Reading Between the Lines: Abstaining from VLM-Generated OCR Errors via Latent Representation Probes

Jihan Yao[1]*, Achin Kulshrestha[2], Nathalie Rauschmayr[2], Reed Roberts[2],
Banghua Zhu[1], Yulia Tsvetkov[1], Federico Tombari[2]
[1]University of Washington [2]Google
jihany2@cs.washington.edu kulac@google.com rauschmayr@google.com

Abstract

As VLMs are deployed in safety-critical applications, their ability to abstain from answering when uncertain becomes crucial for reliability, especially in Scene Text Visual Question Answering (STVQA) tasks. For example, OCR errors like misreading "50 mph" as "60 mph" could cause severe traffic accidents. This leads us to ask: Can VLMs know when they can't see? Existing abstention methods suggest pessimistic answers: they either rely on miscalibrated output probabilities or require semantic agreement unsuitable for OCR tasks. However, this failure may indicate we are looking in the wrong place: uncertainty signals could be hidden in VLMs' internal representations.

Building on this insight, we propose Latent Representation Probing (LRP): training lightweight probes on hidden states or attention patterns. We explore three probe designs: concatenating representations across all layers, aggregating attention over visual tokens, and ensembling single layer probes by majority vote. Experiments on four benchmarks across image and video modalities show LRP improves abstention accuracy by 7.6% over best baselines. Our analysis reveals: probes generalize across various uncertainty sources and datasets, and optimal signals emerge from intermediate rather than final layers. This establishes a principled framework for building deployment-ready AI systems by detecting confidence signals from internal states rather than unreliable outputs.

1. Introduction

As Vision-Language Models (VLMs) transition from benchmark testing to operational environments, such as self-driving systems [1], robotics [26], and assistive augmented reality glasses [8]—ensuring their reliability becomes paramount. In these safety-critical deployments, even minor errors can have severe consequences: misreading "50 mph" as "60 mph" could lead to critical accidents. When models cannot guarantee correctness, *abstention* – the ability to refrain from answering when uncertain – serves as a crucial safeguard for preventing such failures.

Among capabilities essential for reliable deployment, Scene Text Visual Question Answering (STVQA) tasks presents particularly reliability challenges. First, OCR is an inherently difficult task for VLMs. VLMs' lightweight alignment layers introduce information bottlenecks that hinder fine-grained perceptual detail transfer [21], where image details such as small character recognition are discarded. Studies have found that LLMs using OCR-extracted text significantly outperforming end-to-end VLMs [9, 24]. Second, VLMs' OCR capabilities prove especially fragile under real-world conditions: text recognition accuracy deteriorates severely under common corruptions like blur and snow [36], with performance degrading sharply on imperfect charts and dynamic video environments [30, 40]. Give the difficulties, a central question raises: *Can VLMs know when they can't see?*

Existing uncertainty estimation approaches suggest pessimistic answers. Token-level methods [22, 38] estimate confidence from token probabilities, yet mounting evidence shows that output logits in VLMs are often severely miscalibrated [12, 44, 46]. Response-level aggregation methods like self-consistency [37] and semantic entropy [11] are not applicable to free-form answers [5] since OCR requires character-level precision instead of semantical equivalence. However, the failure of these surface-level signals to capture uncertainty does not necessarily mean VLMs lack internal awareness of their perceptual limitations. Instead, it may simply mean we are looking in the wrong place.

We hypothesize that *uncertainty signals exist but are hidden in VLMs' latent representations*. This hypothesis is grounded by two observations: (1) different layers encode qualitatively different information [16, 48] with lower layers capturing surface features and upper layers representing more semantic content which may include confidence information, and (2) instruction tuning concentrates changes

*Work done during an internship at Google.

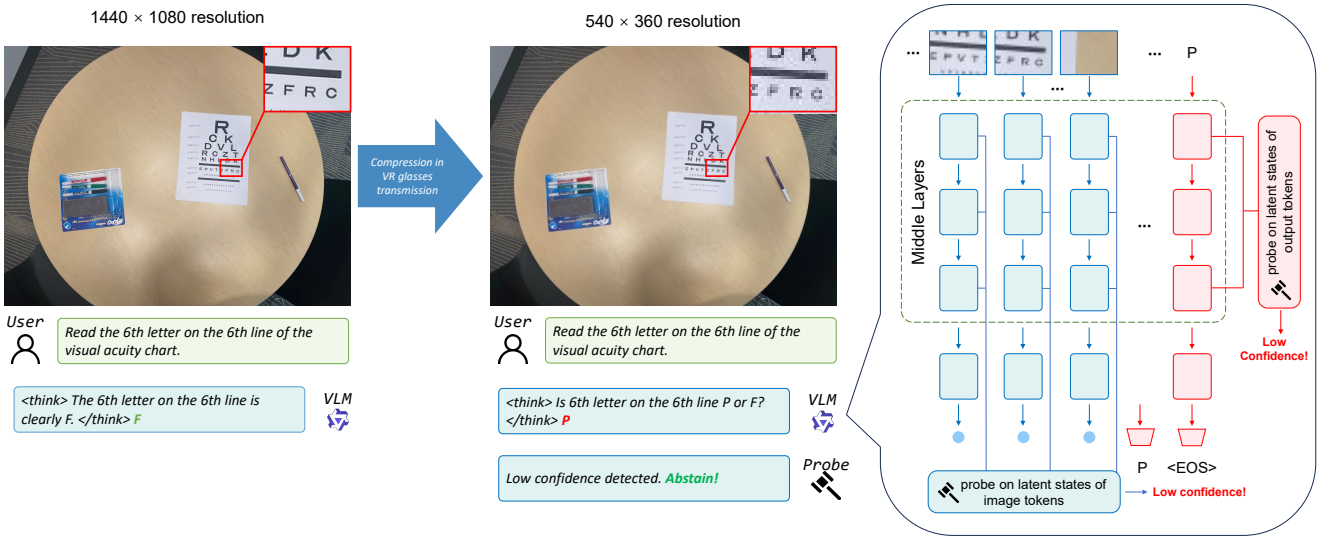

Figure 1. Real-world applications such as assistive AR glasses often encounter low-resolution, corrupted images (middle) rather than the high-quality images (left) typical in laboratory settings. To enable reliable deployment, it is crucial to estimate VLMs' uncertainty and abstain from generating incorrect responses. We propose Latent Representation Probing (LRP) for effective and efficient confidence estimation in VLMs. Our analysis reveals that the most effective probing approaches are: (1) aggregating attention patterns on image token positions across all layers (blue probe, right), and (2) ensembling hidden states from layers that exhibit the strongest confidence signals (red probe, right) through majority vote on output token positions.

in final layers [42] while intermediate layers are better preserved [19, 43]. Rather than relying on miscalibrated output logits, we propose **Probing Latent Representations** – hidden states and attention patterns – to uncover the buried uncertainty signals. We explore three probing strategies: (1) concatenating representations across all layers and training a single probe, (2) aggregating attention specifically over visual tokens before probing, and (3) training separate probes for each layer and ensemble by majority vote.

