

Leveraging Visual Signals for Robust Token-Level Uncertainty in Vision–Language Generation

Joseph Hoche¹ David Brellmann² Gianni Franchi^{1,3}

¹AMIAD, Pôle Recherche, Palaiseau ²Safran Tech ³ENSTA Paris

Abstract

*Uncertainty Quantification (UQ) remains a critical challenge in Large Vision Language Models (LVLMs) for reliable predictions and real-world deployment. However, most existing methods are adapted from the LLM literature and primarily focus on the language modality, leaving the contribution of visual information to LVLM uncertainty largely underexplored. In this paper, we investigate how LVLMs process visual information and whether this process can be used to improve uncertainty estimation. By analyzing hidden representations after the integration of visual features during the generation process, we observe that high-confidence predictions rely more heavily on visual content than uncertain ones. Building on this insight, we propose **Visual-Grounded Token UQ (VIG-TUQ)**, a training-free framework that explicitly incorporates visual grounding into uncertainty estimation by weighting token-level language uncertainty with visual grounding scores. We evaluate VIG-TUQ on multiple datasets and across diverse LVLM architectures, including early-fusion, late-fusion, and native-fusion models. Results indicate that our method often improves upon existing token-level uncertainty approaches. Code and data will be made available upon acceptance.*

1. Introduction

Large Vision-Language Models (LVLMs) extend Large Language Models (LLMs) with visual perception and excel across a wide spectrum of multimodal tasks requiring visual and language understanding [17, 20, 54–56]. Despite their impressive abilities, LVLMs are not yet reliable. In particular, they remain prone to hallucinations, i.e., they may generate incorrect, incomplete, fabricated, or misleading content that appears plausible [26, 29, 47]. Such behavior limits their use in safety-critical applications like autonomous driving [16], robotics [22] or medicine [31]. A promising direction for improving deployment reliability is uncertainty quantification (UQ) to detect when generated

content may be unreliable [2, 15, 24]. Prior works show promising results for UQ methods in the language setting for LLMs [8, 11, 53]. Although these approaches can be extended to LVLMs [38, 62], their extension is not straightforward. LVLMs operate across multiple modalities and their uncertainty may stem from both linguistic and visual sources. Recent studies have highlighted the tendency of LVLMs to prioritize internal textual knowledge over external visual information [14, 41, 48, 49]. For instance, given an image depicting a blue orange and the question “*what color is the orange?*”, an LVLM may confidently answer “*orange*”, basing its answer on its parametric knowledge and language prior rather than on the actual visual input. This mismatch motivates the importance of developing specific approaches designed for LVLMs that explicitly capture and quantify the contribution of visual information to LVLM predictions.

In this paper, we investigate how LVLM process visual information during next-token prediction and whether this process can be leveraged to improve token-level uncertainty estimation. By analyzing hidden representations during generation, we observe that high-confidence predictions rely more heavily on visual content than uncertain ones (Figure 1). Building on this insight, we propose **Visual-Grounding Token Uncertainty Quantification (VIG-TUQ)**, a training-free uncertainty quantification framework for LVLMs. VIG-TUQ explicitly incorporates visual grounding into uncertainty estimation by weighting token-level language uncertainty with visual grounding scores (Figure 2). Specifically, VIG-TUQ quantifies the influence of visual content on next-token prediction using two complementary visual grounding strategies: (1) comparing next-token distributions with and without visual input, and (2) leveraging visual attention weights over image tokens. Our central hypothesis is that token-level uncertainties should not necessarily contribute equally to the uncertainty of the full generated sequence. Visual grounding is therefore used as a token-importance signal: it determines how strongly each token-level uncertainty contributes to the final sequence-level score, rather than acting as an uncertainty measure by itself. Our experimental results support

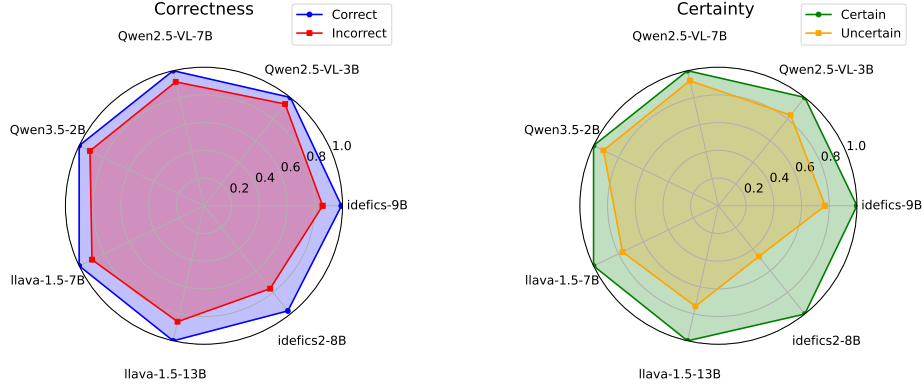


Figure 1. **Correct / confident predictions depend more strongly on visual information than incorrect / uncertain ones.** Radial values indicate the cosine distance between hidden representations from two forward passes: one with visual input and one without. Results are averaged on the **OKVQA** dataset [52]. The cosine distance measures the model reliance on visual information, where larger values indicate a stronger influence of the image on the internal representation.

this hypothesis and highlight the importance of combining multiple visual grounding signals, as their effectiveness can be model-dependent and capture distinct aspects of visual reasoning. Finally, our approach provides a fine-grained into which specific spans or entities contribute to the uncertainty, while remaining computationally efficient by avoiding reliance on external models or sampling.

Contributions. This work makes the following key contributions: (1) We observe that LVLMs tend to be more confident when their predictions rely more strongly on visual input. (2) We introduce a new training-free uncertainty quantification framework that combines different complementary visual grounding scores to weight token-level uncertainty measures. We evaluate VIG-TUQ on multiple datasets and across diverse LVLM architectures, including early-fusion, late-fusion, and native-fusion models. Results show that our method often improves upon existing token-level uncertainty approaches and remain computationally efficient due to the absence of external models or sampling. (3) We analyze whether uncertainty is uniformly distributed across generated tokens or concentrated within a subset of informative tokens. Our findings indicate that visual information helps identify the tokens most relevant for uncertainty estimation, enabling more effective token selection in LVLMs for uncertainty estimation.

2. Related work

Several recent methods have been proposed to enhance reliability by estimating uncertainty in the predictions of LLMs and LVLMs. These methods are broadly categorized as follows [34]: (1) logit-based methods, which directly use token probabilities/logits as uncertainty scores [13, 51]; (2) semantic-based methods, which quantifies uncertainty by assessing the semantic consistency of multiple generated answers [11, 18, 32, 36, 38, 44, 62]; (3) verbalized-based

methods, which estimate uncertainty by directly querying the LLM about the reliability of its own generations [33, 60, 61]; and (4) internal-state-based methods, which analyze internal signals from LLMs after a single generation to estimate uncertainty, such as features at intermediate layers, attention weights, or other latent representations [5, 35, 58]. Methods listed above may rely on external models (e.g., sentence embedding models [1, 18], NLI models [10, 53], instruction-tuned LLMs [11, 30]) or involve training dedicated models to estimate uncertainty [12, 25, 58]. In this paper, we propose a training-free hybrid approach to uncertainty estimation that combines a logit-based component to capture linguistic uncertainty with an internal-state-based component to incorporate visual uncertainty. By explicitly leveraging visual information while maintaining computational efficiency with a single-pass and token-level estimation, we introduce a multimodal-aware approach designed for LVLMs. Additional technical details and relevant results from prior work are discussed throughout the paper.

3. Preliminaries

Notations. Let $\mathcal{X}_v$ denote the visual input space, $\mathcal{X}_l$ the linguistic input space, and $\mathcal{V}$ the LVLM vocabulary. A multimodal input is depicted as $\mathbf{x} = (\mathbf{x}_v, \mathbf{x}_l)$, where the visual component $\mathbf{x}_v \in \mathcal{X}_v$ consists of a sequence of visual patches and the textual component $\mathbf{x}_l \in \mathcal{X}_l$ consists of a sequence of textual tokens.

An LVLM defines a parameterized function $f_\theta : \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathbb{R}^d$ to encode the visual-linguistic context into a latent representation. In practice, $f_\theta(\cdot, \cdot)$ typically includes a transformer decoder architecture, potentially combined with separate encoders for each modality (see Section 3.1). For a given multimodal input $\mathbf{x} = (\mathbf{x}_v, \mathbf{x}_l)$, the model then autoregressively generates

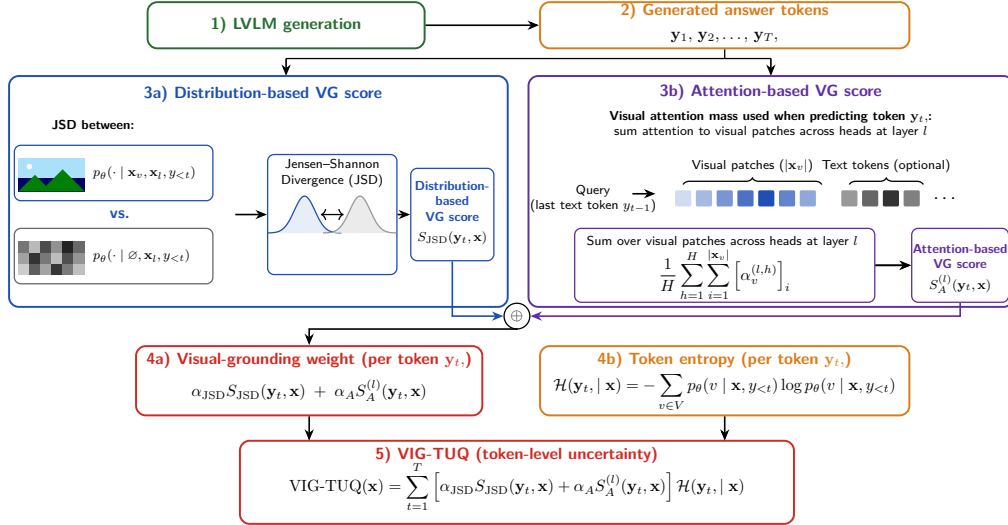


Figure 2. **Overview of the VIG-TUQ pipeline.** VIG-TUQ weights token-level language uncertainty with visual grounding scores using two complementary strategies: a distribution-based score obtained from the Jensen–Shannon divergence between predictions with and without the image, and an attention-based score measuring the attention mass assigned to visual patches.

a token sequence $\mathbf{y} = [y_t]_{t=1}^T$, where each token y_t is sampled according to:

$$p_{\theta}(y_t | \mathbf{x}, \mathbf{y}_{<t}) = \text{softmax}\left(f_h(f_{\theta}(x_v, [x_l | \mathbf{y}_{<t}]))\right), \quad (1)$$

where $f_h : \mathbb{R}^d \rightarrow \mathbb{R}^{|\mathcal{V}|}$ denotes the output head that projects the latent representation to logits over the vocabulary $\mathcal{V}$ and $\mathbf{y}_{<t} = [y_1, \dots, y_{t-1}]$ depicts the previously generated tokens.

3.1. Modality fusion strategies

Given an input $\mathbf{x} = (x_v, x_l)$, an LVL can be described by the approach it uses to combine information from the visual modality $x_v \in \mathcal{X}_v$ and the textual modality $x_l \in \mathcal{X}_l$ within the parameterized function $f_{\theta} : \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathbb{R}^d$.

