BEiTScore++: Reference-free Image Captioning Evaluation with BEiT-3 and Conformalized Quantile Regression

Anonymous ACL submission

Abstract

Reliable uncertainty estimation is crucial for deploying evaluation metrics in real-world settings, where both input ambiguity and domain shifts are common. In this work, we propose an uncertainty-aware extension of the BEiTScore image captioning metric, producing calibrated confidence intervals alongside alignment scores. Our approach models predictive distributions over human judgments, enabling both accurate scoring and principled uncertainty quantification. We evaluate our BEiTScore++ metric across multiple benchmarks with human-annotated ratings, demonstrating strong performance in terms of ranking correlation, calibration, and interval sharpness. BEiTScore++ not only produces accurate predictions, but also captures aleatoric uncertainty arising from ambiguous inputs, and epistemic uncertainty due to limited model knowledge, making it a reliable tool for uncertainty-aware evaluation of image-caption alignment.

1 Introduction

Image captioning evaluation models typically produce a single score that captures the semantic alignment between the image content and a textual description, aiming at scores that are highly correlated with human assessments. However, individual point estimates make it challenging to convey how confident the model is in the produced assessments. This limitation becomes particularly relevant in scenarios where multiple captions receive similar scores while differing substantially in reliability. In these cases, an effective evaluation metric should not only predict caption quality, but also provide an informative estimate of uncertainty.

In this work, we introduce BEiTScore++, an uncertainty-aware extension to the state-of-the-art BEiTScore image captioning metric (Gomes et al., 2026), which predicts both a quality score and a calibrated confidence interval. We formulate the evaluation problem as quantile regression and train the model using a centered interval objective that encourages both coverage and sharpness. To ensure coherent intervals, the quantiles are constructed monotonically from a base quantile prediction, avoiding quantile crossing by design. Moreover, to improve coverage reliability, we further apply post-hoc conformal calibration. Figure 1 provides a general illustration of the proposed procedure.

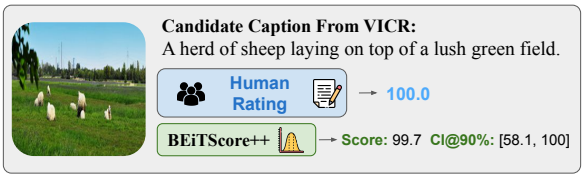

Figure 1: Qualitative example of of BEiTScore++ in an instance from the VICR (Narins et al., 2023) benchmark. The candidate caption is scored as highly relevant by humans, as well as by our BEiTScore++. However, BEiTScore++ also outputs a wide confidence interval, potentially because some of the mentioned elements (sheep) are small and appear slightly distorted, plus the image appears to be stretched.

BEiTScore++ is evaluated on several benchmark datasets with human numeric ratings, allowing us to assess how well the predicted uncertainty correlates with prediction error and human judgments. In particular, we show that correlation with human judgments remains at the level of state-of-the-art evaluation metrics, and that the predicted intervals are well calibrated, sufficiently sharp, and informative for downstream analysis.

Beyond predictive performance, we further investigate whether the estimated uncertainty reflects meaningful sources of variability. We focus on two key factors: input ambiguity and model knowledge gaps. To study input ambiguity, we introduce controlled image degradations, such as Gaussian blur, and observe that uncertainty increases proportionally with the level of visual degradation. To

examine model knowledge gaps, we evaluate the model on out-of-domain artwork datasets and find that uncertainty rises significantly compared to in-domain natural images, reflecting epistemic uncertainty due to domain shift. Overall, BEiTScore++ aims to make image captioning evaluation not only more accurate, but also more trustworthy.[1]

2 Related Work

This section reviews prior work concerning three key areas: image captioning evaluation, uncertainty quantification, and uncertainty calibration.

2.1 Image Captioning Evaluation

CLIPScore (Hessel et al., 2021) was proposed to overcome the limitations of reference-dependent evaluation methods by enabling reference-free scoring through the cosine similarity between CLIP image and text embeddings. However, the CLIP text encoder typically imposes a strict token limit (e.g., 77 tokens), which restricts its use for long-form descriptions. Several extensions and related approaches have been proposed (Sarto et al., 2025; Hu et al., 2023; Narins et al., 2023; Wada et al., 2024; Gomes et al., 2025b), including approaches that adapt CLIP-based methods to longer textual inputs (Chen et al., 2025; Zhang et al., 2024).

