectiveness of our proposed method. POPE, the Polling-based Object Probing Evaluation, is designed to assess object hallucination in VLMs regarding the presence of objects in images. POPE is divided into three settings: random, popular, and adversarial, indicating different methods of sampling hallucination objects. AMBER is a comprehensive benchmark designed to evaluation multiple different types of hallucination including attribute hallucination and relation hallucination. We choose hallucination benchmarks to validate verbalized confidence calibration because calibration tends to be more challenging in scenarios with severe hallucination, making these benchmarks particularly representative for assessing the effectiveness of our approach.

Models We benchmarked the proposed method against several state-of-the-art vision-language models: Qwen-VL-Chat (Bai et al., 2023), Qwen2-VL-7B-Instruct (Wang et al., 2024), InternVL2-8B (Chen et al., 2024), and Phi-3.5-vision-instruct (Abdin et al., 2024). Qwen-VL is a versatile vision-language model adept at understanding, localization, and text reading tasks. Qwen2-VL, an advanced iteration of Qwen-VL, enhances image comprehension across various resolutions and ratios, and extends capabilities to video understanding and multilingual support. InternVL2 is an open-source multimodal large language model designed to bridge the gap between open-source and proprietary commercial models in multimodal understanding. Phi-3.5 is a vision-language model that provides general-purpose AI capabilities, handling both visual and textual inputs efficiently. For each baseline, we utilized the official pre-trained models and followed the recommended evaluation protocols to ensure a fair comparison. For more detailed information on the experimental configuration, please refer to the appendix.

4.2 Evaluation Metrics

We employ five metrics to evaluate our model's verbalized confidence calibration. Let $\hat{y}_i$ be the prediction chosen by the highest confidence $c(\hat{y}_i)$, y_i be the ground-truth, and $p_i \in [0, 1]$ denote the corresponding verbalized confidence score served as the soft probability to compute confidence-aware metrics. We define:

- **Accuracy (Acc)**: $\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i)$.
- **F1 Score**: The harmonic mean of precision and recall.
- **AUC (Area Under ROC Curve)**: The probability that a randomly chosen correct instance is ranked higher (by $c(\hat{y}_i)$) than an incorrect instance.
- **Brier Score (BS)**: $\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$, where $p_i \in [0, 1]$ is the model's predicted probability (verbalized confidence) for the event $y_i = 1$.
- **Expected Calibration Error (ECE)**: Partition samples into K bins of equal confidence range; measure the average gap between mean predicted confidence

and empirical accuracy in each bin:

$$\text{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \big|\text{acc}(B_k) - \text{conf}(B_k)\big|.$$

Here, $\text{acc}(B_k)$ is the average correctness and $\text{conf}(B_k)$ is the average predicted confidence in bin B_k.

Lower BS, ECE and higher Acc, F1, AUC indicate better calibration and overall alignment of confidence with correctness. These metrics are complementary: Acc and F1 evaluate how reliably the model's most confident predictions align with correctness, while AUC, BS, and ECE assess distinct facets of calibration—ranking reliability across all confidence thresholds, probability sharpness, and bin-wise confidence-accuracy alignment, respectively. Together, they provide a holistic view of verbalized confidence calibration, measuring both the trustworthiness of confidence-guided predictions and the statistical alignment between expressed confidence and empirical correctness.

4.3 Experimental Results

To evaluate the effectiveness of our proposed CSP framework, we assess the calibration of verbalized confidence across multiple datasets according to the metrics we introduced above.

4.3.1 Results of Verbalized Confidence Calibration

Table 1 demonstrates significant improvements in calibration across all models and datasets. Our method enhances accuracy and F1 score while reducing ECE, indicating better confidence correctness alignment. Notably, Qwen2-VL, initially suffering from severe miscalibration, improves dramatically post-training. InternVL2, already well-calibrated, still benefits from the proposed approach across challenging settings. These results highlight that our approach not only improve the ability of verbalized confidence prediction in weaker models but also refines confidence estimation in stronger ones, making VLMs more reliable in uncertainty-prone multimodal tasks.