Native fusion. In native fusion, the modalities are directly combined into a single representation before being processed by the transformer decoder architecture. In particular, the parametric function $f_{\theta} : \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathbb{R}^d$ is expressed as a composition $f_{\theta} = f_L \circ f_{L-1} \circ \dots \circ f_1 \circ f_0$, where $f_0 : \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathbb{R}^d$ is the input function and $\{f_i : \mathbb{R}^d \rightarrow \mathbb{R}^d\}_{i=1}^L$ are L decoder transformer layers. The input function $f_0(\cdot, \cdot)$ processes image patches and text tokens independently as:

$$f_0(x_v, x_l) = [E_v(x_v) | E_l(x_l)], \quad (2)$$

where $E_v(\cdot)$ and $E_l(\cdot)$ map visual patches and language tokens to their embeddings via linear transformations. In these architectures, the model is trained end-to-end from scratch to learn multimodal representations natively within the transformer blocks. LVL like Qwen models [59] adopts a native fusion strategy.

Early fusion. In early fusion, features from each modality are first extracted separately using dedicated encoders, and then integrated into a shared representation before being processed by the transformer decoder architecture. In particular, the parametric function $f_{\theta} : \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathbb{R}^d$ is defined as a composition $f_{\theta} = f_L \circ f_{L-1} \circ \dots \circ f_1 \circ f_0$, where $\{f_i : \mathbb{R}^d \rightarrow \mathbb{R}^d\}_{i=1}^L$ are L decoder transformer layers and $f_0 : \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathbb{R}^d$ is the input function defined as:

$$f_0(x_v, x_l) = m_0(E_{\theta_v}(x_v), E_l(x_l)); \quad (3)$$

with $m_0(\cdot, \cdot)$ a multimodal fusion module (e.g., concatenation, linear layer, cross-attention layer); $E_{\theta_v}(\cdot)$ a visual encoder with parameters θ_v (e.g., ViT, CNNs); and $E_l(\cdot)$ a text embedding function. Unlike native fusion, the visual encoder $E_{\theta_v}(\cdot)$ is a pre-trained model that can be optimized independently. LVLs like LLaVA [45] and Qwen-VL [6] adopt an early fusion approach. Further details on early fusion, fusion modules, and unimodal encoders can be found in [4, 43].

Late fusion. Late fusion also starts with the extraction of unimodal features using dedicated encoders. However, unlike early fusion, the integration between modalities occurs within the intermediate representations of the transformer decoder architecture rather than before them. In particular, the parametric function $f_{\theta} : \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathbb{R}^d$ can be defined recursively over L layers as

$$\begin{aligned} f_{\theta}(x_v, x_l) &= f_L(h_{L-1}, x_v), \quad \text{where} \\ h_i &= f_i(h_{i-1}, x_v) = g_i(m_i(E_{\theta_v^{(i)}}(x_v), h_{i-1})) \\ (1 \leq i \leq L) h_0 &= E_l(x_l); \end{aligned} \quad (4)$$

with $E_l(\cdot)$ a text embedding function. For each layer i , $g_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$ depicts a transformer decoder layer, $m_i(\cdot, \cdot)$ denotes a multimodal fusion module (e.g., concatenation, linear layer, cross-attention layer), and $E_{\theta_v^{(i)}}(\cdot)$ is a visual encoder with parameters $\theta_v^{(i)}$ (e.g., ViT, CNNs). Note that the vision modality is not necessarily integrated at every layer and the fusion module can simply preserve the hidden representation, i.e., $m_i(E_{\theta_v^{(i)}}(x_v), h_i) = h_i$. Similarly to early fusion, visual encoders $E_{\theta_v^{(i)}}(\cdot)$ are pre-trained models that can be optimized independently. LVLMs like IDEFICS [40] and Flamingo [3] adopt such approach. Further details on late fusion, fusion modules, and unimodal encoders can be found in [4, 43].

3.2. Token-level UQ methods

Given a multimodal input x , token-level UQ methods assign an uncertainty score s_t to each generated token y_t (equation 1), with a higher score indicating greater uncertainty. The uncertainty score $s(\cdot)$ for a generated sequence $y = [y_t]_{t=1}^T$ can be derived by aggregating the token-level scores of each token y_t , e.g., by averaging across tokens $s(y) = \frac{1}{T} \sum_{t=1}^T s_t$ or by taking their maximum $s(y) = \max_t s_t$.

Negative Log-Likelihood (Log-Perplexity). A straightforward token-level uncertainty score is the token negative log-likelihood [13, 50]:

$$\text{NLL}(y_t, x, y_{<t}) = -\log p_{\theta}(y_t | x, y_{<t}). \quad (5)$$

Token Entropy. Another common token-level score is the token entropy of the autoregressive distribution [13, 50]:

$$\mathcal{H}(y | x, y_{<t}) = -\sum_{y \in \mathcal{V}} p_{\theta}(y | x, y_{<t}) \log p_{\theta}(y | x, y_{<t}). \quad (6)$$

Claim-Conditioned Probability (CCP). CCP [10] estimates token-level uncertainty by measuring how much the top K token alternatives $\{y_t^k\}_{k=1}^K$ contradict the generated token y_t using a Natural Language Inference (NLI) model:

$$\text{CCP}(y_t, x, y_{<t}) = -\log \frac{\sum_{k: \text{NLI}(y_{<t}, y_t^k, y_t) = \text{entail}} p_{\theta}(y_t^k | x, y_{<t})}{\sum_{k: \text{NLI}(y_{<t}, y_t^k, y_t) \in \{\text{entail}, \text{contradict}\}} p_{\theta}(y_t^k | x, y_{<t})}, \quad (7)$$

where $\text{NLI}(\cdot)$ determines whether concatenating y_t^k with the preceding generated context $y_{<t}$ entails (semantic equivalence) or contradicts (different meaning) the generated content $y_{<t} \cup y_t$.

Token-Shifting Attention to more Relevant (Token-SAR). Token-SAR [9] defines a token-level uncertainty score by reweighting token likelihood with a semantic relevance factor as:

$$\text{Token-SAR}(y_t, x, y_{<t}) = -\log p_{\theta}(y_t | x, y_{<t}) \tilde{R}(x, y, y_t), \quad (8)$$

where $\tilde{R}(x, y, y_t) = R(x, y, y_t) / \sum_{i=1}^T R(x, y, y_i)$ is the normalized relevance score with

$$R(x, y, y_t) = 1 - |g(x \cup y, x \cup y \setminus \{y_t\})|$$

the relevance that quantifies how important y_t is in reflecting the semantics of y by comparing the semantic change before and after removing this token and $g(\cdot, \cdot)$ a similarity between two inputs on a scale of -1 to 1 .

4. Method

In this section, we introduce VIsual-Grounding Token Uncertainty Quantification (VIG-TUQ), a training-free UQ framework for LVLMs that weights token-level language uncertainty with visual grounding scores. We first discuss our motivations (Section 4.1), then describe the visual grounding score used (Section 4.2), and finally introduce the proposed method (Section 4.3).

4.1. Motivations

A recent line of works highlights the tendency of LVLMs to prioritize internal textual knowledge over external visual information [14, 41, 48, 49]. This suggests that visual grounding could play a crucial role in identifying unreliable responses. To investigate this phenomenon and better understand the contribution of visual information to token generations, we conduct a pilot study on how LVLMs process visual information. Specifically, we compare hidden representations (via the cosine distance) obtained from two independent forward passes: one with visual evidence and one without it. This analysis is performed on generated sentences from the OKVQA dataset [52] across different LVLm architectures [48] using greedy decoding with visual evidence (further details are provided in Appendix F). Figure 1 summarizes our results and highlights that correct and confident predictions rely more on the visual content than incorrect and uncertain ones. Furthermore, experiments reported in Appendix C and Appendix F indicate that the extent to which visual information is processed depends on both the training setup and the model architecture (see Section 3). These observations motivate the introduction of multiple visual grounding scores to capture different aspects of visual understanding. Overall, our results suggest that tokens whose generation strongly depends on visual input are more informative and should therefore be assigned greater weight in uncertainty estimation. Building on this insight, we introduce different visual grounding token-level

strategies and our VIG-TUQ approach in the following sections.

4.2. Visual grounding scores

We propose two scoring strategies to measure how much an LVLM relies on its visual input when generating a token.

Distribution-based scores. Given a multimodal input $\mathbf{x} = (\mathbf{x}_v, \mathbf{x}_l)$, with visual content $\mathbf{x}_v$ and textual content $\mathbf{x}_l$, this approach quantifies the contribution of the visual input $\mathbf{x}_v$ by measuring how much the next-token distribution changes when the visual information is removed. In particular, for each generated token $\mathbf{y}_t$, we compare the original LVLM next-token distribution $p_\theta(\cdot \mid \mathbf{x}_v, \mathbf{x}_l, \mathbf{y}_{<t})$ with the next-token distribution $p_\theta(\cdot \mid \emptyset, \mathbf{x}_l, \mathbf{y}_{<t})$ when the visual information $\mathbf{x}_v$ is removed using the Jensen-Shannon divergence:

$$d_t(\mathbf{y}, \mathbf{x}) = \text{JSD}(p_\theta(\cdot \mid \mathbf{x}_v, \mathbf{x}_l, \mathbf{y}_{<t}) \parallel p_\theta(\cdot \mid \emptyset, \mathbf{x}_l, \mathbf{y}_{<t})). \quad (9)$$

A larger value of $d_t(\mathbf{y}, \mathbf{x})$ indicates that removing the visual input induces a stronger distributional shift when generating $\mathbf{y}_t$ and thus a greater reliance on visual information. To obtain a relative token-level importance score, we normalize the divergence values across all generated tokens: $S_{\text{JSD}}(\mathbf{y}_t, \mathbf{x})$, where

$$S_{\text{JSD}}(\mathbf{y}_t, \mathbf{x}) = \frac{d_t(\mathbf{y}_t, \mathbf{x})}{\sum_{i=1}^T d_t(\mathbf{y}_i, \mathbf{x})}. \quad (10)$$

This normalized score captures the relative contribution of the visual input to each generated token $\mathbf{y}_t$ with respect to the full sequence.

Attention-based Scores. The attention mechanism provides another way to estimate how much the visual content contributes to the next-token prediction. For instance, given a multimodal input $\mathbf{x} = (\mathbf{x}_v, \mathbf{x}_l)$ with a visual content $\mathbf{x}_v$ and a textual content $\mathbf{x}_l$, we could use the attention weight vector used for predicting the next token in the l -th layer and the h -th attention head that interacts with the visual content $\mathbf{x}_v$:

$$\alpha_t^{(l,h)} = \text{softmax}\left(\frac{\mathbf{K}_t^{(l,h)} \mathbf{q}_t^{(l,h)}}{\sqrt{d_k}}\right) \in [0, 1]^N, \quad (11)$$

where $\mathbf{q}^{(l,h)} \in \mathbb{R}^{d_k}$ is the query vector corresponding to the last textual token in $\mathbf{x}_l$; $\mathbf{K}^{(l,h)} \in \mathbb{R}^{N \times d_k}$ is the key matrix featuring the visual content $\mathbf{x}_v$ and possibly the textual content $\mathbf{x}_l$ (in the case of self-attention); d_k is the dimensionality of the key and query vectors; and N is the number of visual (and possibly textual) elements in $\mathbf{x}$. We denote by $\alpha_{t,v}^{(l,h)}$ the subset of attention weights interacting with visual tokens or patches. For cross-attention heads,

$\alpha_{t,v}^{(l,h)} = \alpha_t^{(l,h)}$, while for self-attention heads, $\alpha_{t,v}^{(l,h)} = [\alpha_{t,1}^{(l,h)}, \dots, \alpha_{t,|\mathbf{x}_v|}^{(l,h)}]$, where $|\mathbf{x}_v|$ denotes the number of visual tokens or patches. We then define the attention-based visual grounding score for token $\mathbf{y}_t$ at layer l as

$$a_t^{(l)}(\mathbf{y}, \mathbf{x}) = \frac{1}{H} \sum_{h=1}^H \sum_{i=1}^{|\mathbf{x}_v|} [\alpha_{t,v}^{(l,h)}]_i, \quad (12)$$

where H is the number of attention heads that interact with the visual content. A higher value of $a_t^{(l)}(\mathbf{y}, \mathbf{x})$ indicates stronger dependency on the visual modality, whereas a lower value suggests that visual information contributes minimally to the next-token prediction at that layer. As for the distribution-based score, we normalize these raw attention scores across all generated tokens to obtain a relative token-level importance score $S_A^{(l)}(\mathbf{y}_t, \mathbf{x})$, where

$$S_A(\mathbf{y}_t, \mathbf{x}) = \frac{a_t(\mathbf{y}_t, \mathbf{x})}{\sum_{i=1}^T a_t(\mathbf{y}_i, \mathbf{x})}. \quad (13)$$

This normalized score captures the relative contribution of token $\mathbf{y}_t$ to the overall visual attention mass in the generated answer.

