Towards Statistical Factuality Guarantee for Large Vision-Language Models

Zhuohang Li[1], Chao Yan[2], Nicholas J. Jackson[1], Wendi Cui[3], Bo Li[4,5], Jiaxin Zhang[6]*, Bradley A. Malin[1,2]

[1]Vanderbilt University [2]Vanderbilt University Medical Center [3]Intuit
[4]University of Illinois Urbana-Champaign [5]Virtue AI [6]Salesforce AI Research
{zhuohang.li, b.malin}@vanderbilt.edu

Abstract

Advancements in Large Vision-Language Models (LVLMs) have demonstrated impressive performance in image-conditioned text generation; however, hallucinated outputs–text that misaligns with the visual input–pose a major barrier to their use in safety-critical applications. We introduce ConfLVLM, a conformal-prediction-based framework that achieves finite-sample distribution-free statistical guarantees to the factuality of LVLM output. Taking each generated detail as a hypothesis, ConfLVLM statistically tests factuality via efficient heuristic uncertainty measures to filter out unreliable claims. We conduct extensive experiments covering three representative application domains: general scene understanding, medical radiology report generation, and document understanding. Remarkably, ConfLVLM reduces the error rate of claims generated by `LLaVa-1.5` for scene descriptions from 87.8% to 10.0% by filtering out erroneous claims with a 95.3% true positive rate. Our results further show that ConfLVLM is highly flexible, and can be applied to any black-box LVLMs paired with any uncertainty measure for any image-conditioned free-form text generation task while providing a rigorous guarantee on controlling hallucination risk.

1 Introduction

Large Vision-Language Models (LVLMs) which combine Large Language Models (LLMs) with computer vision modules, have demonstrated remarkable multi-modal abilities (Liu et al., 2024b; Abdin et al., 2024; Meta; OpenAI). LVLMs are designed to receive both free text and visual content as inputs (e.g., images or videos) and generate text responses to user queries about the visual input, thus enabling a flexible conversational interface for numerous visual perception and multi-modal comprehension tasks. This capability has triggered successful developments across various application domains, including cross-modality agents specializing in graphical user interface understanding and planning (Hong et al., 2024), visual-language foundation models for analyzing pathology slides (Lu et al., 2024), and autonomous driving assistants offering real-time decision-making (Wen et al., 2023). Despite the excitement and new opportunities LVLMs offer, they can produce *hallucinations* (Bai et al., 2024)—text outputs that contain erroneous, or simply fabricated, assertions that deviate from the visual content. This can significantly limit their adoption in safety-critical fields, such as healthcare and autonomous driving.

Related Work. To date, the research on LVLM hallucinations has generally focused on two threads. One line of work aims to build benchmarks (Li et al., 2023; Yin et al., 2023a; Lovenia et al., 2023; Sun et al., 2023; Kaul et al., 2024; Guan et al., 2024; Jing et al., 2023) and metrics (Rohrbach et al., 2018; Li et al., 2023; Jing et al., 2023) for assessing and analyzing hallucinations in popular LVLMs. Most studies in this area concentrate on object hallucination in discriminative question-answering tasks (Li et al., 2023; Yin et al., 2023a; Lovenia et al., 2023), with a lesser focus on other types of hallucination in free-text generation tasks (Jing et al., 2023; Kaul et al., 2024; Guan et al., 2024). The other line of work aims to develop strategies to mitigate LVLM hallucinations. Notable strategies in this line of investigation include refining the training and instruction tuning phases by optimizing the loss function (Jiang et al., 2024) or alignment objective (Zhao et al., 2023; Gunjal et al., 2024; Sun et al., 2023), as well as enhancing the inference phase by designing new decoding algorithms (Leng et al., 2024; Favero et al., 2024; Deng et al., 2024). However, these solutions are often resource-heavy and lack flexibility, typically requiring model retraining or white-box access during decoding. By

*Work done while at Intuit.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11435–11456
November 4-9, 2025

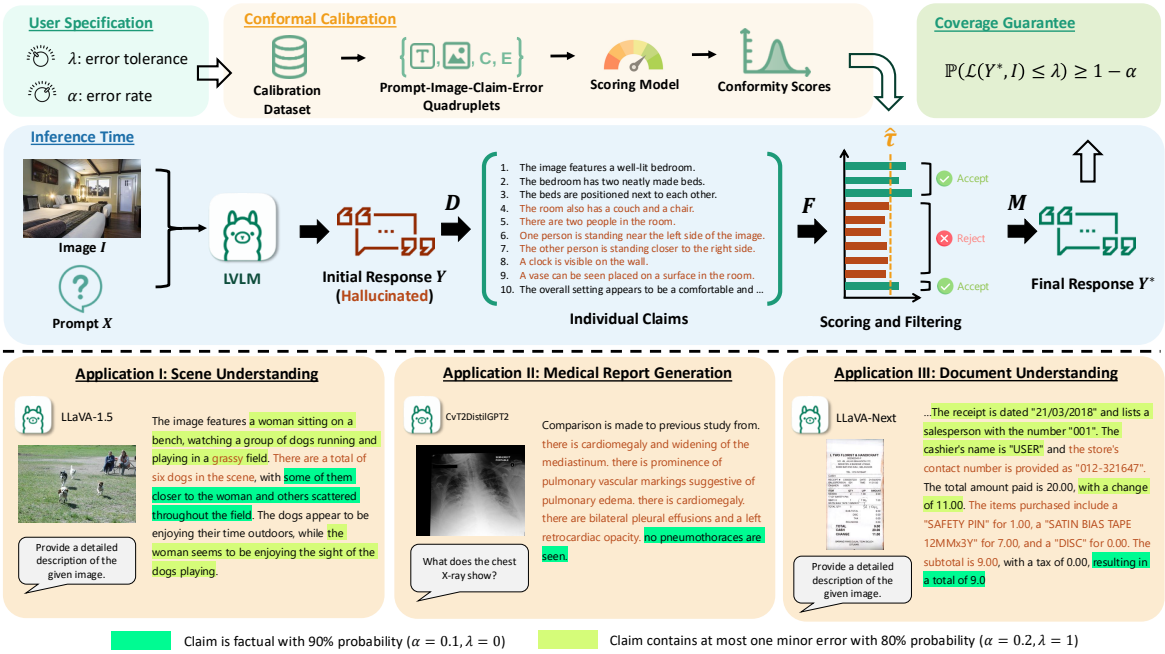


Figure 1: **Overview**: given user-specified error tolerance λ, error rate α, and a calibration dataset, CONFLVLM returns a more reliable response for any new image and prompt at inference time through *sampling*, *decomposing* D, *filtering* F, and *merging* M, to ensure that the risk of the final response Y^* is controlled with high probability. Illustrative examples, one for each application domain, are provided for outcome demonstration, where claims are highlighted to indicate CONFLVLM's confidence using specific conformity score and error tolerance level. Unhighlighted claims correspond to low confidence in factuality check.

contrast, *post hoc* correction methods offer a more flexible approach, which improves responses from any black-box LVLMs through the assistance of external language or vision modules (Yin et al., 2023b), dedicated revisor model (Zhou et al., 2023), or self-revision techniques (Lee et al., 2023). However, the current collection of methods relies solely on heuristics and lacks rigorous statistical guarantees of factuality for the revised output, a necessity in mission-critical application domains.

Our Contribution. To address these challenges, in this study, we introduce CONFLVLM, a framework with statistical guarantees on output factuality (alignment of text response with visual context) that seamlessly integrates with LVLMs of any architecture, complexity, and purpose. CONFLVLM treats LVLM-generated free text as a series of individual claims, each corresponding to a testable hypothesis. Each claim is then evaluated for its factuality using discriminative mechanisms built from the same (or auxiliary) LVLM to filter out unsupported claims and retain those that meet factual standards. To provide statistical guarantees on the factuality of retained claims, CONFLVLM employs a conformal prediction framework that allows control for flexible error rates, as well as error tolerance levels, defined by users to suit specific application needs. By leveraging the generally more stable discriminative capabilities of LVLMs through calculating and ranking predefined conformity scores, CONFLVLM enables quantitative factual assessments. We demonstrate the effectiveness of CONFLVLM across three representative domains, i.e., general scene understanding (Section 4), medical radiology report generation (Section 5), and document understanding (Section 6), to validate its effectiveness in ensuring a desired level of response factuality for various state-of-the-art LVLMs and explore multiple potentially useful conformity scores. Our results, covering over $81,000$ claims generated from eight popular LVLMs, establish CONFLVLM as a general-purpose framework that operates with any LVLM in an assumption- and finetuning-free manner, thus promoting trustworthiness in LVLM applications broadly. Moreover, this work opens up a novel research space, where each component of CONFLVLM invites further innovation to improve the rigorous guarantee of hallucination mitigation for multi-modal models.

2 Preliminaries

2.1 Large Vision-Language Models

Large Vision-Language Models (LVLM) are generative models that are typically composed of a visual model $h(\cdot)$, a language model parameterized by $\boldsymbol{\theta}$, and a fusion model $g(\cdot)$. The most popular implementations of LVLMs, such as Llava (Liu et al., 2024b), combine a pre-trained visual encoder (e.g., CLIP (Radford et al., 2021)) and a pre-trained Large Language Model (LLM) (e.g., Vicuna (Chiang et al., 2023)) by training a projection network as the fusion model to convert extracted visual features into the LLM's embedding space in a process known as visual instruction tuning. During inference, an LVLM takes an input image $\boldsymbol{I}$ and a text prompt $\boldsymbol{X} = [\mathrm{x}_1, ..., \mathrm{x}_l]$, and outputs a text response $\boldsymbol{Y} = [\mathrm{y}_1, ..., \mathrm{y}_m]$, where x_i and y_j are individual tokens. This is achieved by first converting the image into a sequence of visual tokens using the visual model $\boldsymbol{V} = [\mathrm{v}_1, ..., \mathrm{v}_k] = g \circ h(\boldsymbol{I})$ and then sampling the response from the conditional distribution in an autoregressive manner: $p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{V}) = \prod_{j=1}^{m} p_{\boldsymbol{\theta}}(\mathrm{y}_j|\boldsymbol{X}, \boldsymbol{V}, \boldsymbol{Y}_{<j})$.

Hallucination of LVLMs. Hallucination originates from the space of language models, where the generated text response is either non-factual (conflicts with verifiable facts) or unfaithful (does not follow the user's instructions). In the context of LVLMs, hallucination refers to the phenomenon where the generated text response deviates from the provided visual content. Common types of LVLM hallucinations include *object* hallucination (e.g., falsely identifying non-existent objects), *attribute* hallucination (e.g., wrong color, shape, or material), and *relation* hallucination (e.g., human-object interaction, relative position) (Bai et al., 2024).

2.2 Split Conformal Prediction

Split conformal prediction (SCP) (Vovk et al., 2005; Shafer and Vovk, 2008) is a distribution-free method for quantifying the uncertainty of black-box prediction algorithms by constructing prediction sets with finite-sample coverage properties.

Coverage Guarantee. For a black-box prediction function $f : \mathcal{X} \to \mathcal{Y}$, let $\{(X_i, Y_i)\}_{i=1}^{n+1}$ be an exchangeable set of feature and label pairs sampled from the joint distribution on $\mathcal{X} \times \mathcal{Y}$. The goal of split conformal prediction is to use the calibration data $\{(X_i, Y_i)\}_{i=1}^{n}$ and f to construct a prediction set $\hat{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ for the new data point such that it achieves valid *coverage*, i.e., containing the true label with high probability $\mathbb{P}\big(Y_{n+1} \in \hat{C}(X_{n+1})\big) \geq 1 - \alpha$ for any user-specified error rate $\alpha \in (0, 1)$.