Moreover, as the results shown in Figure 3 from AMBER attribute dataset, our approach consistently improves both the Brier Score and the AUC across all tested models, underscoring more accurate confidence estimation and stronger separability between correct and incorrect predictions. From the calibration plots of Brier Score, we observe that each model's calibration curve shifts closer to the diagonal "perfect calibration" line after training, indicating that predicted probabilities better match the actual likelihood of correctness. Correspondingly, the Brier Scores decrease substantially for each model, e.g., Qwen-VL decrease from 0.4731 to 0.2778, reflecting reduced mean squared error between predicted probabilities and binary outcomes. Simultaneously, the ROC Curves show higher AUC, meaning the models after calibration separate true and false positives more effectively for a wide range of confidence thresholds. These joint gains on both Brier Score and AUC confirm that our perturbation-based preference training leads to better verbalized confidence calibration and more reliable confidence judgments, ultimately making the VLMs more trustworthy for multimodal tasks. Additional results for other datasets can be found in the appendix.

4.3.2 Ablation Experiments

Below we provide an ablation study to investigate the individual contributions of each component in our framework. Specifically, we explore four settings: (1) SFT only: using only supervised fine-tuning, (2) SimPO only: applying preference optimization to the base model without SFT, (3) Global Noise: applying perturbation globally on the entire image instead of mask-based perturbations, and (4) Original RLAIF: fine-tuning with the original, unmodified RLAIF dataset. We compare these ablations against our full approach, which incorporates semantic mask perturbation, SFT, and SimPO jointly. Figure 4 summarizes the performance of each ablation across representative metrics for verbalized confidence calibration.

SFT Only vs. Full Method. Fine-tuning the model with our newly constructed perturbation-based dataset without preference optimization does yield moderate improvements. However, it remains notably behind the performance of our full CSP framework. This suggests that while learning from the perturbed images is beneficial, the model also needs preference optimization to robustly align its outputs and confidence estimation.

SimPO Only vs. Full Method. Directly applying SimPO to the base model without SFT on perturbed data shows nearly no gains. Without the exposure to semantic perturbations, preference optimization alone struggles to calibrate the model under visual uncertainty.

Global Noise vs. Mask-Based Perturbation. Replacing our semantic mask perturbation with uniform noise on the entire image do not increase calibration performance. This underscores the importance of masking only the key objects: local, object-centric perturbations more realistically simulate the uncertainty conditions that VLMs encounter, enabling finer control and better confidence calibration.

Original RLAIF vs. Perturbation-Based Data. Finally, using just the original RLAIF dataset for SFT and SimPO does not have improvement. Indeed, the added supervision and varied noise conditions in our semantic perturbation dataset appear crucial for learning robust, multimodal confidence cues.

Altogether, the ablation results emphasize that both mask-based noise perturbation and preference optimization are needed to achieve the best alignment and calibration. Semantic perturbations successfully expose the model to diverse and realistic uncertainty scenarios, while SimPO fine-tunes how the model ranks and expresses confidence about those responses.

Model	POPE									AMBER					
	Random			Popular			Adversarial			Attribute			Relation		
	Acc (↑)	*F1* (↑)	*ECE* (↓)	*Acc* (↑)	*F1* (↑)	*ECE* (↓)	*Acc* (↑)	*F1* (↑)	*ECE* (↓)	*Acc* (↑)	*F1* (↑)	*ECE* (↓)	*Acc* (↑)	*F1* (↑)	*ECE* (↓)
Qwen-VL	0.25	0.21	0.5699	0.3	0.22	0.5249	0.27	0.21	0.5475	0.37	0.47	0.4732	0.1	0.07	0.4421
using CSP	**0.67**	***0.68***	**0.4225**	**0.67**	**0.68**	**0.4289**	**0.61**	**0.64**	**0.4407**	**0.69**	**0.7**	**0.4158**	**0.6**	**0.7**	**0.3674**
Qwen2-VL	0.11	0.01	0.412	0.04	0.01	0.477	0.04	0.01	0.4767	0.13	0.21	0.4694	0.03	0.02	0.442
using CSP	**0.71**	***0.73***	**0.4049**	**0.69**	**0.72**	**0.417**	**0.72**	**0.74**	**0.3948**	**0.78**	**0.81**	**0.3951**	**0.65**	**0.71**	**0.4**
InternVL2	0.78	0.74	0.0698	0.71	0.69	0.1333	0.66	0.65	0.1846	0.41	0.21	0.2501	0.23	0.18	0.309
using CSP	**0.79**	***0.74***	**0.0642**	**0.79**	**0.73**	**0.0888**	**0.78**	**0.73**	**0.1285**	**0.72**	**0.68**	**0.2246**	**0.79**	**0.82**	**0.2932**
Phi3.5-V	0.48	0.35	0.1798	0.28	0.28	0.3768	0.28	0.28	0.3791	0.25	0.22	0.3953	0.19	0.1	0.3424
using CSP	**0.69**	**0.7**	**0.095**	**0.69**	**0.7**	**0.2267**	**0.64**	**0.67**	**0.2399**	**0.54**	**0.57**	**0.2878**	**0.4**	**0.25**	**0.3307**