We evaluate our approach across four diverse STVQA benchmarks spanning image and video modalities. Among all methods, ensemble hidden state probing achieves the best performance, improving average abstention accuracy over best baseline by 7.6%. Analysis reveals that hidden states encode richer semantic uncertainty signals than attention patterns, and LRP is the only method having meaningful abstention performance on video inputs where baselines fail due to their inability to handle long sequences effectively.

To understand the strong performance of hidden state probing, we conduct three analyses. First, we investigate what signals probes capture by evaluating on four synthetic datasets designed to simulate various low model confidence scenarios. The average abstention accuracy of 0.825 suggests they learn general uncertainty representations rather than dataset-specific patterns. Second, layer-wise analysis confirms our hypothesis: the optimal probe operates on intermediate layers instead of the final layer. Finally, to address generalization concerns inherent to training-based methods, cross-dataset evaluation shows that attention pattern probes achieve substantially stronger out-of-distribution performance, improving by 5.8% over best training-based baselines.

In summary, our contributions are threefold: (1) To the best of our knowledge, this work is the first to evaluate VLMs reliability and abstention capabilities in STVQA tasks and systematically compare various abstention baselines; (2) Our proposed method, latent representation probing, consistently outperforms state-of-the-art baselines across diverse benchmarks; (3) Our method also facilitates comprehensive analyses: we demonstrate that latent representation probing capture robust, generalizable uncertainty representations. This work advances the path toward reliable VLM deployment by enabling models to recognize when they should abstain rather than hallucinate, which is a critical step for trustworthy deployment.

2. Related Work

Abstention in VLMs Existing approaches of abstention in VLMs can be categorized into three main paradigms. (1) Token-level methods estimate confidence from output probabilities, often applying temperature scaling [35] or verbal-

ized confidence [15]. However, these approaches face severe limitations: studies reveal that VLMs exhibit systematic miscalibration with prevalent overconfidence [12, 46]. (2) Response-level methods aggregate information across multiple model generations, including self-consistency [37, 50], semantic entropy [11], and self-evaluation frameworks [6]. While effective for constrained tasks, these methods struggle with open-ended text recognition where character-level precision is required and semantic similarity is insufficient. (3) Training-based methods fine-tune models to explicitly refuse uncertain queries through refusal-aware instruction tuning [4, 49] or train a better self-reflection model [41]. Previous approaches primarily operate on final-layer representations and have not been systematically evaluated on text-intensive VQA tasks.

Parallelly, recent work has explored internal representations for hallucination detection on object-focused VQA tasks. Attention-based approaches [17, 23, 45, 51] analyze cross-modal attention patterns to detect hallucinations, offering *interpretability* of hallucination. Duan et al. [10], Phukan et al. [31] leverage contextual embeddings from middle layers. These approaches primarily focus on interpretability, aiming to understand *why* hallucinations occur rather than maximizing detection performance. Moreover, some methods have high computational cost, making it prohibitive for video inputs. In contrast, our lightweight probes require only a single forward pass.

STVQA Benchmarks STVQA encompasses diverse tasks across both image and video. Image-based benchmarks range from basic scene text understanding (TextVQA [32], ST-VQA [3]) to specialized domains including document analysis (DocVQA [28]), chart interpretation (ChartQA [27]), and book covers (OCRVQA [29]). Video benchmarks extend these challenges with temporal reasoning requirements (M4-ViteVQA [52], EgoTextVQA [53]). Comprehensive evaluations like OCRBench [24] and SeedBench Plus 2 [20] assess models across key information extraction, text location, etc. These benchmarks collectively evaluate models' abilities in spatial and temporal reasoning, and fine-grained perception. However, STVQA presents unique challenges: text appears in arbitrary orientations and fonts, and usually come up with corruptions, demanding robust performance under diverse real-world conditions.

3. Baselines

Despite growing interest in VLM reliability, existing abstention methods for STVQA tasks lack systematic comparison. To establish a comprehensive evaluation of these approaches, we define the abstention problem as: given a visual input $\mathbf{v}$ (image or video) and a question $\mathbf{q}$, a generated answer $\mathbf{a} = \text{VLM}(\mathbf{v}, \mathbf{q})$, our goal is to determine when the model should abstain from answering. We aim to learn a confidence score $s(\mathbf{v}, \mathbf{q}, \mathbf{a}, \mathbf{e}) \rightarrow [0, 1]$ that estimates the model's uncertainty, where $\mathbf{e}$ is optional external confidence evidence. The model should abstain when $s(\mathbf{v}, \mathbf{q}, \mathbf{a}, \mathbf{e}) < \tau$ and answer otherwise, where τ is a selected threshold on the validation set.

Token Probability A well-calibrated VLM's token probabilities should reflect response reliability [14], with more probable answers more likely correct. We use the average token probability as the confidence indicator. Specifically, for an answer $\mathbf{a} = [t_1, t_2, \ldots, t_n]$ with token probabilities $[p_1, p_2, \ldots, p_n]$, the abstention function is:

$$s = \exp(\frac{1}{n}\sum_{i=1}^{n} log(p_i)) \tag{1}$$

Ask for Calibration Verbalized confidence has been proven to be more calibrated than token probability [34]. After generating answer $\mathbf{a}$, we elicit a numerical confidence $c \in [0, 1]$ by prompting *"Rate how likely the answer is correct from 1 to 5, where 5 is the most likely."*:

$$s = \text{VLM}([\mathbf{v}, \mathbf{q}, \mathbf{a}, \text{PROMPT}]) \tag{2}$$

, where $[\cdot]$ denotes concatenation.

