A Conformal Risk Control Framework for Granular Word Assessment and Uncertainty Calibration of CLIPScore Quality Estimates

Anonymous ACL submission

Abstract

This study explores current limitations of learned image captioning evaluation metrics, specifically the lack of granular assessment for individual word misalignments within captions, and the reliance on single-point quality estimates without considering uncertainty. To address these limitations, we propose a simple yet effective strategy for generating and calibrating CLIPScore distributions. Leveraging a model-agnostic conformal risk control framework, we calibrate CLIPScore values for task-specific control variables, to tackle the aforementioned two limitations. Experimental results demonstrate that using conformal risk control, over the distributions produced with simple methods such as input masking, can achieve competitive performance compared to more complex approaches. Our method effectively detects misaligned words, while providing formal guarantees aligned with desired risk levels, and improving the correlation between uncertainty estimations and prediction errors, thus enhancing the overall reliability of caption evaluation metrics.

1 Introduction

Image Captioning (IC) evaluation is a crucial task in vision-and-language research, aiming to assess how accurately textual descriptions represent visual contents. Reference-free metrics such as CLIPScore (Hessel et al., 2021; Gomes et al., 2025), which measure quality by computing the cosine similarity between image and text embeddings, have been shown to correlate strongly with human judgments. However, simply scoring captions is often insufficient, as these quality assessments can be hard to interpret or unreliable.

In many cases, effective evaluation requires not only an overall score of caption quality, but also the detection of specific errors within the caption. Without this granular information, the assessment can seem incomplete or less useful. Beyond the lack of granularity, existing metrics provide IC quality assessments relying on single-point estimates, without incorporating any indication of confidence over their predictions. This absence of uncertainty quantification can be problematic, as even high-performing models may produce erroneous and misleading scores, reducing user trust.

To address these challenges, we propose a conformal risk control framework, to obtain task-specific, calibrated predictions, in conjunction with a simple yet effective strategy for generating distributions over CLIPScore predictions. This provides us with a principled way to adapt IC evaluation both to fine-grained analysis for each caption, and to a broader view of performance over a dataset, allowing for user-defined criteria to determine risk.

First, we enhance interpretability by detecting misalignments between images and texts, identifying specific words that are incorrect. Second, we overcome the limitations of single-point evaluation by introducing well-calibrated intervals, providing a trustworthy measure of caption reliability.

Experimental findings demonstrate that using conformal risk control, over the distributions produced with simple methods for expressing uncertainty, such as masking parts of the input, can achieve competitive performance on word error detection compared to more complex and specialized approaches. Conformal risk control can also provide improvements in correlation between uncertainty estimations and prediction errors, enhancing the overall reliability of the caption evaluation metrics. Furthermore, we emphasize that other existing state-of-the-art methods can also benefit from our conformal calibration framework, gaining formal guarantees over their results. The proposed methodology is model-agnostic, and our work underscores risk control's adaptability and broad applicability, offering a compelling case for its integration into vision and language research.

2 Related Work

Recently, there has been a paradigm shift toward the use of reference-free evaluation metrics for assessing image captioning models. One of the pioneering metrics in this new approach is CLIPScore (Hessel et al., 2021), which evaluates captions without ground-truth references. Built on the Contrastive Language-Image Pretraining (CLIP) model (Radford et al., 2021), CLIPScore calculates a modified cosine similarity between representations of the image and the caption under evaluation. This approach has shown high correlation with human judgments, outperforming established reference-based metrics like BLEU and CIDEr (Vedantam et al., 2015). CLIPScore has become a widely adopted metric for image caption evaluation, inspiring the development of numerous new learned evaluation metrics that build on CLIP (Sarto et al., 2023; Hu et al., 2023; Kim et al., 2022; Gomes et al., 2025).

However, scoring alone is insufficient for comprehensive evaluation, leading to an increasing amount of recent studies focused on identifying specific misalignments between images and texts. Shekhar et al. (2017) introduced the FOIL-it benchmark, featuring data with misalignments by replacing nouns in MS-COCO (Lin et al., 2014) captions with semantically similar alternatives. Building on this foundation, ALOHa (Petryk et al., 2024) expanded the scope by addressing misalignments involving a broader range of objects, particularly visual concepts under-represented in training data for captioning models (Agrawal et al., 2019).

In terms of recent methods for detecting misalignments, Rich-HF (Liang et al., 2024) employs human-annotated datasets of mismatched keywords and implausible image regions, to train a multimodal language model capable of providing dense alignment feedback. In turn, Nam et al. (2024) introduced a novel approach for detecting dense misalignments using pre-trained CLIP models. Their method refines gradient-based attribution computations, leveraging negative gradients of individual text tokens as indicators of misalignment.

3 From Point Estimates to Distributions

Recent studies with similar goals in other fields, such as machine translation evaluation, have employed techniques like deep ensembles or Monte Carlo (MC) dropout to construct output distributions using instance regressor systems (Lakshminarayanan et al., 2017; Kendall and Gal, 2017; Glushkova et al., 2021; Zhan et al., 2023). Unfortunately, neither approach is fully model-agnostic or fits our specific objectives. Deep ensembles are unsuitable since we aim to measure the uncertainty of individual publicly available models, without further training, and MC dropout is impractical since CLIP models generally lack dropout layers.

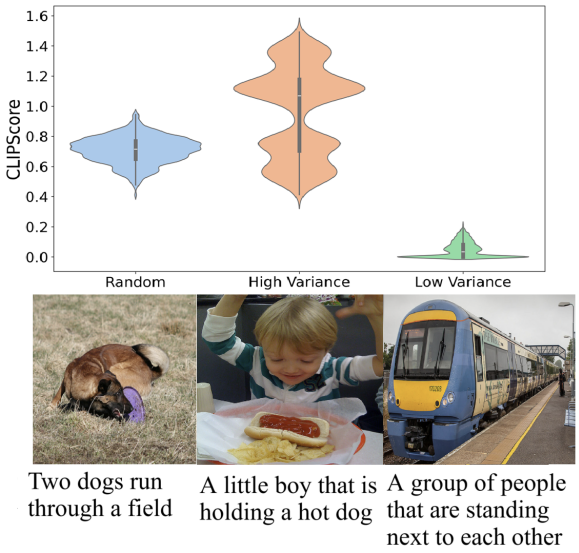

Figure 1: Violin plots of the CLIPScore distributions.

We propose an alternative strategy for producing score distributions that express uncertainty, leveraging the attention masks of the CLIP vision and text encoders to generate output distributions by randomly masking portions of the input data. We create I samples for images by randomly masking $\xi_i\%$ of the attention patches. For captions, we generate T samples by randomly masking $\xi_t\%$ of the attention tokens, corresponding to specific parts of speech, namely nouns, proper nouns, numerals, verbs, adjectives, and adverbs. This strategy allows us to produce I image embeddings and T text embeddings, which can be combined to compute $I \times T$ different CLIPScore values, following the procedure outlined in Appendix A. Figure 1 presents violin plots illustrating the CLIPScore distributions for three cases from the VICR dataset: a random image-caption pair, a high-variance instance identified by our method, and a low-variance instance according to our method.

4 Conformal Detection of Caption Errors

In this section, we describe the application of conformal risk control for detecting caption errors in misaligned image-text pairs. Leveraging the attention mask sampling method described in Section

3, we can calibrate a control variable λ that acts as a threshold to identify wrong words in the caption. Empirical results show that this method provides a good performance across several well-established benchmarks in the field (Shekhar et al., 2017; Petryk et al., 2024; Liang et al., 2024). Furthermore, we compare the results of our simple yet robust and well-calibrated method, against more complex, specialized, and state-of-the-art approaches, underscoring its advantages and effectiveness.

4.1 Deriving Per-Word Error Estimates

The proposed attention mask sampling method generates the CLIPScore distribution output by systematically masking parts of the input. This process inherently facilitates the evaluation of each word's contribution to the overall CLIPScore value.