4.3. VIG-TUQ

Our approach leverages visual grounding scores defined in Section 4.2 and combines them with token-level uncertainty scores. Our approach assumes that tokens more strongly grounded in visual content contribute more significantly to uncertainty estimation, as they are more likely to support factual claims in the generated output (see Section 5).

Visual-grounding token UQ. We define VIG-TUQ as a UQ framework for LVLMs that weights token-level uncertainty scores with visual grounding scores. Specifically, given a multimodal input $\mathbf{x}$ and a generated answer $\mathbf{y} = [\mathbf{y}_t]_{t=1}^T$, we define a visual-grounded uncertainty score by combining the distribution-based visual grounding score $S_{\text{JSD}}(\cdot, \cdot)$ (equation 10) and the attention-based visual grounding score $S_A^{(l)}(\cdot, \cdot)$ (equation 13) with the token entropy $\mathcal{H}(\mathbf{y}_t \mid \mathbf{x})$ (equation 6), and summing over all generated tokens:

$$\text{VIG-TUQ}(\mathbf{x}) = \sum_{t=1}^T \left[\alpha_{\text{JSD}} S_{\text{JSD}}(\mathbf{y}_t, \mathbf{x}) + \alpha_A S_A^{(l)}(\mathbf{y}_t, \mathbf{x}) \right] \mathcal{H}(\mathbf{y}_t \mid \mathbf{x}), \quad (14)$$

where $\alpha_{\text{JSD}} > 0$ and $\alpha_A > 0$ are weighting hyperparameters, and l denotes the layer used to compute $S_A^{(l)}(\cdot, \cdot)$. Importantly, the grounding scores do not represent uncertainty themselves: they determine how strongly each token-level entropy contributes to the uncertainty of the full generated

sequence. Tokens that are weakly grounded in the visual input contribute less to the final uncertainty score, while strongly grounded tokens contribute more through their entropy. The sequence-level uncertainty is therefore obtained by combining the entropy of all generated tokens according to their weighted by their degree of visual grounding.

5. Experiments

5.1. Experimental setup

Datasets and models. We evaluate our method across a range of visual question answering (VQA) tasks, datasets, and model architectures. Specifically, we conduct experiments on four benchmark datasets: **ADVQA** [42], **VQARAD** [39], **OKVQA** [52], and **VizWiz** [21]. Our evaluation includes several recent LVLM architectures, namely Qwen2.5-VL-3B and Qwen2.5-VL-7B [7], Qwen3.5-2B [59], llava1.5-7B, llava1.5-13B [46], idfics-9B [40] and idfics2-8B [40]. Additional details regarding the datasets, models, and experimental setup are provided in Appendix A.

Evaluation metrics. Following prior work [11, 27, 37, 38], we evaluate uncertainty estimation methods using the AUROC metric [23] and ECE [19]. AUROC (Area Under the Receiver Operating Curve) measures the ability of an uncertainty score to distinguish between correct and incorrect model predictions. Higher values indicate better discrimination, with a perfect uncertainty estimator achieving a score of 1, while a random uncertainty measure yields an expected score of 0.5. ECE (Expected Calibration Error) quantifies how well predicted probabilities align with empirical accuracies, where lower values indicate better calibration. Additional details on the evaluation metrics are provided in Appendix A.

Ground truth label. To evaluate the correctness of generated answers, we adopt an *LLM-as-a-Judge* approach, following the recommendation of [27]. For each generated answer, the dataset provides a set of valid ground-truth responses. We use Llama-3.1-8B to determine whether the generated answer is semantically consistent with any of the provided ground-truth answers. This setup has been shown to be reliable and to exhibit strong agreement with human judgments [28].

Baselines. We compare our method against several widely used token-level (TL) uncertainty baselines as well as two sample-diversity approaches, that are semantic-level (SL). For TL methods, we include Token Entropy (equation 6), Log-Perplexity (equation 5), Max-Prob [13, 50], CCP 7, and Token-SAR (equation 8). We also include results of SL methods like the Semantic Entropy (SE) [37]

and the Kernel Language Entropy (KLE) with a Matern kernel [53]. Importantly, our primary comparison is with token-level methods, which estimate uncertainty from a single generated answer. We also report SE and KLE as stronger but more expensive reference methods, since they require multiple generated answers to estimate the uncertainty. Hyperparameter details are provided in Appendix E.

5.2. Results

Overall performance comparison. Tables 1 and 2 report the performance of the evaluated uncertainty estimation methods. Table 1 presents AUROC scores aggregated across LVLM architectures for each dataset, while Table 2 reports results aggregated across datasets for each model architecture. Across all datasets, VIG-TUQ achieves the best or tied-best AUROC among token-level methods, improving over Token-Entropy by up to 4.2 AUROC points on VQARAD. Remarkably, despite using only a single generated answer, VIG-TUQ also remains competitive with the more expensive semantic-level methods: it outperforms SE on VizWiz and VQARAD, nearly matches it on ADVQA, and even outperforms KLE on VQARAD. Consequently, they incur a significantly higher inference cost compared to token-level approaches. Finally, we compute standard deviations of our AUROCs in Tab 1, 3 and Tab A.6 that confirm the statistical significance of our results. Overall, these results show that VIG-TUQ improves token-level UQ consistently while retaining a computational cost close to lightweight single-answer baselines, and approaches the performance of more expensive semantic-level methods.

Visual grounding & informative tokens. Our approach is based on the assumption that visual grounding scores help identifying the most relevant tokens for uncertainty estimation. We further hypothesize that the effectiveness of these scores is model-dependent and that they capture distinct aspects of visual reasoning. To test this hypothesis, we split generated tokens into a top half and a bottom half according to a given selector, and compute the mean token entropy within each half. Tokens are ranked using either the distribution-based score (equation 10), the attention-based score (equation 13), or their weighted combination (equation 14, as used in VIG-TUQ). Importantly, visual grounding is used only to rank tokens in this experiment and does not enter the uncertainty score itself. While Token-SAR estimates token importance from semantic relevance using an external semantic model, and CCP relies on an external NLI model to assess semantic consistency among token alternatives, our results show that selecting tokens according to their visual grounding yields more informative token entropies for sequence-level uncertainty estimation. This indicates that, in LVLMs, visual grounding provides a more task-relevant and trustworthy notion of token importance

Table 1. **VIG-TUQ improves AUROC compared to all token-level (TL) baselines and remains competitive with semantic-level (SL) baselines.** Evaluation metrics for datasets averaged across all LVLM architectures. Values in parentheses are 95% confidence intervals. **Bold** values indicates best performance, underline values indicate the second-best among TL.

		OKVQA		VIZWIZ		VQARAD		ADVQA		Compute s
		AUROC $\uparrow$	ECE $\downarrow$	AUROC $\uparrow$	ECE $\downarrow$	AUROC $\uparrow$	ECE $\downarrow$	AUROC $\uparrow$	ECE $\downarrow$	
TL	<i>Token-Sar</i>	0.591 (0.008)	0.137 (0.007)	0.588 (0.012)	0.126 (0.010)	0.616 (0.027)	0.183 (0.021)	0.564 (0.015)	0.173 (0.013)	1.56
	<i>CCP</i>	0.625 (0.008)	0.112 (0.007)	0.640 (0.012)	0.101 (0.010)	0.621 (0.027)	0.163 (0.021)	0.631 (0.015)	0.126 (0.013)	3.35
	<i>Max-Prob</i>	0.621 (0.008)	0.125 (0.007)	0.641 (0.012)	0.119 (0.010)	0.626 (0.027)	0.194 (0.021)	0.621 (0.015)	0.180 (0.013)	1.02
	<i>Token-Entropy</i>	0.627 (0.008)	0.109 (0.007)	0.650 (0.012)	0.093 (0.010)	0.642 (0.027)	0.200 (0.021)	0.612 (0.015)	0.168 (0.013)	1.01
	<i>VIG-TUQ-A</i> (ours)	<u>0.648</u> (0.008)	<u>0.100</u> (0.007)	<u>0.676</u> (0.012)	0.086 (0.010)	<u>0.670</u> (0.027)	0.194 (0.021)	<u>0.629</u> (0.015)	0.161 (0.013)	1.21
	<i>VIG-TUQ-JSD</i> (ours)	0.628 (0.008)	0.078 (0.007)	0.664 (0.012)	0.097 (0.010)	0.638 (0.027)	0.210 (0.021)	0.605 (0.015)	0.164 (0.013)	1.32
	VIG-TUQ (ours)	0.653 (0.008)	0.100 (0.007)	0.679 (0.012)	<u>0.088</u> (0.010)	0.684 (0.027)	<u>0.180</u> (0.021)	0.631 (0.015)	<u>0.159</u> (0.013)	1.37
SL	<i>SE</i> (Gold)	0.664 (0.008)	0.074 (0.007)	0.649 (0.012)	0.061 (0.010)	0.652 (0.027)	0.174 (0.021)	0.632 (0.015)	0.194 (0.013)	4.69
	<i>KLE</i> (Gold)	0.690 (0.008)	0.072 (0.007)	0.682 (0.012)	0.063 (0.010)	0.656 (0.027)	0.174 (0.021)	0.646 (0.015)	0.192 (0.013)	5.68

Compute is the mean inference time in second for generation + uncertainty quantification.

Table 2. **VIG-TUQ improves AUROC compared to all token-level (TL) baselines and remains competitive with semantic-level (SL) baselines.** Evaluation metrics for LVLM architectures averaged across all vision datasets. **Bold** values indicates best performance. underline values indicate the second-best among TL.

		Qwen2.5-VL-3B		Qwen2.5-VL-7B		llava-1.5-13B		llava-1.5-7B		idefics2-8B		idefics-9B		Qwen3.5-2B	
Method		AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
TL	<i>Token-Sar</i>	0.554	0.175	0.588	0.183	0.639	0.136	0.609	0.136	0.599	0.153	0.550	0.170	0.587	0.130
	<i>CCP</i>	0.644	0.120	0.620	0.148	0.627	0.137	0.624	0.120	0.627	0.137	0.659	0.118	0.604	0.099
	<i>Max-Prob</i>	0.643	0.150	0.616	0.172	0.625	0.163	0.623	0.159	0.617	0.162	0.656	0.164	0.611	0.112
	<i>Token-Entropy</i>	0.651	0.126	0.622	0.164	0.650	0.140	0.632	0.133	0.628	0.172	0.629	0.154	0.616	0.109
	<i>VIG-TUQ-A</i> (ours)	0.689	0.116	0.651	0.154	0.660	0.147	0.640	0.139	0.659	0.139	0.660	0.144	0.632	0.107
	<i>VIG-TUQ-JSD</i> (ours)	0.652	0.112	0.611	0.142	0.657	0.123	0.652	0.124	0.651	0.150	0.600	0.197	0.614	0.113
	<i>VIG-TUQ</i> (ours)	0.691	0.112	0.653	0.153	0.671	0.135	0.663	0.126	0.661	0.142	0.660	0.144	0.633	0.106
SL	<i>SE</i> (Gold)	0.662	0.114	0.651	0.148	0.668	0.127	0.680	0.121	0.655	0.125	0.674	0.121	0.555	0.123
	<i>KLE</i> (Gold)	0.673	0.113	0.658	0.147	0.683	0.127	0.686	0.121	0.668	0.124	0.687	0.120	0.626	0.127

VIG-TUQ is as described in Eq14, *VIG-TUQ-A* and *VIG-TUQ-JSD* are variants using only one of the visual grounding criteria.