Self Consistency Self-consistency [37] generates multiple responses and uses majority voting for abstention. For OCR tasks with free-form responses, we measure answer similarity using character accuracy (CA). Given k sampled answers $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$, the confidence is estimated by:

$$s = \frac{1}{k}\sum_{i=1}^{k} \text{CA}(\mathbf{a}, \mathbf{a}_i) \tag{3}$$

Prompt to Abstain Inspired by none-of-the-above options in multiple-choice QA [18], we adapt this approach for free-form OCR tasks by instructing the model to abstain when uncertainty. We prepend the prompt with: *"Answer 'I don't know' if you are uncertain about your answer."* The confidence function is then:

$$s = \mathbb{I}(\text{"I don't know"} \notin \mathbf{a}) \tag{4}$$

, where $\mathbb{I}[\cdot]$ is the indicator function.

VLM Judge We can train a VLM to judge the answer correctness [18, 47]. Given the visual input $\mathbf{v}$, question $\mathbf{q}$, and answer $\mathbf{a}$, along with the instruction *"Is this answer correct? Answer True or False."*, the VLM judge should output a special token whose probability indicates correctness:

$$s = \mathbb{I}(p(\text{"True"}|\mathbf{v}, \mathbf{q}, \mathbf{a}) > p(\text{"False"}|\mathbf{v}, \mathbf{q}, \mathbf{a})) \tag{5}$$

R-Tuning Following R-tuning approaches [49], we fine-tune the model to express confidence alongside its answers. During training, we augment correct answers with the suffix *"I am sure"* and incorrect answers with *"I am unsure"*. At inference, the model generates both answer and confidence, and we extract the confidence score:

$$s = \mathbb{I}(\text{"I am sure"} \in \mathbf{a}) \tag{6}$$

Summed Visual Attention Ratio (SVAR) Inspired by the fact that hallucinated answers receive lower overall attention to visual tokens, SVAR [17] uses the image token attention weights from middle layers as confidence signals. The confidence is estimated by:

$$s = \sum_{\ell=\ell_{\text{start}}}^{\ell_{\text{end}}} \frac{1}{H} \sum_{h=1}^{H} \sum_{i=1}^{n} A_{i,o_1}^{(\ell,h)} \tag{7}$$

, where $A_i^{(\ell,h)}$ is the attention weight of the first output token's o_1 attention on i-th visual token at layer ℓ and head h, n is the number of visual tokens, H is the number of attention heads.

Contextual Lens Contextual Lens [31] leverages hidden states from middle layers instead of attention weights. Given image token hidden states $\{\mathbf{h}_i^{(\ell)}\}_{i=1}^{n}$ and answer token hidden states $\{\mathbf{h}_j^{(\ell)}\}_{j=1}^{m}$ at layer ℓ, the confidence is computed as the maximum cosine similarity between them:

$$s = \max_{\ell} \max_{i} \frac{\mathbf{h}_i^{(\ell)} \cdot \left(\sum_{j=1}^{m} \mathbf{h}_j^{(\ell)}\right)}{\|\mathbf{h}_i^{(\ell)}\| \| \sum_{j=1}^{m} \mathbf{h}_j^{(\ell)}\|} \tag{8}$$

where m is the number of answer tokens at layer ℓ.

4. Latent Representation Probing (LRP)

Our approach trains lightweight probes on VLMs' internal representations to detect OCR errors. We investigate two types of latent representations during inference and three different probe designs.

4.1. Latent Representations

Hidden State Representations We extract hidden state representations from intermediate transformer layers. For a VLM with L transformer blocks, let $\mathbf{h}_j^{(\ell)}$ denote the latent state of the j-th token at layer ℓ, where $\mathbf{h}^{(\ell)} = \text{TransformerBlock}_\ell(\mathbf{h}^{(\ell-1)})$. Given an answer $\mathbf{a} = [t_1, t_2, \ldots, t_m]$, we average the hidden states of all output tokens at each layer:

$$\mathbf{h}_{\text{out}}^{(\ell)} = \frac{1}{m} \sum_{j=1}^{m} \mathbf{h}_{t_j}^{(\ell)} \in \mathbb{R}^D \tag{9}$$

, where D is the hidden dimension.

Attention Pattern Representations For each attention head h at layer ℓ, the attention weight from output token j to input token i is computed as $A_{j,i}^{(\ell,h)}$. We average the attention weights across all output tokens:

$$\mathbf{a}_{\text{out}}^{(\ell)} = \frac{1}{m} \sum_{j=1}^{m} A_{j,:}^{(\ell,:)} \in \mathbb{R}^{n \times H} \tag{10}$$

, where m is the number of output tokens, n is the number of input tokens, and H is the number of attention heads.

4.2. Probe Design

Concatenated Probe For latent states, we concatenate and flatten representations across all L layers, then train a 4-layer MLP $f_\theta : \mathbb{R}^{L \cdot D} \to \{0, 1\}$:

$$s = f_\theta\left([\mathbf{h}_{\text{out}}^{(1)}, \mathbf{h}_{\text{out}}^{(2)}, \ldots, \mathbf{h}_{\text{out}}^{(L)}]_{\text{flatten}}\right) \tag{11}$$

For attention patterns, we flatten across layers and heads for each input token position, then train a 4-layer transformer encoder $g_\theta : \mathbb{R}^{n \times (L \cdot H)} \to [0, 1]$:

$$s = g_\theta\left([\mathbf{a}_{\text{out}}^{(1)}, \mathbf{a}_{\text{out}}^{(2)}, \ldots, \mathbf{a}_{\text{out}}^{(L)}]_{\text{flatten}}\right) \tag{12}$$

Visual-Focused Probe Motivated by prior work [17] showing that attention to visual tokens is most informative for hallucination detection, we propose using multi-head activation as confidence features. Let $\mathcal{V}$ denote the set of visual token positions. We average attention weights over visual tokens: $\mathbf{a}_{\text{vis}}^{(\ell)} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathbf{a}_{\text{out},i}^{(\ell)} \in \mathbb{R}^H$. We then apply a 4-layer MLP $f_\theta : \mathbb{R}^{L \cdot H} \to \{0, 1\}$:

$$s = f_\theta\left([\mathbf{a}_{\text{vis}}^{(1)}, \mathbf{a}_{\text{vis}}^{(2)}, \ldots, \mathbf{a}_{\text{vis}}^{(L)}]_{\text{flatten}}\right) \tag{13}$$