Other approaches leverage Large Language Models (LLMs) for caption evaluation (Chan et al., 2023; Lee et al., 2024; Kim et al., 2025). These models benefit from strong zero-shot capabilities, long context windows, and the ability to capture nuanced semantic differences. However, they incur high computational costs and slow inference times, making them impractical for large-scale benchmarking or integration into decoding pipelines.

More recent work revisits learned encoder-based metrics using more expressive cross-encoder architectures. Gomes et al. (2026) proposed BEiTScore, with basis on a cross-encoder initialized from a BEiT-3 visual question answering checkpoint, this way combining strong pretrained representations with computational efficiency. BEiTScore employs cross-attention between the visual and textual modalities, enabling fine-grained sensitivity to visual-linguistic inconsistencies. Its training strategy incorporates a diverse data mixture designed to expose the model to long descriptions and complex visual-linguistic relationships, along with adversarial augmentations that enhance compositional coverage and relational reasoning.

[1]The corresponding source code and the different model checkpoints will be made available upon acceptance.

Empirical results show that BEiTScore maintains robust compositional understanding across both short and long descriptions, while matching or surpassing specialized LLM-based metrics at a fraction of the computational cost, making it a viable and cost-effective alternative to LLM-based caption evaluation approaches.

2.2 Uncertainty Quantification

Applications that rely on machine learning predictions often require an estimate of confidence alongside the prediction itself, in order to mitigate overconfident decisions. The process of estimating this confidence is commonly referred to as uncertainty quantification. From a philosophical perspective, uncertainty can be framed by asking *where does the randomness lie?* This leads to two distinct viewpoints, namely Bayesian approaches which attribute randomness to model parameters, and frequentist approaches which assume randomness originates from the data.

In Bayesian methods, neural network weights are modeled as probability distributions rather than fixed values (Mackay, 1992; Welling and Teh, 2011; Tran et al., 2019). Training focuses on learning these distributions and, during inference, weights are sampled accordingly. While theoretically appealing, these approaches are computationally expensive, particularly for modern models with millions or billions of parameters, leading to the widespread adoption of frequentist approximations.

A common approximation approach corresponds to deep ensembles, where multiple models are trained independently and the empirical variance of their predictions is used as a measure of uncertainty (Garmash and Monz, 2016; Kendall and Gal, 2017; Pearce et al., 2020; Zhan et al., 2023). Another widely used technique is Monte Carlo dropout (Gal and Ghahramani, 2016; Glushkova et al., 2021), which applies dropout at test time to perform multiple stochastic forward passes, estimating uncertainty through variations in the outputs. However, both approaches require multiple forward passes at inference time to construct uncertainty estimates, which can become increasingly slow as the number of samples grows.

Beyond variance-based approaches, quantile regression has emerged as a compelling alternative, particularly in settings where the model can be fully trained and controlled. Quantiles offer a flexible and non-parametric representation of uncertainty, are directly interpretable in the target space,

and enable straightforward construction of prediction intervals. Traditionally, quantile regression is trained using the pinball loss (Koenker, 2006). However, recent work (Chung et al., 2021) shows that the trade-off between sharpness and coverage that is induced by the pinball loss is implicit and often suboptimal, proving that, in highly expressive models, this trade-off tends to favor sharpness at the expense of proper calibration. To address this limitation, Chung et al. (2021) propose alternative loss functions that explicitly balance sharpness and coverage calibration during training.

2.3 Uncertainty Calibration

Uncertainty calibration naturally complements uncertainty quantification, as confidence estimates are only meaningful if they reflect the true frequency with which events occur. In the context of prediction intervals, calibration requires that nominal confidence levels match empirical coverage, for example aiming toward a 90% interval that contains the true value approximately 90% of the time.

For quantile-based methods, a key prerequisite for well-calibrated intervals is to avoid *quantile crossing*, where lower quantiles exceed higher ones and the induced prediction sets become incoherent. Narayan et al. (2021) address this by enforcing monotonicity in the quantile level within deep lattice networks, guaranteeing non-crossing quantiles while retaining model flexibility.

Conformal Risk Control (CRC) provides a general, distribution-free framework for calibrating arbitrary user-defined risk functions. Angelopoulos et al. (2022) extend split conformal prediction by choosing a scalar parameter that controls the expected value of any bounded loss that is monotone in the size of the prediction set. When the loss is chosen as an indicator of miscoverage, CRC recovers classical conformal prediction with finite-sample coverage guarantees, while more general loss choices enable direct control of application-specific risks, such as the false negative rate or task-specific distortion measures.