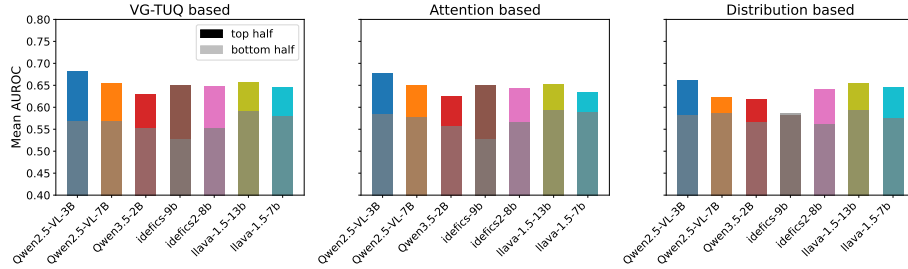


Figure 3. **Visual grounding scores help identify the tokens most relevant for uncertainty estimation.** AUROC performance reported for the mean token entropy of the top half vs. bottom half of generated tokens, ranked by a given selector, across different LVLM architectures. Tokens are ranked according to the distribution-based score (equation 10), the attention-based score (equation 13), or their weighted combination as used in VIG-TUQ (details in Appendix B). Colored bars correspond to the top half of tokens and gray bars to the bottom half. All results are averaged across all datasets.

for uncertainty estimation, while avoiding the additional external models required by Token-SAR and CCP. Results presented in Figure A.4 lead to three main observations. (1) For all three selectors, the top half consistently yields higher mean AUROC than the bottom half across models and datasets, confirming that visual grounding scores identify tokens carrying more informative uncertainty signal; (2) The attention-based and distribution-based criteria perform better on different models and datasets, which con-

firms their complementary nature and motivates their combination in VIG-TUQ; (3) The weighted combination used in VIG-TUQ matches or exceeds the top-half performance of either individual criterion, showing that combining both signals yields a more robust token selector across models. Overall, these findings suggest that uncertainty is not uniformly distributed across the generated sequence and that only a subset of tokens carries the most informative signals for uncertainty quantification. Additional experiments

about the validity of our grounding score are in Appendix B.

Computational cost. We evaluate the efficiency of VIG-TUQ by comparing its average inference time with existing token-level UQ methods on OKVQA. Table 1 and A.5 reports the average runtime for answer generation and uncertainty estimation under identical backbone models and hardware settings. Since all token-level methods share the same decoding procedure, generation time remains constant and any runtime differences arise only from the post-hoc UQ computation. Unlike CCP and Token-SAR, which depend on external models and thus introduce additional computational overhead, VIG-TUQ is more efficient because it leverages only token probabilities and attention weights that are computed by the generation model itself. Although lightweight baselines such as Token-Entropy, Log-Perplexity, and Max-Prob remain slightly faster, VIG-TUQ incurs only minimal overhead while being substantially more efficient than other token-level UQ approaches.

5.3. Global Model-Level Hyperparameters

Most UQ methods require some hyperparameter tuning. In our comparison, only simple baselines such as Token-Entropy and Max-Prob are parameter-free, while methods such as CCP, Token-SAR, SE, and KLE also require selecting thresholds, relevance parameters, or other settings. VIG-TUQ therefore follows the same general setting. In the main experiments, we tune its hyperparameters for each model–dataset pair using training data. However, for easier deployment, it is useful to know whether one configuration can be shared across datasets. We therefore evaluate a fixed configuration per model, shared across all four datasets.

Table 3 reports AUROC values as bootstrapped means (standard deviations) over 1,000 resamples, macro-averaged across the four datasets (OKVQA, VIZWIZ, VQARAD, ADVQA). We compare three settings: (i) token entropy, (ii) VIG-TUQ (model only), using a single hyperparameter configuration shared across datasets (model-level), and (iii) VIG-TUQ (model/data), using per-dataset tuned hyperparameters.

For every model, the gap between VIG-TUQ (model only) and Token-Entropy exceeds the standard deviation of either estimate, indicating that the observed gains are statistically meaningful rather than noise. Notably, even when hyperparameters are shared across all four datasets rather than tuned per dataset, VIG-TUQ still yields a significant and consistent improvement over the baseline for every model. These results indicate that a single, easy-to-deploy hyperparameter setting per model is sufficient to achieve strong and statistically robust gains, supporting the practicality of our approach beyond the per-dataset tuning

Table 3. AUROC (bootstrapped mean, std. dev. over 1,000 resamples) macro-averaged across OKVQA, VIZWIZ, VQARAD, and ADVQA, comparing Token-Entropy, VIG-TUQ with shared model-level hyperparameters, and VIG-TUQ with per-dataset tuned hyperparameters.

Model	Token-Entropy AUROC↑	VIG-TUQ (model only) AUROC↑	VIG-TUQ (model/data) AUROC↑
Qwen2.5-VL-3B	0.651 (0.015)	<u>0.690</u> (0.015)	0.691 (0.015)
Qwen2.5-VL-7B	0.622 (0.017)	<u>0.642</u> (0.017)	0.653 (0.016)
llava-1.5-13b	0.650 (0.016)	<u>0.668</u> (0.015)	0.671 (0.011)
llava-1.5-7b	0.632 (0.016)	<u>0.655</u> (0.016)	0.663 (0.016)
idefics2-8b	0.627 (0.016)	<u>0.647</u> (0.016)	0.661 (0.015)
idefics-9b	0.629 (0.016)	<u>0.658</u> (0.015)	0.660 (0.015)
Qwen3.5-2B	0.616 (0.015)	<u>0.627</u> (0.015)	0.633 (0.015)

setup used in the main experiments.

6. Conclusion

In this paper, we present a token-level analysis of how reliance on visual evidence in LVMs affects uncertainty quantification. Our findings indicate that visual information helps identify the most relevant token for uncertainty quantification and underscore the importance of combining multiple visual grounding signals, as their effectiveness can vary across models and capture complementary aspects of visual reasoning. Building on these insights, we introduce Visual-Grounded Token UQ (VIG-TUQ), a training-free and computationally efficient framework that explicitly incorporates visual grounding into uncertainty estimation by weighting token-level language uncertainty with visual grounding scores. Experiments on multiple datasets and across diverse LVLM architectures (including early-fusion, late-fusion, and native-fusion models) indicate that our method often improves upon existing token-level uncertainty approaches. We hope our work will encourage future research on vision-aware and fine-grained uncertainty quantification for LVLMs.

Limitations. While our approach improves upon existing token-level uncertainty methods, semantic uncertainty methods can still perform better in some cases, but they are more costly. Combining them with visual grounding could be explored in future work.

References

- [1] Samir Abdaljalil, Hasan Kurban, Parichit Sharma, Erchin Serpedin, and Rachad Atat. Sindex: Semantic inconsistency index for hallucination detection in llms. *arXiv preprint arXiv:2503.05980*, 2025. 2
- [2] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *IF*, 2021. 1
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 4
- [4] Jisu An, Junseok Lee, Jeoungeun Lee, and Yongseok Son. Towards llm-centric multimodal fusion: A survey on integration strategies and techniques. *arXiv preprint arXiv:2506.04788*, 2025. 3, 4
- [5] Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it’s lying. In *EMNLP*, 2023. 2
- [6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2024. 3
- [7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 1
- [8] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs’ internal states retain the power of hallucination detection. In *ICLR*, 2024. 1
- [9] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 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, 2024. 4
- [10] Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385, 2024. 2, 4
- [11] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024. 1, 2, 6, 4
- [12] Laura Fieback, Jakob Spiegelberg, and Hanno Gottschalk. Metatoken: Detecting hallucination in image descriptions by meta classification. *arXiv preprint arXiv:2405.19186*, 2024. 2
- [13] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555, 2020. 2, 4, 6
- [14] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *Computer Vision – ECCV 2024*, pages 148–166, Cham, 2025. Springer Nature Switzerland. 1, 4
- [15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 1
- [16] Haoxiang Gao, Zhongruo Wang, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A survey for foundation models in autonomous driving. In *2025 6th International Conference on Computer Vision and Data Mining (ICCVDM)*, pages 63–71. IEEE, 2025. 1
- [17] Gemini. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 1
- [18] Yashvir S Grewal, Edwin V Bonilla, and Thang D Bui. Improving uncertainty quantification in large language models via semantic embeddings. *arXiv preprint arXiv:2410.22685*, 2024. 2
- [19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 6, 1
- [20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 2025. 1
- [21] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018. 6, 1
- [22] Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, et al. Multimodal fusion and vision-language models: A survey for robot vision. *Information Fusion*, page 103652, 2025. 1
- [23] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 6, 1
- [24] Jose Miguel Hernandez-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *ICML*, 2015. 1
- [25] Joseph Hoche, Andrei Bursuc, David Brellmann, Gilles Louppe, Pavel Izmailov, Angela Yao, and Gianni Franchi. Improving semantic uncertainty quantification in llms with semantic gaussian processes. *arXiv preprint arXiv:2512.14177*, 2025. 2
- [26] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *TIS*, 2025. 1

- [27] Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in LLMs. In *EMNLP*, 2025. 6
- [28] Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34716–34733, 2025. 6
- [29] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *CS*, 2023. 1
- [30] Ziwei Ji, Lei Yu, Yeskendir Koishkenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3769–3793, Suzhou, China, 2025. Association for Computational Linguistics. 2
- [31] Yixing Jiang, Jesutofunmi A Omiye, Cyril Zakka, Michael Moor, Haiwen Gui, Shayan Alipour, Seyed Shahabeddin Mousavi, Jonathan H Chen, Pranav Rajpurkar, and Roxana Daneshjou. Evaluating general vision-language models for clinical medicine. *MedRxiv*, pages 2024–04, 2024. 1
- [32] Minsuh Joo and Hyunsoo Cho. Cleanse: Uncertainty estimation approach using clustering-based semantic consistency in LLMs. In *ACL Workshops*, 2025. 2
- [33] 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. 2
- [34] Sungmin Kang, Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, and Salman Avestimehr. Uncertainty quantification for hallucination detection in large language models: Foundations, methodology, and future directions, 2025. 2
- [35] Hazel Kim, Tom A. Lamb, Adel Bibi, Philip Torr, and Yarin Gal. Detecting LLM hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32310–32322, Suzhou, China, 2025. Association for Computational Linguistics. 2
- [36] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023. 2
- [37] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023. 6
- [38] Gregory Kang Ruey Lau, Hieu Dao, Nicole Kan Hui Lin, and Bryan Kian Hsiang Low. Uncertainty quantification for multimodal large language models with coherence-adjusted semantic volume, 2025. 1, 2, 6
- [39] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 2018. 6, 1
- [40] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023. 4, 6, 1
- [41] Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. VLind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4129–4144, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. 1, 4
- [42] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *ICCV*, 2021. 6, 1
- [43] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*, 2024. 3, 4
- [44] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *TMLR*, 2024. 2
- [45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 3
- [46] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 6, 1
- [47] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1
- [48] Lin Long, Changdae Oh, Seongheon Park, and Sharon Li. Understanding language prior of LLMs by contrasting chain-of-embedding. In *The Fourteenth International Conference on Learning Representations*, 2026. 1, 4
- [49] Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in VLMs. In *Forty-second International Conference on Machine Learning*, 2025. 1, 4
- [50] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *ICLR*, 2021. 4, 6
- [51] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. 2
- [52] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 2, 4, 6, 1
- [53] Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Martinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. In *NeurIPS*, 2024. 1, 2, 6, 4
- [54] Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey

Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025. [1](#)

- [55] OpenAI. Gpt-4 technical report, 2024.
- [56] OpenAI. gpt-oss-120b and gpt-oss-20b model card, 2025. [1](#)
- [57] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. [2](#)
- [58] Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Koteck, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *ICLR*, 2025. [2](#)
- [59] Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. [3](#), [6](#), [1](#)
- [60] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. [2](#)
- [61] Gal Yona, Roei Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2405.16908*, 2024. [2](#)
- [62] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. VI-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation, 2024. [1](#), [2](#)

Leveraging Visual Signals for Robust Token-Level Uncertainty in Vision–Language Generation

– Supplementary Material –

Overview of the Supplementary Material

The supplementary material provides additional analyses and complete experimental details supporting the main findings of the paper.