Ensemble Probe Instead of using all layers jointly, we train individual probes for each layer separately. For latent states, we train L independent MLPs $\{f_\theta^{(\ell)}\}_{\ell=1}^{L}$, where each $f_\theta^{(\ell)} : \mathbb{R}^d \to \{0, 1\}$ operates on a single layer's representation. We evaluate each layer's performance on a validation set and select the top-K layers with highest abstention accuracy. At inference, we aggregate predictions through majority voting:

$$s = \mathbb{I}[\frac{1}{K} \sum_{\ell \in \mathcal{L}_{\text{top-K}}} f_\theta^{(\ell)}(\mathbf{h}_{\text{out}}^{(\ell)}) \geq 0.5] \tag{14}$$

where $\mathcal{L}_{\text{top-K}}$ denotes the selected layers. In our experiments, we set $K = 5$. The same approach applies to attention pattern probes.

4.3. Training

The probes are trained using binary cross-entropy loss with soft labels:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(s_i) + (1-y_i)\log(1-s_i)] \quad (15)$$

where $y_i \in [0,1]$ is the soft correctness label computed by each dataset's evaluation script and N denotes the total number of samples. Since the probes we train are classifiers, LRP are threshold-free during inference.

5. Experiments

5.1. Models and Datasets

We employ two open-source VLMs for our experiments, QWEN-VL-2.5 [2] and GEMMA3-12B [33] respectively. For single-pass methods, we employ greedy decoding and max generation length 128 to ensure reproducible results. We use temperature 1.0 and generate 5 responses per prompt when sampling is required. Since GEMMA3-12B doesn't natively support video inputs, we process videos into 1 fps sequences and inputs as multiple images. For more detailed experiment setup, please refer to Appendix.

We conduct experiments on four text-rich VQA datasets spanning both image and video modalities, with the first two focused on general OCR tasks and last two specifically on in-the-wild Text VQA tasks. For all datasets, we split the data into training, validation, and test sets with a 6:2:2 ratio. We select the abstention threshold τ on the validation set and report all results on the test set.

- **OCRBench v2** [13] is by far the most comprehensive OCR evaluation benchmark comprising 10K human-verified QA pairs across text recognition, text location, scene text-centric VQA, document-oriented VQA, text element parsing, and knowledge-intensive text reasoning.
- **SEED-Bench-2-Plus** [20] contains 2.3K multiple-choice questions with human annotations spanning three online visual displays (charts, maps, and webs). We evaluate models using free-form generation instead of the original multiple choices to increase task difficulty.
- **HierText** [25] features hierarchical annotations of text in the wild scenes and documents, containing 11K images with word, line, and paragraph level annotations. Since HierText provides only OCR annotations, we use GEMINI-2.5-PRO [7] to synthesize question-answer pairs based on the annotations, followed by human validation to ensure quality and relevance.
- **EgoTextVQA** [53] is a egocentric video QA benchmark containing 1.5K outdoor driving and 4.4K indoor robotics housekeeping activities. Due to computational limitation, we sample the videos clips with less 1 minute duration.

5.2. Evaluation Metrics

We adopt four complementary metrics to evaluate abstention performance following previous work [12, 38]. Since OCR labels are continuous rather than binary, we modify the standard metrics to handle soft correctness labels.

Effective Reliability (ER) evaluates the overall system reliability of VLMs with abstention mechanism. It rewards correct answers while penalizing incorrect ones, thus penalizing both over-abstaining and under-abstaining:

$$\text{ER} = \frac{1}{N}\sum_{i=1}^{N}(2y_i - 1)\cdot \mathbb{I}[s_i \geq \tau] \quad (16)$$

Abstention Accuracy (A-Acc) evaluates whether the model makes correct abstention decisions: abstaining on incorrect answers and answering on correct answers:

$$\text{A-Acc} = \frac{1}{N}\sum_{i=1}^{N}[\mathbb{I}[s_i < \tau]\cdot(1-y_i) + \mathbb{I}[s_i \geq \tau]\cdot y_i] \quad (17)$$

Reliable Accuracy (R-Acc) penalizes under-abstention where the model should abstain when they are likely to make critial mistakes:

$$\text{R-Acc} = \frac{\sum_{i=1}^{N} y_i \cdot \mathbb{I}[s_i \geq \tau]}{\sum_{i=1}^{N}\mathbb{I}[s_i \geq \tau]} \quad (18)$$

Abstention Precision (A-Pre) penalizes over-abstention where the model unnecessarily abstains on samples it could answer correctly:

$$\text{A-Prec} = \frac{\sum_{i=1}^{N}(1-y_i)\cdot \mathbb{I}[s_i < \tau]}{\sum_{i=1}^{N}\mathbb{I}[s_i < \tau]} \quad (19)$$

5.3. Results

We present the abstention performance of baseline methods and LRP in Table 1.

- **Ensemble hidden state probe and visual-focused attention pattern probe are the best.** For hidden states, the ensemble probe yields the best average effective reliability and abstention accuracy, improving upon the best baseline by 33.8% and 7.6% respectively. For attention patterns, the visual-focused probe achieves the second-best performance, improving by 23.3% and 6.1%. This suggests attention patterns, which are scalar relevance scores, lack the rich semantic information encoded in hidden states that is necessary for detecting OCR errors.