Recent work has, for instance, leveraged this framework to improve both granularity and calibration of image captioning evaluation metrics. Gomes et al. (2025a) generate distributions over CLIPScore values via simple perturbation strategies (e.g., input masking), and then apply conformal risk control to calibrate these score distributions for word-level error detection and to construct calibrated CLIPScore intervals. This approach yields formal guarantees on user-specified risk bounds, detects foil words in captions, and improves correlation between uncertainty estimates and residuals, thus enhancing CLIPScore's reliability as an image captioning evaluation metric.

3 Properties of Uncertainty Estimates

Uncertainty estimates should satisfy several desirable properties that make them both statistically valid and practically useful. In this work, we focus on four central criteria: calibration, sharpness, monotonicity, and conditional coverage. We also emphasize interpretability and robustness to distribution shift, since uncertainty is only useful if it can be understood and trusted in deployment.

Calibration means that nominal confidence levels match empirical frequencies: for example, a 90% prediction interval should contain the true target about 90% of the time. In turn, sharpness refers to the informativeness of the estimate, typically measured by the width of the prediction interval or the dispersion of the predictive distribution. Following standard principles in probabilistic forecasting, one should first aim to achieve valid coverage and then, among calibrated estimators, maximize sharpness (Gneiting and Raftery, 2007; Chung et al., 2021). In other words, narrow intervals are desirable only when they remain reliable.

For quantile-based uncertainty estimates, the predicted quantiles should be monotone in the quantile level. This prevents quantile crossing, where a lower quantile exceeds a higher one, rendering the resulting interval incoherent. Enforcing non-crossing is important not only for mathematical consistency, but also for producing interpretable confidence intervals that can be reliably used downstream (Narayan et al., 2021).

A calibrated uncertainty estimate may satisfy marginal coverage, meaning that the desired coverage holds on average over the full data distribution. However, marginal validity alone can hide systematic failures on particular regions of the input space or on minority subgroups (Angelopoulos and Bates, 2021; Zerva and Martins, 2024). Conditional coverage is therefore a stronger and more desirable property, requiring reliable coverage within subpopulations or for individual types of inputs. In practice, it is often the difference between an uncertainty estimate that is globally valid and one that is actually dependable for decision-making in heterogeneous settings. However, conditional coverage is something that, in most cases, is very hard if

Algorithm 1 Monotone quantile construction for uncertainty-aware predictions in BEiTScore++

Require: Image I and text description T
$\mathbf{h} \leftarrow \text{BEiT3}(I, T)$
Predict bounded deltas:
  $\boldsymbol{\delta}_{\text{low}} \leftarrow \sigma(f_{\text{low}}(\mathbf{h}))$
  $\delta_{\text{mid}} \leftarrow \sigma(f_{\text{mid}}(\mathbf{h}))$
  $\boldsymbol{\delta}_{\text{high}} \leftarrow \sigma(f_{\text{high}}(\mathbf{h}))$
Initialize lower quantile: $q_0 \leftarrow \delta_{\text{low}}^{(0)}$
for $i = 1 < K_{\text{low}}$ **do**
  $q_i \leftarrow q_{i-1} + \delta_{\text{low}}^{i}(1 - q_{i-1})$
end for
Append median: $q_{K_{\text{low}}} \leftarrow q_{K_{\text{low}}-1} + \delta_{\text{mid}}(1 - q_{K_{\text{low}}-1})$
for $j = 1 < K_{\text{high}} + 1$ **do**
  $q_{K_{\text{low}}+j} \leftarrow q_{K_{\text{low}}+(j-1)} + \delta_{\text{high}}^{(j-1)}(1 - q_{K_{\text{low}}+(j-1)})$
end for
return $\mathbf{y} \leftarrow [q_0, q_1, \ldots, q_{K_{\text{low}}+K_{\text{high}}}]$

not impossible to achieve (Vovk, 2012), and thus a relaxation of this property is often targeted.

Uncertainty estimates should also be easy to interpret by users and practitioners. This is especially important when the output is a confidence interval, quantile set, or calibrated score, since the representation should communicate not only how uncertain the model is but also what the uncertainty means in the scale of the task. Interpretable uncertainty is more likely to support human review, debugging, and downstream risk-aware decisions.

Finally, a useful uncertainty estimate should be sensitive to different types of inputs. In particular, uncertainty should typically increase when the model encounters inputs that are highly ambiguous, or that differ from the training distribution, rather than having outputs that remain overconfident.