Conformal Calibration. Suppose there is a *conformity score* function $S(X, Y) \in \mathbb{R}$ that measures how well a given sample *conforms* to the observed data. The split conformal procedure uses the calibration data set $\{(X_i, Y_i)\}_{i=1}^{n}$ to derive conformity scores $\{S(X_i, Y_i)\}_{i=1}^{n}$, where a larger value indicates the model is more confident about the prediction being true. To calibrate the prediction set to the desired level of coverage, we then compute a threshold $\hat{\tau}$ that is approximately the $1-\alpha$ quantile of the conformity scores. At the time of inference, given a new data point X_{n+1}, we construct the prediction set as $\hat{C}(X_{n+1}) = \{y \in \mathcal{Y} : S(X, y) \geq \hat{\tau}\}$. If the data are exchangeable, then this prediction set will satisfy the desired coverage property.

3 Ensuring Factuality for LVLMs

3.1 Problem Formulation

Given a pair of image and text prompt (I_{n+1}, X_{n+1}) and a set of n calibration data points, our goal is to generate a reliable response Y^*_{n+1} using the LVLM, such that it contains a low error with high probability; i.e.,

$$\mathbb{P}(\mathcal{L}(Y^*_{n+1}, I_{n+1}) \leq \lambda) \geq 1 - \alpha, \qquad (1)$$

where $\mathcal{L} : \mathcal{Y} \times \mathcal{I} \to \mathbb{R}^+_0$ is a monotonic *loss* function that measures the level of misalignment between the statement and the image (e.g., the occurrence of object hallucination, or inaccuracy in item attribute or quantity) and λ is a user-specified tolerance. A larger loss indicates a greater amount of error, while a loss of zero indicates that the statement made in the response is *factual* with respect to the provided image. We define $\mathcal{L}(\varnothing, \cdot) = 0$, which indicates that we do not penalize the model for abstaining from responding when it is uncertain, as here we only focus on the assurance of factuality while neglecting other aspects of reliability such as omission of information. The choice among different loss functions offers the flexibility to provide a statistical guarantee with respect to a *user-defined*, *quantifiable* measure of factuality.

3.2 Error Control in LVLMs

Due to the open-set and free-form nature of the natural language output, directly attempting to construct prediction sets is not attainable for generative models like LVLMs. Instead, we adapt the

recently proposed *conformal factuality* (Mohri and Hashimoto; Cherian et al., 2024) framework to the multi-modal setting as the central tool for achieving statistical guarantees on the factuality of LVLM outputs. The key idea is to exploit the connection between linguistic entailments and uncertainty sets to back off from the original statement (uncensored response from LVLM) by gradually removing unreliable claims with high uncertainty until the desired level of correctness is achieved.

Error Control Procedure. We start by defining a scoring function $r(C^j_{n+1}, I_{n+1}) \in \mathbb{R}$ that captures the system's confidence about the claim concerning the provided visual context, where a larger score indicates that the claim is more aligned with the provided image and, thus, is more likely to be true. Given I_{n+1} and X_{n+1}, we execute the following steps to generate a more reliable response Y^*_{n+1}:

(1) *Initial Hypothesis Generation*: Sample an initial response Y_{n+1} from the LVLM $p_{\boldsymbol{\theta}}(Y|X_{n+1}, V_{n+1})$, where $V_{n+1} = g \circ h(I_{n+1})$.

(2) *Decomposition*: Apply a *decomposition* operator D to breakdown the initial response into a set of individual claims: $\boldsymbol{C}_{n+1} = D(Y_{n+1}) = \{C^j_{n+1}\}_{j=1}^{s_i}$.

(3) *Individual Hypothesis Testing*: Define the *filtering* operator as $F(\boldsymbol{C}_{n+1}; \tau) := \{C^j_{n+1} : r(C^j_{n+1}, I_{n+1}) > \tau\}$. Generate a filtered set of claims $F(\boldsymbol{C}_{n+1}; \tau) \subseteq \boldsymbol{C}_{n+1}$. This step can be thought of as testing each individual hypothesis C^j_{n+1} and accepting the hypothesis only if the test statistic $r(C^j_{n+1}, I_{n+1})$ is greater than the chosen threshold τ.

(4) *Combination*: Apply a *merge* operator M to combine the filtered claims into the final response $Y^*_{n+1} = M\big(F(\boldsymbol{C}_{n+1}; \tau)\big)$.

Calibration. Next, to calibrate the filtering operator F, we set the conformity score S for each set of claims to be the minimum threshold that ensures the loss of the filtered set of claims is controlled to be within tolerance:

$$S(\boldsymbol{C}_{n+1}, I_{n+1}) = \inf\{\tau : \mathcal{L}\big(F(\boldsymbol{C}_{n+1}; \tau), I_{n+1}\big) \leq \lambda\}.$$

Finally, we implement the calibrated filtering operator as $\hat{F}(\boldsymbol{C}_{n+1}) := F(\boldsymbol{C}_{n+1}; \hat{\tau}) = \{C^j_{n+1} : r(I, C^j_{n+1}) > \hat{\tau}\}$, where $\hat{\tau}$ is set to be the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$-th quantile of the conformity scores $\{S(\boldsymbol{C}_i, I_i)\}_{i=1}^n$ estimated on the calibration dataset.

The following theorem indicates that if the data are exchangeable, the response produced using the calibrated filtering operator will satisfy the error control objective in Ineq. 1.

Theorem 3.1 (SCP Coverage Guarantee (Shafer and Vovk, 2008))**.** *Define the error scores* $\boldsymbol{E}_i := \{\mathcal{L}(C^j_i, I_i) : C^j_i \in \boldsymbol{C}_i\}$. *Let* $\{(X_i, I_i, \boldsymbol{C}_i, \boldsymbol{E}_i)\}_{i=1}^{n+1}$ *be exchangeable, then the following lower bound holds for any* $\alpha \in (\frac{1}{n+1}, 1)$*:*

$$\mathbb{P}\Big(\mathcal{L}\big(\hat{F}(\boldsymbol{C}_{n+1}), I_{n+1}\big) \leq \lambda\Big) \geq 1 - \alpha.$$

If the loss function is monotonic, meaning that $\mathcal{L}(\hat{F}_1(\boldsymbol{C}_i, I_i)) \leq \mathcal{L}(\hat{F}_2(\boldsymbol{C}_i, I_i))$ *for any* $\hat{F}_1(\boldsymbol{C}_i, I_i) \subseteq \hat{F}_2(\boldsymbol{C}_i, I_i)$*, then the following upper bound also holds:*

$$1 - \alpha + \frac{1}{n+1} \geq \mathbb{P}\Big(\mathcal{L}\big(\hat{F}(\boldsymbol{C}_{n+1}), I_{n+1}\big) \leq \lambda\Big).$$

Proof. Without loss of generality, let us assume that the conformity scores are sorted as $s_1 < s_2 < ...s_n$, where $s_i = S(\boldsymbol{C}_i, I_i)$. Notice that under the definition of S, the event $\{s_{n+1} \leq \hat{\tau}\}$ implies $\{\mathcal{L}\big(\hat{F}(\boldsymbol{C}_{n+1}), I_{n+1}\big) \leq \lambda\}$. By exchangeability, $\mathbb{P}(s_{n+1} \leq s_{\lceil (n+1)(1-\alpha) \rceil}) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \geq 1 - \alpha$, which implies the result. To prove the upper bound, notice that the two events $\{s_{n+1} \leq \hat{\tau}\}$ and $\{\mathcal{L}\big(\hat{F}(\boldsymbol{C}_{n+1}), I_{n+1}\big) \leq \lambda\}$ are now equivalent if the loss function is monotone. The result can then be obtained through $\mathbb{P}(s_{n+1} \leq s_{\lceil (n+1)(1-\alpha) \rceil}) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \leq \frac{(1-\alpha)(n+1)+1}{n+1} = 1 - \alpha + \frac{1}{n+1}$. □

3.3 Deriving Conformity Scores

In practice, the decomposition operator can be implemented by prompting the language model part of the LVLM, which does not rely on any external resources. To derive the conformity scores, we will need to find a suitable scoring function $r(C_i, I_i)$. Built on conformal prediction, our framework should maintain valid coverage with any arbitrary heuristic scores. However, in practice, a score that better captures the relevance between a claim C_i and the given image I_i can enable a better tradeoff (i.e., allowing the same coverage guarantee while filtering out less content). We primarily consider the following two types of scores.

Internal Scores. We consider the following scores to capture the internal confidence of an LVLM regarding a statement. (1) *Log Probability of Text Tokens*: we compute the log probability of text tokens from the claim given only the text prompt as $r(C_i, I_i) = \log p_{\boldsymbol{\theta}}(C_i|X_i)$, which does not make use of the visual context and thus serves as a language prior baseline. (2) *Log Probability of Text Tokens Conditioned on Image*: we compute the log probability of text tokens from the claim conditioned on both the visual and the text prompt as $r(C_i, I_i) = \log p_{\boldsymbol{\theta}}(C_i|X_i, V_i)$, which is the visual instruction tuning objective (Liu et al., 2024b) of most LVLMs. (3) *Log Probability Ratio*: finally, we consider the ratio between the two probabilities, i.e., $r(C_i, I_i) = \log \frac{p_{\boldsymbol{\theta}}(C_i|X_i,V_i)}{p_{\boldsymbol{\theta}}(C_i|X_i)}$. This is motivated by the observation that most hallucinations in LVLMs occur because their language prior tend to dominate visual perception during decoding (Favero et al., 2024; Leng et al., 2024; Liu et al., 2024c), and the probability ratio can be an informative measure of the true influence of the visual prompt regardless of the language prior.

External Scores. In addition to internal scores, we also consider capturing the confidence in a statement regarding an image using a (lightweight) external model. *Energy-based models* (LeCun et al., 2006) $E : \mathcal{Y} \times \mathcal{I} \rightarrow \mathbb{R}$ are particularly suited for this task as they are trained to map image-text pairs to a scalar energy score so that $e^{-E(Y,I)} \propto \mathbb{P}(Y, I)$. As such, we can simply set the scoring function to return the negative energy score, $r(C_i, I_i) = -E(C_i, I_i)$.

4 Case Study I: General Scene Understanding

4.1 Setup

We first evaluate the LVLM factuality framework for general scene understanding tasks. To do so, we use 500 randomly selected images from the MSCOCO (Lin et al., 2014) validation set with more than three objects (same as the POPE dataset (Li et al., 2023)).

LVLMs. We use four start-of-the-art LVLMs for the evaluation of the scene understanding task, including three open-sourced models `LLaVA-1.5` (Liu et al., 2024b), `Phi-3.5-vision-instruct` (Abdin et al., 2024), `Llama-3.2-11B-Vision` (Meta), and one close-sourced model `GPT-4o-mini` (OpenAI). We prompt each LVLM to generate a detailed description for each image and decompose the original response into independent claims. Our prompts are provided in the Appendix.

Error Annotation and Loss Function. We consider the following five types of errors for scene understanding: (1) *Object identification*: The claim involves hallucinated or wrongly identified objects; (2) *Attribute (in)accuracy*: The claim involves incorrect attributes (e.g., color, size, shape); (3) *Spatial relations*: The claim involves incorrect spatial relationships between objects; (4) *Interaction/Action (in)accuracy*: The claim involves incorrect or hallucinated action or interaction; and (5) *Quantitative information*: The claim involves incorrect numeric details (e.g., the wrong object count). We prompt `GPT-4o` to label each claim as either correct or belonging to one or more error categories. In practice, the loss function can be tailored according to specific use cases. For our experiments, we consider a cumulative loss function that assigns a loss score to each response based on the total error its claims contain. Specifically, all claims start with a loss score equal to 0. For each "Object identification" error contained in a claim, the loss is increased by 3; for every other type of error contained, the loss is increased by 1. A correct claim thus receives a loss of 0. This choice of loss structure reflects the common consensus that hallucinating non-existing objects is typically more harmful compared to other types of hallucinations.