First we perform T iterations of the text encoder mask sampling process. For each iteration, we mask a set of words in the caption, W_t, using the attention mask in the text encoder to produce a text mask embedding (E_t^M). For each masked word w_j we keep track of its index j in the original caption. We define E_C as the text embedding of the original caption. Then, we compute the CLIPScore difference between the resulting text mask embedding and the original caption text embedding, with respect to I image embeddings generated by randomly masking patches of the image (E_i^M) (see Section 3). The degree of contribution of W_t to the original CLIPScore can be quantified as the average of this difference over the I images, as formally described in Equation 1.

$$v_t = \frac{1}{I}\sum_{i=1}^{I}\left(\text{CLIPS}(E_t^M, E_i^M) - \text{CLIPS}(E_C, E_i^M)\right) \quad (1)$$

Note that a positive difference indicates that the masked words negatively contributed to the CLIPScore value in the original caption. Consequently, these words are more likely to act as misaligned words, which diminish the overall relevance or coherence of the caption in relation to the image.

Next, we aggregate the results of Equation 1 over the indexes j of the masked words, obtaining the average error scores $V[j]$, as follows:

$$V[j] = \frac{1}{\sum_t^T \mathbf{1}_{\{w_j \in W_t\}}} \sum_t^T v_t \cdot \mathbf{1}_{\{w_j \in W_t\}}. \quad (2)$$

To create the error score vector f_v, we apply a sigmoid transformation, $\sigma(\cdot)$, to V, such that

$$f_v[j] = \sigma(V[j]). \quad (3)$$

While the application of the sigmoid function does not enhance performance, it confines the error scores to a finite range, facilitating the implementation of the conformal risk control framework.

4.2 Risk Control on Word Error Detection

Our aforementioned method can already help identify the most likely inadequate word, as the one with the highest score in f_v from Equation 3. However, the approach of simply taking the word with the highest score falls short in two scenarios: multi-class cases where captions may contain no errors and multi-label cases where captions may have multiple inadequate words. To address this, we introduce a threshold-based approach to determine which words should be classified as errors. Specifically, we aim to obtain prediction sets $\mathcal{S}_\lambda(x)$ of misaligned words, defined as follows:

$$\mathcal{S}_\lambda(x) = \{x : f_v(x) > \lambda\}, \quad (4)$$

where the control variable λ acts as a threshold.

Ideally, we aim to optimize the selection of λ so that our prediction sets meet specific user requirements regarding caption quality and error detection. For example, in some tasks, we may prioritize minimizing the false positive rate to ensure that only highly reliable captions are included, while in others, we may focus on reducing the false negative rate to avoid missing potentially useful captions. The choice of λ can alternatively be calibrated to strike the right balance between precision and recall, depending on the task's objectives. To be able to account for these requirements, we rely on conformal risk control (Angelopoulos et al., 2022), since it allows control over different performance criteria, providing statistical guarantees on their bounds. Specifically, let us assume $R(\lambda)$ is a non-increasing and monotonic function of λ, corresponding to our preferred quality criteria. This function serves as a performance metric for $\mathcal{S}_\lambda$, offering an interpretable assessment of its quality.

We can then use a calibration set to get the optimal parameter $\hat{\lambda}$ while ensuring formal guarantees about the risk level. Specifically, for a user-defined risk tolerance α and error rate δ, we aim to satisfy:

$$\mathbb{P}\big(R(\hat{\lambda}) < \alpha\big) \geq 1 - \delta. \quad (5)$$

The procedure that we use to find $\hat{\lambda}$, in order to satisfy the Inequality 5, assumes that we have access to a pointwise Upper Confidence Bound

(UCB) for the risk function for each value of λ:

$$\mathbb{P}\big(R(\lambda) \leq \underbrace{\hat{R}^{+}(\lambda)}_{\text{UCB}}\big) \geq 1 - \delta. \tag{6}$$

We can then choose $\hat{\lambda}$ as the smallest value of λ such that the entire confidence region to the right of $\hat{\lambda}$ falls below the target risk tolerance α:

$$\hat{\lambda} = \inf\Big\{\lambda \in \Lambda : \hat{R}^{+}(\lambda) \leq \alpha,\ \forall \lambda' \geq \lambda\Big\}. \tag{7}$$

As mentioned by Bates et al. (2021), the bound guarantees that act as foundations to obtain the conformal risk-controlling prediction sets, work as long as we have access to a concentration result. In other words, they work as long as we have a mathematical guarantee that the risk is tightly bounded (controlled), and does not deviate too much from its expected value. Therefore, we can construct the UCB for the risk using concentration inequalities. This approach leverages the empirical risk, which is computed by averaging the loss of the set-valued predictor $\mathcal{S}_\lambda$ over a calibration set. The empirical risk is defined as:

$$\hat{R}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(Y_i, \mathcal{S}_\lambda(X_i)), \tag{8}$$

where n is the size of the calibration set, $\mathcal{L}(Y_i, \mathcal{S}_\lambda(X_i))$ represents the loss for each pair (Y_i, X_i), and $\mathcal{S}_\lambda(X_i)$ is the prediction generated by the set-valued predictor for input X_i.

A concentration inequality provides bounds on the tail probabilities of a random variable, and it is typically expressed in the following form:

$$\mathbb{P}\Big(|\hat{R}(\lambda) - R(\lambda)| \geq \epsilon\Big) \leq h(\epsilon; R(\lambda)), \tag{9}$$

where $h(\epsilon; R(\lambda))$ is a non-increasing function of $\epsilon > 0$ and depends on the parameter $R(\lambda)$. By appropriately rearranging this inequality, we can control either the lower or upper tail probability.

In general, a UCB can be obtained if the lower tail probability for $\hat{R}(\lambda)$ of the concentration inequality can be controlled in the following sense:

Proposition 1. Suppose $g(t; R)$ is a non-decreasing function in $t \in \mathbb{R}$ for every R:

$$\mathbb{P}\Big(\hat{R}(\lambda) \leq t\Big) \leq g(t; R(\lambda)). \tag{10}$$

Then, $\hat{R}^{+}(\lambda) = \sup\Big\{R : g(\hat{R}(\lambda); R) \geq \delta\Big\}$ satisfies the Inequality 6. The proof of Proposition 1 can be found in Appendix B.

There are numerous concentration inequalities to choose from. In this work, we opted for a combination of Hoeffding and Bentkus bounds (Bentkus, 2004)[1]. We can obtain a tighter lower tail probability bound for $\hat{R}(\lambda)$, combining Propositions 2 and 3, described in Appendix C. We thus have

$$g^{HB}(t; R(\lambda)) = \min\Big(g^{H}(t; R(\lambda)), g^{B}(t; R(\lambda))\Big),$$

where $g^{H}(t; R(\lambda))$ and $g^{B}(t; R(\lambda))$ refer to the Hoeffding and Bentkus lower tail probability bounds, respectfully.

Applying Proposition 1, we obtain a $(1 - \delta)$ upper confidence bound for $R(\lambda)$ as:

$$\hat{R}^{+}_{HB}(\lambda) = \sup\Big\{R : g^{HB}(\hat{R}(\lambda); R) \geq \delta\Big\}.$$

We can now determine the optimal threshold $\hat{\lambda}$ for calibrating the prediction sets $\mathcal{S}_\lambda(x)$, as defined in Equation 4, by using the upper bound risk from $\hat{R}^{+}_{HB}(\lambda)$ and applying it in Equation 7. This selection for the control variable ensures a formal guarantee that the user-defined risk remains controlled within the specified tolerance, as described in Equation 5, even if the test data deviates slightly from the calibration distribution. However, this guarantee holds only as long as the distribution shift is not too severe, preserving the validity of the concentration result assumption.

4.3 Experimental Results

This section presents the datasets, the evaluation metrics, and the results for misaligned word recognition using the proposed method. For all experiments, we apply our methods on the multilingual LAION ViT-B/32 and LAION ViT-H/14 models as they have shown robust performance on English data (Schuhmann et al., 2022; Gomes et al., 2025).

4.3.1 Datasets and Evaluation Metrics

To ensure a fair and comprehensive evaluation, we used three well-established test benchmarks:

- **FOIL-it**: 198,960 pairs (Shekhar et al., 2017);
- **FOIL-nocaps**: 5,000 pairs (Petryk et al., 2024);
- **Rich-HF**: 955 pairs (Liang et al., 2024).