4 An Uncertainty Aware BEiTScore

We formulate uncertainty estimation for BEiTScore as a quantile regression problem. The proposed model, which we named BEiTScore++, combines a BEiT-3 backbone with a quantile regression head comprising 11 output nodes, which predict five centered prediction intervals (75%, 80%, 85%, 90%, 95%) together with a central value corresponding to the 50th percentile. These intervals provide a nested family of "typical" uncertainty bands around the median. Jointly learning this grid of quantiles encourages the model to capture a smooth, coherent approximation of the full conditional distribution of scores, which acts as a form of distributional regularization compared to training independent models for individual quantiles (Narayan et al., 2021).

Quantile Regression Loss: Rather than relying on the standard pinball loss, we use the Centered Interval Loss proposed by Chung et al. (2021), which explicitly balances coverage calibration and interval sharpness. We denote a $(1-\alpha)$ confidence interval by $\hat{l} = \mathbb{Q}_{\frac{\alpha}{2}}$ and $\hat{u} = \mathbb{Q}_{1-\frac{\alpha}{2}}$, and define the centered interval loss as:

$$\mathcal{L}_{\text{CI}}^{\alpha} = (\hat{u}_\alpha - \hat{l}_\alpha) + \tfrac{2}{\alpha}(\hat{l}_\alpha - y)\mathbb{I}\{y < \hat{l}_\alpha\} + \tfrac{2}{\alpha}(y - \hat{u}_\alpha)\mathbb{I}\{y > \hat{u}_\alpha\} \quad (1)$$

This loss explicitly penalizes the width of the intervals, thus promoting sharpness. Chung et al. (2021) show that the minimum of its expected value is attained at the true conditional quantiles, thereby also encouraging well-calibrated coverage. The considered design reflects the goal of probabilistic forecasting: intervals should be not only narrow but also correctly calibrated. Accordingly, the lower and upper quantiles are trained jointly to improve interval quality, while the median estimate is supervised with an L_1 loss. The full training objective is thus the sum of the centered interval losses across the five intervals, plus the L_1 loss on the median, defined as follows:

$$\mathcal{L}_{\text{total}} = \sum_{\alpha \in \mathcal{A}} \mathcal{L}_{\text{CI}}^{\alpha} + \left| y - \mathbb{Q}_{0.5} \right|, \quad (2)$$

where $\mathcal{A} = \{0.05, 0.10, 0.15, 0.20, 0.25\}$ corresponds to the $(95\%, 90\%, 85\%, 80\%, 75\%)$ prediction intervals. Figure 2 illustrates the proposed architecture for uncertainty modeling, and the corresponding post-hoc calibration procedure.

Monotone Quantile Construction. The lowest quantile in BEiTScore++, i.e., the 2.5 percentile, is predicted directly by the respective node, and the remaining quantiles are constructed iteratively from the first by predicting positive deltas towards the upper tail. Concretely, the model first estimates the lowest score, and then progressively expands toward higher quantiles using bounded multiplicative offsets. This design enforces quantile monotonicity by construction, thus preventing quantile crossing and ensuring that the predicted intervals remain ordered. In addition, because all quantile outputs are constrained to lie in $[0, 1]$, they remain compatible with the score range of BEiTScore and preserve interpretability. Algorithm 1 summarizes the monotone quantile construction used in BEiTScore++.

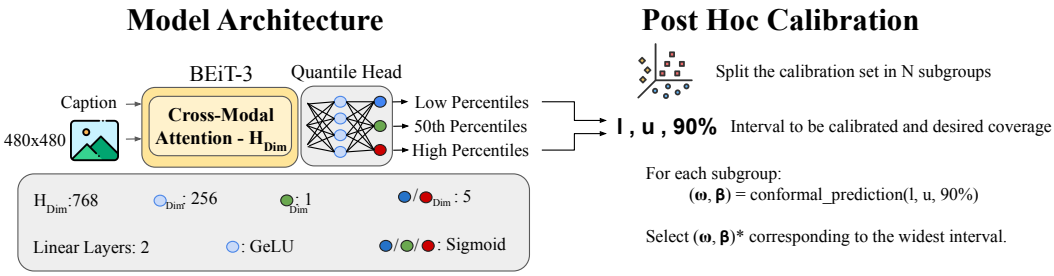

Figure 2: The proposed architecture for uncertainty modeling and the respective post-hoc calibration method.