Section B further studies the role of visual grounding in uncertainty estimation. We show that selecting tokens according to their visual grounding scores provides more informative uncertainty estimates than random token selection (Figure A.4). We also perform an intervention-based experiment by masking the most visually relevant tokens, showing a stronger effect than random masking and providing additional evidence that the proposed attention-based grounding score identifies visual information that is important for the model’s prediction. Section C presents additional experiments that further analyze the behavior of our method under different settings. Section D reports the complete quantitative results for all evaluated models, datasets, and uncertainty estimation methods. Section E provides the hyperparameters and implementation details used in our experiments to facilitate reproducibility. Finally, Section F analyzes how visual information influences the hidden representations of the model across different layers.

Overall, these additional analyses further support our main finding that **visual grounding provides useful information for identifying which generated tokens should contribute most to uncertainty estimation**.

A. Experimental protocol

A.1. Datasets

We evaluate VIG-TUQ on several Visual Question Answering (VQA) datasets: **ADVQA** [42], which consists of a training set of 5,000 samples and a test set of 2,000 samples, both drawn from the original training set; **VQARAD** [39], which includes 2,000 samples from the original training set and 500 samples from the original test set; **OKVQA** [52], which contains 5,000 training samples drawn from the original training set and 2,000 test samples from the validation set; and **VizWiz** [21], which contains 5,000 training samples and 2,000 test samples drawn from the original training set.

A.2. Models

We evaluate VIG-TUQ on widely used Large Vision Language Models (LVLMs). In particular, we conduct experi-

ments on Qwen2.5-VL-3B and Qwen2.5-VL-7B [7], Qwen3.5-2B [59], LLaVa1.5-7B, LLaVa1.5-13B [46], idefics-9B [40] and idefics2-8B [40]. We include models with different architectures to cover the range of modality fusion strategies discussed in Section 3.1 and to analyze how these fusion strategies affect the visual grounding. In particular, llava, idefics2 and Qwen-VL models adopt a *early fusion* approach, while idefics-9B model rely on an *late fusion* approach and Qwen3.5 on a *native* fusion one.

A.3. Metrics

AUROC. We assign a probability score to each example for belonging to the positive class (positive reflects that the LVLM is certain about its answer). Area Under the ROC Curve (AUROC) [23] measures how well these scores rank true positives above true negatives. It is the chance that a randomly chosen positive example receives a higher score than a randomly chosen negative one. Higher AUROC means better class separation.

ECE. Expected Calibration Error (ECE) [19] measures how well a model’s predicted confidence aligns with its empirical accuracy. A low ECE indicates that the model is well calibrated, meaning that predictions made with confidence c are correct approximately c fraction of the time. Conversely, a high ECE indicates miscalibration, for example when the model is systematically overconfident or underconfident. Since ECE is defined for confidence scores in $[0, 1]$, we first normalize all UQ scores to this range before computing ECE.

B. Token selection analysis

In addition to Section 5.2, we conducted additional experiments to evaluate whether visual grounding scores help identify the tokens most relevant for uncertainty estimation; we conducted a token selection experiment.

Effect of grounding-based token selection. For each generated answer $\mathbf{y} = [y_1, \dots, y_T]$, we compute the entropy of each token, $\mathcal{H}(y_t \mid \mathbf{x})$ (equation 6). We then rank the tokens using either the distribution-based grounding score $S_{\text{JSD}}(\mathbf{y}_t, \mathbf{x})$ (equation 10) or the attention-based grounding score $S_A^{(l)}(\mathbf{y}_t, \mathbf{x})$ (equation 13). For each ranking, we keep only the top $k\%$ of tokens and compute the

Table A.4. Effect of visual-token masking on the ability of the mean token probability to predict answer correctness (AUROC) on VQA-RAD with Qwen2.5-VL-3B. Both interventions remove the same amount of visual context.

Score	AUROC
Mean token probability (unperturbed image)	0.6872
Mean token probability (random masking)	0.6294
Mean token probability (attention-guided masking)	0.5688

final uncertainty score as the sum of the entropies of the selected tokens. As a baseline, we repeat the same experiment by randomly selecting $k\%$ of the generated tokens. To reduce the effect of randomness, this random selection is repeated 10 times for each model, dataset, and value of k , and we report the average AUROC. Figure A.4 shows the results for different LVLM architectures.

Intervention-based validation of the visual grounding score. Attention weights do not provide a complete causal explanation of how much the model relies on the image. For this reason, in VIG-TUQ, the attention-based grounding score is not used alone as an uncertainty score or as a saliency map. Instead, it is used only to weight token-level entropy and is combined with the complementary distribution-based JSD score.

To further evaluate whether the attention-based score identifies visual regions that are important for the model prediction, we perform an intervention experiment on VQA-RAD with Qwen2.5-VL-3B. For each generated token, we compare three settings:

1. **Original image:** the token probability is obtained using the unchanged input image.
2. **Attention-guided masking:** we mask the 10% of visual tokens with the highest attention-based grounding scores and recompute the token probability.
3. **Random masking:** we mask the same number of visual tokens, but select them uniformly at random.

For each generated answer, we average the token-level probabilities over the whole sequence to obtain a sentence-level score. These scores are then used to predict whether the answer is correct, and performance is evaluated using AUROC.

As reported in Table A.4, masking the visual tokens selected by our grounding score degrades AUROC substantially more than masking an equal number of randomly chosen visual tokens (0.5688 versus 0.6294, against 0.6872 without perturbation). Since both interventions remove the same quantity of visual context, this gap indicates that the grounding score identifies visual information that is particularly influential for the model’s prediction. We stress that this experiment does not constitute a pixel-level causal ex-

planation, and that masking may itself induce a distribution shift; it nonetheless provides considerably stronger evidence than a purely correlational analysis.

C. Additional experiments

Impact of LVLM architecture. Figure A.5 aims to analyze how different LVLM architectures use visual information across their layers. To do so, we compute the Centered Kernel Alignment (CKA) similarity between the hidden representations of each LVLM (on OKVQA) and the final-layer features of a reference vision encoder, DINOv2 [57]. We use DINOv2 as a reference because it is trained for visual representation learning and captures strong image features across many vision tasks. As a result, it provides a useful proxy for purely visual information. We therefore measure how similar LVLM representations are to these visual features. For each LVLM, we extract hidden states across all transformer layers and compare them to the reference visual features. Because the evaluated models have different numbers of layers, we normalize layer indices to a relative depth scale from 0.0, corresponding to the embedding layer, to 1.0, corresponding to the final transformer layer. This enables a direct comparison of where visual information is most strongly reflected across architectures.

The resulting plot shows how closely each model’s internal representations align with the original visual features as generation progresses through the network. High CKA similarity indicates that the model representation remains strongly related to the visual encoder representation, whereas lower similarity suggests a shift toward language- or task-specific representations. We observe that `idefics-9B`, the only late-fusion architecture in this comparison, exhibits the strongest similarity to the reference vision encoder. This behavior is consistent with its architectural design: unlike early-fusion models, which inject visual features mainly before the language-model layers, late-fusion models continue to integrate visual components throughout the network, leading to stronger visual alignment at intermediate or deeper layers.

C.1. Computational cost

We assess the efficiency of VIG-TUQ by comparing its average inference time against existing token-level UQ methods on OKVQA. Table A.5 reports the mean runtime for answer generation and uncertainty estimation, measured with the same backbone models and hardware. Because every token-level method follows an identical decoding procedure, generation time is fixed across methods, so all runtime gaps stem solely from the post-hoc UQ computation. CCP and Token-SAR rely on external models and therefore incur extra computational cost, whereas VIG-TUQ avoids this by reusing only the token probabilities and attention weights already produced by the generation model. Lightweight

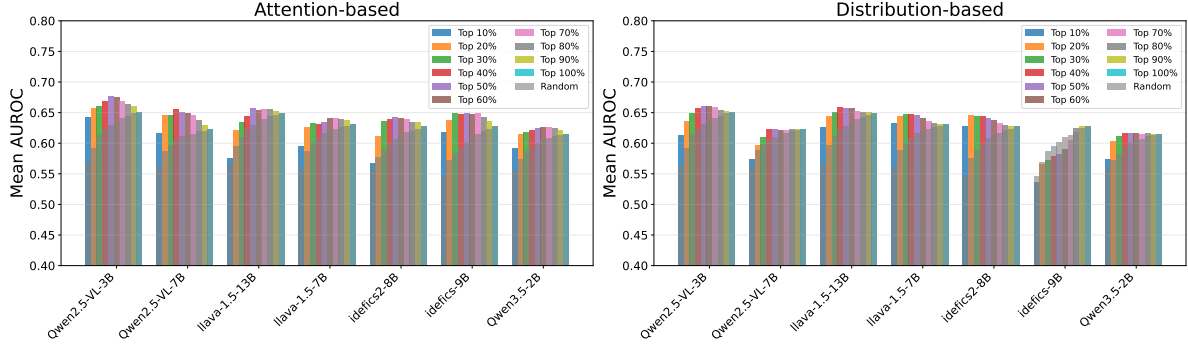


Figure A.4. **Visual grounding scores help identify the tokens most relevant for uncertainty estimation.** AUROC performance reported for the sum of top k% token entropies across different LVLM architectures. Top k% token entropies are selected according to either the attention-based score or the distribution-based score (equation 10), attention-based score (equation 13), or randomly from generated tokens. Gray bars correspond to a random token-selection baseline and averaged over 10 runs. All results are averaged across all datasets.

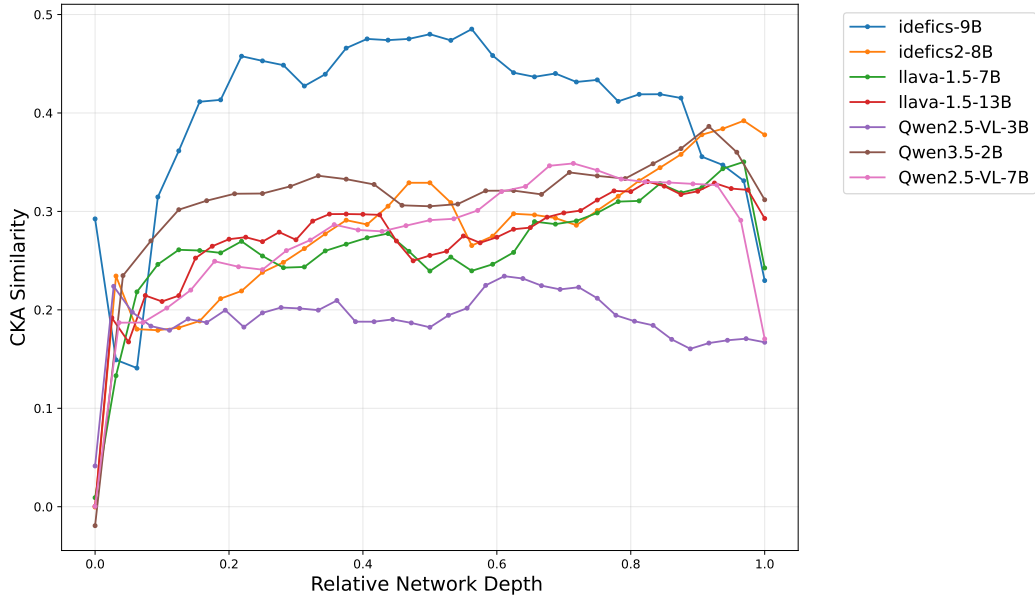


Figure A.5. CKA similarity between LVLM hidden states and reference vision features across normalized network depth. For each model, we compare layer-wise hidden representations to the final-layer representation of a frozen vision encoder. Higher CKA values indicate stronger alignment with the original visual representation. The late-fusion model *idefics-9B* exhibits the highest visual alignment across much of the network, consistent with architectures that continue to inject visual information throughout the transformer layers.

baselines such as Token-Entropy, Log-Perplexity, and Max-Prob are still marginally faster, but VIG-TUQ adds only negligible overhead and is considerably more efficient than the remaining token-level UQ approaches.