Scoring Function. We use two configurations of pretrained CLIP (Radford et al., 2021) models, `CLIP-ViT-Base` with $32px$ patch size and `CLIP-ViT-Large` with $14px$ patch size, to derive the normalized image and text embeddings and compute the dot product as external confidence scores.

4.2 Results

ConfLVLM Achieves Any Desired Level of Coverage. We examine the validity of ConfLVLM by measuring the empirical coverage (ratio of responses that satisfy $\mathcal{L}\big(\hat{F}(\boldsymbol{C}_{n+1}), I_{n+1}\big)\big) \leq \lambda$ over total number of responses) under different levels of desired (theoretical) coverage determined by $1 - \alpha$. We set the error tolerance λ to 0 (most restrictive) and report the results over 50 random splits of calibration (400 data points for establishing conformal prediction) and test data (100 data points for computing the empirical coverage). The

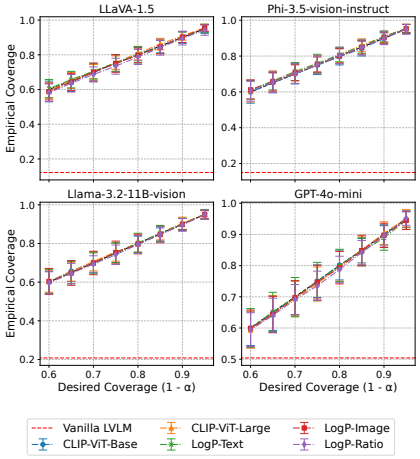


Figure 2: Alignment between empirical and desired (theoretical) coverage in *scene understanding* (with $\lambda = 0$). Vanilla LVLM (red dashed line) refers to the base setting where the LVLM-generated responses are returned to users without using ConfLVLM.

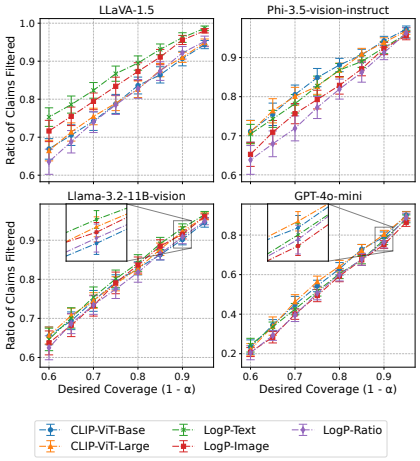


Figure 3: Average ratio of claims filtered with varying coverage using different scoring functions in *scene understanding* (with $\lambda = 0$). Standard errors are marked.

results in Fig. 2 confirm that ConfLVLM can achieve the desired level of coverage with all types of scoring functions and for all LVLMs considered. By contrast, Vanilla LVLM (i.e., responses without filtration) leads to significantly low coverage that signals a failure in the model's reliability.

ConfLVLM Achieves Higher Filtering Efficiency Than Baselines. Approaching the desired coverage level of $1 - \alpha$ involves flagging and filtering out low-confidence claims and, in some cases, abstaining from providing a response. Next, we analyze the ratio of claims being filtered out by ConfLVLM (Fig. 3) and the associated rate of abstention (Fig. 4). We observe a general trend where the ratio of filtered claims and the abstention rate increase as the desired coverage level rises. This is expected as the unfiltered responses

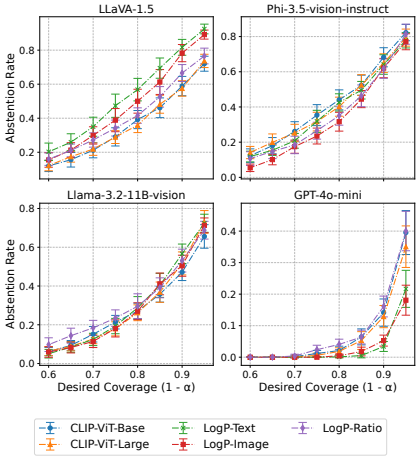


Figure 4: Abstention rate with varying coverage using different scoring functions in *scene understanding* (with $\lambda = 0$).

from LVLMs contain a large number of non-factual claims (e.g., 87.8% responses from `LLaVA-1.5` are erroneous), and thus ensuring a lower error rate requires the framework to be more conservative and filter more content. When comparing across different scoring functions, it can be seen that, in the case of `LLaVA-1.5`, the external confidence scores based on `CLIP` models achieve a lower ratio of filtered claims and also a lower abstention rate than internal scores. This is notable because `LLaVA-1.5` uses `CLIP` as the visual encoder, which implies that there may be certain deficiencies in the visual instruction tuning process. However, such a trend is weaker or non-existent with other LVLMs. This may be because claims generated by other LVLMs are typically longer and contain more details, which is more difficult to capture using `CLIP`. When comparing across LVLMs, we see that `GPT-4o-mini` requires filtering out much fewer claims to achieve the desired coverage, which is because `GPT-4o-mini` has better empirical factuality performance compared to other models.

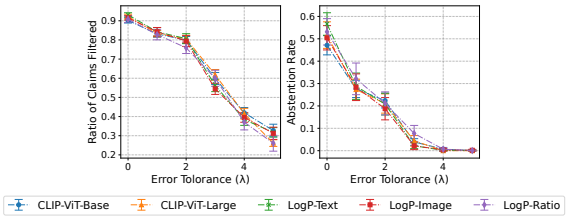
Now we analyze how accurate ConfLVLM is in filtering out nonfactual claims. To our knowledge, there are no existing baselines with statistical factuality guarantees. Thus, we consider a simple *Random Filtering* baseline beyond Vanilla LVLM, which drops claims uniformly at random with probability α. As shown in Table 1, for `LLaVA-1.5` with the setting of $\alpha = 0.1, \lambda = 0$, ConfLVLM achieves a high true positive rate (TPR, or Recall) of 0.953 and a relatively high F1 score of 0.504. Vanilla LVLM, in contrast, has a TPR of 0 and F1 of 0, as it does not perform any filtration. When randomly dropping 10% claims,

Table 1: Claim-level results on the *scene understanding* task using `CLIP-ViT-Large` as scoring function (ConfLVLM with $\alpha = 0.1, \lambda = 0$).

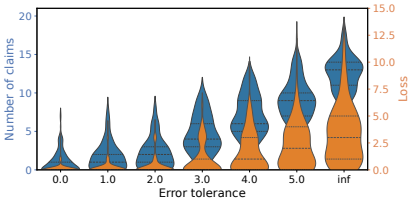
LVLM	Configuration	TPR↑	F1↑
`LLaVA-1.5`	Vanilla LVLM	0.0	0.0
	Random Filtering	0.104	0.158
	ConfLVLM	**0.953**	**0.504**
`Phi-3.5-vision -instruct`	Vanilla LVLM	0.0	0.0
	Random Filtering	0.102	0.145
	ConfLVLM	**0.945**	**0.401**
`Llama-3.2-11B-vision`	Vanilla LVLM	0.0	0.0
	Random Filtering	0.099	0.121
	ConfLVLM	**0.936**	**0.269**
`GPT-4o-mini`	Vanilla LVLM	0.0	0.0
	Random Filtering	0.106	0.070
	ConfLVLM	**0.850**	**0.100**

Random Filtering reaches a TPR of 0.104 and an F1 of 0.158, which are approximately 9× and 3.2× lower than those achieved by ConfLVLM. We have similar observations for other LVLMs and these highlight the utility of ConfLVLM in identifying nonfactual claims. It is also evident that other LVLMs with ConfLVLM implemented demonstrate lower values of TPR and F1 compared to `LLaVA-1.5`, e.g., `GPT-4o-mini` achieves a TPR of 0.850 and an F1 of 0.100. This occurs because models like `GPT-4o-mini` typically generate more comprehensive descriptions of an image, leading to a higher number of claims, and, at the same time, a smaller proportion of nonfactual claims (examples in Appendix). In other words, the task of identifying nonfactual claims itself is much harder for `GPT-4o-mini` than for `LLaVA-1.5`. Nonetheless, ConfLVLM still outperforms the baselines.

Error Tolerance Allows Flexible Control Over Coverage-Utility Tradeoff. To study the impact of error tolerance, we plot the ratio of claims filtered and abstention rate using `Llama-3.2-11B-Vision` with varying λ while keeping $\alpha = 0.1$ in Fig. 5a. We observe the expected behavior that ConfLVLM will filter out less content and abstain less as the error tolerance increases. In Fig. 5b, we plot the model's response distributions with different error tolerances and using `CLIP-ViT-Large` as the scoring function. When the error tolerance is set to ∞, i.e., using the raw LVLM output without censoring, the model can generate detailed responses (75% responses have ≥ 10 claims) but also have high risks of hallucination (more than 50% responses have loss ≥ 3). As the error tolerance decreases, the loss of responses gradually reduces with the number of claims. For example, an error tolerance is set to 3, which reduces more than 75% of responses to



(a) Ratio of claims filtered and abstention rate



(b) Response distributions using `CLIP-ViT-Large` as scoring function

Figure 5: Comparison of `Llama-3.2-11B-Vision`'s response with different error tolerances (λ) while fixing $\alpha = 0.1$ in *scene understanding*.

below 2, with the median of the number of claims being 4. This shows that error tolerance can serve as an additional tuning knob that allows the user to flexibly choose the desired level of error within the acceptable range of utility.

5 Case Study II: Medical Report Generation

5.1 Setup

Next, we evaluate ConfLVLM on the radiology report generation task. We sample 500 chest X-ray images from the MIMIC-CXR (Johnson et al., 2019) dataset, each from a distinct patient.

LVLMs. We consider the following three medical-domain LVLMs for this task: (1) `LlaVa-Med` (Li et al., 2024a) is a biomedical LVLM instruction-tuned on several corpora in the biomedicine domain. We use the latest v1.5 `Mistral` 7B version. (2) `CvT2DistilGPT2` (Nicolson et al., 2023) is LVLM based on the encoder-to-decoder architecture developed for chest X-ray report generation. The originally released model weights are trained on the MIMIC-CXR dataset. To avoid data leakage, we retrain the model on a disjoint subset of MIMIC-CXR that does not contain any patients involved in our evaluation. (3) `MAIRA-2` (Bannur et al., 2024) is the latest radiology-specific LVLM developed by Microsoft Research. It is based on a similar architecture as `LlaVa`, featuring a `Rad-DINO` visual encoder and a language model based on `Vicuna` 7B v1.5 for grounded report generation.

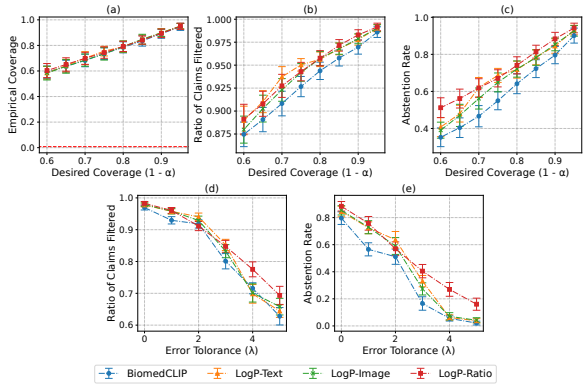


Figure 6: `MAIRA-2` results in *medical report generation*.