Table 1: Evaluation of verbalized confidence alignment with correctness across POPE and AMBER datasets. Accuracy (Acc) and F1 Score (F1) are higher-the-better (↑), while Expected Calibration Error (ECE) is lower-the-better (↓). These metrics do not measure dataset performance but rather assess the model's ability to express confidence in alignment with correctness. Our proposed method consistently improves confidence calibration across all models and settings.

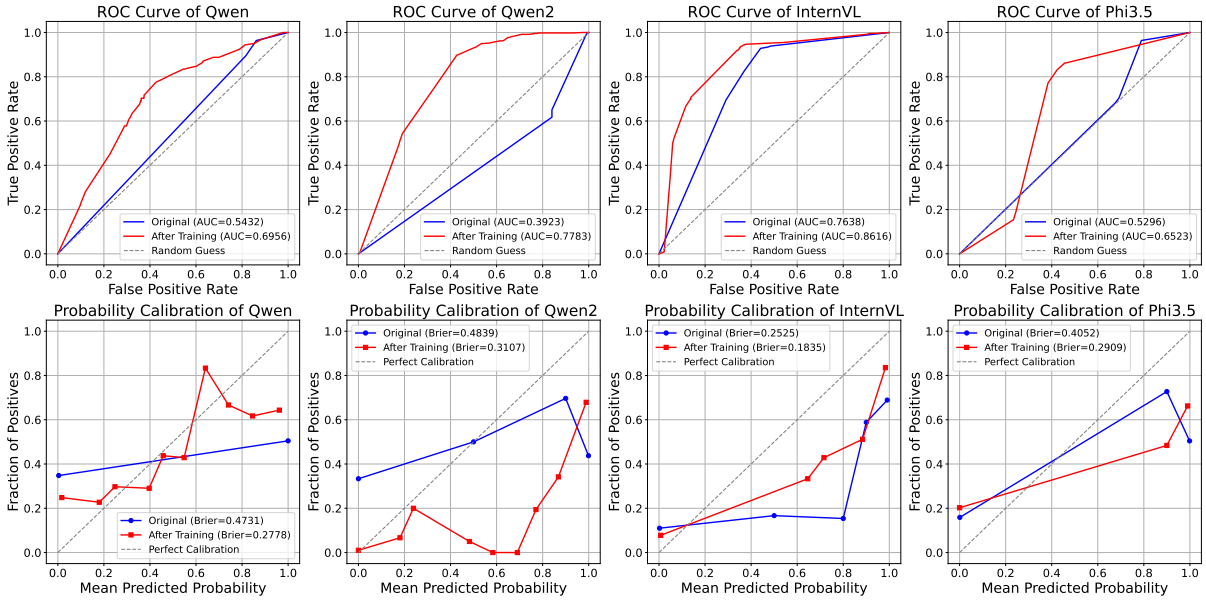

Figure 3: ROC curves (top row) and probability calibration plots (bottom row) on the AMBER attribute dataset, comparing their performance before and after applying our proposed confidence calibration method. The ROC curves illustrate improved true positive rates (higher AUC values) after training, while the probability calibration plots indicate better alignment between predicted confidence and correctness (lower Brier Scores).