D. Complete results

In this section, we present the detailed performance of all UQ methods across datasets and models described in Appendix A. Table A.6 reports the AUROC scores achieved

by each method on every dataset and LVLM architecture.

Comparison with token-level UQ methods. VIG-TUQ consistently outperforms conventional token-level uncertainty metrics, including Log-Perplexity, Max-Prob, Token-Entropy, CCP, and Token-SAR. For instance, on the **OKVQA** dataset, VIG-TUQ achieves the highest AUROC score. Similar trends are observed for **VQARAD** and **ADVQA**, where VIG-TUQ ranks among the top-performing

Table A.5. **VIG-TUQ is computationally efficient.** This table shows the average inference time for: answer generation and uncertainty quantification on the *OKVQA* dataset.

Model	Max-Prob	Token-Entropy	CCP	Token-SAR	VIG-TUQ	KLE
Qwen2.5-3B	1.01s	1.00s	3.26s	1.71s	1.29s	4.78s
Qwen2.5-7B	1.58s	1.60s	5.28s	2.39s	2.00s	8.42s
llava-1.5-7B	0.81s	0.84s	2.81s	1.09s	1.01s	4.83s
llava-1.5-13B	1.39s	1.41s	4.64s	2.10s	1.75s	7.40s
idefics2-8B	0.65s	0.66s	2.18s	0.99s	0.82s	3.47s
idefics-9B	0.72s	0.68s	2.30s	1.06s	1.20s	4.60s
Qwen3.5-2B	0.98s	0.88s	2.98s	1.58s	1.52s	6.26s
Mean	1.02s	1.01s	3.35s	1.56s	1.37s	5.68s

methods. These results confirm that integrating multiple importance scores and visual grounding scores improves the performance of token-level UQ methods.

Comparison with consistency-based methods. Semantic Entropy (SE) [11] and Kernel Language Entropy (KLE) [53] achieve slightly higher AUROC on certain datasets. This behavior is expected as these methods estimate uncertainty by measuring the semantic consistency of generated responses, which provides a more comprehensive assessment of uncertainty across the entire generated content. However, these methods require multiple sampling, which also makes them significantly more expensive at inference time (see Table A.5). In contrast, our approach requires only a single or two samplings and therefore remains much more efficient while still maintaining competitive performance.

Effect across datasets. We also observe that the performance of UQ methods varies across datasets. For instance, VIG-TUQ clearly outperforms on **OKVQA** and **VQARAD**, which require stronger reasoning and better exploitation of the visual content, which makes reliable uncertainty estimation more challenging. On **VizWiz**, the improvements are still visible but slightly smaller. This dataset contains real-world images with visual noise, which can make the visual grounding more difficult for all models.

Effect across model architectures. The benefits of VIG-TUQ are consistent across different LVLM architectures. This suggests that VIG-TUQ is not tied to a specific architecture and can be effectively applied across a wide range of multimodal architectures.

E. Hyperparameters

The hyperparameters used in VIG-TUQ are reported in Table A.8. They are selected on a small training split for each dataset-model combination. To limit the search space and avoid overfitting to the validation setting, we impose a strict constraint on the weighting coefficients: both α_{JSD}

and α_A are restricted to integer values in $[0, 5]$. This deliberately coarse search is intended to ensure that the performance gains of VIG-TUQ do not rely on extensive hyperparameter tuning. The selected hyperparameters show that some model- and dataset-specific adaptation is beneficial. This is consistent with the token selection analysis in Figure A.4, where the attention-based and distribution-based visual grounding scores do not perform best in the same settings. For example, for **llava-1.5-7B** on **VQARAD**, VIG-TUQ relies only on the distribution-based score, whereas for **Qwen3.5-2B** on **OKVQA**, it relies only on the attention-based score. Nevertheless, in most model-dataset combinations, both scores receive nonzero weights, suggesting that the two grounding criteria provide complementary information for uncertainty quantification.

Hyperparameter sensitivity and tuning requirements.

As discussed above, VIG-TUQ requires only a lightweight hyperparameter search. To avoid relying on extensive tuning, we restrict the weighting coefficients α_{JSD} and α_A to integer values in $\{0, \dots, 5\}$, and tune them together with the attention layer l on a small training split. Table A.7 reports the selected hyperparameters when varying the number of training examples K . We observe that stable and competitive hyperparameters can already be obtained with as few as $K = 100$ examples. In several cases, such as **Idefics-9B** on **OKVQA** and **VQARAD**, the selected configuration remains identical across different values of K . Even when the selected coefficients vary, as for **Qwen2.5-VL-3B** on **OKVQA**, the resulting AUROC remains nearly unchanged. These results suggest that VIG-TUQ is not overly sensitive to the exact hyperparameter choice and does not require large validation sets or fine-grained tuning to achieve strong performance.

F. Analysis of visual influence on hidden representations

In this section, we examine how the internal representations of an LVLM change when visual input is removed. In particular, we compare the hidden representations obtained during normal generation with those obtained when the image is replaced by a blank input.

LVLM. Let $\mathbf{x} = (\mathbf{x}_v, \mathbf{x}_l)$ be a multimodal input with a visual content $\mathbf{x}_v$ and a textual content $\mathbf{x}_l$. As described in Section 3, an LVLM defines a parameterized function $f_\theta : \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathbb{R}^d$ to encode the visual-linguistic context into a latent representation. In particular, the parametric function $f_\theta : \mathcal{X}_v \times \mathcal{X}_l \rightarrow \mathbb{R}^d$ can be defined recursively over L layers

Table A.6. **VIG-TUQ outperforms token-level UQ baselines.** Evaluation metrics for LVLm architectures across all datasets and models.