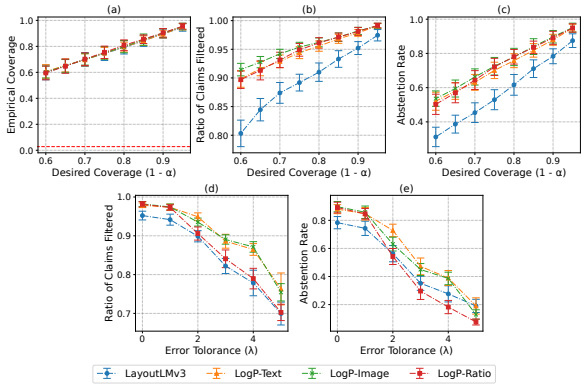


Figure 7: `LLaVA-NeXT` results in *document understanding*.

Error Annotation and Loss Function. We leverage `GPT-4o` for error annotation. To ensure quality, we provide `GPT-4o` with the ground truth report written by qualified physicians in addition to the chest X-ray images. This eliminates the requirement for `GPT-4o` to understand the actual medical image as it can verify the veracity of each claim by just checking if it is entailed by the ground truth report. Specifically, the errors are categorized as: (1) *Conflicting error*: The claim directly contradicts information provided in the ground truth report; (2) *Implausible error*: The claim does not directly conflict with or align with the ground truth report, and is implausible within the given context; and (3) *Plausible error*: The claim does not directly conflict with or align with the ground truth report, but remains plausible. Similar to scene understanding, we assign each occurrence of errors (1)-(3) a loss of 3, 2, and 1, respectively, and compute the accumulated loss as the final loss for each response.

Scoring Function. We use `BiomedCLIP` (Zhang et al., 2023) to compute similarity between claim and image pairs as the external confidence score.

5.2 Results

Given limited space, we present the results of `MAIRA-2` and defer other LVLMs to the Appendix.

We plot the empirical coverage versus desired coverage in Fig. 6a. The results verify that ConfLVLM can achieve tight error control in the revised responses after filtering. This is particularly important considering that medical report generation is a much more challenging task that requires precise control over the risk of output hallucination.

We plot the average ratio of claims filtered and abstention rate at various desired levels of coverage in Fig. 6b and Fig. 6c, respectively, given fixed $\lambda = 0$. We observe a trend similar to the scene understanding task, where the ratio of filtered claims and abstention rate increases with the desired coverage level. Notably, given a fixed coverage level, `BiomedCLIP` archives the lowest ratio of filtered claims and abstention rate among all scoring functions. Fig. 6d and Fig. 6e present the results with varying error tolerance while keeping $\alpha = 0.1$. In particular, changing λ from 0 to 3 reduces the abstention rate by more than half, while still maintaining a relatively low error (compared to the median of the loss of unfiltered responses, which is 7).

6 Case Study III: Document Understanding

6.1 Setup

Finally, we evaluate ConfLVLM on the document understanding task, where we randomly select 500 invoice scan/images from the SROIE (Huang et al., 2019) dataset.

LVLMs. We consider two LVLMs, `LLaVA-Next` (Liu et al., 2024a), which is the latest model in the `LLaVA` family with enhanced visual reasoning and OCR capabilities, and `Phi-3.5-vision-instruct`.

Error Annotation and Loss Function. We consider the following error types for this task: (1) *Field misinterpretation*: Incorrectly identify important fields such as mistaking "Subtotal" for "Total Amount", or misrecognizing non-existing fields. (2) *Numerical and quantitative errors*: Incorrect amounts, totals, or quantity values, as well as calculation discrepancies (e.g., subtotal, tax, and total relationship). (3) *Date error*: Misrecongizing date or misinterpreting date formats. (4) *Item error*: Misrecongizing item or item details, or falsely identifying non-existing items. (5) *Other errors*:

Other errors such as misspelling or misrecognizing character, layout, and alignment issues. Each occurrence of Numerical and Date Errors gets an additional loss of 3, each occurrence of field or item error gets an additional loss of 2, and the occurrence of other errors gets an additional loss of 1. Similar to other tasks, we compute the accumulated loss for each response.

Scoring Function. We use `LayoutLMv3` (Huang et al., 2022) which is a pre-trained multi-modal transformer to derive embeddings for the claim text and document images and compute their cosine similarity as the external confidence score.

6.2 Results

We now show the results of `LLaVA-NeXT` and defer the results of other models to the Appendix.

We first verify the alignment of empirical coverage and desired coverage in Fig. 7a. The results show that CONFLVLM can achieve a precise level of coverage on the document understanding task.

Next, we investigate the ratio of filtered claims and abstention rate with varying α and fixed $\lambda = 1$ in Fig. 7b and Fig. 7c, respectively, and with varying λ and fixed $\alpha = 0.1$ in Fig. 7d and Fig. 7e, respectively. Besides the general trend that increasing desired coverage or reducing the error tolerance would result in filtering out more content and more frequently abstaining, we additionally observe that `LayoutLMv3` achieves significantly lower rates compared to other scoring functions based on the LVLM's internal confidence, e.g., preserving approximately 10% more content when $\alpha = 0.4$. This shows that a small dedicated model is more accurate than LVLMs in terms of verifying the factuality of claims regarding document images.

7 Discussion

Impact of Calibration Data Size. We have considered a fixed calibration data size of 400. To study the impact of calibration data size, we use `Llama-3.2-11B-Vision` as an example and vary the calibration data from 50 to 400 samples, each with 50 random train-test splits, and plot the empirical coverage and ratio of claims filtered of three sets of (α, λ) parameters in Fig. 8. We observe that CONFLVLM can consistently achieve the desired level of coverage while maintaining the same ratio of filtered claims regardless of the calibration data size, though a larger calibration dataset could help reduce the result variance.

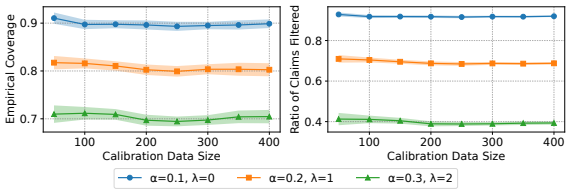


Figure 8: Impact of calibration data size on the empirical coverage and utility of `Llama-3.2-11B-Vision` on the *scene understanding* task (with 95% CI).

Discriminative vs. Generative Models. In our experiments, we observe that in many cases small discriminative vision-language models (e.g., `BiomedCLIP` for medical report generation and `LayoutLMv3` for document understanding) outperform LVLMs in terms of capturing the relevance between a text claim and an image. One potential reason is that, compared to large generative models, small discriminative models are easier to optimize and can learn useful representations more efficiently. This hints that besides serving as the image encoder, these models can be used as critics to censor LVLM outputs and improve factuality with lower computational costs.

8 Conclusion

In this work, we propose CONFLVLM, a framework for achieving statistical factuality guarantee of LVLM output through decomposing responses into individual verifiable hypotheses and filtering out those with low confidence given the image content. We demonstrate with three application domains that by choosing the desired error rate and tolerance, CONFLVLM offers users flexible control over the hallucination risk of LVLM output.

Acknowledgments

This research was funded in part by the Intuit University Collaboration Program and NIH grant U54HG012510 (Bridge2AI).

Limitations

Our study, while demonstrating the effectiveness of CONFLVLM in providing statistical guarantees on the factuality of LVLM-generated content, has two limitations. First, the current framework assumes exchangeability of the calibration and test datasets. Although this is a mild assumption (more relaxed than the common I.I.D. assumption), it may not always hold in applications where distribution shifts frequently occur. Addressing this limitation would require extending the framework to handle

such distributional changes robustly. Discussions on potential solutions can be found in Appendix Section C. Second, while ConfLVLM is flexible and broadly applicable, it inherently involves a trade-off between factuality guarantees and content utility: an increase in the desired level of coverage naturally results in more aggressive filtering and reduced response informativeness. Future work may investigate methods to better balance this trade-off (e.g., train a better scoring module), enhancing factual accuracy without compromising on utility.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

John J Cherian, Isaac Gibbs, and Emmanuel J Candès. 2024. Large language model validity via enhanced conformal prediction methods. *arXiv preprint arXiv:2406.09714*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org *(accessed 14 April 2023)*, 2(3):6.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. 2023. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and 1 others. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto.

2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27228–27238.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, and 1 others. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zhuohang Li, Jiaxin Zhang, Chao Yan, Kamalika Das, Sricharan Kumar, Murat Kantarcioglu, and Bradley Malin. 2024b. Do you know what you are talking about? characterizing query-knowledge relevance for reliable retrieval augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6130–6151.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, and 1 others. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874.

Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models — ai.meta.com. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. [Accessed 07-11-2024].

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Forty-first International Conference on Machine Learning*.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633.

OpenAI. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. [Accessed 07-11-2024].

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15840–15853.

Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, and 1 others. 2023. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023a. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023b. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

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

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

A Additional Results

A.1 Comparison to Heuristic-based Mitigation

Woodpecker. Heuristic-based mitigation, such as *Woodpecker* (Yin et al., 2023b), relies on a series of external models, including `BLIP-2`, `GroundingDINO`, and proprietary models such as `GPT`, to reduce object hallucination, and thus it is not readily applicable to the specialized domains considered in our paper (i.e., medicine and finance). Even if Woodpecker is applicable, unlike our method which is driven by a confidence score, Woodpecker is driven by the matching of textual claims to objects extracted from the image. In this sense, when Woodpecker fails to match an object, it filters out the claim. As such, our method offers a continuous confidence score that is tunable, whereas Woodpecker only offers a binary match/no match strategy. We conducted additional experiments on the scene understanding task by randomly selecting 100 images and measuring Woodpecker's claim filtering efficiency (in terms of TPR) and final response accuracy, with results shown in Table 2. We observe that Woodpecker suffers from low TPR in claim filtering and low final response accuracy, partially due to `GroundingDINO`'s high FPR in open-set object detection.

Table 2: Comparison to heuristic-based mitigation.

Method	Claim Filtering Efficiency (TPR) ↑	Response Accuracy ↑
Woodpecker	59.1%	41%
CoVe	37.0%	23%
VCD ($\beta = 0.1$)	35.5%	20%
VCD ($\beta = 0.2$)	34.8%	24%
VCD ($\beta = 0.3$)	33.0%	22%
ICD ($\beta = 0.1$, **P**)	41.1%	26%
ICD ($\beta = 0.1$, **N**)	41.5%	26%
ICD ($\beta = 0.2$, **P**)	41.5%	27%
ICD ($\beta = 0.2$, **N**)	41.5%	27%
ConfLVLM	**95.3%**	**90%**

Other Mitigation. We additionally compare ConfLVLM with other prominent hallucination mitigation methods that do not rely on extensive external resources, including Chain-of-Verification (CoVe) and enhanced self-revision approaches, including Visual Contrastive Decoding (VCD) and Instruction Contrastive Decoding (ICD).

- *Chain-of-Verification (CoVe)*: The CoVe method, originally introduced by Dhuliawala et al. (2024), was primarily proposed for text-only data. To ensure a fair comparison in our visual setting, we adapted CoVe by first decomposing the LVLM's response into individual claims (atomic factoids) and then invoking the model to verify each claim independently using separate prompts.

- *Visual Contrastive Decoding (VCD)*: VCD (Leng et al., 2024) aims to reduce the model's over-reliance on statistical bias and unimodal priors that lead to object hallucinations, by modifying the decoding process to contrast output distributions from original and distorted visual inputs.

- *Instruction Contrastive Decoding (ICD)*: Similarly to VCD, ICD (Wang et al., 2024) improves decoding by contrasting between two distributions induced by the original instruction and the disturbance instruction.

We evaluated ConfLVLM against CoVe, VCD, and ICD using `LLaVA-1.5` on the scene understanding task. For VCD and ICD, we used their original code implementations open-sourced by the authors.