Method	OCRBench v2				SeedBench Plus 2				HierText				EgoTextVQA				Avg
	ER	A-Acc	R-Acc	A-Pre	ER	A-Acc	R-Acc	A-Pre	ER	A-Acc	R-Acc	A-Pre	ER	A-Acc	R-Acc	A-Pre	
Qwen2.5-VL-7B																	
Token Prob.	0.108	0.665	0.732	0.645	0.169	0.667	0.651	0.687	0.161	0.663	0.644	0.685	0.049	0.681	0.622	0.695	0.669
Ask for Calib.	0.035	0.592	0.560	0.606	0.156	0.654	0.784	0.604	0.034	0.536	0.521	0.607	0.001	0.635	0.501	0.731	0.604
Self Consist.	0.120	0.677	**0.868**	0.640	0.138	0.636	0.658	0.619	0.124	0.625	**0.722**	0.642	0.010	0.644	**0.727**	0.642	0.646
Prompt to Abs.	-0.164	0.418	0.418	**1.000**	0.057	0.529	0.529	**1.000**	-0.146	0.427	0.427	**1.000**	-0.328	0.336	0.336	**1.000**	0.427
VLM Judge	0.211	0.768	0.755	0.777	0.316	0.737	0.742	0.728	**0.192**	**0.749**	0.689	0.811	0.000	0.634	1.000	0.634	0.722
R-Tuning	0.122	0.680	0.596	0.828	0.316	0.737	0.709	0.821	-0.090	0.467	0.453	0.833	-	-	-	-	0.628
SVAR	0.030	0.587	0.550	0.604	0.184	0.605	0.644	0.537	-0.093	0.464	0.448	0.595	0.003	0.636	0.562	0.638	0.573
Context Lens	0.016	0.573	0.516	0.627	0.162	0.583	0.583	0.625	-0.003	0.554	0.498	0.684	0.000	0.634	1.000	0.634	0.586
Concat (Attn.)	0.036	0.594	0.522	0.901	0.300	0.721	0.689	0.849	0.071	0.628	0.553	0.783	0.069	0.703	0.599	0.757	0.662
Concat (Hid.)	0.224	0.781	0.774	0.786	0.314	0.735	0.712	0.798	0.121	0.678	0.611	0.757	0.085	0.718	0.611	0.785	0.728
Visual (Attn.)	0.210	0.768	0.760	0.772	0.314	0.735	**0.785**	0.673	0.102	0.659	0.575	0.837	0.088	0.722	0.653	0.750	0.721
Ensemble (Attn.)	0.171	0.729	0.678	0.775	0.246	0.667	0.650	0.744	0.093	0.650	0.660	0.646	0.043	0.676	0.713	0.672	0.680
Ensemble (Hid.)	**0.231**	**0.789**	0.781	0.794	**0.342**	**0.763**	0.775	0.744	0.152	0.709	0.652	0.765	**0.104**	**0.738**	0.674	0.765	**0.750**
Gemma3-12B																	
Token Prob.	0.094	0.700	0.605	0.777	0.136	0.656	0.674	0.644	0.059	0.690	0.583	0.750	0.098	0.595	0.597	0.594	0.660
Ask for Calib.	-0.169	0.436	0.409	0.806	-0.024	0.496	0.488	0.889	-0.263	0.368	0.368	**1.000**	0.023	0.518	0.512	**0.867**	0.455
Self Consist.	0.137	0.743	0.718	0.754	0.215	0.735	0.693	0.787	0.139	**0.771**	0.680	0.828	0.113	0.608	0.596	0.626	0.714
Prompt to Abs.	-0.148	0.453	0.421	0.951	0.132	0.568	0.566	**1.000**	-0.111	0.449	0.444	**1.000**	0.044	0.546	0.523	0.866	0.504
VLM Judge	0.166	0.760	**0.840**	0.735	0.230	0.667	**0.862**	0.576	0.068	0.632	0.547	0.846	0.112	0.607	0.576	0.698	0.667
R-Tuning	-0.117	0.477	0.437	**0.955**	0.268	0.704	0.675	0.796	-0.127	0.437	0.437	**1.000**	-	-	-	-	0.539
SVAR	0.005	0.599	0.514	0.619	0.129	0.566	0.570	0.514	0.003	0.567	**1.000**	0.565	0.009	0.505	0.505	**1.000**	0.559
Context Lens	-0.123	0.471	0.429	0.734	0.101	0.537	0.557	0.389	-0.025	0.539	0.423	0.561	0.017	0.512	0.508	0.733	0.515
Concat (Attn.)	-0.016	0.578	0.490	0.928	0.300	0.737	0.784	0.684	0.146	0.709	0.658	0.753	-	-	-	-	0.674
Concat (Hid.)	0.153	0.747	0.669	0.810	0.281	0.717	0.705	0.743	0.046	0.610	0.529	0.924	0.108	0.604	0.573	0.689	0.669
Visual (Attn.)	0.187	0.781	0.769	0.788	0.272	0.708	0.731	0.676	**0.158**	0.721	0.669	0.767	-	-	-	-	**0.737**
Ensemble (Attn.)	0.140	0.734	0.724	0.739	0.268	0.704	0.705	0.703	0.149	0.712	0.718	0.709	-	-	-	-	0.717
Ensemble (Hid.)	**0.203**	**0.797**	0.792	0.800	**0.311**	**0.748**	0.728	0.792	0.152	0.715	0.637	0.813	**0.152**	**0.647**	**0.669**	0.629	0.727

Table 1. We evaluate the abstention performance of eight methods and our proposed method on four STVQA datasets across image and video modalities. Best results are in **bold**, second best are in underline, incompatible or unavailable results are denoted as "-" and our methods are in blue background .. "Avg" is the micro-average of abstention accuracy of four datasets. Too high reliable accuracy or abstention precision usually indicates over-abstention or under-abstention, while effective reliability and abstention accuracy are comprehensive metrics considering both over abstention or under-abstention. Ensemble probe on hidden states yield the best average abstention accuracy while visual-focused probe on attention pattern yield the second best.

- **Baseline methods fail for various reasons.** Among training-free methods, self-consistency achieves the best performance with 68.0% average abstention accuracy. However, its high reliable accuracy paired with low abstention precision on Qwen2.5-VL-7B reveals a tendency to over-abstain due to threshold sensitivity. Conversely, prompt-to-abstain exhibits severe under-abstention with extremely high abstention precision, indicating that small VLMs lack self-reflection capabilities. Among training-based methods, LRP surpasses VLM Judge in 7 of 8 (model, dataset) combinations, demonstrating that training lightweight probes on latent representations is more effective and efficient than end-to-end fine-tuning. R-tuning's failure in 2 of 6 combinations further reveals its sensitivity to dataset composition (ratio of certain vs. uncertain examples) and training instability.
- **LRP is the only abstention method working for video modality.** On video datasets, LRP is the only method providing meaningful abstention performance, while all baseline methods achieve an average effective reliability of merely 0.051. This advantage arises because video inputs span hundreds of frames, but latent representations compress this information into compact embeddings, making LRP naturally more robust to long sequences compared to baselines.