		Qwen2.5-AUROC↑	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-72B	llava-1.5-13B	llava-1.5-7B	llava-1.5-7B	llava-1.5-7B	idefics2-8B	idefics2-8B	idefics2-8B	idefics-9B	idefics-9B	Qwen3.5-2B	Qwen3.5-2B			
		ECE↓	ECE↓	ECE↓	ECE↓	ECE↓	ECE↓	ECE↓	ECE↓	ECE↓	ECE↓	ECE↓	ECE↓	ECE↓	AUROC↑	ECE↓			
OKVQA	Log Perplexity	0.610 (0.008)	0.110 (0.007)	0.595 (0.008)	0.109 (0.007)	0.635 (0.008)	0.119 (0.007)	0.634 (0.008)	0.118 (0.006)	0.637 (0.008)	0.116 (0.006)	0.637 (0.008)	0.116 (0.006)	0.637 (0.008)	0.605 (0.008)	0.104 (0.007)	0.649 (0.008)	0.071 (0.006)	
	Token-Sar	0.551 (0.008)	0.159 (0.007)	0.581 (0.008)	0.115 (0.007)	0.639 (0.008)	0.145 (0.007)	0.637 (0.008)	0.114 (0.007)	0.604 (0.008)	0.159 (0.007)	0.524 (0.008)	0.158 (0.007)	0.600 (0.008)	0.158 (0.007)	0.600 (0.008)	0.110 (0.007)	0.607 (0.008)	
	CCP	0.607 (0.008)	0.111 (0.007)	0.611 (0.008)	0.088 (0.007)	0.614 (0.008)	0.168 (0.007)	0.612 (0.008)	0.138 (0.007)	0.647 (0.008)	0.172 (0.006)	0.642 (0.008)	0.172 (0.006)	0.644 (0.008)	0.172 (0.006)	0.644 (0.008)	0.070 (0.006)	0.644 (0.008)	
	Mac-Prob	0.592 (0.008)	0.153 (0.007)	0.608 (0.008)	0.113 (0.007)	0.611 (0.008)	0.150 (0.007)	0.638 (0.008)	0.116 (0.006)	0.642 (0.008)	0.145 (0.007)	0.608 (0.008)	0.145 (0.007)	0.608 (0.008)	0.145 (0.007)	0.608 (0.008)	0.070 (0.006)	0.608 (0.008)	
	Token-Entropy	0.610 (0.008)	0.109 (0.007)	0.601 (0.008)	0.102 (0.007)	0.641 (0.008)	0.139 (0.007)	0.638 (0.008)	0.116 (0.006)	0.642 (0.008)	0.145 (0.007)	0.608 (0.008)	0.145 (0.007)	0.608 (0.008)	0.145 (0.007)	0.608 (0.008)	0.069 (0.006)	0.608 (0.008)	
	VIG-TUQ-A (ours)	0.657 (0.008)	0.077 (0.006)	0.620 (0.008)	0.092 (0.007)	0.644 (0.008)	0.136 (0.007)	0.651 (0.008)	0.106 (0.006)	0.664 (0.008)	0.143 (0.006)	0.645 (0.008)	0.143 (0.006)	0.645 (0.008)	0.143 (0.006)	0.645 (0.008)	0.073 (0.006)	0.645 (0.008)	
	VIG-TUQ-JSD (ours)	0.652 (0.008)	0.019 (0.007)	0.606 (0.008)	0.020 (0.007)	0.642 (0.008)	0.049 (0.006)	0.646 (0.008)	0.065 (0.007)	0.662 (0.008)	0.065 (0.007)	0.662 (0.008)	0.065 (0.007)	0.662 (0.008)	0.065 (0.007)	0.662 (0.008)	0.021 (0.006)	0.662 (0.008)	
	VIG-TUQ (ours)	0.664 (0.008)	0.021 (0.006)	0.620 (0.008)	0.092 (0.007)	0.657 (0.008)	0.137 (0.007)	0.657 (0.008)	0.109 (0.006)	0.674 (0.008)	0.144 (0.007)	0.645 (0.008)	0.144 (0.007)	0.674 (0.008)	0.144 (0.007)	0.674 (0.008)	0.021 (0.006)	0.674 (0.008)	
	SE	0.669 (0.008)	0.041 (0.006)	0.647 (0.008)	0.080 (0.007)	0.690 (0.008)	0.076 (0.007)	0.708 (0.008)	0.075 (0.006)	0.702 (0.008)	0.110 (0.006)	0.706 (0.008)	0.110 (0.006)	0.706 (0.008)	0.110 (0.006)	0.706 (0.008)	0.099 (0.006)	0.706 (0.008)	
OKVQA	KLE	0.682 (0.008)	0.035 (0.007)	0.654 (0.008)	0.078 (0.007)	0.700 (0.008)	0.074 (0.006)	0.717 (0.008)	0.069 (0.006)	0.715 (0.008)	0.110 (0.007)	0.721 (0.008)	0.110 (0.007)	0.721 (0.008)	0.110 (0.007)	0.721 (0.008)	0.112 (0.007)	0.721 (0.008)	
VIZWIZ	Log Perplexity	0.690 (0.012)	0.088 (0.010)	0.634 (0.013)	0.131 (0.011)	0.652 (0.012)	0.081 (0.009)	0.667 (0.012)	0.076 (0.010)	0.635 (0.012)	0.108 (0.010)	0.612 (0.013)	0.121 (0.010)	0.634 (0.012)	0.108 (0.010)	0.612 (0.013)	0.121 (0.010)	0.634 (0.012)	0.108 (0.010)
	Token-Sar	0.525 (0.013)	0.185 (0.011)	0.560 (0.013)	0.181 (0.012)	0.636 (0.012)	0.082 (0.010)	0.628 (0.012)	0.075 (0.010)	0.597 (0.013)	0.132 (0.010)	0.576 (0.013)	0.126 (0.011)	0.591 (0.013)	0.132 (0.010)	0.576 (0.013)	0.126 (0.011)	0.591 (0.013)	0.132 (0.010)
	CCP	0.665 (0.012)	0.121 (0.010)	0.612 (0.013)	0.155 (0.011)	0.639 (0.012)	0.096 (0.010)	0.655 (0.013)	0.100 (0.011)	0.639 (0.012)	0.132 (0.011)	0.615 (0.012)	0.132 (0.011)	0.615 (0.012)	0.132 (0.011)	0.615 (0.012)	0.132 (0.011)	0.615 (0.012)	0.132 (0.011)
	Mac-Prob	0.663 (0.012)	0.108 (0.010)	0.614 (0.013)	0.132 (0.011)	0.637 (0.012)	0.136 (0.010)	0.640 (0.013)	0.130 (0.010)	0.632 (0.012)	0.121 (0.010)	0.652 (0.012)	0.129 (0.011)	0.631 (0.012)	0.129 (0.011)	0.631 (0.012)	0.129 (0.011)	0.631 (0.012)	0.129 (0.011)
	Token-Entropy	0.686 (0.012)	0.088 (0.010)	0.637 (0.013)	0.127 (0.011)	0.662 (0.012)	0.080 (0.010)	0.673 (0.012)	0.072 (0.010)	0.639 (0.012)	0.108 (0.010)	0.618 (0.013)	0.123 (0.011)	0.632 (0.012)	0.123 (0.011)	0.632 (0.012)	0.123 (0.011)	0.632 (0.012)	0.123 (0.011)
	VIG-TUQ-A (ours)	0.722 (0.011)	0.085 (0.010)	0.680 (0.012)	0.103 (0.011)	0.674 (0.012)	0.076 (0.010)	0.686 (0.012)	0.059 (0.010)	0.687 (0.012)	0.089 (0.011)	0.641 (0.013)	0.113 (0.011)	0.669 (0.012)	0.089 (0.011)	0.641 (0.013)	0.113 (0.011)	0.669 (0.012)	0.089 (0.011)
	VIG-TUQ-JSD (ours)	0.721 (0.012)	0.026 (0.010)	0.680 (0.012)	0.112 (0.011)	0.645 (0.012)	0.086 (0.010)	0.679 (0.012)	0.058 (0.009)	0.659 (0.012)	0.111 (0.011)	0.603 (0.013)	0.157 (0.011)	0.660 (0.012)	0.111 (0.011)	0.603 (0.013)	0.157 (0.011)	0.660 (0.012)	0.111 (0.011)
	VIG-TUQ (ours)	0.727 (0.011)	0.085 (0.010)	0.685 (0.012)	0.105 (0.011)	0.673 (0.012)	0.076 (0.010)	0.692 (0.012)	0.059 (0.010)	0.668 (0.012)	0.110 (0.011)	0.641 (0.013)	0.113 (0.011)	0.669 (0.012)	0.110 (0.011)	0.641 (0.013)	0.113 (0.011)	0.669 (0.012)	0.110 (0.011)
	SE	0.694 (0.012)	0.035 (0.010)	0.675 (0.012)	0.118 (0.011)	0.657 (0.012)	0.039 (0.010)	0.676 (0.012)	0.020 (0.010)	0.644 (0.012)	0.067 (0.010)	0.673 (0.012)	0.062 (0.011)	0.651 (0.012)	0.067 (0.010)	0.673 (0.012)	0.062 (0.011)	0.651 (0.012)	0.067 (0.010)
VIZWIZ	KLE	0.716 (0.012)	0.036 (0.010)	0.687 (0.012)	0.118 (0.011)	0.689 (0.012)	0.042 (0.010)	0.699 (0.012)	0.028 (0.010)	0.672 (0.012)	0.066 (0.010)	0.699 (0.012)	0.065 (0.011)	0.614 (0.013)	0.066 (0.010)	0.699 (0.012)	0.065 (0.011)	0.614 (0.013)	0.066 (0.010)
VQARAD	Log Perplexity	0.666 (0.027)	0.176 (0.021)	0.646 (0.030)	0.255 (0.019)	0.660 (0.028)	0.217 (0.021)	0.660 (0.029)	0.219 (0.021)	0.599 (0.027)	0.179 (0.020)	0.687 (0.025)	0.157 (0.022)	0.612 (0.026)	0.179 (0.020)	0.687 (0.025)	0.157 (0.022)	0.612 (0.026)	0.179 (0.020)
	Token-Sar	0.611 (0.027)	0.180 (0.020)	0.637 (0.030)	0.254 (0.019)	0.656 (0.028)	0.216 (0.022)	0.589 (0.030)	0.221 (0.022)	0.599 (0.027)	0.173 (0.022)	0.595 (0.027)	0.173 (0.022)	0.589 (0.027)	0.173 (0.022)	0.595 (0.027)	0.173 (0.022)	0.589 (0.027)	0.173 (0.022)
	CCP	0.661 (0.027)	0.148 (0.022)	0.653 (0.030)	0.233 (0.020)	0.608 (0.030)	0.196 (0.022)	0.595 (0.030)	0.222 (0.022)	0.587 (0.027)	0.173 (0.022)	0.657 (0.026)	0.148 (0.022)	0.589 (0.027)	0.173 (0.022)	0.657 (0.026)	0.148 (0.022)	0.589 (0.027)	0.173 (0.022)
	Mac-Prob	0.674 (0.026)	0.164 (0.021)	0.649 (0.029)	0.252 (0.019)	0.618 (0.030)	0.222 (0.020)	0.602 (0.029)	0.222 (0.020)	0.568 (0.028)	0.202 (0.022)	0.664 (0.026)	0.192 (0.022)	0.611 (0.026)	0.202 (0.022)	0.664 (0.026)	0.192 (0.022)	0.611 (0.026)	0.202 (0.022)
	Token-Entropy	0.669 (0.027)	0.171 (0.021)	0.639 (0.031)	0.258 (0.019)	0.659 (0.028)	0.225 (0.020)	0.603 (0.030)	0.217 (0.021)	0.605 (0.027)	0.177 (0.021)	0.606 (0.025)	0.154 (0.022)	0.624 (0.026)	0.177 (0.021)	0.606 (0.025)	0.154 (0.022)	0.624 (0.026)	0.177 (0.021)
	VIG-TUQ-A (ours)	0.712 (0.026)	0.168 (0.021)	0.676 (0.030)	0.256 (0.019)	0.672 (0.028)	0.260 (0.021)	0.598 (0.029)	0.253 (0.021)	0.663 (0.027)	0.176 (0.024)	0.634 (0.026)	0.176 (0.024)	0.634 (0.026)	0.176 (0.024)	0.634 (0.026)	0.176 (0.024)	0.634 (0.026)	0.176 (0.024)
	VIG-TUQ-JSD (ours)	0.616 (0.027)	0.194 (0.021)	0.540 (0.030)	0.205 (0.019)	0.622 (0.028)	0.245 (0.021)	0.676 (0.029)	0.230 (0.021)	0.683 (0.027)	0.246 (0.021)	0.636 (0.026)	0.230 (0.021)	0.631 (0.026)	0.246 (0.021)	0.636 (0.026)	0.230 (0.021)	0.631 (0.026)	0.246 (0.021)
	VIG-TUQ (ours)	0.672 (0.026)	0.168 (0.021)	0.676 (0.030)	0.256 (0.019)	0.693 (0.028)	0.215 (0.021)	0.670 (0.028)	0.201 (0.021)	0.665 (0.027)	0.149 (0.022)	0.736 (0.024)	0.154 (0.022)	0.644 (0.026)	0.149 (0.022)	0.736 (0.024)	0.154 (0.022)	0.644 (0.026)	0.149 (0.022)
	SE	0.685 (0.026)	0.168 (0.021)	0.632 (0.030)	0.198 (0.019)	0.630 (0.029)	0.258 (0.021)	0.683 (0.029)	0.246 (0.021)	0.611 (0.027)	0.125 (0.021)	0.669 (0.026)	0.145 (0.021)	0.651 (0.026)	0.145 (0.021)	0.651 (0.026)	0.145 (0.021)	0.651 (0.026)	0.145 (0.021)
VQARAD	KLE	0.679 (0.026)	0.169 (0.021)	0.630 (0.030)	0.198 (0.019)	0.644 (0.029)	0.259 (0.021)	0.678 (0.029)	0.247 (0.021)	0.620 (0.027)	0.126 (0.021)	0.666 (0.026)	0.144 (0.021)	0.679 (0.026)	0.144 (0.021)	0.679 (0.026)	0.144 (0.021)	0.679 (0.026)	0.144 (0.021)
ADVQA	Log Perplexity	0.568 (0.015)	0.143 (0.013)	0.614 (0.015)	0.168 (0.012)	0.638 (0.014)	0.108 (0.012)	0.617 (0.015)	0.138 (0.012)	0.624 (0.015)	0.171 (0.012)	0.599 (0.016)	0.253 (0.011)	0.565 (0.015)	0.174 (0.012)	0.565 (0.015)	0.174 (0.012)	0.565 (0.015)	0.174 (0.012)
	Token-Sar	0.531 (0.016)	0.193 (0.013)	0.574 (0.015)	0.186 (0.012)	0.629 (0.014)	0.125 (0.011)	0.581 (0.015)	0.140 (0.011)	0.598 (0.015)	0.182 (0.011)	0.506 (0.015)	0.249 (0.011)	0.531 (0.015)	0.199 (0.011)	0.531 (0.015)	0.199 (0.011)	0.531 (0.015)	0.199 (0.011)
	CCP	0.642 (0.015)	0.119 (0.012)	0.619 (0.015)	0.147 (0.012)	0.609 (0.014)	0.181 (0.012)	0.617 (0.015)	0.140 (0.012)	0.613 (0.015)	0.191 (0.012)	0.591 (0.016)	0.237 (0.011)	0.582 (0.015)	0.237 (0.011)	0.582 (0.015)	0.237 (0.011)	0.582 (0.015)	0.237 (0.011)
	Mac-Prob	0.642 (0.015)	0.185 (0.013)	0.594 (0.015)	0.197 (0.012)	0.636 (0.014)	0.142 (0.013)	0.624 (0.015)	0.146 (0.012)	0.633 (0.015)	0.194 (0.013)	0.665 (0.015)	0.237 (0.011)	0.554 (0.015)	0.237 (0.011)	0.554 (0.015)	0.237 (0.011)	0.554 (0.015)	0.237 (0.011)
	Token-Entropy	0.642 (0.015)	0.148 (0.012)	0.612 (0.015)	0.175 (0.012)	0.639 (0.014)	0.128 (0.012)	0.613 (0.015)	0.143 (0.012)	0.624 (0.015)	0.176 (0.012)	0.594 (0.016)	0.250 (0.011)	0.556 (0.015)	0.250 (0.011)	0.556 (0.015)	0.250 (0.011)	0.556 (0.015)	0.250 (0.011)
	VIG-TUQ-A (ours)	0.668 (0.014)	0.123 (0.013)	0.626 (0.015)	0.163 (0.012)	0.651 (0.014)	0.116 (0.012)	0.624 (0.015)	0.137 (0.012)	0.644 (0.015)	0.167 (0.012)	0.617 (0.016)	0.237 (0.011)	0.527 (0.014)	0.237 (0.011)	0.527 (0.014)	0.237 (0.011)	0.527 (0.014)	0.237 (0.011)
	VIG-TUQ-JSD (ours)	0.642 (0.015)	0.150 (0.013)	0.580 (0.015)	0.143 (0.012)	0.613 (0.014)	0.133 (0.012)	0.613 (0.015)	0.142 (0.012)	0.590 (0.015)	0.159 (0.012)	0.589 (0.015)	0.242 (0.011)	0.567 (0.015)	0.242 (0.011)	0.567 (0.015)	0.242 (0.011)	0.567 (0.015)	0.242 (0.011)
	VIG-TUQ (ours)	0.662 (0.014)	0.131 (0.012)	0.632 (0.015)	0.159 (0.012)	0.662 (0.014)	0.112 (0.012)	0.623 (0.015)	0.133 (0.012)	0.642 (0.015)	0.165 (0.012)	0.617 (0.016)	0.237 (0.011)	0.567 (0.014)	0.237 (0.011)	0.567 (0.014)	0.237 (0.011)	0.567 (0.014)	0.237 (0.011)
	SE	0.61 (0.015)	0.214 (0.012)	0.651 (0.015)	0.195 (0.012)	0.695 (0.014)	0.135 (0.012)	0.682 (0.015)	0.141 (0.012)	0.662 (0.015)	0.197 (0.012)	0.659 (0.015)	0.243 (0.011)	0.517 (0.015)	0.243 (0.011)	0.517 (0.015)			