[1] Exploring other alternatives could lead to the discovery of even tighter bounds for this use case, but it was considered out of scope for this work.

4.3.2 Datasets and Evaluation Metrics

The three datasets associate images with either correct captions or captions containing intentional errors. Among them, FOIL-it and FOIL-nocaps are constructed using the same underlying methodology: one object is replaced by a conceptually similar word (i.e., *dog* can be replaced by *cat*). FOIL-nocaps, built on the nocaps dataset (Agrawal et al., 2019), includes a broader range of visual concepts not typically found in standard training or evaluation datasets, which are often limited to the object classes defined in MS-COCO (Lin et al., 2014). It combines in-domain and out-of-domain captions, with the latter containing novel-class words that captioning models are unlikely to encounter in conventional evaluation datasets, testing our method's ability to generalize beyond familiar concepts.

Since the aforementioned datasets are word-level multi-class benchmarks primarily focused on objects, errors are restricted to nouns. We use the Rich-HF dataset to broaden our evaluation to include multi-label scenarios and a more diverse range of word-level errors. This dataset comprises both AI-generated and human-written prompts resembling captions, collected from the Pick-a-Pic dataset (Kirstain et al., 2023). The creators of Rich-HF carefully selected photo-realistic images for their broader applicability while ensuring a balanced representation across image categories.

Based on these three datasets, we conduct two types of assessments across two different classification tasks: a multi-class task and a multi-label task for detecting misaligned words in captions. The assessments are as follows:

Caption Classification – Determining whether a caption is misaligned. We evaluate this task using average precision (AP) and instance-level F1 score.

Word Error Detection – Identifying specific misaligned words within a caption. For multi-class benchmarks, we measure location accuracy (LA), while for multi-label tasks, we use word-level precision, recall, and F1 score.

To calibrate the threshold in Equation 7, we must define the risk function. Our goal is to detect misaligned words without resorting to trivial solutions of over-detecting most words as misaligned. To achieve this, we control the False Discovery Rate (FDR) for multi-class tasks, and the False Positive Rate (FPR) for multilabel scenarios. In Appendix D, a more detailed explanation of each metric is provided. Those metrics serve as the target risk, enabling us to effectively evaluate the performance of the prediction sets $\mathcal{S}_\lambda$ in Equation 7.

	All Instances					Foil Only	
	Calib. Set		Test Set			Test Set	
α	FDR	F1	FDR	AP	F1	LA	LA_{Set}
10%	9,69	61,74	10,10	60,75	61,93	33,68	34,39
15%	14,62	63,12	15,02	60,34	63,31	37,33	38,53
20%	19,58	63,55	20,20	59,68	63,76	40,15	41,92
25%	24,55	63,21	25,13	58,92	63,56	42,33	44,69
30%	29,52	62,77	30,24	58,04	62,81	44,07	47,06
35%	34,50	61,90	35,25	57,24	61,81	45,60	49,31
40%	39,49	60,65	40,16	56,44	60,49	46,82	51,18
45%	44,48	58,86	45,11	55,58	58,76	47,88	53,11
50%	49,47	56,72	50,27	54,71	56,68	48,81	54,88

Table 1: Calibration results for risk control using the multilingual LAION ViT-B/32 CLIP model, with the FOIL-it dataset as the calibration and test set. The highlighted row corresponds to the best calibration F1 score.

4.3.3 Assessing Multi-Class Guarantees

To assess conformal guarantees on the word level multi-class task, we calibrate the threshold λ using 10% of the FOIL-it validation set and evaluate performance on the FOIL-it and FOIL-nocaps benchmarks. Table 1 presents results for different risk tolerance levels. The findings show that the proposed inequality bounds are able to efficiently align the user-defined tolerance with the observed values for the chosen quality metric (i.e., the FDR), which are consistently below but close to the chosen α.

Increasing the risk tolerance level makes the method more permissive, classifying more words as errors. This improves word-level accuracy but reduces instance-level average precision, as more instances are classified as misaligned. To balance the trade-off between instance-level precision and recall, we rely on the best F1 score, on the calibration set, to select a proper risk tolerance, thus selecting $\alpha = 20\%$. We then use the calibrated outputs at the selected α to compare against state-of-the-art methods for both FOIL-it and the more challenging FOIL-nocaps benchmarks in Table 2. Note that for this table, we only calibrate on FOIL-it (but not FOIL-nocaps) data. By evaluating on benchmarks with different data distributions, we can also assess the validity of the concentration result assumption.

Indeed, empirical results on the FOIL-nocaps dataset indicate a more conservative estimation, as there is a slight deviation between the controlled metric (i.e., the FDR) and the desired tolerance (Table 2). We attribute this to distribution differences between the calibration and test sets. Nevertheless,

	FOIL-it			FOIL-nocaps											
				Overall			In Domain			Near Domain			Out Domain		
Model	**FDR**	**AP**	**LA**	**FDR**	**AP**	**LA**	**FDR**	**AP**	**LA**	**FDR**	**AP**	**LA**	**FDR**	**AP**	**LA**
CHAIR (Rohrbach et al., 2018)	–	92,5	79	–	58,3	14,4	–	57,8	13,5	–	59,1	17,6	–	58,1	12,2
Aloha (Petryk et al., 2024)	–	61,4	40	–	69,5	45,2	–	71,8	47,4	–	66,7	47,3	–	70,9	48,8
GAE_B (Nam et al., 2024)	–	71,4	73,2	–	69,0	60,3	–	67,3	54,7	–	68,4	59,7	–	71,3	63,2
GAE_H (Nam et al., 2024)	–	80,6	83,6	–	79,4	71,6	–	78,9	66,1	–	79,3	70,8	–	80,2	74,8
Our Method with ML LAION ViT-B/32	20,2	59,7	40,2	18,6	64,4	54,9	20,4	70,0	53,5	19,6	72,2	56,3	16,2	74,4	52,6
Our Method with ML LAION ViT-H/14	19,8	63,4	51,4	19,1	65,7	60,3	19,2	70,4	56,7	19,4	72,5	63,0	18,5	74,0	56,2

Table 2: Results of the calibrated sampling method on the FOIL-it and FOIL-nocaps benchmarks.

	Calib. Set		Test Set					
α	**FPR**	**F1**	**FPR**	**AP**	**F1**	**PREC**	**REC**	**F1**
10%	8,09	56,87	7,13	78,22	52,95	21,03	39,29	27,40
15%	12,65	59,68	10,74	79,29	58,97	26,14	43,76	32,73
20%	17,41	61,40	16,92	80,73	65,42	30,43	50,49	37,97
25%	22,16	61,24	24,03	80,44	66,02	31,02	56,80	40,12
30%	27,02	58,75	31,42	80,41	66,76	31,22	62,55	41,65
35%	31,85	56,66	36,84	80,06	66,40	31,15	66,47	42,42
40%	36,74	56,30	40,04	79,48	65,28	31,76	69,41	43,58
45%	41,75	55,07	45,02	78,95	64,15	31,84	72,86	44,32
50%	46,67	54,17	48,64	78,25	62,37	31,80	76,00	44,84

Table 3: Results for risk control using the multilingual LAION ViT-B/32 model, with the Rich-HF validation set for calibration and the test set for evaluation. Highlighted row corresponds to the best calibration F1 score.

our method successfully controls the risk, suggesting that the distribution shift is not too severe, and that the concentration result assumption remains valid. Additionally, our approach achieves performance comparable to ALOHa (Petryk et al., 2024) on the FOIL-it benchmark, and both ALOHa and CHAIR (Rohrbach et al., 2018), on FOIL-nocaps. Notably, both CHAIR and ALOHa are more complex methods, with ALOHa leveraging large language models to detect erroneous words.