Following the setup in their papers, we employed the VCD/ICD method to predict the factuality of each claim. As recommended in the original papers, we set $\alpha = 1$ for both VCD and ICD. For VCD, we used a noise step of 500 and explored different β values. For ICD, we additionally considered its combination with both positive (P) and negative (N) prefixes. Similarly to the experiments for comparison with Woodpecker, we used a randomly selected set of 100 images (containing 788 claims in total) and measured claim filtering efficiency (in terms of True Positive Rate) and final response accuracy. The results are presented in Table 2. As the results demonstrate, CONFLVLM significantly outperforms CoVe, VCD, and ICD in both claim filtering efficiency and final response accuracy.

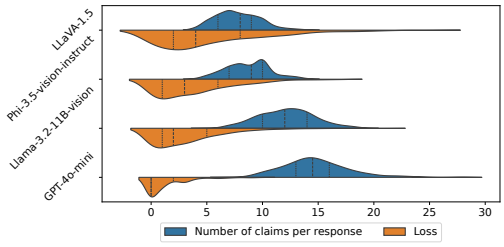
A.2 Annotation Reliability

Human Raters. We used `GPT-4o` for assisting with annotating error types. To verify its annotation quality, we randomly select a subset of 50 images and $1,182$ associating claims generated by `LLaVA-1.5` on the scene understanding task, and recruit two human annotators to generate independent error type annotations. The instruction provided to the human raters is analogous to the prompt shown in Table 4. The measured averaged Intraclass Correlation Coefficient (ICC) between `GPT-4o` and human annotations is 0.85, with the 95% CI being $[0.82, 0.87]$. This result confirmed that the annotations show high inter-rater reliability (by convention, any ICC value above 0.75 is considered to be good reliability).

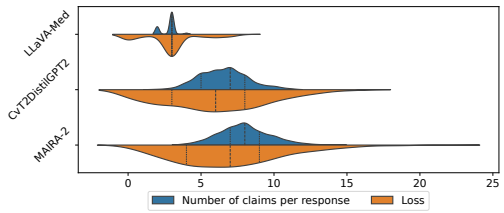
LVLM Raters. We measured the averaged ICC between `GPT-4o` and `Gemini-1.5-pro` on the scene understanding task to be 0.81, with the 95% CI being $[0.77, 0.85]$, which shows high inter-rater reliability.

A.3 LVLM Output Distribution

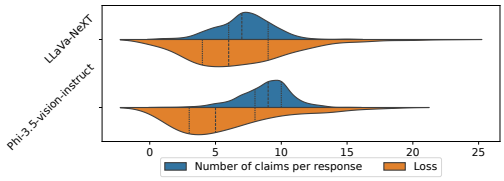
In Fig. 9, we compare the quality of vanilla LVLM outputs (i.e., raw responses without filtering any claim) by visualizing the distribution of the number of claims and loss per response. An ideal LVLM should be expressive (output more claims) while maintaining a low risk of hallucination (yield low loss values). On the scene understanding task, `GPT-4o-mini` clearly outperforms other models but is still prone to errors. On the medical report generation task, `CvT2DistilGPT2` and `MAIRA-2` both outperform `LLaVA-Med` by a large margin, but still have very high loss values in most responses. Similarly, on the document understanding tasks, both model show a long tail in loss distribution. These observations necessitate the adoption of error control methods with statistical guarantees.



(a) Scene understanding



(b) Medical report generation



(c) Document understanding

Figure 9: Comparison of the quality of raw LVLM responses.

A.4 Omitted Results from Main Paper

Here we provide the omitted results from main paper for medical report generation and document understanding.

Medical Report Generation. Fig. 10 compares the empirical and desired coverage of `LlaVa-Med`, `CvT2DistilGPT2`, and `MAIRA-2` on the medical report generation task. The same conclusion is drawn as in the general scene understanding setting: CONFLVLM achieves the desired level of coverage across all types of scoring functions, whereas Vanilla LVLM (i.e., responses without any filtration) produces significantly low coverage. Fig. 11 shows the average ratio of filtered claims across a range of desired coverage, whereas Fig. 12 presents the abstention rate as a function of desired coverage.

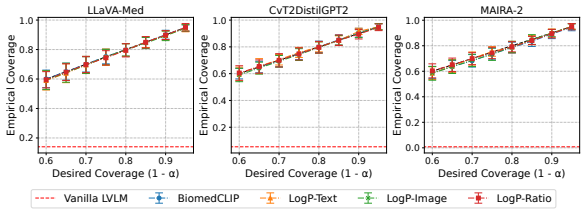


Figure 10: Alignment between the empirical and desired (theoretical) coverage on the *medical report generation* task (with $\lambda = 0$). Vanilla LVLM (red dashed line) refers to the base setting where the LVLM-generated responses are returned to users without using ConfLVLM.

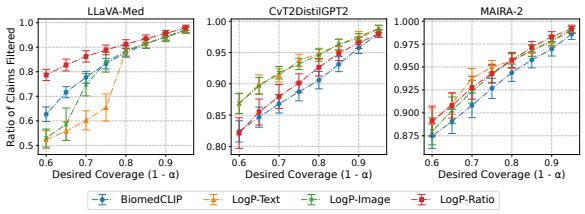


Figure 11: Average ratio of claims filtered with varying coverage using different scoring functions on the *medical report generation* task (with $\lambda = 0$).

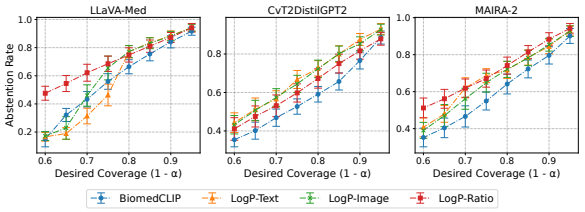


Figure 12: Abstention rate with varying coverage using different scoring functions on the *medical report generation* task (with $\lambda = 0$).

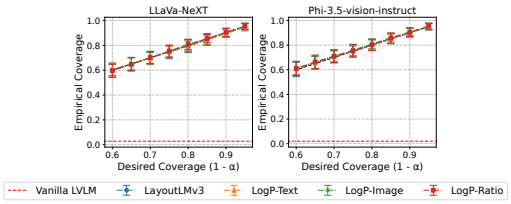


Figure 13: Alignment between the empirical and desired (theoretical) coverage on the *document understanding* task (with $\lambda = 0$). Vanilla LVLM (red dashed line) refers to the base setting where the LVLM-generated responses are returned to users without using ConfLVLM.

Document Understanding. Fig. 13 shows the alignment between the empirical and desired coverage of `LLaVA-Next` and `Phi-3.5-vision-instruct` using ConfLVLM on the document understanding task. The same conclusions regarding model coverage are reached as in the other two image understanding settings. Fig. 14 shows the average ratio of filtered claims across a range of desired coverage, whereas Fig. 15 presents the abstention rate as a function of desired coverage. In Fig. 16, we compare `LLaVa-NeXT`'s response with different error tolerances and a fixed error rate.

A.5 Response-level and Claim-level Results

In Table 3, we show the response-level (rate of responses containing at least one error and the average loss per response) and claim-level (TPR, FNR, and F1 for detecting erroneous responses) results under various α, λ configurations on the scene understanding task. Results are averaged over 50 random data splits.

A.6 Empirical Coverage and Utility with $\lambda > 0$

In addition to the plots in the main paper with $\lambda = 0$, we plot the empirical coverage, ratio of claims filtered, and abstention rate with error tolerance $\lambda = 1, 2$ for scene understanding in Fig. 17, Fig. 18, and Fig. 19, for medical report generation in Fig. 20, Fig. 21, and Fig. 22, and for document understanding in Fig. 23, Fig. 24, and Fig. 25, respectively.

A.7 Examples

We include additional examples of applying ConfLVLM to control error in responses from LVLMs for scene understanding (Fig. 26), medical report generation (Fig. 27), and document understanding (Fig. 28).

B Implementation Details

Prompts for Image-conditioned Free-text Generation. We use the following prompt for evaluating LVLMs on the scene and document understanding tasks: "⟨`Image`⟩ *Provide a detailed description of the given image.*". To evaluate medical (radiology) report generation, we use the following prompt for `LLaVA-Med`: "⟨`Image`⟩ *What does the chest X-ray show?*", whereas `CvT2DistilGPT2` and `MAIRA-2` do not require any text prompt for generating reports.

Prompts for Error Annotation. We include our prompts for LLM-assisted error annotation of the scene understanding, medical report generation,

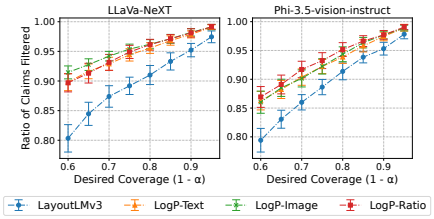


Figure 14: Average ratio of claims filtered with varying coverage using different scoring functions on the *document understanding* task (with $\lambda = 0$).

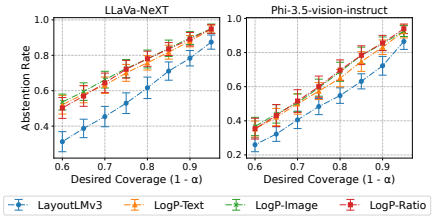


Figure 15: Abstention rate with varying coverage using different scoring functions on the *document understanding* task (with $\lambda = 0$).

and document understanding tasks in Table 4, 5, and 6, respectively.

Decomposition and Merge Operators. We implement the decomposition and merging operations by prompting the language model part of the LVLM. For rare cases where the LVLM does not support or cannot correctly implement the decomposing operation (e.g., dedicated models such as `MAIRA-2` and `CvT2DistilGPT2`), we use `GPT-4o-mini` as the substitute model to implement the decomposition and combination operations. The prompt for decomposing claims is *"Breakdown the above statement into a set of indepedent and self-contained claims. Each claim should be a short sentence. Output only a numbered list of claims."*. The prompt for merging claims is *"Merge the above claims about an image into a cohesive statement. Reuse the words from the original claims and do not generate any new claims."*.

C Additional Discussions

Coverage vs. Reliability. Although our method can achieve the precise coverage as specified by the user, the statistical guarantee only holds marginally with split conformal prediction. However, conditional guarantees may be required for certain applications, e.g., to ensure health equity among groups of patients in healthcare. Future work could consider the integration with advanced conformal

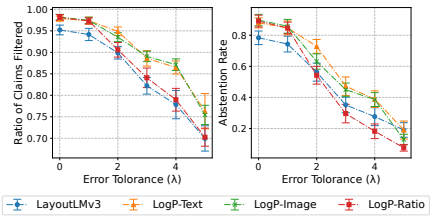


Figure 16: Comparison of `LLaVa-NeXT`'s response with different error tolerances (λ) while fixing $\alpha = 0.1$ on the *document understanding* task.

Table 3: Response- and claim-level results on the *scene understanding* task (averaged over 50 random splits).