Table A.8

	Qwen2.5-VL-3B			Qwen2.5-VL-7B			llava-1.5-13B			llava-1.5-7B			idefics2-8B			idefics-9B			Qwen3.5-2B		
Dataset	α_{JSD}	α_A	l	α_{JSD}	α_A	l	α_{JSD}	α_A	l	α_{JSD}	α_A	l	α_{JSD}	α_A	l	α_{JSD}	α_A	l	α_{JSD}	α_A	l
OKVQA	1	4	7	0	1	4	2	3	6	2	3	16	1	4	29	0	1	16	0	1	5
VIZWIZ	3	1	9	2	3	15	1	4	16	3	4	8	1	1	20	0	1	14	1	3	5
VQARAD	0	1	6	0	1	6	3	1	3	1	0	18	3	2	0	0	1	17	1	1	5
ADVQA	1	4	7	1	3	4	1	4	6	1	3	4	1	4	1	0	1	14	1	3	0

when the image is removed. In other words, confident predictions tend to rely more on visual information. Another observation is that the distance generally increases in deeper layers of the network. This suggests that the influence of visual information becomes stronger in later stages of the Transformer.

Analysis wrt prediction correctness. Figure A.7 compares the distance between representations for correct and incorrect predictions. For all models, we also observe that the distance between the two representations is larger for correct predictions than for incorrect ones. This indicates that when the model produces the right answer, its internal representation is more strongly affected by removing the image. In other words, correct predictions tend to rely more on visual information. The separation between correct and incorrect predictions is generally slightly smaller than the one observed between certain and uncertain predictions, suggesting that visual reliance is more directly reflected in model confidence than in correctness. As in the certainty analysis, the distance also tends to increase in deeper layers of the network, indicating that visual information has a stronger influence in later Transformer layers.

Additional details on the spider plot. Figure 1 provides a compact aggregation of the layer-wise analyses reported in Figures A.6 and A.7. To summarize each comparison in the spider plot, we select the layer at which the absolute gap between the two group-wise mean curves is maximal. This selected layer corresponds to the dashed vertical line in Figures A.7 and A.6. We then extract the two group values at this layer, jointly normalize them within the corresponding comparison, and use them as the radial values for the associated model in Figure 1. Thus, each spoke in Figure 1 summarizes, for one model, the strongest layer-wise separation between the two groups, with larger polygon separation indicating that the image-induced representation shift is more discriminative of correctness or certainty.

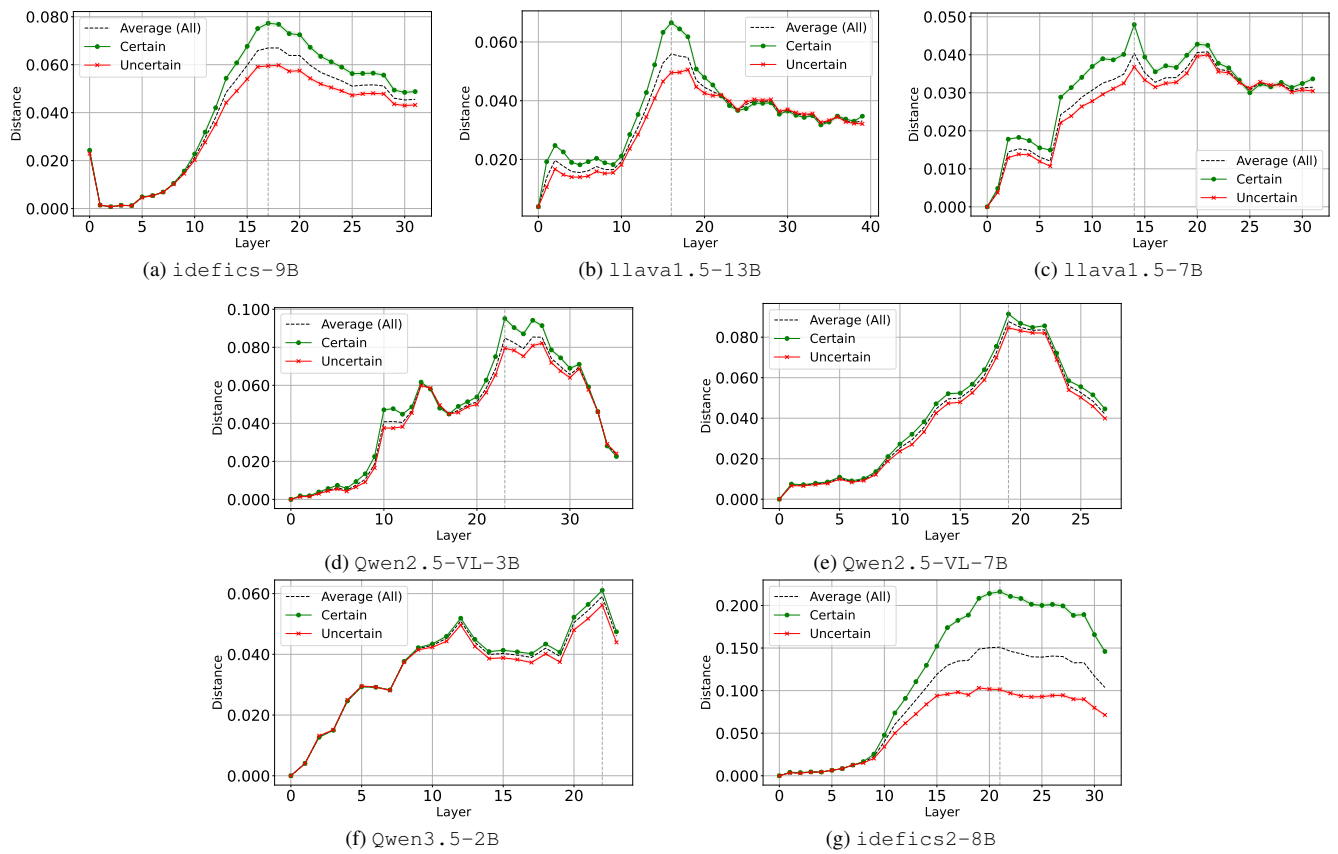


Figure A.6. **Layer-wise distance between hidden states from the original generation pass and the teacher-forced pass with a masked image.** The curves show the average similarity across layers for certain and uncertain samples for different LLM architectures.

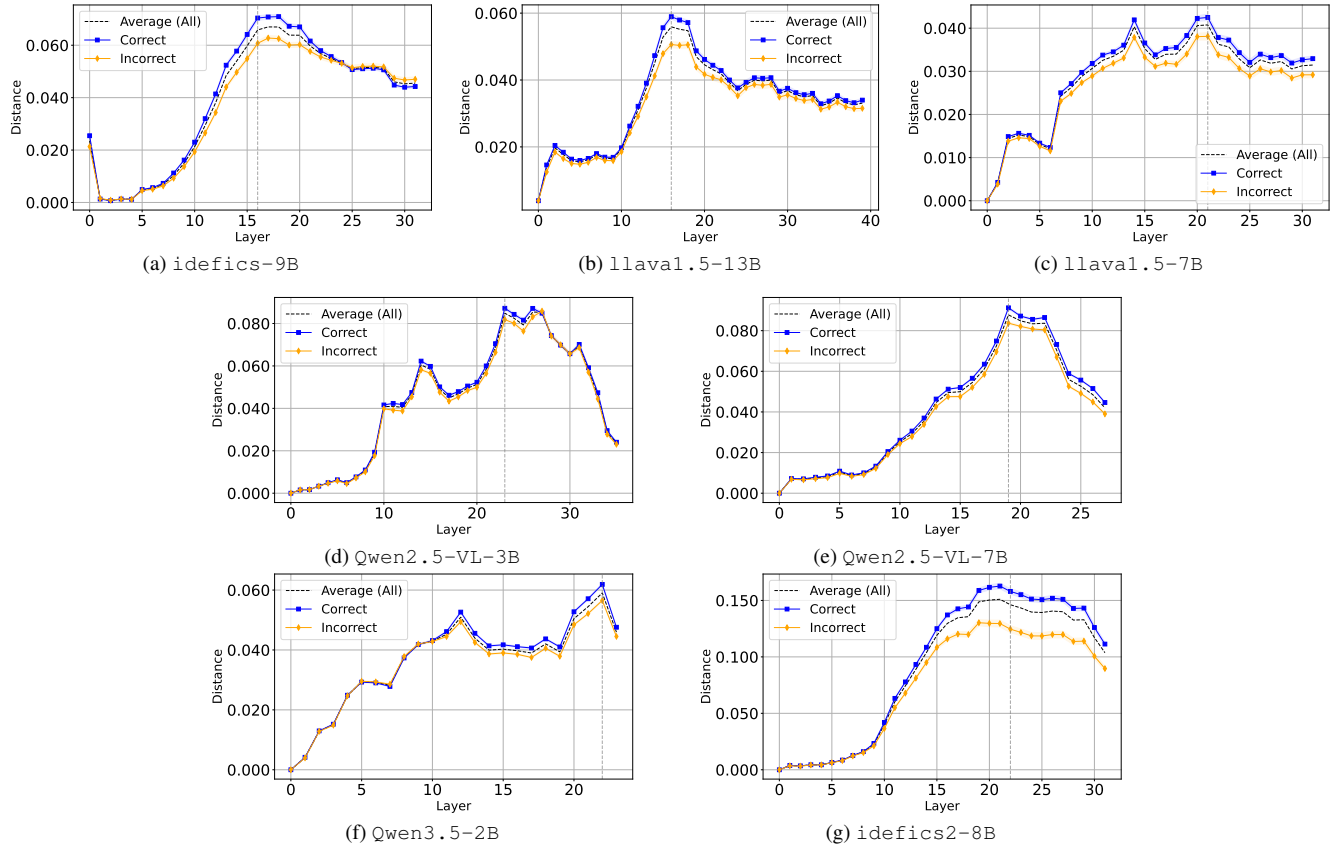


Figure A.7. **Layer-wise distance between hidden states from the original generation pass and the teacher-forced pass with a masked image.** The curves show the average similarity across layers for correct and incorrect samples for different LLM architectures.