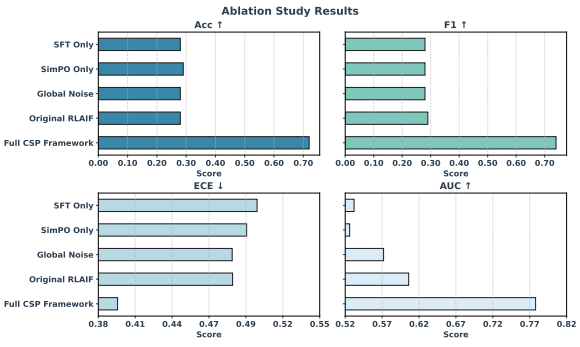

Figure 4: **Ablation results** for different variants of our method under POPE adversarial of model Qwen2

4.3.3 Analysis Experiments

Our proposed CSP framework not only enhances verbalized confidence calibration but also preserves overall all performance. Below, we analyze two key aspects of our results: (1) the positive impact of benchmark overall performance and (2) the improvement in confidence-probability alignment.

Preserving and Enhancing Model Performance A key concern in confidence calibration is whether adjustments to verbalized confidence lead to unintended trade-offs in model accuracy and general task performance. Our results show that this is not the case, i.e. our approach does not degrade overall model performance and even enhances key evaluation metrics. As illustrated in Figure 5, the bar chart compares key metrics across three configurations: the base vision-language model without additional calibration steps, the model after supervised fine-tuning on the perturbation-augmented

Model	POPE						AMBER			
	Random		Popular		Adversarial		Attribute		Relation	
	Spearman ρ	*Kendall* τ	*Spearman* ρ	*Kendall* τ	*Spearman* ρ	*Kendall* τ	*Spearman* ρ	*Kendall* τ	*Spearman* ρ	*Kendall* τ
Qwen-VL	0.06	0.05	0.11	0.09	0.07	0.06	0.16	0.12	-0.02	-0.02
using CSP	**0.16**	***0.11***	**0.14**	**0.09**	**0.14**	**0.1**	**0.29**	**0.2**	**0.09**	**0.06**
Qwen2-VL	0.26	0.21	0.13	0.11	0.11	0.09	-0.1	-0.08	-0.08	-0.07
using CSP	**0.33**	***0.25***	**0.29**	**0.22**	**0.37**	**0.27**	**0.53**	**0.39**	**0.22**	**0.16**
InternVL2	0.78	0.63	0.75	0.6	0.7	0.56	0.49	0.39	0.43	0.35
using CSP	**0.85**	***0.68***	**0.83**	**0.66**	**0.82**	**0.65**	**0.76**	**0.6**	**0.61**	**0.45**
Phi3.5-V	0.55	0.45	0.39	0.31	0.36	0.29	0.17	0.13	0.24	0.19
using CSP	**0.6**	**0.48**	**0.57**	**0.45**	**0.54**	**0.42**	**0.38**	**0.28**	**0.38**	**0.29**

Table 2: Spearman's (ρ) and Kendall's (τ) correlations between internal and verbalized confidence across models and datasets. Higher values indicate better alignment. Our calibration method consistently improves performance.

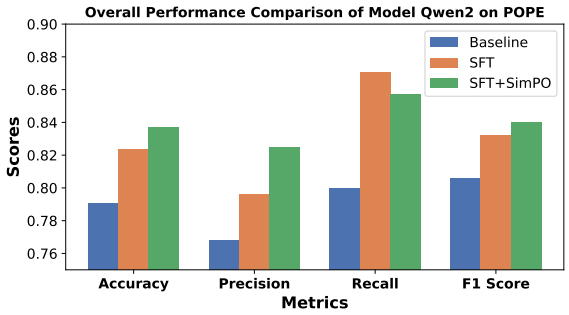

Figure 5: Comparison of Accuracy, Precision, Recall, and F1 Score across different model configurations.

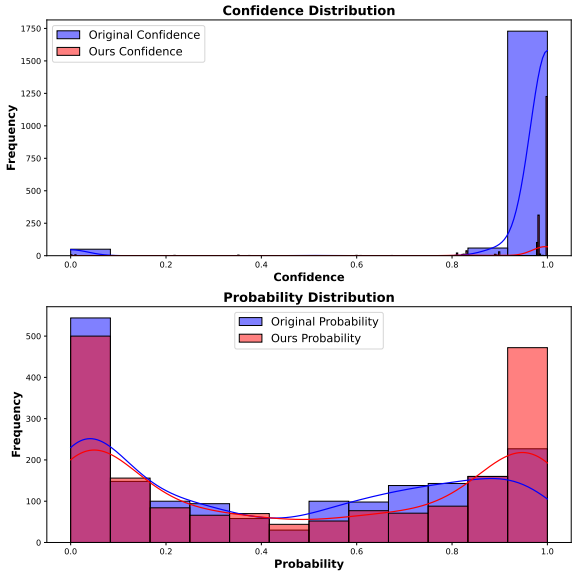

Figure 6: Comparison of token-level probability distributions before (left) and after (right) applying our method.