LVLM	Configuration	Response-level		Claim-level		
		Error Rate	Average Loss	TPR	FNR	F1
`LLaVA-1.5`	Vanilla	0.8782	5.5028	0.0	1.0	0.0
	$\alpha = 0.1, \lambda = 0$	0.102	0.206	0.953	0.047	0.504
	$\alpha = 0.1, \lambda = 2$	0.212	0.529	0.886	0.114	0.507
	$\alpha = 0.3, \lambda = 0$	0.291	0.780	0.836	0.164	0.503
	$\alpha = 0.3, \lambda = 2$	0.536	1.771	0.637	0.363	0.499
`Phi-3.5-vision-instruct`	Vanilla	0.850	3.895	0.0	1.0	0.0
	$\alpha = 0.1, \lambda = 0$	0.094	0.147	0.945	0.055	0.401
	$\alpha = 0.1, \lambda = 2$	0.288	0.611	0.829	0.171	0.392
	$\alpha = 0.3, \lambda = 0$	0.306	0.655	0.818	0.182	0.390
	$\alpha = 0.3, \lambda = 2$	0.648	1.911	0.480	0.520	0.345
`Llama-3.2-11B-vision`	Vanilla	0.793	3.129	0.0	1.0	0.0
	$\alpha = 0.1, \lambda = 0$	0.105	0.205	0.936	0.064	0.269
	$\alpha = 0.1, \lambda = 2$	0.232	0.543	0.831	0.169	0.266
	$\alpha = 0.3, \lambda = 0$	0.306	0.714	0.772	0.228	0.264
	$\alpha = 0.3, \lambda = 2$	0.628	1.805	0.408	0.592	0.227
`GPT-4o-mini`	Vanilla	0.493	1.265	0.0	1.0	0.0
	$\alpha = 0.1, \lambda = 0$	0.097	0.168	0.850	0.150	0.100
	$\alpha = 0.1, \lambda = 2$	0.285	0.552	0.544	0.456	0.099
	$\alpha = 0.3, \lambda = 0$	0.300	0.589	0.510	0.490	0.0988
	$\alpha = 0.3, \lambda = 2$	0.493	1.265	0.0	1.0	0.0

methods to achieve conditional validity (Gibbs et al., 2023). Besides coverage, investigating other important aspects of LVLM reliability, such as omission, and designing better scoring functions to achieve the same level of coverage while preserving more content are also interesting avenues for future research.

Distribution Shift. Our LVLM factuality framework follows the standard conformal prediction setting by assuming data exchangeability (a weaker notion than IID). In cases where this may not hold, there are alternative strategies that one could invoke, such as periodically updating the calibration set or applying a discount factor to older samples. For significant shifts, such as out-of-domain data, incorporating an additional OOD detection layer (Li et al., 2024b) can help to empirically preserve coverage.

D License of Artifacts

Datasets. The MSCOCO and SROIE datasets are released under a Creative Commons Attribution 4.0 License. MIMIC-CXR is released under a PhysioNet Credentialed Health Data License 1.5.0.

Models. `LLaVA-1.5` is under the Llama-2 community license. `Phi-3.5-vision-instruct` is

Table 4: Data annotation prompt for scene understanding.

Error annotation prompt for scene understanding

System *"You are an expert annotator tasked with evaluating statements generated by a vision-language model (VLM). Given an image and a claim, your task is to verify the factuality of the claim based on how well it aligns with the provided image. You should focus only on significant or material correctness, ignoring minor differences or non-essential details, especially in spatial relationships or specific object types.*
*The errors are categorized as follows: 1. **Object Identification (Object)**: The claim involves hallucinated or wrongly identified objects. Ignore minor distinctions between similar objects (e.g., slotted spoon vs regular spoon) unless it fundamentally changes the meaning of the claim. 2. **Attribute Accuracy (Attribute)**: The claim involves incorrect attributes (e.g., color, size, shape). Only flag attributes if they are critical to the understanding of the claim. 3. **Spatial Relations (Spatial)**: The claim involves incorrect spatial relationships between objects. Only flag spatial errors if they significantly change the scene (e.g., "above the water" vs. "in the water" can be ignored unless the context requires precision). 4. **Interaction/Action Accuracy (Interaction)**: The claim involves incorrect or hallucinated action or interaction. 5. **Quantitative Information (Quantitative)**: The claim involves incorrect numeric details (e.g., wrong object count).*
For each claim, generate a JSON object with four fields: - "reasoning": a brief explanation of why the claim is correct or incorrect. - "label": a boolean value (True or False) where True means the claim is factually correct, and False means it is incorrect. - "error_type": a list of error types (e.g., ["Object", "Attribute"]) if the claim contains errors, or an empty list if the claim is fully correct.
Example: Given an image of two orange cats, and the following list of claims: 1. This image features several cute cats. 2. There are a total number of three cats. 3. One cat is orange, the others are black. 4. There is also a dog behind the cats.
Return: [{"reasoning": "The claim is general and no significant error can be found.", "label": true, "error_type": []}, {"reasoning": "There are two cats in the image, not three.", "label": false, "error_type": ["Quantitative"]}, {"reasoning": "One cat is orange, but the other is not black.", "label": false, "error_type": ["Quantitative", "Attribute"]}, {"reasoning": "There is no dog in the image.", "label": false, "error_type": ["Object"]}] "
User *"List of claims: {claims} For each claim, return a JSON object with "reasoning", "label" (true or false), "error_type" (might contain multiple types from ["Object", "Attribute", "Spatial", "Interaction", "Quantitative"]). "*

under the MIT license. `Llama-3.2-11B-vision` is under the Llama-3.2 community license. `LlaVa-Med` is under the Apache license 2.0 license. `CvT2DistilGPT2` is under the GNU General Public License v3.0. `MAIRA-2` is under the Microsoft research license. `LLaVA-Next` is under the Apache-2.0 license.

Table 5: Data annotation prompt for medical report generation.

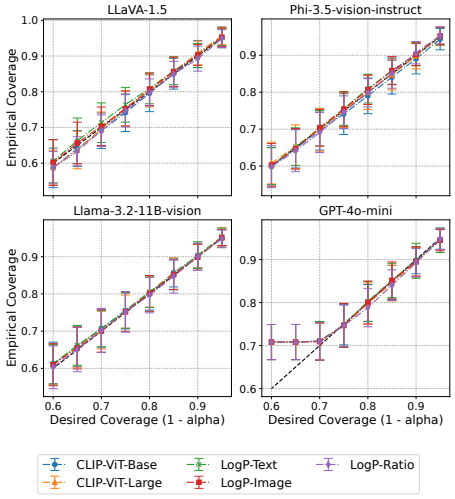
Error annotation prompt for medical report generation

System *"You are an experienced radiologist tasked with evaluating statements generated by a Medical AI model. Given a chest x-ray image, a ground truth report generated by expert human radiologist, and a claim generated by the AI model, your task is to verify the factuality of the claim based on how well it aligns with the provided ground truth report. IMPORTANT: A claim should be deemed correct only if it is directly entailed by the ground truth report.*
*The errors are categorized as follows: 1. **Conflicting Error (Conflicting)**: The claim directly contradicts information provided in the ground truth report. 2. **Implausible Error (Implausible)**: The claim does not directly conflict with or align with the ground truth report, and is implausible within the given context. 3. **Plausible Error (Plausible)**: The claim does not directly conflict with or align with the ground truth report, but remains plausible within the given context.*
For each claim, generate a JSON object with four fields: - "reasoning": a brief explanation of why the claim is correct or incorrect. - "label": a boolean value (True or False) where True means the claim is factually correct, and False means it is incorrect. - "error_type": a list of error types (e.g., ["Conflicting", "Plausible"]) if the claim contains errors, or an empty list if the claim is fully correct.
Example: Given a chest x-ray report and the following list of claims: 1. There is no evidence of lung consolidation. 2. The heart size is mildly enlarged. 3. There are signs of a pleural effusion.
Return: ["reasoning": "The ground truth report confirms no lung consolidation.", "label": true, "error_type": [], "reasoning": "The ground truth report describes the heart size as normal.", "label": false, "error_type": ["Conflicting"], "reasoning": "The ground truth report does not mention a pleural effusion, but it is a plausible interpretation in some cases.", "label": false, "error_type": ["Plausible"]] "
User *"List of claims: {claims} For each claim, return a JSON object with "reasoning", "label" (true or false), "error_type" (might contain multiple types from ["Conflicting", "Implausible", "Plausible"]). "*

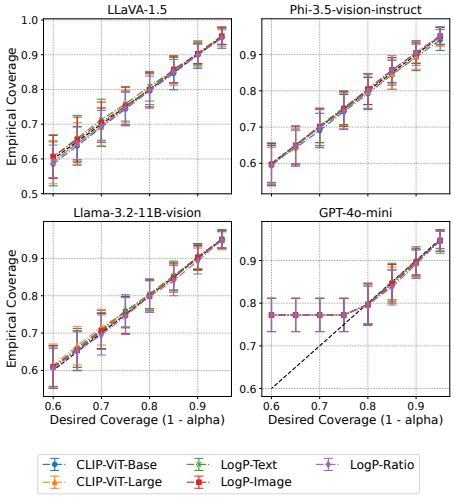
Table 6: Data annotation prompt for document understanding.

Error annotation prompt for document understanding

System *"You are an expert annotator tasked with evaluating statements generated by a Document AI model. Given a ground truth document (image and text) and a claim, your task is to verify the factuality of the claim based on how well it aligns with the provided document.*
*The errors are categorized as follows: 1. **Field Misinterpretation (Field)**: Incorrectly identify important fields such as mistaking "Invoice Date" for "Due Date", "Subtotal" for "Total Amount", or misrecognize non-existing field. 2. **Numerical and Quantitative Errors (Numerical)**: Incorrect amounts, totals, or quantity values, as well as calculation discrepancies (e.g., subtotal, tax, and total relationship). 3. **Date Error (Date)**: Misrecongizing date or misinterpreting date formats. 4. **Item Error (Item)**: Misrecongizing item or item details, or falsely identifying non-existing item. 5. **Other Errors (Other)**: Other errors such as misspell or misrecognize character, layout and alignment issues.*
For each claim, generate a JSON object with four fields: - "reasoning": a brief explanation of why the claim is correct or incorrect. - "label": a boolean value (True or False) where True means the claim is factually correct, and False means it is incorrect. - "error_type": a list of error types (e.g., ["Numerical", "Item"]) if the claim contains errors, or an empty list if the claim is fully correct.
Example: Given an invoice of buying a Chopping Board at a shop named Ẅalmart, and the following list of claims: 1. This image is a printed invoice. 2. The merchant name is Ẅallmart. 3. The items listed on the receipt include two Chopping Board, and a Knife.
Return: ["reasoning": "The image shows a printed invoice.", "label": true, "error_type": [], "reasoning": "The merchant name is spelled incorrectly.", "label": false, "error_type": ["Other"], "reasoning": "Only one Chopping Board, and no Knife purchased.", "label": false, "error_type": ["Numerical", "Item"],] "
User *"List of claims: {claims} For each claim, return a JSON object with "reasoning", "label" (true or false), "error_type" (might contain multiple types from ["Field", "Numerical", "Date", "Item", "Other"]). "*

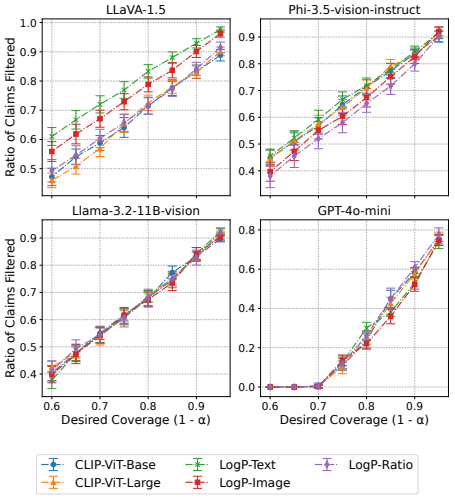


(a) $\lambda = 1$

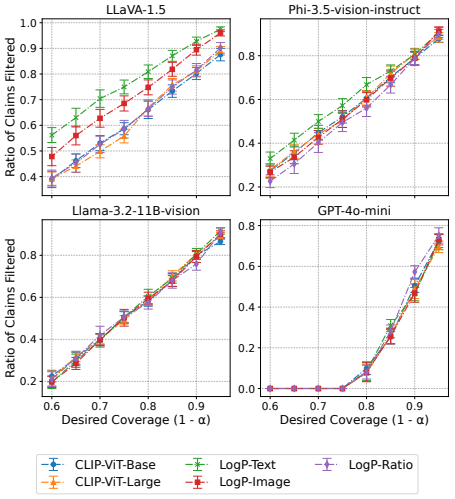


(b) $\lambda = 2$

Figure 17: Comparison of empirical and desired (theoretical) coverage on the scene understanding task with different error tolerances (λ).