Quality Score Estimation. The regression head that predicts the median estimate is initialized from the original BEiTScore model, which was trained with a diverse data mixture designed to expose the model to long descriptions and complex visual-linguistic relations (Gomes et al., 2026). We further added the nodes that estimate the lower and upper deltas, considering a bias over the weights that pushes the lower quantiles toward values near 0, and the upper quantiles toward values near 1. This provides an initial state where the model produces wide intervals with high coverage, from which training can progressively refine calibration and sharpness. This initialization follows the principle that probabilistic forecasts should first achieve calibration and then optimize sharpness, as emphasized by Gneiting and Raftery (2007). Empirically, we found that initializing all quantile heads identically (i.e., maximum sharpness), without this bias, led to a suboptimal local minimum, producing intervals that were very sharp but poorly calibrated, with weaker correlation between prediction error and estimated uncertainty.

Conformal Calibration. We apply a post-hoc calibration step using conformal risk control. We define the calibrated prediction interval as follows:

$$\hat{l} = \max\big(0, (l - m) \cdot \omega + m - \beta\big), \quad (3)$$

$$\hat{u} = \max\big(0, (u - m) \cdot \omega + m + \beta\big), \quad (4)$$

where l, m, and u denote the lower, median, and upper quantile predictions that are being targeted for conformalization, respectively, and where ω and β are calibration parameters optimized within the conformal framework to match a desired coverage level. Our procedure follows a general conformal risk control formulation (Angelopoulos et al., 2022; Gomes et al., 2025a), where the calibration objective is to satisfy a user-specified risk constraint, in our case corresponding to miscoverage.

Approximating Conditional Coverage: As noted by Angelopoulos and Bates (2021), conformal calibration on a held-out calibration set provides formal guarantees only for marginal coverage. While this is valuable, it may hide poor behavior on specific subsets of the data. To obtain a more robust calibration that better approximates conditional coverage, we group training and calibration examples according to image similarity. Specifically, we extracted image embeddings and apply k-means clustering to partition the data into homogeneous subgroups. The number of clusters (i.e., $k = 63$ in our tests) was selected using a combination of the elbow method and the silhouette score. We then partitioned the training and calibration set according to these clusters.

With the training dataset partitioned into subgroups, we organize mini-batches so that, at each iteration, all samples in a batch are drawn from a single subgroup selected at random. This design is inspired by the group-batching strategy used to induce adversarial group calibration in Chung et al. (2021), where datapoints are grouped based on input features and gradient updates are computed on group-specific batches to improve adversarial group calibration. In our setting, we additionally use the partitioned calibration set to perform conformal calibration independently within each subgroup, obtaining N candidate pairs of calibration parameters (ω, β). We then select the most conservative pair (i.e., the one producing the widest calibrated interval) so that the resulting calibration is valid across all clusters. This strategy is intended to improve robustness across subpopulations by prioritizing coverage preservation, even at the cost of slightly wider intervals.

5 Experimental Evaluation

This section starts by presenting the evaluation methodology, describing implementation details, datasets, and evaluation metrics. It then describes the experimental results, starting with tests that assessed the correlation to human ratings, and then analyzing the uncertainty estimates.

5.1 Implementation Details

We implement BEiTScore++ using a simple regression head on top of BEiT-3 encoders of size base and large, composed of two hidden layers with 256 units each and GeLU activation functions. The output nodes, together with the encoder weights, are initialized from the original BEiTScore model. The output layer consists of 11 nodes with sigmoid activations. The first five nodes correspond to lower percentiles and are initialized with an additional bias of -8, while the last five nodes correspond to higher percentiles and are initialized with an additional bias of 8, as described in Section 4. The model is trained with a batch size of 16 using early stopping with a patience of 15 epochs. All experiments were conducted on a single NVIDIA H200 GPU, although comparable results can be obtained on lesser-memory GPUs by reducing the batch size and using gradient accumulation accordingly.

5.2 Datasets and Evaluation Metrics

Model training uses a mixture of datasets with human-annotated image-caption alignment scores, including VICR (Narins et al., 2023), Polaris (Wada et al., 2024), Nebula (Matsuda et al., 2024), and EvalMuse-40k (Han et al., 2024). These datasets provide continuous human judgments which we normalize to the range [0,1].