dataset, and the model further enhanced through preference optimization. From the results, we see that both SFT and SFT+SimPO outperform the baseline in all metrics. This performance gain is largely due to the semantic perturbation embedded in our dataset. By applying perturbations specifically to key semantic regions, the model learns to associate visual uncertainty with corresponding confidence levels while preserving task-relevant features. Consequently, the model not only becomes better at estimating verbalized confidence but also maintains or even improves its core predictive ability.

Strengthening Verbalized Confidence-Probability Alignment Although our training process does not explicitly constrain internal confidence, Table 2 and Figure 6 reveal a notable improvement in confidence-probability correlation. Before calibration, the model exhibited strong overconfidence, assigning excessively high probabilities even to uncertain responses. After fine-tuning with semantic perturbations and preference optimization, the probability distribution becomes more balanced, reducing misleadingly high-certainty predictions. We hypothesize that this improvement arises as an indirect effect of our approach that semantic perturbations expose the model to controlled uncertainty, forcing it to modulate confidence more realistically. Besides, preference optimization reinforces correct confidence ranking, indirectly refining probability estimates. In addition, reduced multimodal bias helps mitigate overconfidence, as the model learns to rely more on visual cues rather than textual priors. While our method was designed to calibrate verbalized confidence, these results suggest that better uncertainty modeling also aligns internal probabilities, making the model more trustworthy overall.

5 Conclusion

We introduced a semantic mask perturbation framework that simulates visual uncertainties to improve verbalized calibration in vision-language models. By pairing each perturbation level with a corresponding confidence score and further refining model behaviors through preference optimization, our method significantly enhances verbalized confidence calibration of correctness. Experimental results across various benchmarks and model architectures demonstrate consistent gains in evaluation metrics, decreased calibration errors, and improved reliability without sacrificing overall performance. Our approach thus represents a practical step toward more trustworthy multimodal systems.

Limitations

While our approach demonstrates strong empirical improvements in verbalized confidence calibration for VLMs, several important limitations remain:

- **Model Scale Constraints.** Due to limited computational resources, our experiments primarily focused on models with moderate parameter sizes. It remains unclear whether these gains will hold or even amplify when applied to significantly larger vision-language models.
- **Object-Level Perturbation.** Our semantic mask perturbation currently targets object-level masks, which effectively captures uncertainty around core entities. However, many real-world scenarios involve more nuanced uncertainties tied to contextual and knowledge-based cues (e.g., subtle background details, temporal coherence, or commonsense inferences). Incorporating additional perturbation mechanisms that account for these richer modalities is left for future work.
- **LoRA vs. Full-Parameter Fine-Tuning.** In this work, we primarily apply full-parameter fine-tuning, which may risk partial forgetting of previously acquired knowledge. It remains unclear if a parameter-efficient strategy such as LoRA can match or surpass our results while avoiding catastrophic forgetting. Further research comparing different fine-tuning techniques would offer deeper insights into this trade-off.
- **Generality Beyond Current Benchmarks.** Although our method shows consistent improvements on widely adopted benchmarks, its generalizability to more diverse or specialized domains (e.g., medical imaging, remote sensing) has not been fully established. Subsequent research could explore the adaptability of our framework in domain-specific settings with potentially unique uncertainty factors.

Addressing these limitations could further enhance the robustness, scalability, and versatility of our semantic perturbation-based calibration framework for vision-language models.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595.

Tobias Groot and Matias Valdenegro-Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of TrustNLP Workshop@ NAACL 2024*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Zaid Khan and Yun Fu. 2024. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10854–10863.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 315–334, Bangkok, Thailand. Association for Computational Linguistics.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2025. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

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

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*.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. Sayself: Teaching llms to express confidence with self-reflective rationales. *arXiv preprint arXiv:2405.20974*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, Qi Guo, and Yunji Chen. 2024a. Towards analyzing and mitigating sycophancy in large vision-language models. *arXiv preprint arXiv:2408.11261*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. 2024b. Swift: a scalable lightweight infrastructure for fine-tuning. *arXiv preprint arXiv:2408.05517*.