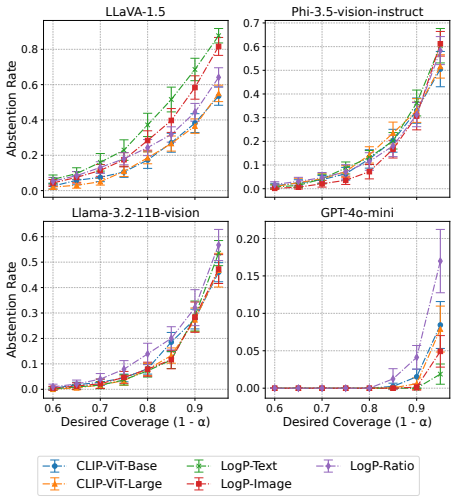


(a) $\lambda = 1$

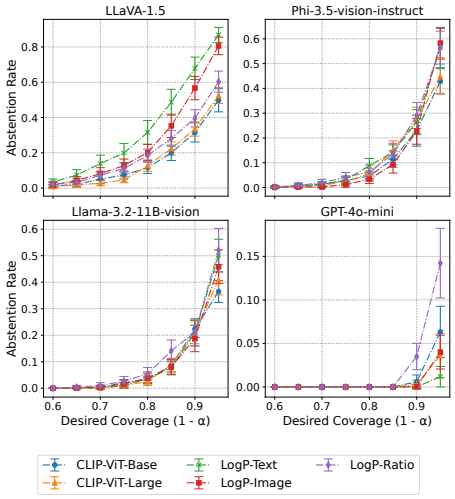


(b) $\lambda = 2$

Figure 18: Average ratio of claims filtered with varying coverage using different scoring functions on the scene understanding task with different error tolerances (λ).

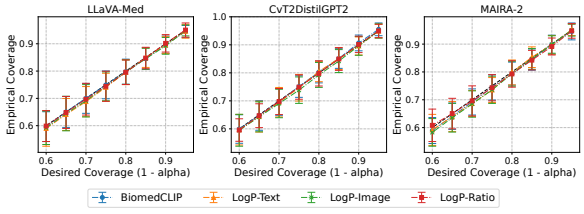


(a) $\lambda = 1$

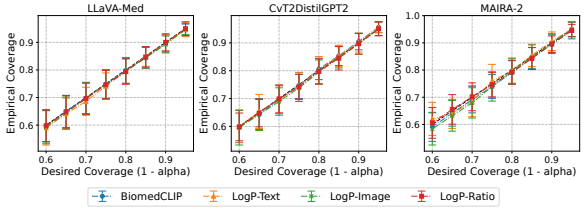


(b) $\lambda = 2$

Figure 19: Abstention rate with varying coverage using different scoring functions on the scene understanding task with different error tolerances (λ).

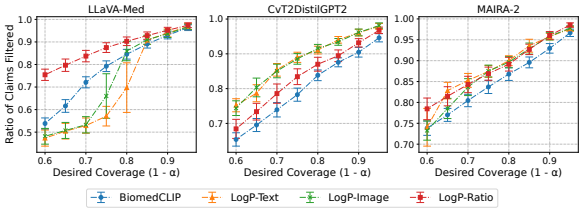


(a) $\lambda = 1$

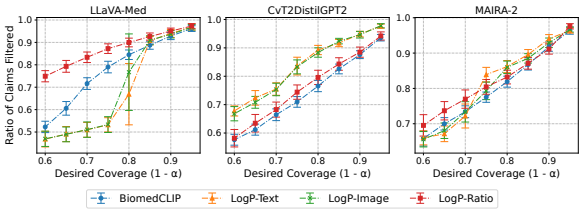


(b) $\lambda = 2$

Figure 20: Comparison of empirical and desired (theoretical) coverage on the medical report generation task with different error tolerances (λ).

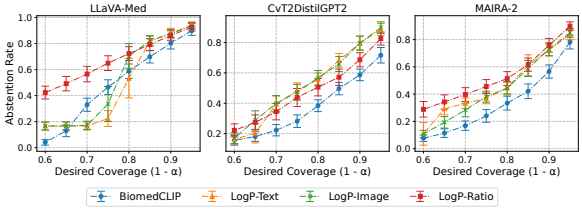


(a) $\lambda = 1$

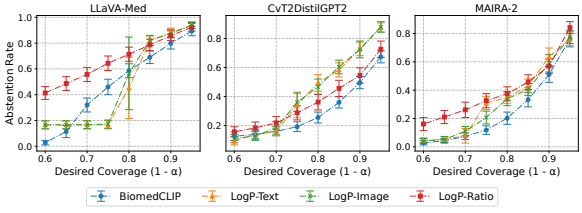


(b) $\lambda = 2$

Figure 21: Average ratio of claims filtered with varying coverage using different scoring functions on the medical report generation task with different error tolerances (λ).

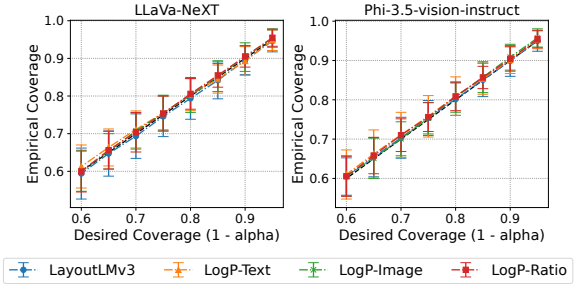


(a) $\lambda = 1$

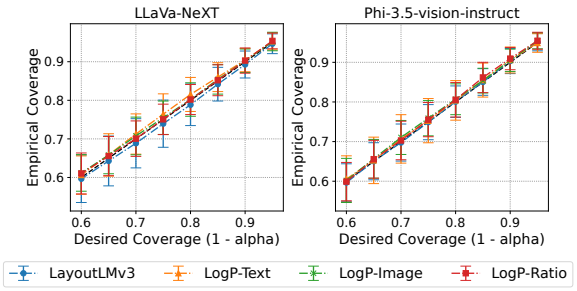


(b) $\lambda = 2$

Figure 22: Abstention rate with varying coverage using different scoring functions on the medical report generation task with different error tolerances (λ).

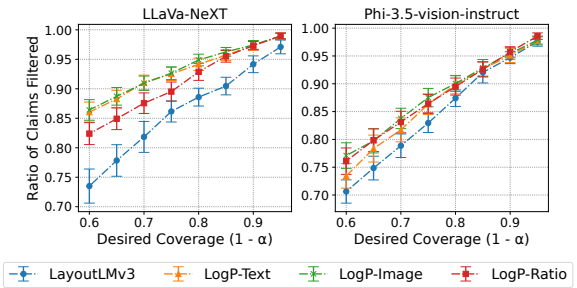


(a) $\lambda = 1$

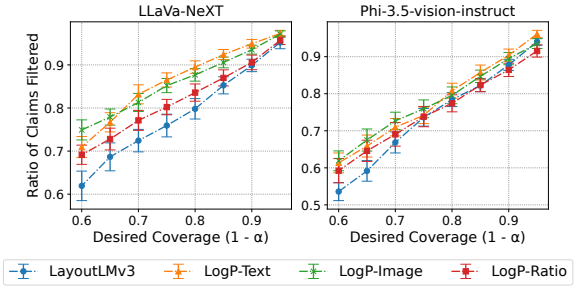


(b) $\lambda = 2$

Figure 23: Comparison of empirical and desired (theoretical) coverage on the document understanding task with different error tolerances (λ).

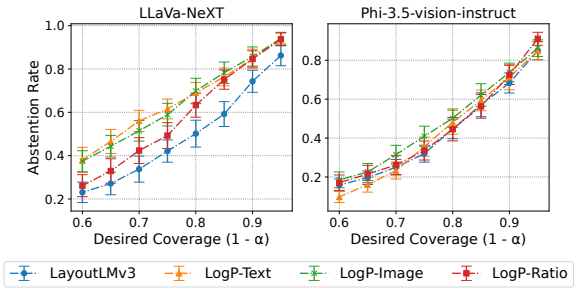


(a) $\lambda = 1$

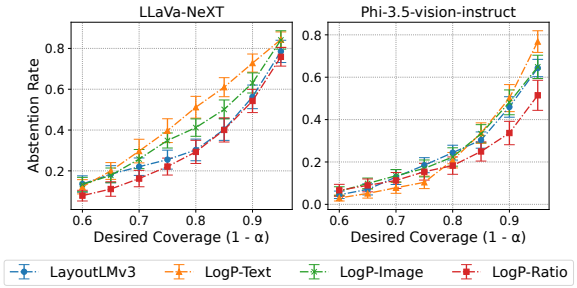


(b) $\lambda = 2$

Figure 24: Average ratio of claims filtered with varying coverage using different scoring functions on the document understanding task with different error tolerances (λ).



(a) $\lambda = 1$



(b) $\lambda = 2$

Figure 25: Abstention rate with varying coverage using different scoring functions on the document understanding task with different error tolerances (λ).

Image	LVLM	Initial Response	Claims	Final Response ($\alpha = 0.1, \lambda = 0$)	Final Response ($\alpha = 0.2, \lambda = 1$)	Final Response ($\alpha = 0.3, \lambda = 2$)
	LLaVA-1.5	The image features a bathroom with a large bathtub sitting next to a toilet. The bathtub is filled with water, and there is a potted plant placed nearby. The bathroom also has a sink and a mirror, which is located above the sink. The overall atmosphere of the bathroom appears to be clean and well-maintained.	1. The image features a bathroom. 2. The bathroom has a large bathtub filled with water. 3. There is a toilet in the bathroom. 4. There is a potted plant placed nearby. 5. The bathroom has a sink. 6. The bathroom has a mirror located above the sink. 7. The overall atmosphere of the bathroom appears to be clean and well-maintained.	The overall atmosphere of the bathroom appears to be clean and well-maintained. (Claim 7)	The image showcases a clean and well-maintained bathroom. (Claim 1, 7)	The image showcases a bathroom with a large bathtub filled with water, a sink, and a clean, well-maintained atmosphere. (Claim 1, 2,5,7)
	Phi-3.5-vision-instruct	The image captures a serene bathroom scene. Dominating the space is a bathtub, its gray tiles contrasting with the white sink that sits atop a marble countertop. The bathtub is equipped with a showerhead, ready to provide a relaxing bath. A small plant adds a touch of greenery to the room, placed on the floor next to the bathtub. The floor is adorned with tiles in shades of blue and green, adding a pop of color to the otherwise neutral palette. A towel is casually draped over the bathtub, suggesting recent use. The walls, covered in gray tiles, reflect the light, enhancing the brightness of the room. A window punctuates one of the walls, allowing natural light to filter in and illuminate the space. The overall layout and design of the bathroom suggest a modern and stylish interior.	1. The bathroom features a gray tiled bathtub and a white sink. 2. A showerhead is installed on the bathtub. 3. A small plant is placed on the floor next to the bathtub. 4. The floor is covered with blue and green tiles. 5. A towel is draped over the bathtub. 6. The walls are covered in gray tiles. 7. A window is present on one of the walls.	[Abstained]	The bathroom has a gray tiled bathtub and a white sink, with a small plant placed on the floor next to the bathtub. (Claim 1, 3)	The bathroom features a gray tiled bathtub and a white sink. A showerhead is installed on the bathtub, and a small plant is placed on the floor next to it. The floor is covered with blue and green tiles, and a towel is draped over the bathtub. The walls are covered in gray tiles. (Claim 1, 2, 3, 4, 5, 6)

	Llama-3.2-11B-vision	The image depicts a bathroom with a large bathtub, sink, and toilet. The room is tiled in gray and brown marble-patterned tiles, with a white sink and toilet. A potted plant sits in the corner of the room, and a white towel is placed on the floor next to the bathtub. The bathtub is white and has a decorative border around it. The floor is made up of large, square tiles in a checkerboard pattern, with alternating gray and brown colors. The walls are also tiled in the same pattern as the floor.
		