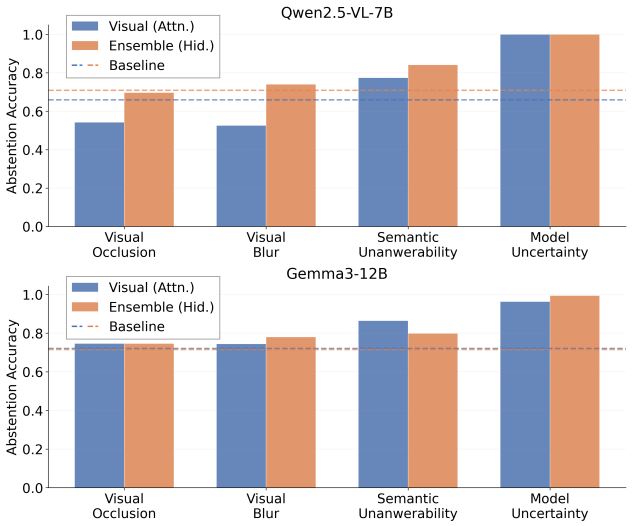

Figure 2. Abstention accuracy evaluated by probes trained on HIERTEXT. Baseline represents probes' abstention accuracy on original HIERTEXT dataset. LRP demonstrates high generalization capability to various unconfident scenarios, indicating it learns general confidence representation.

6. Analysis

6.1. Various Abstention

To verify that our probes learn general uncertainty signals, we create four unseen synthetic dataset from HierText, and measure the zero-shot prediction accuracy of probes trained on HIERTEXT. Each dataset simulates a distinct uncertainty source described in [38]: (1) **Visual occlusion**: masking text regions with white boxes; (2) **Visual blur**: adding Gaussian noise to the image ($\sigma = 5$); (3) **Semantic unanwerability**: using GEMINI-2.5-PRO to replace a few words in the questions such that they ask about non-existent text or regions; (4) **Model uncertainty**: prepending a random low-confidence prefix (e.g., "I'm not sure") to model outputs before generation. The intuition behind is: if probes decode genuine uncertainty signals from latent representations, they should generalize across different uncertainty sources, as these perturbations all fundamentally affect model confidence despite having distinct forms.

Results in Figure 2 shows that visual-focused attention probes achieve higher abstention accuracy than HIERTEXT where the probes are trained on in 6 out of 8 (model, dataset) combinations, while ensemble hidden state probes exceed in 7 out of 8 combinations. This cross-uncertainty generalization indicates that probes capture fundamental confidence signals rather than dataset-specific patterns. Furthermore, these results confirm that uncertainty signals are indeed encoded in the model's internal representations and can be effectively extracted through probing.

Original Dataset	Method	OCR	SEED	HIER	EGO	Avg
QWEN2.5-VL-7B						
OCR	Self Consist.	0.677	0.623	0.625	0.644	0.631
	VLM Judge	0.768	**0.656**	0.632	0.628	0.639
	Visual (Attn.)	0.768	0.647	0.613	0.677	0.646
	Ensemble (Hid.)	**0.789**	0.654	**0.669**	**0.680**	**0.668**
SEED	Self Consist.	**0.709**	0.636	**0.669**	0.662	**0.680**
	VLM Judge	0.601	0.737	0.551	0.632	0.595
	Visual (Attn.)	0.553	0.735	0.604	**0.645**	0.601
	Ensemble (Hid.)	0.501	**0.763**	0.563	0.547	0.537
HIER	Self Consist.	0.677	0.623	0.625	0.644	**0.648**
	VLM Judge	0.672	**0.645**	**0.749**	0.628	**0.648**
	Visual (Attn.)	0.676	0.579	0.659	**0.668**	0.641
	Ensemble (Hid.)	**0.680**	0.568	0.709	0.645	0.631
EGO	Self Consist.	**0.677**	0.623	**0.625**	0.644	**0.642**
	VLM Judge	0.619	**0.715**	0.560	0.634	0.632
	Visual (Attn.)	0.623	0.689	0.604	0.722	0.639
	Ensemble (Hid.)	0.587	0.658	**0.625**	**0.738**	0.623
GEMMA3-12B						
OCR	Self Consist.	0.743	**0.726**	**0.771**	0.587	0.694
	VLM Judge	0.760	0.607	0.690	0.615	0.637
	Visual (Attn.)	0.781	0.684	0.721	-	**0.703**
	Ensemble (Hid.)	**0.797**	0.697	0.697	**0.622**	0.672
SEED	Self Consist.	0.720	0.735	**0.724**	0.610	**0.685**
	VLM Judge	**0.733**	0.667	0.666	0.619	0.673
	Visual (Attn.)	0.536	0.708	0.613	-	0.575
	Ensemble (Hid.)	0.695	**0.748**	0.508	**0.629**	0.611
HIER	Self Consist.	**0.743**	**0.726**	**0.771**	0.587	**0.685**
	VLM Judge	0.657	0.623	0.632	**0.613**	0.631
	Visual (Attn.)	0.692	0.601	0.721	-	0.647
	Ensemble (Hid.)	0.633	0.583	0.715	0.542	0.586
EGO	Self Consist.	**0.741**	**0.741**	**0.737**	**0.608**	**0.740**
	VLM Judge	0.543	0.607	0.495	0.607	0.549
	Visual (Attn.)	-	-	-	-	-
	Ensemble (Hid.)	0.590	0.590	0.551	0.647	0.577

Table 2. Cross-dataset abstention accuracy. Each row shows results when training on the original dataset (left) and testing on different datasets (columns). "OCR", "Seed", "Hier", "Ego" are short for OCRBENCH V2, SEEDBENCH PLUS 2, HIERTEXT and EGOTEXTVQA respectively. Best results are in **bold**, our methods are in blue background. "Avg" is the average abstention accuracy on unseen datasets (self-excluded). Self-consistency, a training-free abstention mechanism shows the best overall generalization capability, while visual-focused attention pattern probe generalizes best among training-based methods.

6.2. Generalization

Prior work has noted that training-based abstention methods often suffer from limited generalization to unseen dis-

tributions [12]. To evaluate this concern, we conduct cross-dataset evaluation by training probes on one dataset and testing on another, comparing our approach against the best-performing training-based baseline (VLM Judge) and threshold-based baseline (Self-Consistency) from Table 1.