A Prompt Templates

In this section, we present the full prompts in various aspects of our experiments.

A.1 Key Object Region Extraction Prompt

we use the following prompts when generating the key object region description for the constructed datasets.

> Extract the single most important keyword (a noun or object) from each of the following question-answer pairs. Provide only one keyword.
> Example 1: Question: What kind of potato chips are on the plate? Answer: There are some light yellow thin slice-shaped potato chips in this plate, which look very crispy. Keyword: potato chips
> Example 2: Question: What color is the car parked outside the house? Answer: The car parked outside is a bright red sedan. Keyword: car
> Example 3: Question: What kind of fruits are in the basket? Answer: The basket contains fresh green apples and ripe yellow bananas. Keyword: fruits
> Now, using the following question and answer, extract one most important keyword. Just output the keyword directly.
> Question: ***Question***
> Answer: ***Answer***
> Keyword:

A.2 Prompts for Evaluating Verbalized Confidence

For the following prompt, a few-shot prompt may be used in some models.

> "A Language model was asked: {***Query***}, Model's answer was {***Response***} Analyze its answer. How certain are you about model's answer from 1%-100% ? "

B Additional Experiment Details

B.1 Implementation Details

We use SWIFT (Zhao et al., 2024b) framework for the training procedure of our method. Details of hyperparameters can be seen below. For hyperparameters not explicitly mentioned, we use the default settings.

B.1.1 Hyperparameters used for supervised fine-tuning

- **SFT type:** Full
- **Batch size:** 2
- **gradient checkpointing:** True
- **gradient accumulation steps:** 8
- **Number of epochs:** 1

B.1.2 Hyperparameters used for preference optimization

- **RLHF type:** SimPO
- **Batch size:** 1
- **gradient checkpointing:** True
- **gradient accumulation steps:** 16
- **Number of epochs:** 1
- β**:** 2.0
- γ_{simpo}**:** 1.0
- α_{cpo}**:** 0.0
- **warm-up ratio** 0.03

B.2 Dataset License

In this section, we list the licenses of the datasets we used in this paper. We used the datasets for research purposes as allowed by the corresponding licenses and consistent with the intended use.

POPE (Li et al., 2023): MIT License. We donwloaded the data from POPE.

AMBER (Wang et al., 2023): Apache License. We donwloaded the data from AMBER.

B.3 Computation Requirements

We ran our experiments on a server with 2× AMD EPYC 7513 32-Core Processor and 4× NVIDIA A100-SXM4-80GB and 1T RAM.

C Additional Results

C.1 Additional Calibration Results

We illustrate the additional results of Brier Score and ECE of POPE dataset and AMBER relation dataset.