To evaluate predictive performance and the quality of uncertainty estimates, we conduct experiments on several benchmark datasets with human ratings. These include VICR (Narins et al., 2023), Polos (Wada et al., 2024), Nebula (Matsuda et al., 2024), Flickr8K-Expert (Hodosh et al., 2013), Flickr8K-CF (Hodosh et al., 2013), and Composite (Aditya et al., 2015). In addition, for the analysis of different sources of uncertainty, we consider additional datasets spanning in-domain and out-of-domain distributions. In-domain datasets include IIW (Garg et al., 2024), DOCCI (Onoe et al., 2024), Localized Narratives (Pont-Tuset et al., 2020), and MSCOCO (Lin et al., 2014). In turn, out-of-domain datasets focus on artwork images, namely with Docent (Ananthram et al., 2025) and ArtCaps (Lu et al., 2022). Appendix B provides a detailed characterization of all the datasets that were used.

We assess model performance using four complementary metrics: coverage, Kendall's τ_c, sharpness, and UPS. Coverage evaluates whether predicted intervals achieve the desired confidence level. Kendall's τ_c measures ordinal agreement between median predictions and human ratings, being widely used in prior work focused on image captioning evaluation metrics. Sharpness quantifies the tightness of the predicted intervals, with narrower intervals indicating more informative uncertainty estimates. Finally, the Uncertainty Pearson Score (UPS) measures the correlation between predicted uncertainty and prediction error, capturing whether higher uncertainty aligns with larger deviations from human judgments.

5.3 Assessments Using Human Ratings

Our main goal is to assess whether the estimated uncertainty, measured by the width of the predicted confidence intervals, correlates well with the model's prediction error, defined as the absolute difference between the human rating and the predicted 50th percentile. In parallel, we measure Kendall's τ_c between the resulting median scores and the human ratings. Ideally, we want BEiTScore++ to match the state-of-the-art correlation with human judgment results of the original BEiTScore metric. We also analyze sharpness, to determine whether improved uncertainty estimation comes at the cost of wider intervals. Together, these metrics allow us to evaluate whether the model achieves a useful trade-off between calibration, sharpness, correlation with human judgments, and the extent to which higher uncertainty is associated with larger deviations from human ratings.

Table 1 presents a quantitative comparison across multiple datasets and model configurations, including a MC dropout baseline over the original BEiTScore, the proposed uncertainty-aware BEiTScore++ variants, and a set of ablations that justify our design choices. Specifically, the table includes: (i) MC dropout as a uncertainty baseline, evaluated with the BEiTScore backbone in both base and large variants; (ii) a variant where the quantile regression head predicts quantiles without the proposed chained mechanism to avoid quantile crossing, trained with the standard pinball loss; and (iii) a variant that does not use our subgroup-based batching strategy for adversarial group calibration, nor the subgroup-wise calibration procedure for relaxed conditional coverage calibration.

We observe consistent improvements when replacing the standard loss and the regression head with our centered-interval loss paired with chained quantile heads, particularly in UPS and sharpness across all benchmarks. After post-hoc calibration, coverage also improves across datasets. Furthermore, comparing models trained with and with-