Although our method falls short compared to the recent approach by Nam et al. (2024), which employs a sophisticated gradient-based attribution technique where the negative gradient of individual text tokens signals misalignments, we emphasize the simplicity of our attention sampling method to produce CLIPScore distributions, and the model-agnostic nature of our calibration framework. Unlike these more complex approaches that rely on specific architectures or gradient-based computations, our method can be applied to a wide range of models, including the current state-of-the-art systems for further calibration to user-requirements and formal guarantee assessments. Appendix F provides additional qualitative analyses for the FOIL-it and FOIL-nocaps benchmarks.

4.3.4 Assessing Multi-Label Guarantees

To evaluate conformal guarantees in the word-level multi-label task, we calibrate our system on the validation set of Rich-HF, and assess its performance on the corresponding test set.

Model	**ft.**	**PREC**	**REC**	**F1**
ALOHa (Petryk et al., 2024)		34,4	31,1	38,5
Rich-HF (MH) (Liang et al., 2024)	✓	43,3	62,9	33,0
Rich-HF (AP) (Liang et al., 2024)	✓	43,9	61,3	34,1
GAE_B (Nam et al., 2024)		39,8	32,8	50,4
GAE_H (Nam et al., 2024)		42,7	36,5	51,6
Our Method with ML LAION ViT-B/32		31,2	62,6	41,7
Our Method with ML LAION ViT-H/14		32.0	64.2	42.7

Table 4: Results of the calibrated sampling method on the Rich-HF benchmark.

Table 3 presents results over increasing risk tolerance levels. Similarly to the multi-class results, we consistently control the risk to align with the target tolerance level in the calibration set. However, a notable discrepancy emerges between the tolerance level and the risk metric (i.e., the false positive rate) on the calibration set. This discrepancy arises primarily due to the limited size of the Rich-HF calibration set, which contains only 955 samples. The small sample size increases the margin of error for the upper confidence bound, which is an intentional overestimation in order to achieve more general and robust guarantees of risk control, leading to more conservative threshold estimates.

Variability in caption characteristics further affects the applicability of the thresholds. For instance, calibrating on datasets with longer captions but testing on shorter ones will lead to higher thresholds, giving rise to an undesired strict behaviour when classifying misaligned words. In turn, the reverse scenario, i.e., calibrating on shorter captions and testing on longer ones, can produce overly lenient thresholds. Together, these factors influence the ability to reliably control risk across diverse scenarios. Appendix E presents a visualization highlighting the differences between the calibration and test sets of Rich-HF, supporting a better understanding of these differences.

Table 4 compares our calibrated method with current state-of-the-art systems. Despite its simplicity and general-purpose design, our method outperforms both the LLM-based ALOHa approach and the specialized fine-tuned model used in the

Rich-HF benchmark, achieving superior F1 performance. Similarly to the multi-class experiments, our simple method achieved lower F1 scores than the more complex and recent approach by Nam et al. (2024), although in this case we achieved significantly higher recall.

5 Conformalized Intervals for CLIPScore

We now test a second application of risk control over CLIPScore, to address the limitations of single-point evaluation metrics in IC assessments to get reliable and interpretable confidence intervals for each IC score. Leveraging the uncertainty quantification method described in Section 3, we fit a truncated Gaussian distribution to construct intervals. These intervals help quantify model uncertainty more effectively, providing a nuanced and trustworthy assessment of caption quality.

The choice of truncated Gaussian distributions is motivated by CLIPScore being inherently bounded, as it is defined as a modified cosine similarity. In addition, it allows us to define a more meaningful rescaling of initially estimated uncertainties, effectively reordering confidence intervals to align with the deviation from ground truth, as described in the following sections.

5.1 Risk Control on Human Correlation

Calibrating confidence intervals for CLIPScore assessments is particularly challenging because CLIPScore was not trained to predict human judgment scores, but rather to correlate with them. As a result, we cannot rely on typical risk functions such as coverage (Zerva and Martins, 2024), which measures the proportion of times the ground truth falls within the computed confidence intervals. A suitable risk function must account for this indirect relationship, ensuring meaningful calibration.

We propose a new risk function to calibrate our intervals that does not depend on the match of scale between the output distributions and the ground truth, specifically defined as follows:

$$R(\lambda) = 1 - \text{ReLU}(r(|\hat{\mu}(\lambda) - y|, \hat{\sigma}(\lambda)). \quad (11)$$

This risk function leverages the Uncertainty Pearson Score (UPS), denoted as:

$$\text{UPS} = r(|\hat{\mu}(\lambda) - y|, \hat{\sigma}(\lambda)), \quad (12)$$

where r is the Pearson correlation coefficient and y the ground truth (human score) (Glushkova et al., 2021). This metric quantifies the correlation between prediction errors and uncertainty estimates. The values $\hat{\mu}(\lambda)$ and $\hat{\sigma}(\lambda)$ are derived by fitting a truncated Gaussian distribution, using the original mean μ, and scaled standard deviation $\lambda\sigma$. The values for μ and σ are obtained empirically from the CLIPScore distribution obtained via masking.

Notably, the risk function is not monotonically non-increasing. The direct application of the framework described in Section 4 involves the assumption of monotonicity of the risk function, otherwise we cannot extend the pointwise convergence result, from Equation 7, into a result on the validity of a data-driven choice of λ. To address this, we propose a strategy based on the Learn Then Test (LTT) technique (Angelopoulos and Bates, 2021), which leverages the duality between tail probability bounds in concentration inequalities and conservative p-values. This approach enables us to identify $\hat{\lambda}$ that satisfies Equation 5, extending the concentration result assumption to more general and complex risks. The procedure outputs a subset $\hat{\Lambda} \subseteq \Lambda$, ensuring all selected sets $\hat{\Lambda}$ of λ values control the user-defined risk. We describe the process below.

Step 1: We first define the risk tolerance α. Our objective is to calibrate λ such that the resulting risk level is lower than the initial one. Looking at Equation 11, this implies maximizing a positive correlation between estimated uncertainties and deviation from the ground truth which naturally leads to more reliable and interpretable uncertainties. Thus, we set α as the risk $R(\lambda)$ at $\lambda = 1$:

$$\alpha = 1 - \text{ReLU}(r(|\hat{\mu}(1) - y|, \hat{\sigma}(1)). \quad (13)$$

Step 2: For each $\lambda \in \Lambda$, in which Λ refers to the set of acceptable values, we associate the null hypothesis $\mathcal{H}_\lambda : R(\lambda) > \alpha$. Note that rejecting $\mathcal{H}_\lambda$ means the selection of a value for λ that controls the user-defined risk.

Step 3: As noted by Bates et al. (2021), the upper bound $g(\hat{R}(\lambda); R)$, derived from Proposition 1, can be interpreted as a conservative p-value for testing the one-sided null hypothesis $\mathcal{H}_0 : R(\lambda) > R$. Therefore, for each null hypothesis $\mathcal{H}_\lambda$, we can compute conservative p-values p_λ using $g(\hat{R}(\lambda); \alpha)$ to test the hypothesis $\mathcal{H}_\lambda : R(\lambda) > \alpha$.

Step 4: Return $\hat{\Lambda} = \mathcal{A}(\{p_\lambda\}_{\lambda \in \Lambda})$, where $\mathcal{A}$ is an algorithm designed to control the Family-Wise Error Rate (FWER). This is important because, when conducting multiple hypothesis tests, the probability of making at least one Type I error

increases as the number of tests grows. Each individual test has a small chance of being a false positive (e.g., $p_\lambda < 0.05$), but as more tests are performed, these small probabilities accumulate, raising the overall risk of an error. For the case where $\Lambda = \{\lambda : p_\lambda < \delta\}$, the FWER is given by:

$$\text{FWER}(\Lambda) = 1 - (1 - \delta)^{|\Lambda|}. \quad (14)$$

We will use throughout the experiments the Bonferroni correction, which tests each hypothesis at level $\delta/|\Lambda|$, ensuring that the probability of at least one failed test is no greater than δ by the union bound.

$$\hat{\Lambda} = \{\lambda : p_\lambda < \frac{\delta}{|\Lambda|}\}. \quad (15)$$

Step 5: With the set $\hat{\Lambda}$ containing all the λ values that successfully control the user-defined risk with statistical significance, we can further refine the selection using other specific metrics on the calibration set. In this case, we aim to identify $\hat{\lambda}$, which maximizes the UPS. Given our chosen risk (Equation 11), this corresponds naturally to the λ value with the lowest p-value.