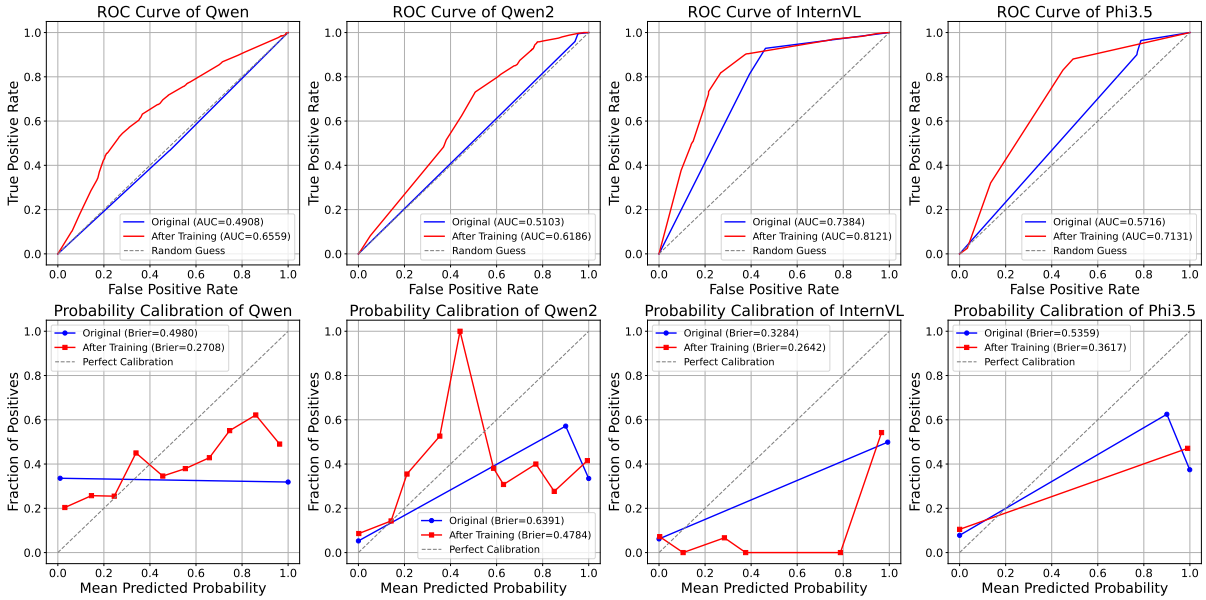

Figure 7: ROC curves (top row) and probability calibration plots (bottom row) on the POPE adversarial dataset.

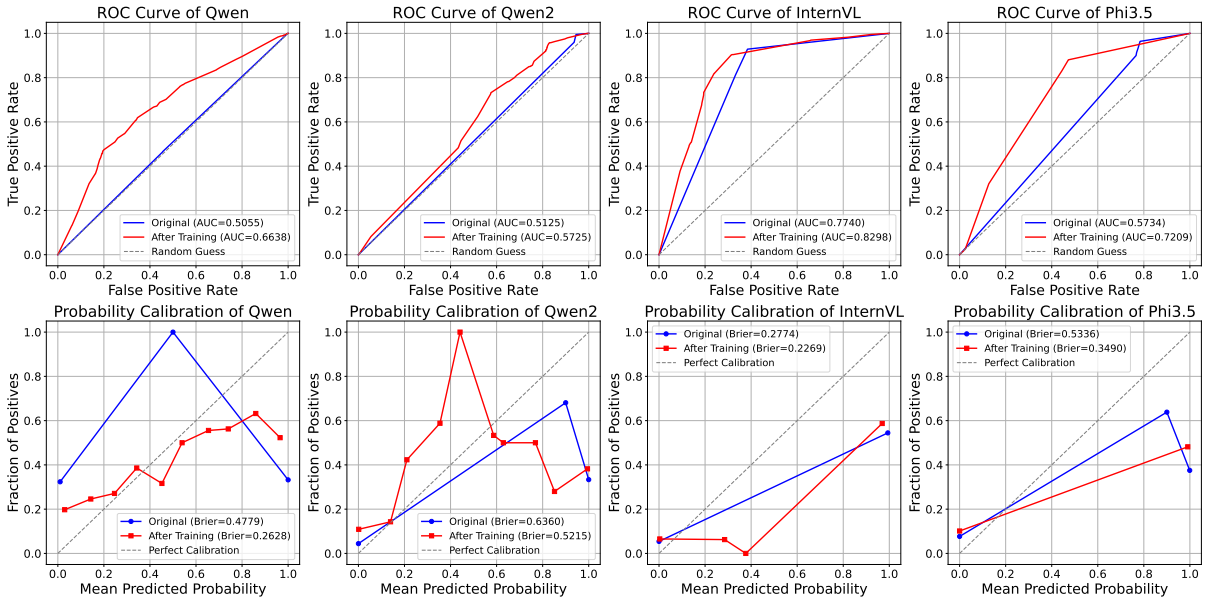

Figure 8: ROC curves (top row) and probability calibration plots (bottom row) on the POPE popular dataset.