Results in Table 2 show that self-consistency demonstrates better generalization across datasets. Among training-based methods, visual-focused attention probe achieves the best generalization with 0.636 average abstention accuracy, outperforming best training baseline VLM Judge. Attention probes generalize better than hidden state probes due to their simpler, more universal representations.

However, training data diversity significantly impacts generalization. When trained on OCRBench v2, the most comprehensive dataset, visual-focused attention probe achieves 0.675 average cross-dataset abstention accuracy, surpassing self-consistency by 1.8% relative improvement. This demonstrates that with sufficient training data covering diverse uncertainty patterns, learned probes can achieve strong generalization capabilities.

6.3. Optimal Layer

To understand why latent representations encode richer confidence information, we analyze how confidence signals evolve across transformer layers. We plot the abstention accuracy of probes trained on each layer's hidden states across four datasets in Figure 3.

We observe two key patterns. For SeedBench Plus 2, HierText, and EgoTextVQA, abstention accuracy exhibits a clear rise-then-fall trend, peaking around layers 15-20 for Qwen2.5-VL-7B and around layers 20-30 for Gemma3-12B. For OCRBench v2, accuracy remains consistently high across most layers after initial rise. Notably, Gemma3-12B shows noisier patterns with confidence peaks appearing in shallower layers compared to Qwen2.5-VL-7B, which may reflect insufficient training or over-aggressive alignment given its lower performance.

The rise-then-fall pattern confirms that confidence signals are strongest in intermediate layers, validating prior findings that alignment processes degrade calibration near output layers [19, 42]. However, OCRBench v2's consistent performance demonstrates that with sufficient training data diversity, probes can extract effective confidence signals even from later layers where signals are weaker. This reduces the pressure of layer selection when adequate training data is available.

7. Conclusion

In this work, we address a critical gap in VLM deployment: the ability to abstain when uncertain in Scene Text Visual Question Answering tasks. We propose Latent Representation Probing, a lightweight approach that trains probes on latent representations across transformer layers to detect

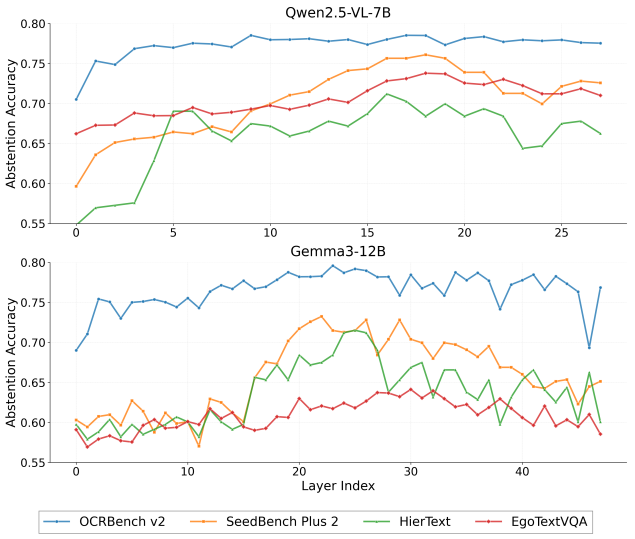

Figure 3. The abstention accuracy of probes trained on each layer's hidden states across four datasets. For SeedBench, HierText and EgoTextVQA, the abstention accuracy first increases and then decreases, indicating that confidence signals are richer in the intermediate instead of final layers.

low confidence. We provide the first comprehensive evaluation of state-of-the-art abstention mechanisms on VLMs. Experiments demonstrate that LRP substantially outperforms baselines by 7.6% in abstention accuracy, while requiring only a single forward pass. Our analysis reveals three key insights: (1) LRP learn generalizable uncertainty representations that transfer across distinct unconfident scenarios, achieving 0.825 average abstention accuracy; (2) LRP exhibits superior cross-dataset generalization, outperforming training-based baselines by 5.8% in abstention accuracy; (3) optimal confidence signals emerge in intermediate layers rather than final layers. Our paper reveals a broader principle: model confidence is not merely an output property but emerges progressively through internal computation, offering a principled path toward deployment-ready AI systems.

References

[1] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M Paixao, Filipe Mutz, et al. Self-driving cars: A survey. *Expert systems with applications*, 165:113816, 2021. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In

Proceedings of the IEEE/CVF international conference on computer vision, pages 4291–4301, 2019. 3

[4] Sungguk Cha, Jusung Lee, Younghyun Lee, and Cheoljong Yang. Visually dehallucinative instruction generation: Know what you don't know. *arXiv preprint arXiv:2402.09717*, 2024. 3

[5] Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023. 1

[6] Zijun Chen, Wenbo Hu, Guande He, Zhijie Deng, Zheng Zhang, and Richang Hong. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. *arXiv preprint arXiv:2412.14660*, 2024. 3

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5

[8] Oscar Danielsson, Magnus Holm, and Anna Syberfeldt. Augmented reality smart glasses in industrial assembly: Current status and future challenges. *Journal of Industrial Information Integration*, 20:100175, 2020. 1

[9] Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*, 2024. 1

[10] Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. *arXiv preprint arXiv:2503.10602*, 2025. 3

[11] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. 1, 3

[12] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024. 1, 3, 5, 8

[13] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024. 5, 1

[14] Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*, 2023. 3

[15] Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. *arXiv preprint arXiv:2405.02917*, 2024. 3

[16] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019. 1

[17] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025. 3, 4

[18] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. 3

[19] Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024. 2, 8

[20] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 3, 5, 1

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[22] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022. 1

[23] Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*, 2025. 3

[24] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. 1, 3

[25] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2022. 5, 1

[26] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024. 1

[27] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022. 3

[28] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 3

[29] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 3

[30] Sankalp Nagaonkar, Augustya Sharma, Ashish Choithani, and Ashutosh Trivedi. Benchmarking vision-language models on optical character recognition in dynamic video environments. *arXiv preprint arXiv:2502.06445*, 2025. 1

[31] Anirudh Phukan, Harshit Kumar Morj, Apoorv Saxena, Koustava Goswami, et al. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms. *arXiv preprint arXiv:2411.19187*, 2024. 3, 4

[32] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3

[33] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 5

[34] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023. 3

[35] Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. *arXiv preprint arXiv:2402.07417*, 2024. 2