5.2 Experimental Results

This section presents the datasets, the evaluation metrics, and the results for conformalizing CLIPScore intervals using the proposed method.

5.2.1 Datasets and Evaluation Metrics

To ensure a fair and comprehensive evaluation, we used four well-established datasets designed to evaluate the correlation between vision-and-language model outputs and human judgments:

- **VICR**: $3,161$ instances (Narins et al., 2024);
- **Polaris**: $8,726$ instances (Narins et al., 2024);
- **Ex-8k**: $5,664$ instances (Hodosh et al., 2013);
- **COM**: $13,146$ instances (Aditya et al., 2015).

We will use the validation set of VICR to calibrate the CLIPScore confidence intervals described in the previous section and assess both the human judgment correlation (Kendal-τ_C), and the correlation between prediction errors and uncertainty estimates (UPS). As mentioned in Section 5.1, to calibrate the scaling factor of the standard deviation, we use the Uncertainty Pearson Risk (UPR) function shown in Equation 11. To evaluate our results, we use UPS and Accuracy. In Appendix D, we provide a more detailed explanation of each metric.

	VICR		Polaris		EX-8K		COM	
Method	**UPS**	τ_c	**UPS**	τ_c	**UPS**	τ_c	**UPS**	τ_c
B-PRE	$22,1$	$63,1$	$38,1$	$50,1$	$2,8$	$53,1$	$18,3$	$47,2$
B-POS	$36,4$	$61,5$	$44,1$	$49,4$	$13,4$	$51,9$	$26,1$	$46,9$
H-PRE	$42,6$	$67,8$	$60,2$	$51,0$	$24,0$	$56,9$	$18,3$	$54,6$
H-POS	$49,6$	$66,4$	$70,1$	$50,6$	$23,1$	$55,8$	$27,1$	$53,6$

Table 5: Performance before (PRE) and after (POS) calibration of the CLIPScore confidence intervals across two model sizes: B (ViT-B/32) and H (ViT-H/14).

5.2.2 Guarantees on Maximal Correlation

In this section, we evaluate the performance gains achieved through the risk control calibration process applied to CLIPScore distributions obtained by fitting a truncated Gaussian to the output distributions of the attention sampling method. Our primary objective is to improve the correlation between prediction errors and uncertainty estimates (i.e., the standard deviation), which is measured by the UPS metric, while preserving overall system performance on external metrics, specifically by maintaining a strong correlation between the interval's mean value and human judgments.

Table 5 presents results before (PRE) and after (POS) calibration of the CLIPScore confidence intervals. For both model sizes, we achieve a significant improvement in performance in terms of UPS across all datasets without significantly compromising the correlation with human ratings. Hence, our findings align with our original objective, providing a lightweight, model-agnostic methodology for obtaining more reliable confidence intervals over caption scores.

6 Conclusions

We proposed a method for producing and calibrating distributions on CLIPScore assessments, enabling granular caption evaluation and uncertainty representation. We leverage conformal risk control to address word-level misalignment detection and confidence estimation, allowing for flexible, task-specific risk-control with formal guarantees. The experimental results demonstrate competitive performance against more complex models on several well-established benchmarks while allowing for a more controllable and trustworthy performance in detecting misaligned words and improved correlation between uncertainty estimates and prediction errors without compromising human rating alignment. Our work highlights the potential of conformal calibration in enhancing the robustness and reliability of vision-and-language evaluation metrics.

Limitations and Ethical Considerations

The research reported on this paper aims to enhance transparency and explainability, given that we advanced methods that can shed new light into the evaluation process of image captioning models.

Our research aimed to enhance the transparency and explainability of image captioning model evaluations by introducing methods that offer uncertainty intervals and identify misaligned words within captions. It is nonetheless important to notice that our research does not specifically tackle potential biases in the CLIPScore evaluation metric (or biases existing in the popular benchmark datasets that also supported our experiments), neither does it address specific known limitations associated to CLIP models. Additionally, our experiments were conducted exclusively in English, leaving open questions about the generalizability of our conformal risk control framework and word-level assessment to other languages, especially those with distinct morphological structures or syntactic complexities. Previous work has shown that uncertainty quantification methods are broadly applicable across languages, but often require language-specific calibration to ensure fair, balanced performance (Zerva and Martins, 2024). Expanding our approach to linguistically diverse datasets is an important direction for future work.

While our method improves interpretability and provides well-calibrated CLIPScore intervals, human evaluation remains indispensable for ensuring the reliability of model assessments. Automated metrics should complement, not replace, human judgment, especially in sensitive applications, where misinterpretations can have significant consequences. Caution is essential when calibrating uncertainty, as miscalibrated intervals may foster unwarranted confidence, particularly in high-risk contexts. Future research should prioritize expanding linguistic diversity, refining uncertainty quantification techniques, and integrating large-scale human validation to improve the robustness and reliability of our approach.

We also note that we used GitHub Copilot during the development of our research work, and we used ChatGPT for minor verifications during the preparation of this manuscript.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Anastasios N Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2022. Conformal risk control. *arXiv preprint arXiv:2208.02814*.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6).

Vidmantas Bentkus. 2004. On hoeffding's inequalities. *The Annals of Probability*, 32(2).

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. PaLI: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations*.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938.

Gonçalo Gomes, Chrysoula Zerva, and Bruno Martins. 2025. Evaluation of multilingual image captioning: How far can we get with clip models? *arXiv preprint arXiv:2502.06600*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*.

Wassily Hoeffding. 1994. Probability inequalities for sums of bounded random variables. *The Collected Works of Wassily Hoeffding*.

Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. InfoMetIC: An informative metric for reference-free image caption evaluation. *arXiv preprint arXiv:2305.06002*.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. 2022. Mutual information divergence: A unified metric for multimodal generative models. In *Proceedings of the Annual Meeting on Neural Information Processing Systems*.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Proceesings of the Annual Meeting on Neural Information Processing Systems*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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 *Proceeding of the European Conference on Computer Vision*.

JeongYeon Nam, Jinbae Im, Wonjae Kim, and Taeho Kil. 2024. Extract free dense misalignment from CLIP. *arXiv preprint arXiv:2412.18404*.

Lothar D Narins, Andrew Scott, Aakash Gautam, Anagha Kulkarni, Mar Castanon, Benjamin Kao, Shasta Ihorn, Yue-Ting Siu, James M Mason, Alexander Blum, et al. 2024. Validated image caption rating dataset. In *Proceedings of the Annual Meeting on Neural Information Processing Systems*.

Suzanne Petryk, David M Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E Gonzalez, and Trevor Darrell. 2024. ALOHa: A new measure for hallucination in captioning models. *arXiv preprint arXiv:2404.02904*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

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

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chrysoula Zerva and André F. T. Martins. 2024. Conformalizing machine translation evaluation. *Transactions of the Association for Computational Linguistics*, 12:1460–1478.

Runzhe Zhan, Xuebo Liu, Derek F Wong, Cuilian Zhang, Lidia S Chao, and Min Zhang. 2023. Test-time adaptation for machine translation evaluation by uncertainty minimization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

A The CLIPScore Metric

We now formally describe the CLIPScore metric (Hessel et al., 2021). In brief, CLIPScore is based on a modified cosine similarity between representations for the input image and the caption under evaluation. The image and the caption are both passed through the respective feature extractors from a given CLIP model. Then, we compute the cosine similarity of the resultant embeddings, adjusting the resulting value through a re-scaling operation. For an image with visual CLIP embedding $\mathbf{v}$ and a candidate caption with textual CLIP embedding

$\mathbf{c}$, a re-scaling parameter is set as $w = 2.5$ and we compute the corresponding CLIPScore as follows:

$$\text{CLIPScore}(\mathbf{c}, \mathbf{v}) = w \times \max(\cos(\mathbf{c}, \mathbf{v}), 0). \quad (16)$$

Since CLIPScore is derived from a modified cosine similarity, it naturally inherits its bounded nature. As a result, CLIPScore values always fall within the interval $[0, 2.5]$.