		* **Bathtub:**
		  * White
		  * Decorative border around it
		  * Located in the corner of the room
		* **Sink:**
		  * White
		  * Located on the left side of the room
		  * Has a marble-patterned countertop
		* **Toilet:**
		  * White
		  * Located on the left side of the room
			1. The image depicts a bathroom.
			2. The bathroom has a large bathtub.
			3. The bathroom has a sink.
			4. The bathroom has a toilet.
			5. The room is tiled in gray and brown marble-patterned tiles.
			6. The sink is white.
			7. The toilet is white.
			8. A potted plant sits in the corner of the room.
			9. A white towel is placed on the floor next to the bathtub.
			10. The bathtub is white.
			11. The bathtub has a decorative border around it.
			12. The floor is made up of large, square tiles in a checkerboard pattern.
			13. The floor tiles are alternating gray and brown colors.
				[Abstained]	The image depicts a bathroom with a large bathtub, tiled in gray and brown marble-patterned tiles, and a white towel placed on the floor next to the bathtub.
					
					(Claim 1, 2, 5, 9)
						The image depicts a bathroom with a large bathtub, a sink, and a toilet. The bathroom is tiled in gray and brown marble-patterned tiles. A potted plant sits in the corner of the room. A white towel is placed on the floor next to the bathtub. The bathtub is white and has a decorative border around it. The floor is made up of large, square tiles in a checkerboard pattern, with the tiles alternating gray and brown colors. The room has a white bathtub, a sink, and a toilet. The bathroom has a white towel, a potted plant, and a decorative border around the bathtub.
						
						(Claim 1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13)

	GPT-4o-mini	The image depicts a spacious bathroom with a modern and stylish design. The walls are covered with large, gray-green tiles that create a cohesive look throughout the room. The flooring features a diamond pattern with alternating squares of beige and teal, adding a touch of color and visual interest. On the left side, there is a round white sink mounted on a marble countertop, which has a sharp, angular design that complements the contemporary aesthetic. Below the sink, there is a white toilet, neatly positioned. Adjacent to the sink is a bathtub, which is built into the wall. It has a sleek, rectangular shape and features decorative tiles along its edge, giving it an elegant touch. Above the bathtub, there is a faucet, and the space is well-lit, likely by natural light coming from the windows positioned near the ceiling. In the corner, a potted plant adds a hint of greenery, enhancing the bathroom's fresh and inviting atmosphere. On the floor, there are several white bath mats, providing comfort and warmth underfoot. To the right, there's a yellow cabinet with a textured surface, offering storage space and contributing a pop of color to the otherwise muted palette. Additionally, a small basket with flowers is placed on top of the cabinet, adding a decorative element to the overall decor. The overall ambiance of the bathroom is serene and well-organized, making it a relaxing space.	1. The bathroom has a modern and stylish design. 2. The walls are covered with large, gray-green tiles. 3. The flooring features a diamond pattern with alternating squares of beige and teal. 4. There is a round white sink mounted on a marble countertop. 5. The marble countertop has a sharp, angular design. 6. Below the sink, there is a white toilet. 7. Adjacent to the sink is a bathtub built into the wall. 8. The bathtub has a sleek, rectangular shape. 9. Decorative tiles are featured along the edge of the bathtub. 10. There is a faucet above the bathtub. 11. The space is well-lit, likely by natural light from windows near the ceiling. 12. A potted plant is placed in the corner of the bathroom. 13. The potted plant adds a hint of greenery to the space. 14. There are several white bath mats on the floor. 15. The bath mats provide comfort and warmth underfoot. 16. To the right, there is a yellow cabinet with a textured surface. 17. The yellow cabinet offers storage space. 18. The yellow cabinet contributes a pop of color to the bathroom. 19. A small basket with flowers is placed on top of the cabinet. 20. The basket adds a decorative element to the overall decor. 21. The overall ambiance of the bathroom is serene and well-organized. 22. The bathroom is a relaxing space.	The bathroom has a modern and stylish design, with decorative tiles featured along the edge of the bathtub and a potted plant placed in the corner. (Claim 1, 9, 12)	The bathroom has a modern and stylish design, with walls covered in large, gray-green tiles and flooring featuring a diamond pattern with alternating squares of beige and teal. A round white sink is mounted on a marble countertop, which has a sharp, angular design. Below the sink is a white toilet, and adjacent to the sink is a sleek, rectangular bathtub built into the wall, with decorative tiles featured along its edge. A faucet is placed above the bathtub. A potted plant is placed in the corner of the bathroom, while several white bath mats on the floor provide comfort and warmth underfoot. The yellow cabinet contributes a pop of color, and the basket adds a decorative element to the overall decor. The overall ambiance is serene, well-organized, and relaxing. (Claim 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 18, 20, 21, 22)	The bathroom has a modern and stylish design, with large gray-green tiles covering the walls and a diamond-patterned floor featuring alternating squares of beige and teal. A round white sink is mounted on a marble countertop with a sharp, angular design, positioned above a white toilet. Adjacent to the sink is a sleek, rectangular bathtub built into the wall, accented with decorative tiles along its edge and topped with a faucet. The space is well-lit, likely from natural light filtering through windows near the ceiling, enhancing its serene and relaxing ambiance. A potted plant in the corner adds a hint of greenery, while several white bath mats provide comfort and warmth underfoot. To the right, a textured yellow cabinet adds a pop of color and offers storage space, topped with a small basket of flowers for a decorative touch. The overall atmosphere is serene, well-organized, and inviting. (Claim 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22)

Figure 26: Examples of responses on the scene understanding task.

Image	LVLM	Initial Response	Claims	Final Response ($\alpha = 0.1, \lambda = 0$)	Final Response ($\alpha = 0.2, \lambda = 1$)	Final Response ($\alpha = 0.3, \lambda = 2$)
	LLaVA-Med	The chest X-ray shows bilateral infiltrates, which are abnormal substances or collections of cells that have entered the lung tissue. These infiltrates can be caused by various factors, such as infections, inflammation, or other lung conditions.	1. The chest X-ray shows bilateral infiltrates. 2. Bilateral infiltrates are abnormal substances or collections of cells that have entered the lung tissue. 3. These infiltrates can be caused by various factors, such as infections, inflammation, or other lung conditions.	[Abstained]	The chest X-ray shows bilateral infiltrates. (Claim 1)	The chest X-ray shows bilateral infiltrates. (Claim 1)
	CvT2DistilGPT2	in comparison with the study of there is little change and no evidence of acute cardiopulmonary disease. cardiac silhouette is at the upper limits of normal in size. no vascular congestion pleural effusion or acute focal pneumonia.	1. There is little change in the study results. 2. There is no evidence of acute cardiopulmonary disease. 3. The cardiac silhouette is at the upper limits of normal in size. 4. There is no vascular congestion. 5. There is no pleural effusion. 6. There is no acute focal pneumonia.	[Abstained]	There is no evidence of acute cardiopulmonary disease or pleural effusion. (Claim 2, 5)	There is no evidence of acute cardiopulmonary disease, the cardiac silhouette is at the upper limits of normal in size, and there is no pleural effusion. (Claim 2, 3, 5)
	MAIRA-2	The heart size is top normal. The hilar and mediastinal contours are normal. The lungs are clear without evidence of focal consolidations concerning for pneumonia. There is no pleural effusion or pneumothorax. The visualized osseous structures are unremarkable.	1. The heart size is top normal. 2. The hilar contours are normal. 3. The mediastinal contours are normal. 4. The lungs are clear. 5. There is no evidence of focal consolidations concerning for pneumonia. 6. There is no pleural effusion. 7. There is no pneumothorax. 8. The visualized osseous structures are unremarkable.	There is no evidence of focal consolidations concerning for pneumonia. (Claim 5)	There is no evidence of focal consolidations concerning for pneumonia, and there is no pleural effusion. (Claim 5, 6)	The lungs are clear, with no evidence of focal consolidations concerning for pneumonia or pleural effusion. (Claim 4, 5, 6)

Figure 27: Examples of responses on the medical report generation task.

Image	LVLM	Initial Response	Claims	Final Response ($\alpha = 0.1, \lambda = 0$)	Final Response ($\alpha = 0.2, \lambda = 1$)	Final Response ($\alpha = 0.3, \lambda = 2$)
	LLaVA-NeXT	The image displays a tax invoice from a company named "Gardenia Bakery (KL) Sdn Bhd" with a reference number "139986 X". The invoice is addressed to a customer named "Makassar Fresh Market SDN BHD" and is dated "15/10/2017". The invoice is for a purchase made at a location in Selangor, Malaysia. The invoice lists several items with their respective quantities, unit prices, and total prices. The items include "0.C. White", "Wholemeal", "0.C. Jumbo", and "CR-chocolate". The total amount due for the purchase is 23.56 Malaysian Ringgit. The invoice also includes a note stating that the customer is required to make necessary adjustments to the invoice based on the actual quantity received.	1. The image displays a tax invoice. 2. The invoice is from a company named "Gardenia Bakery (KL) Sdn Bhd". 3. The invoice is addressed to a customer named "Makassar Fresh Market SDN BHD". 4. The invoice is dated "15/10/2017". 5. The invoice lists several items with their respective quantities, unit prices, and total prices. 6. The items include "0.C. White", "Wholemeal", "0.C. Jumbo", and "CR-chocolate". 7. The total amount due for the purchase is 23.56 Malaysian Ringgit. 8. The invoice includes a note stating that the customer is required to make necessary adjustments to the invoice based on the actual quantity received.	[Abstained]	The invoice is dated "15/10/2017". (Claim 4)	The image displays a tax invoice dated '15/10/2017'. (Claim 1, 4)
	Phi-3.5-vision-instruct	The image you've shared is a tax invoice from a gardenia bakery. The invoice is printed in black and white, with the text clearly legible. The bakery's contact information is prominently displayed at the top of the page. The invoice is addressed to a customer named Rizwan, located in Jaya, Selangor. The invoice number is 7015F710, and it was issued on the 15th of October, 2017. The total amount due is 23.56. The invoice details the items purchased, their quantities, and their prices. It also includes the total cost of the supplies and the total payable amount. The invoice is signed by the bakery's representative, E&E.	1. The image is a tax invoice from a gardenia bakery. 2. The invoice is printed in black and white. 3. The bakery's contact information is prominently displayed at the top of the page. 4. The invoice is addressed to a customer named Rizwan, located in Jaya, Selangor. 5. The invoice number is 7015F710. 6. The invoice was issued on the 15th of October, 2017. 7. The total amount due is 23.56. 8. The invoice details the items purchased, their quantities, and their prices. 9. It also includes the total cost of the supplies and the total payable amount. 10. The invoice is signed by the bakery's representative, E&E.	The invoice number is 7015F710. (Claim 5)	The invoice number is 7015F710 and the total amount due is 23.56. (Claim 5, 7)	The invoice is printed in black and white and has the number 7015F710. The total amount due is 23.56. (Claim 2, 5, 7)

Figure 28: Examples of responses on the document understanding task.