[36] Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. Analysing the robustness of vision-language-models to common corruptions. *arXiv preprint arXiv:2504.13690*, 2025. 1

[37] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 1, 3

[38] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13: 529–556, 2025. 1, 5, 7

[39] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. 1

[40] Philip Wootaek Shin, Jack Sampson, Vijaykrishnan Narayanan, Andres Marquez, and Mahantesh Halappanavar. Losing the plot: How vlm responses degrade on imperfect charts. *arXiv e-prints*, pages arXiv–2509, 2025. 1

[41] Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. *arXiv preprint arXiv:2504.13169*, 2025. 3

[42] Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. *arXiv preprint arXiv:2310.00492*, 2023. 2, 8

[43] Zeguan Xiao, Diyang Dou, Boya Xiong, Yun Chen, and Guanhua Chen. Enhancing uncertainty estimation in llms with expectation of aggregated internal belief. *arXiv preprint arXiv:2509.01564*, 2025. 2

[44] Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. Do language models mirror human confidence? exploring psychological insights to address overconfidence in llms. *arXiv preprint arXiv:2506.00582*, 2025. 1

[45] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2025. 3

[46] Jihan Yao, Wenxuan Ding, Shangbin Feng, Lucy Lu Wang, and Yulia Tsvetkov. Varying shades of wrong: Aligning llms with wrong answers only. *arXiv preprint arXiv:2410.11055*, 2024. 1, 3

[47] Jihan Yao, Yushi Hu, Yujie Yi, Bin Han, Shangbin Feng, Guang Yang, Bingbing Wen, Ranjay Krishna, Lucy Lu Wang, Yulia Tsvetkov, et al. Mmmg: a comprehensive and reliable evaluation suite for multitask multimodal generation. *arXiv preprint arXiv:2505.17613*, 2025. 3, 1

[48] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 1

[49] Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'i don't know'. 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 7106–7132, 2024. 3, 4

[50] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024. 3

[51] Yudong Zhang, Ruobing Xie, Xingwu Sun, Yiqing Huang, Jiansheng Chen, Zhanhui Kang, Di Wang, and Yu Wang. Dhcp: Detecting hallucinations by cross-modal attention pattern in large vision-language models. *arXiv preprint arXiv:2411.18659*, 2024. 3

[52] Minyi Zhao, Bingjia Li, Jie Wang, Wanqing Li, Wenjing Zhou, Lan Zhang, Shijie Xuyang, Zhihang Yu, Xinkun Yu, Guangze Li, et al. Towards video text visual question answering: Benchmark and baseline. *Advances in Neural Information Processing Systems*, 35:35549–35562, 2022. 3

[53] Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3363–3373, 2025. 3, 5, 1

Reading Between the Lines: Abstaining from VLM-Generated OCR Errors via Latent Representation Probes

Supplementary Material

8. Experiment Details

Model Details We employ 2 LLMs in the experiments, the detailed information of LLMs we used in this paper is as follows: 1) QWEN2.5-VL-7B, through the QWEN/QWEN2.5-VL-7B-INSTRUCT checkpoint on Huggingface [39]; 2) GEMM3-12B, through the GOOGLE/GEMMA-3-12B-IT checkpoint on Huggingface.

Dataset Details We employ 4 datasets in the experiments. The detailed information is:

- **OCRBench v2 (OCR)** [13]. We collect the dataset through the LMMS-LAB/OCRBENCH-V2 checkpoint on Huggingface. We randomly split the official test split (10,000 questions) to train, validation and test subsets with 6:2:2 ratio.
- **SeedBench Plus 2 (Seed)** [20]. We collect the dataset through the AILAB-CVC/SEED-BENCH-2-PLUS checkpoint on Huggingface. We randomly split the official test split (2,277 questions) to train, validation and test subsets with 6:2:2 ratio.
- **HierText (Hier)** [25]. We collect the dataset through the GOOGLE-RESEARCH-DATASETS/HIERTEXT repository on Github. We randomly split the official test split (1,612 questions) to train, validation and test subsets with 6:2:2 ratio. For each image, GEMINI-2.5-PRO will first randomly choose from paragraph, line or word perspective to synthesize the question. Then it will generate a QA pair with reference to the ground-truth OCR annotation results. Last, it will replace a few words in the original to make up an unanswerable question which ask about a non-existing region or text. The full prompt we use for generating the questions are in Table 3:
- **EgoTextVQA (Ego)** [53]. We collect the dataset through the ZHOUSHENG97/EGOTEXTVQA repository on Github. We randomly split the official outdoor split (4,848 questions) to train, validation and test subsets with 6:2:2 ratio. The indoor split is not inlcuded due to comptuttaion limitation.

Inference and Training Details For all experiments, we use greedy decoding (temperature = 0.0 and top_p == 1.0) and set max_new_tokens=128. For evaluation, we first utilize GEMINI 2.5 PRO to parse the responses into containing only the answer part to get rid of the possible redundant introduction like "Sure! The answer is" and then leverage the official evaluation suite get the response accuracy as labels.

We use the standard SFTTrainer in Huggingface for VLM fine-tuning. We use grid search on learning rate, weight decay and learning rate scheduler and early stop mechanism to find the best hyperparameter for probe-based methods. [47]

Generate TWO questions based on this image and ground-truth OCR results:

1. **ANSWERABLE QUESTION**: A question that CAN be answered using the visible text in the image.
   - The question must reference text location using spatial descriptions. Be diversified and creative about the spatial descriptions (including but not limited to, "the paragraph at the top-left", "the second line in the middle section", "the word next to the logo", etc.).
   - The answer must be exactly one OCR result (paragraph, line, word) from the ground-truth OCR results with the exact corresponding vertices.
   - Make sure the answer is unique and unambiguous from the question with enough constraints.
2. **UNANSWERABLE QUESTION**: A question that CANNOT be answered from the visible text alone.
   - Created by simply replacing a few words from the answerable question.
   - The question is unanswerable because it clearly references an invalid text location (region with no text or doesn't exist) in the image.

Output Format:

```
{
  "answerable_question": {
    "question": "Your question here",
    "answer": "Ground-truth answer here",
    "vertices": [[x1, y1], [x2, y2], ..., [xn, yn]]
  },
  "unanswerable_question": {
    "question": "Your question here"
  }
}
```

Table 3. System prompt for synthesizing HIERTEXT questions.