Note that CLIPScore does not depend on the availability of underlying references for each of the images in an evaluation dataset, hence corresponding to a reference-free image captioning evaluation metric.

B Proof of Proposition 1

The proof for Proposition 1 uses the theorem of probability of subset events.

Theorem 1. If A and B are events in a probability space such that $A \subseteq B$, then:

$$\mathbb{P}(A) \leq \mathbb{P}(B). \quad (17)$$

This is true because probability is additive over disjoint sets and satisfies:

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \backslash A), \quad (18)$$

where $B \backslash A$ represents the part of B not in A.

Using the previous theorem, the proof of Proposition 1 will be divided in three steps, which we describe next.

Step 1. Proof of the following equation:

$$\mathbb{P}\Big(R(\lambda) > \hat{R}^{+}(\lambda)\Big) \leq \mathbb{P}\Big(g(\hat{R}(\lambda); R) < \delta\Big).$$

By construction, $R(\lambda) > \hat{R}^{+}(\lambda)$ implies that $g(R(\lambda); R) < \delta$, because $\hat{R}^{+}(\lambda)$ was chosen as the supremum of R in the following set:

$$\Big\{R : g(\hat{R}(\lambda); R(\lambda)) \geq \delta\Big\}.$$

This establishes that the event $R(\lambda) > \hat{R}^{+}(\lambda)$ necessarily leads to $g(R(\lambda); R) < \delta$. However, the converse does not hold. In other words, the event $R(\lambda) > \hat{R}^{+}(\lambda)$ is strictly contained within the event $g(R(\lambda); R) < \delta$. Applying Theorem 1, we can conclude that:

$$\mathbb{P}\Big(R(\lambda) > \hat{R}^{+}(\lambda)\Big) \leq \mathbb{P}\Big(g(\hat{R}(\lambda); R) < \delta\Big).$$

Next, let G be the CDF of $\hat{R}(\lambda)$:

$$G(t) = \mathbb{P}(\hat{R}(\lambda) \leq t). \quad (19)$$

This implies that $G(t) \leq g(t; R(\lambda))$.

Step 2. Proof of the following equation:

$$\mathbb{P}\Big(g(\hat{R}(\lambda); R) < \delta\Big) \leq \mathbb{P}\Big(G(\hat{R}(\lambda)) < \delta\Big).$$

By definition, $g(t; R)$ serves as an upper bound of $G(t)$. Therefore, the event $g(\hat{R}(\lambda); R) < \delta$, necessarily leads to $G(\hat{R}(\lambda)) < \delta$. However, the converse does not hold. Applying Theorem 1, we can conclude that:

$$\mathbb{P}\Big(g(\hat{R}(\lambda); R) < \delta\Big) \leq \mathbb{P}\Big(G(\hat{R}(\lambda)) < \delta\Big).$$

Step 3. Proof of the following equation:

$$\mathbb{P}\Big(G(\hat{R}(\lambda)) < \delta\Big) \leq \mathbb{P}\Big(\hat{R}(\lambda) < G^{-1}(\delta)\Big).$$

By definition, $G^{-1}(\lambda) = \sup\Big\{x : G(x) \leq \delta\Big\}$, which means that $G^{-1}(\lambda)$ is the highest value satisfying $G(x) \leq \delta$. Therefore, this will always imply $x \leq G^{-1}(\lambda)$. However, the converse is not always guaranteed. Because the event $G(\hat{R}(\lambda)) < \delta$ is strictly contained within the event $\hat{R}(\lambda) < G^{-1}(\delta)$, we can apply Theorem 1, proving:

$$\mathbb{P}\Big(G(\hat{R}(\lambda)) < \delta\Big) \leq \mathbb{P}\Big(\hat{R}(\lambda) < G^{-1}(\delta)\Big).$$

Finally, since the event $\hat{R}(\lambda) < G^{-1}(\delta)$ is strictly contained in $\hat{R}(\lambda) \leq G^{-1}(\delta)$, by applying Theorem 1 we have:

$$\mathbb{P}\Big(\hat{R}(\lambda) < G^{-1}(\delta)\Big) \leq \mathbb{P}\Big(\hat{R}(\lambda) \leq G^{-1}(\delta)\Big).$$

Next, using the definition of $G(x)$, we have that:

$$\mathbb{P}\Big(\hat{R}(\lambda) \leq G^{-1}(\delta)\Big) = G(G^{-1}(\delta)),$$

which by definition leads to $G(G^{-1}(\delta)) \leq \delta$.

Combining all the inequalities proved in each step, we have that:

$$\mathbb{P}\Big(R(\lambda) > \hat{R}^{+}(\lambda)\Big) \leq \delta. \quad (20)$$

Inverting the probability expression yields:

$$\mathbb{P}\Big(R(\lambda) \leq \hat{R}^{+}(\lambda)\Big) \geq 1 - \delta, \quad (21)$$

thus completing the proof.

C Concentration Inequalities

Concentration inequalities provide probabilistic bounds on the deviation of a random variable from its expected value, playing a crucial role in statistical learning theory and probability analysis. This section presents key concentration inequalities, including Hoeffding's and Bentkus' inequalities.

Proposition 2 (Hoeffding's inequality, tighter version (Hoeffding, 1994))**.** Suppose that $g(t; R)$ is a nondecreasing funtion in $t \in \mathbb{R}$ for every R. Then, for any $t < R(\lambda)$, we have that:

$$\mathbb{P}\Big(\hat{R}(\lambda) \leq t\Big) \leq \exp\{-n \cdot f(t; R(\lambda))\},$$

where:

$$f(t; R) = t \cdot \log\left(\frac{t}{R}\right) + (1-t) \cdot \log\left(\frac{1-t}{1-R}\right).$$

The weaker Hoeffding inequality is implied by Proposition 2, noting that $f(t; R) \geq 2(R-t)^2$.

Proposition 3 (Bentkus inequality (Bentkus, 2004))**.** Supposing the loss is bounded above by one, we have that:

$$\mathbb{P}\Big(\hat{R}(\lambda) \leq t\Big) \leq e\mathbb{P}\Big(\text{Bi}(n, R(\lambda)) \leq \lceil nt \rceil\Big),$$

where $\text{Bi}(n, p)$ denotes a binomial random variable with sample size n and success probability p.

D Details on Metrics

This section provides a detailed overview of the metrics used for calibrating the controlled variable and the evaluation metrics applied throughout the two different types of experiments.

D.1 Metrics Used as Risks

The following metrics were used to calibrate the threshold on the experiments regarding detecting misaligned words in the caption.

False Discovery Rate (FDR): This metric is a statistical concept used to control the expected ratio of the number of False Positive classifications (FP) over the total number of positive classifications, including True Positives, (FP + TP). Mathematically, the False Discovery Rate is defined as:

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}}. \tag{22}$$

False Positive Rate (FPR): This metric is a statistical measure used to evaluate the proportion of actual negative instances that are incorrectly classified as positive by a model. It represents the likelihood of a false alarm, where the model predicts a positive outcome when the true outcome is negative. Mathematically, the False Positive Rate is defined as follows:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{23}$$

where FP denotes the number of False Positives, and TN represents the number of True Negatives.

D.2 Evaluation Metrics

The following metrics were applied throughout the experiments to evaluate our methods.

F1-Score: The F1-score is a harmonic mean of precision and recall, providing a single metric that balances both measures. It is particularly useful in scenarios where class imbalance exists, as it considers both False Positives (FP) and False Negatives (FN). Mathematically, the F1-score is defined as:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{24}$$

In turn, Precision is defined as TP/(TP + FP), and Recall is defined as TP/(TP + FN). The F1-score ranges from 0 to 1, where a higher value indicates better model performance in terms of balancing precision and recall.

Average Precision (AP): The Average Precision is a metric commonly used in information retrieval and classification tasks, particularly for evaluating models with imbalanced datasets. It summarizes the precision-recall curve by calculating the weighted mean of precision achieved at each recall threshold, with the increase in recall serving as the weight. Mathematically, it is defined as:

$$\text{AP} = \sum_n (R_n - R_{n-1}) \cdot P_n, \tag{25}$$

where P_n and R_n are the precision and recall at the n-th threshold. Average Precision (AP) provides a single score that reflects the model's ability to correctly rank positive instances, with values closer to 1 indicating better performance.