Prior to Calibration				VICR				Polaris				Nebula				Flickr8K-Expert				Composite			
Model	Head	Loss	Clust.	τ	UPS	Sharp	Cov	τ	UPS	Sharp	Cov	τ	UPS	Sharp	Cov	τ	UPS	Sharp	Cov	τ	UPS	Sharp	Cov
MC Dropout-B	-	-	No	78.2	40.2	0.23	**68.3**	61.5	57.9	0.15	62.4	55.3	39.1	0.25	**65.8**	59.4	38.5	0.17	**64.6**	58.9	15.3	0.3	46.3
MC Dropout-L	-	-	No	**78.8**	42.1	**0.18**	63.2	63.2	60.8	**0.11**	**68.4**	56.7	42.9	**0.21**	61.0	60.2	48.0	**0.12**	58.8	**59.5**	13.7	**0.24**	42.6
BEiTS++-B	Unc.	PinB.	No	78.5	40.4	0.30	63.2	61.9	63.5	0.16	34.0	56.4	47.8	0.31	50.3	60.2	50.9	0.25	44.9	57.6	28.6	0.40	43.8
BEiTS++-B	Ch.	CI	No	77.5	71.2	0.32	65.1	60.9	88.4	0.19	32.4	54.7	63.1	0.35	62.0	58.7	73.6	0.24	36.1	53.9	35.1	0.45	61.9
BEiTS++-B	Ch.	CI	Yes	77.6	69.6	0.33	65.4	61.8	89.4	0.20	33.6	55.6	64.6	0.36	63.7	59.7	73.2	0.24	36.2	55.3	34.2	0.46	**63.5**
BEiTS++-L	Ch.	CI	Yes	78.2	**72.5**	0.30	62.7	**63.4**	**91.0**	0.19	32.9	**57.6**	**65.4**	0.34	62.0	**61.2**	**77.6**	0.22	33.3	57.6	**35.4**	0.41	60.7
After Calibration																							
MC Dropout-B	-	-	No	78.2	43.5	0.34	85.5	61.5	60.6	0.22	97.3	55.3	42.1	0.38	84.7	59.4	45.8	0.23	76.8	58.9	20.0	0.45	68.7
MC Dropout-L	-	-	No	**78.8**	45.6	**0.29**	83.1	63.2	60.2	**0.17**	97.0	56.7	43.4	0.34	85.2	60.2	50.4	**0.19**	73.4	**59.5**	21.5	**0.39**	65.0
BEiTS++-B	Unc.	PinB.	No	78.5	40.5	0.32	82.7	61.9	63.6	**0.17**	95.5	56.4	48.0	**0.33**	76.0	60.2	51.1	0.27	76.6	57.6	28.5	0.42	50.3
BEiTS++-B	Ch.	CI	No	77.5	71.2	0.33	89.0	60.9	88.4	0.20	95.5	54.7	63.1	0.36	88.2	58.7	73.6	0.25	78.5	53.9	35.1	0.46	79.4
BEiTS++-B	Ch.	CI	Yes	77.6	70.3	0.36	**92.6**	61.8	89.6	0.22	**98.4**	55.6	64.7	0.40	**92.9**	59.7	73.7	0.27	**82.9**	55.3	34.7	0.51	**84.1**
BEiTS++-L	Ch.	CI	Yes	78.2	**72.7**	0.32	88.9	**63.4**	**91.0**	0.20	97.3	**57.6**	**65.5**	0.36	90.6	**61.2**	**77.6**	0.24	78.4	57.6	**35.6**	0.44	80.5

Table 1: Quantitative results across multiple datasets and model configurations. **Bold** indicates the best value and underline indicates the second-best value for each metric within each dataset. For Sharpness, lower values are better.

out clustering and applying subgroup-wise calibration shows incremental gains in the target metrics before calibration and substantial gains in coverage afterward, especially on the out-of-distribution benchmarks (i.e., the Flickr8K-Expert and Composite datasets). Compared with Monte Carlo dropout, our method achieves better uncertainty quality while remaining considerably more efficient at inference, since it does not rely on repeated forward passes (i.e., 25 passes in the case of our tests). Importantly, the median-score correlation of BEiTScore++ matches that of the original BEiTScore across the evaluated benchmarks, indicating that introducing uncertainty estimation does not degrade the underlying scoring capability. In other words, BEiTScore++ preserves the state-of-the-art human-correlation performance of BEiTScore, while providing calibrated and more informative uncertainty estimates.

In complement to Table 1, Appendix A presents a detailed comparison of human correlation between BEiTScore++ and several alternative methods in the literature, confirming that BEiTScore++ preserves state-of-the-art performance while providing calibrated uncertainty estimates.

5.4 Different Sources of Uncertainty

This section investigates whether the uncertainty estimates produced by BEiTScore++ meaningfully reflect different underlying sources of uncertainty. We focus on two complementary factors: input ambiguity, arising from degradation or corruption of the input images, and model knowledge gaps, stemming from exposure to ambiguous and out-of-domain data. Through controlled experiments, we assess whether the uncertainty estimator responds in a predictable and interpretable manner. Results are reported for the large variant, although the base-sized model yields similar conclusions, which are provided in Appendix D.

5.4.1 Input Ambiguity

To examine whether BEiTScore++ responds appropriately to input ambiguity, we designed an experiment in which images from the Polaris test split were subjected to varying levels of Gaussian blur. The underlying hypothesis is that increasing blur degrades the visual information available to the model, and consequently model uncertainty should increase proportionally to the level of blur applied.

A key preprocessing step was taken to avoid confounding effects arising from the score distribution. The original Polaris test split is heavily skewed towards low scores, with a large proportion of instances receiving a human rating of zero. For already poor image-caption pairs, the addition of blur could plausibly reinforce the model's confidence in assigning low scores, rather than inducing uncertainty through wider intervals. This behavior would obscure the relationship we aim to measure. To mitigate this, we downsampled instances with a human score of zero, and built a subset of 3624 instances with a more uniform distribution across the full spectrum of scores, aiming for the effect of blur on model uncertainty to be observed clearly across instances of varying quality.