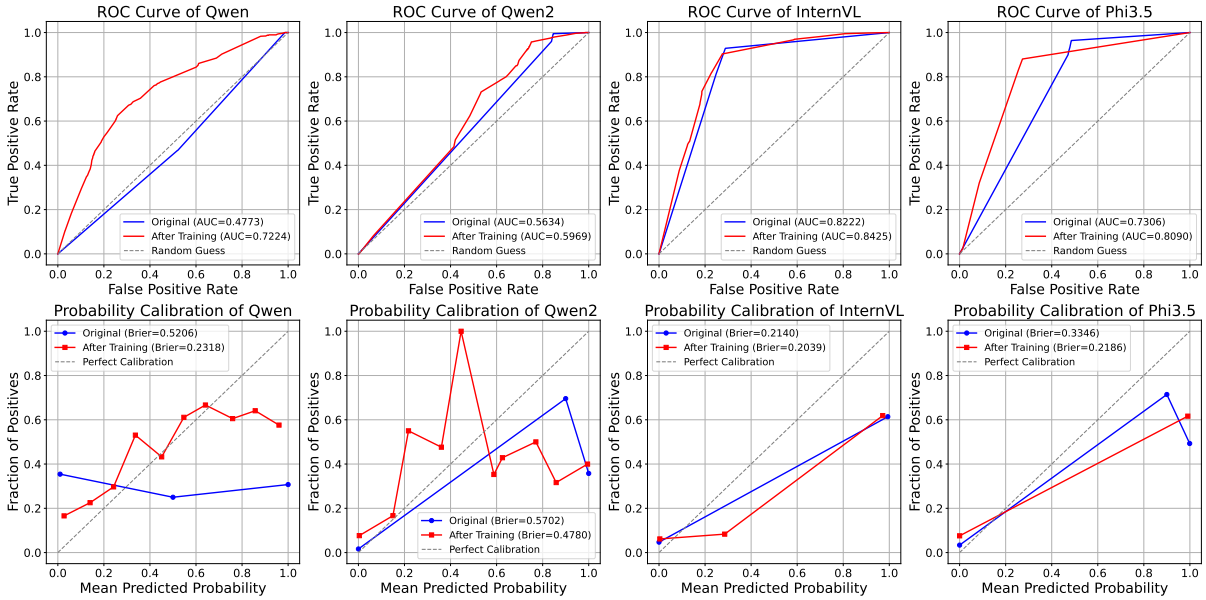

Figure 9: ROC curves (top row) and probability calibration plots (bottom row) on the POPE random dataset.

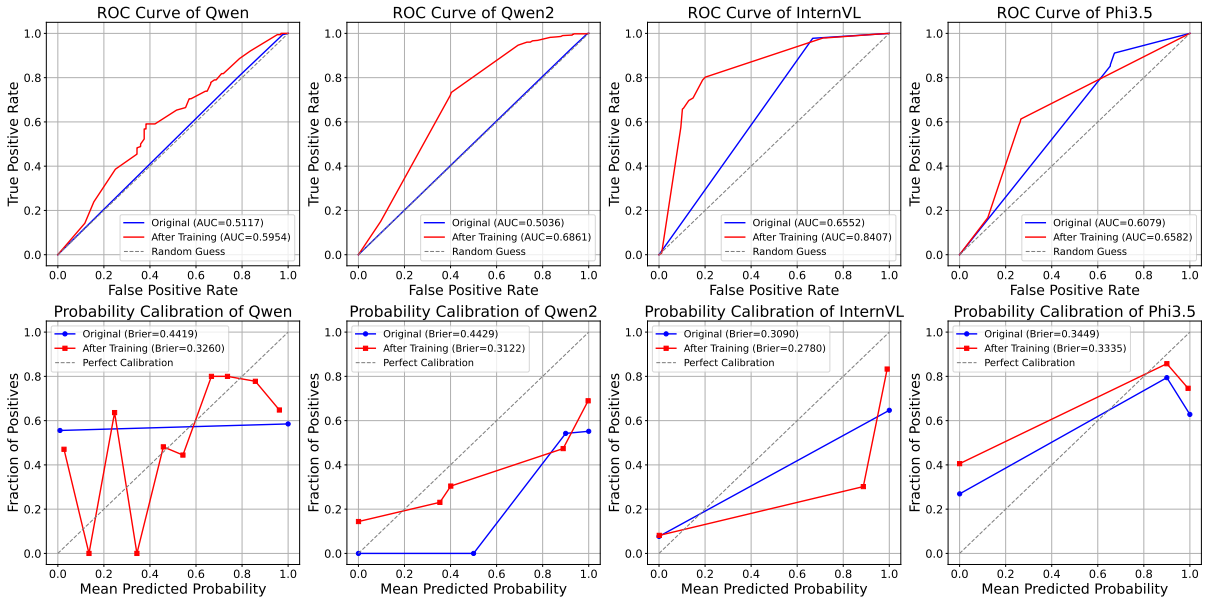

Figure 10: ROC curves (top row) and probability calibration plots (bottom row) on the AMBER relation dataset.