Location Accuracy (LA): Localization Accuracy measures the fraction of samples where we can correctly identify a hallucinated object, among samples that are known to contain hallucinated objects. A sample receives LA_{set} of 1 if at least one of the predicted hallucinated objects was correct,

and an LA of 1 if the minimum matching score was a true hallucination.

Uncertainty Pearson Score (UPS): This metric is a statistical measure used to evaluate the correlation between the absolute error of predictions and their associated uncertainty estimates. It quantifies how well the model's uncertainty estimates aligns with the actual prediction errors, providing insight into the reliability of the uncertainty quantification. Mathematically, the Uncertainty Pearson Score (UPS) is defined as follows:

$$\text{UPS} = \text{P}_C\left(|\mu(\lambda) - y|, \sigma(\lambda)\right), \qquad (26)$$

where $|\mu(\lambda) - y|$ represents the absolute error between the predicted value $\mu(\lambda)$ and the true value y, and $\sigma(\lambda)$ is the estimated uncertainty. A higher UPS indicates better calibration of uncertainty estimates, as it reflects a stronger correlation between prediction errors and uncertainty.

Kendall Tau C: Seeing each of our evaluation datasets as a set of n observations with the form $(\hat{y}_1, y_1), \ldots, (\hat{y}_n, y_n)$, for predicted scores $\hat{y}_i$ and reference ratings y_i, the Kendall Tau C correlation coefficient assesses the strength of the ranking association between the predicted scores and the reference ratings. Unlike Kendall Tau B, which accounts for ties, Kendall Tau C is specifically designed to handle cases where the underlying scales of the scores are different, such as when the number of possible ranks for the predicted scores and the reference ratings differ.

A pair of observations $(\hat{y}_i, y_i)$ and $(\hat{y}_j, y_j)$, where $i < j$, is considered concordant if the sort order of the instances agrees (i.e., if either both $\hat{y}_i > \hat{y}_j$ and $y_i > y_j$ hold, or both $\hat{y}_i < \hat{y}_j$ and $y_i < y_j$ hold). Otherwise, the pair is discordant. The Kendall Tau C coefficient is defined as:

$$\tau_c = \frac{n_c - n_d}{n_0} \times \frac{n-1}{n} \times \frac{m}{m-1}, \qquad (27)$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, $n_0 = n(n-1)/2$ is the total number of possible pairs, and m is the number of distinct values in the ranking scale for the reference ratings. The term $\frac{m}{m-1}$ adjusts for the difference in scale between the predicted scores and the reference ratings, making Kendall Tau C particularly suitable for datasets that feature unequal ranking scales in the predictions and the references.

E Description of the Datasets

The following datasets were used in the calibration and evaluation of our method for detecting misaligned words in captions.

- **Foil-it (Shekhar et al., 2017):** The Foil-it dataset is a synthetic hallucination dataset based on samples from the MS-COCO (Lin et al., 2014) dataset. In this dataset, for each candidate-image pair, a "foil" caption is created which swaps one of the objects (in the MS-COCO detection set), in the caption, with a different and closely related neighbour (chosen by hand to closely match, but aiming to be visually distinct). In our experiments, we used the test split of the Foil-it dataset, which includes $198,8814$ unique image-caption pairs. For calibration, we used 10% of the validation split, which comprises a total of $395,300$ unique image-caption pairs.
- **Foil-nocaps (Petryk et al., 2024):** The FOIL-nocaps dataset was introduced to address limitations of the FOIL-it dataset, which is overly biased towards object-classes present in the MS-COCO dataset. The FOIL-nocaps dataset is based on the nocaps dataset (Agrawal et al., 2019), which consists of images from the OpenImages dataset annotated with captions in a style similar to MS-COCO. The nocaps dataset is divided into three subsets (i.e., in-domain, near-domain, and out-of-domain) based on the relationship of the objects in the images to those in the MS-COCO dataset. Compared to Foil-it, this new dataset aims to provide a more general benchmark for evaluating hallucination detection methods, by including a broader range of object categories and contexts. In our tests, we used the test split of the Foil-nocaps dataset, which includes $5,000$ unique image-caption pairs.
- **Rich-HF (Liang et al., 2024):** The Rich-HF dataset is a comprehensive benchmark for evaluating text-to-image alignment, comprising $18K$ image-text pairs with rich human feedback. It was constructed by selecting a diverse subset of machine generated photo-realistic images from the Pick-a-Pic (Kirstain et al., 2023) dataset, ensuring balanced use of categories such as 'human', 'animal', 'object', 'indoor scene', and 'outdoor scene'. The dataset is annotated using the PaLI (Chen

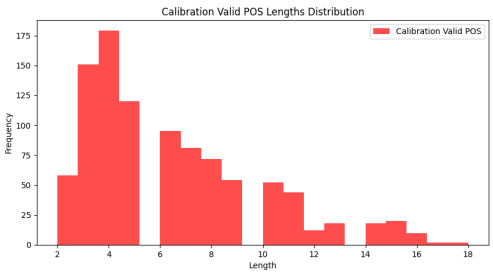

Figure 2: Frequency of sequences, with a given length, featuring words with valid parts of speech used for attention mask sampling in the Rich-HF calibration set.

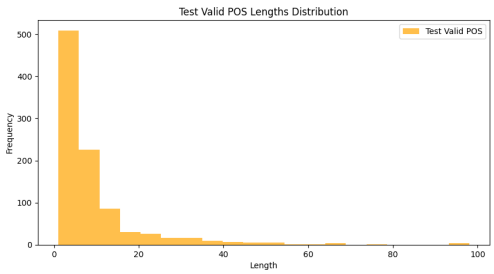

Figure 3: Frequency of sequences, with a given length, featuring words with valid parts of speech used for attention mask sampling in the Rich-HF test set.

et al., 2022) visual question answering model to extract basic features and ensure diversity. Rich-HF includes $16K$ training samples, 955 validation samples, and 955 test samples, with additional human feedback collected on unique prompts and their corresponding images. The dataset provides word-level misalignment annotations and overall alignment scores, making it a valuable resource for evaluating fine-grained text-to-image alignment and hallucination detection methods. Additionally, Rich-HF includes 955 prompt-image pairs with detailed word-level misalignment annotations, covering a wide range of caption lengths, styles, and contents, due to its collection from real users. In our tests, we used the test split of the Rich-HF dataset, and for calibration, we used the validation split.

While calibrating our methods using the Rich-HF dataset, we observed a significant difference in the distribution of the number of words per caption, between the calibration and test sets. Specifically, this disparity applies to words corresponding to valid parts of speech used in our attention mask sampling method, namely, nouns, proper nouns, numerals, verbs, adjectives, and adverbs. As noted in the main manuscript, this variation directly impacts the applicability of the thresholds. Figures 2 and 3 show histograms illustrating the frequency of sequences, with a given length, featuring words with valid parts of speech used for attention mask sampling in the calibration and test sets, respectively. The figures illustrate the significant differences.

The following datasets were used in the calibration and evaluation experiments that assessed the Uncertainty Pearson Score (UPS), and correlation with human judgments.

- **Flickr8K-Expert (Hodosh et al., 2013):** This dataset comprises $16,992$ expert human judgments for $5,664$ image-caption pairs from the Flickr8K dataset. Human assessors graded captions on a scale of 1 to 4, where 4 indicates a caption that accurately describes the image without errors, and 1 signifies a caption unrelated to the image.
- **Composite (Aditya et al., 2015):** This dataset contains $13,146$ image-caption pairs taken from MS-COCO (2007 images), Flickr8K (997 images), and Flickr30K (991 images). Each image originally had five reference captions. One of these references was chosen for human rating and subsequently removed from the reference set that is to be used when assessing evaluation metrics.
- **VICR (Narins et al., 2024):** The Validated Image Caption Rating (VICR) dataset features 68,217 ratings, collected through a gamified approach, for $15,646$ image-caption pairs involving $9,990$ distinct images. The authors of the dataset demonstrated that it exhibits a superior inter-rater agreement compared to other alternatives (e.g., an improvement of 19% in Fleiss' κ when compared to the agreement for the Flickr8K-Expert dataset), and it features a more balanced distribution across various levels of caption quality. In our tests, we used the test split of the VICR dataset, which includes $3,161$ unique image-caption pairs, with $2,000$ images from the MS-COCO 2014 validation dataset and $1,161$ images from the Flickr8K dataset. For calibration, we used the validation split, which comprises $2,310$ unique image-caption pairs.
- **Polaris (Wada et al., 2024):** The Polaris dataset comprises $131,020$ human judgments on image-caption pairs, collected from 550

evaluators. It surpasses existing datasets in scale and diversity, offering an average of eight evaluations per caption, significantly more than Flickr8K (three) and CapEval1K (five). Polaris includes captions generated by ten standard image captioning models, covering both modern and older architectures to ensure output diversity. In our tests, we used the test split of the Polaris dataset, which includes $8,726$ unique image-caption pairs. For calibration, we used the validation split, which comprises $8,738$ unique image-caption pairs.

F Qualitative Results

We conducted a small qualitative study on the multi-class classification task of detecting misaligned words in the Foil-it (Figure 4) and Foil-nocaps (Figure 5) benchmarks, as well as the multi-label classification task using the Rich-HF benchmark (Figure 6). Throughout these qualitative experiments, captions associated with each image follow a color-coded scheme to indicate model performance in detecting misaligned words. Specifically, green highlights true positives, where our model correctly identified a misaligned word. Yellow indicates false negatives, meaning the model failed to detect an incorrect word. Lastly, red denotes false positives, where the model mistakenly flagged a word as misaligned when it was actually correct. Captions without coloured words are entirely correct according to the respective benchmark. This visual coding allows for an intuitive assessment of our model's strengths and weaknesses in the different benchmarks.

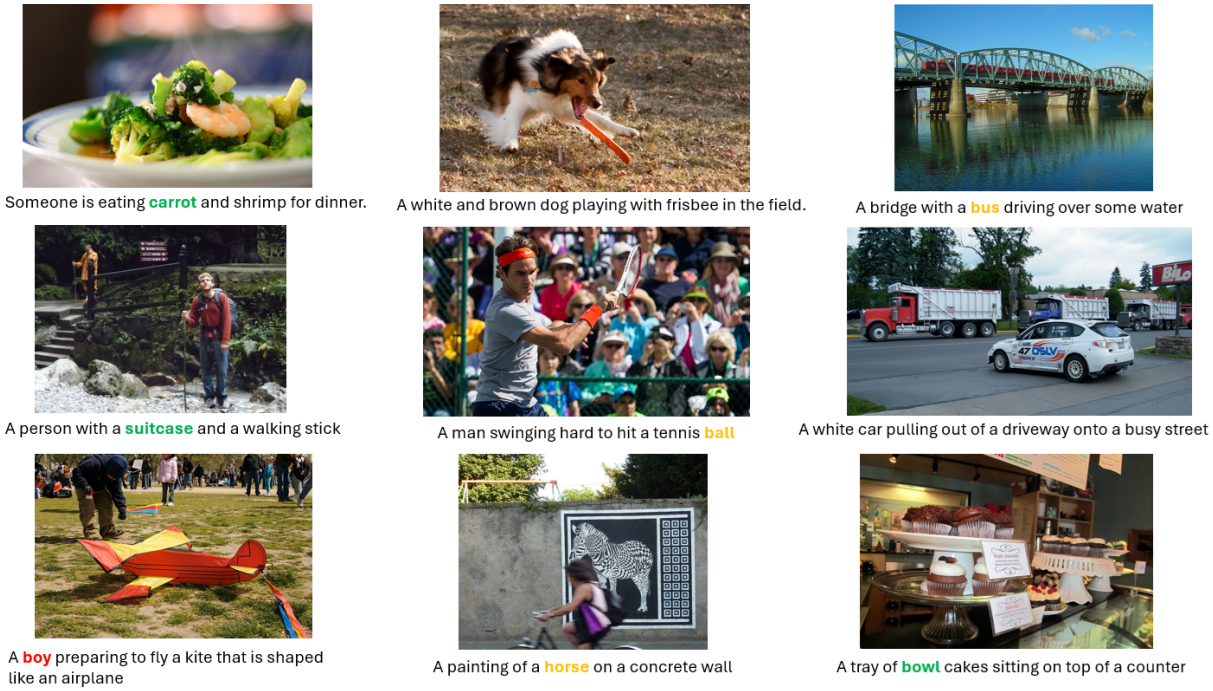

Figure 4: Qualitative results of the calibrated sampling method using the multilingual LAION ViT-H/14 model on the Foil-it test set. For this results, our method was calibrated to 20% False Discovery Rate using the Foil-it validation set for calibration.

Figure 5: Qualitative results of the calibrated sampling method using the multilingual LAION ViT-H/14 model on the Foil-nocaps test set. For this results, our method was calibrated to 20% False Discovery Rate using the Foil-it validation set for calibration.

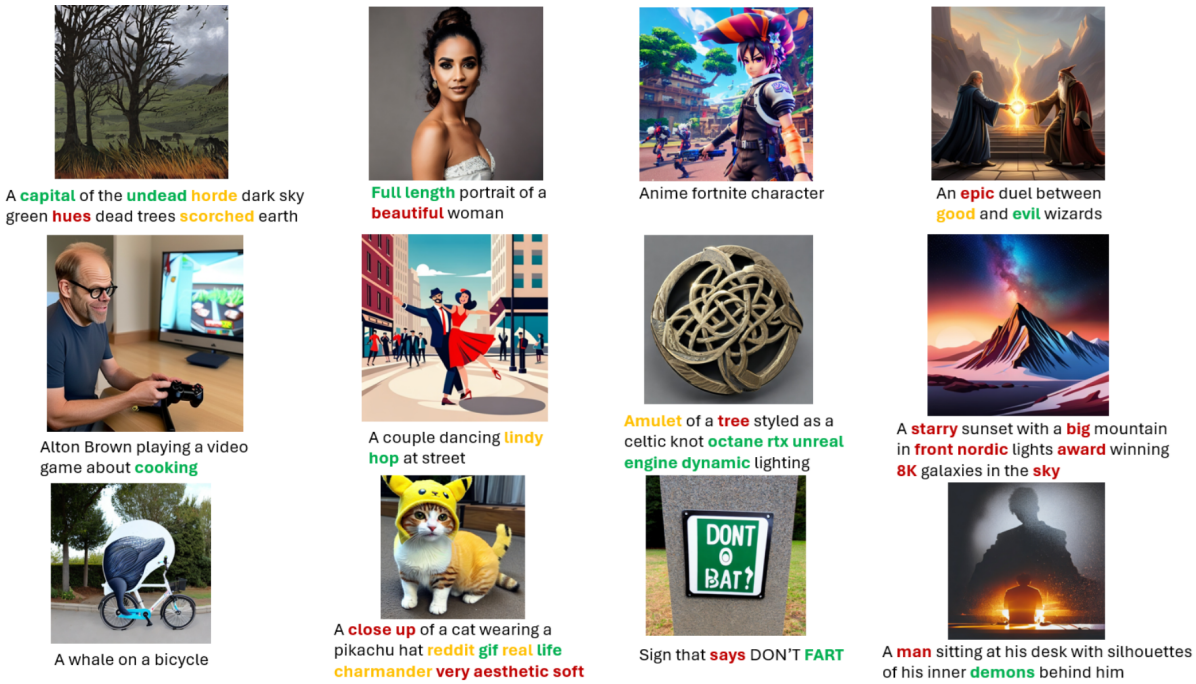

Figure 6: Qualitative results of the calibrated sampling method using the multilingual LAION ViT-H/14 model on the Rich-HF test set. For this results, our method was calibrated to 20% False Discovery Rate using the Rich-HF validation set for calibration.