With the balanced subset, we applied progressively stronger levels of blur to each image and measured the corresponding model uncertainty at each level. Figure 3 shows BEiTScore++ interval width distributions under different blur perturbation levels. The results show that our method behaves as expected. The width of the confidence intervals increases with blur intensity set at 10, and this result is statistically significant. An additional interesting behavior emerges at blur level 20, where a reduction in interval width is observed. We attribute this to the images becoming so severely distorted that the model grows confident that the image and

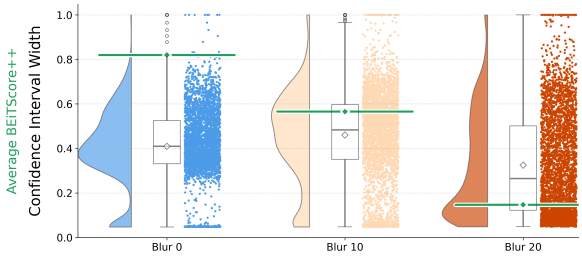

Figure 3: Distribution of 90% confidence interval widths generated by the large variant of BEiTScore++, as Gaussian blur increases on images from the Polaris test split. The horizontal lines indicates the average median score produced by BEiTScore++ at each blur level.

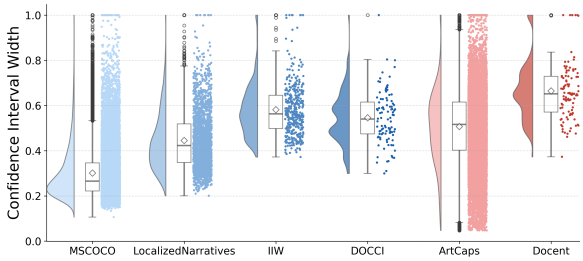

Figure 4: Distribution of 90% confidence interval widths produced by the large variant of the BEiTScore++ metric across in-domain images and out-of-domain artwork datasets, and also across short- and long-caption settings. The color gradient indicates the caption length, with pastel tones for short captions, more saturated tones for detailed descriptions.

caption are misaligned, effectively starting to collapse uncertainty in favor of a low-score prediction. In addition to interval width distribution, Figure 3 also presents the average BEiTScore++ ratings for each blur level, confirming that the model produces lower scores for increased input degradation.

We additionally evaluated model behavior under a broader set of image distortions, including salt-and-pepper noise, JPEG compression artifacts, random patch occlusions, elastic warping (via spatial smoothing and edge curling), and contrast adjustments (see Appendix C for full details). Across most distortion types, we observed a consistent pattern analogous to Gaussian blur. As the severity of the corruption increased, the model's predictive uncertainty grew proportionally, until we reach a very high level of distortion and the model starts to confidently reject the image-caption alignment. This further supports the hypothesis that the estimator is sensitive to input ambiguity, and that the model's confidence is properly reported.

5.4.2 Model Knowledge Gaps

To study uncertainty arising from model knowledge gaps, we analyze the model's confidence on in-domain datasets with natural images, including long-caption datasets IIW (Garg et al., 2024), DOCCI (Onoe et al., 2024), and Localized Narratives (Pont-Tuset et al., 2020), as well as the short-caption dataset MSCOCO (Lin et al., 2014), and compare them against artwork datasets Docent (Ananthram et al., 2025) and ArtCaps (Lu et al., 2022). These artwork datasets are not only out-of-domain but also more ambiguous: their images are paintings rather than realistic photographs, with stylized shapes, smoother edges, and greater abstraction, while the captions, especially the detailed descriptions in Docent, often involve a higher degree of subjective interpretation.

Since BEiTScore++ supports both detailed and shorter captions, we divide the evaluation into two groups: long captions, we compare IIW, DOCCI, and Localized Narratives against Docent; short captions, we compare ArtCaps against MSCOCO.

Figure 4 indicates that in both the short- and long-caption settings, the model exhibits higher uncertainty when evaluating artwork datasets. This behavior reflects the fact that artwork imagery can be more challenging to interpret, and also an increased epistemic uncertainty due to a model knowledge gap, as the model was not fine-tuned on artistic imagery. Model uncertainty also increases with the average size of the captions.

6 Conclusions

We introduced an uncertainty-aware extension of BEiTScore that predicts caption quality and generates calibrated confidence intervals. By combining quantile regression with a monotone construction of quantiles, together with a centered interval loss, BEiTScore++ achieves a balance between calibration and sharpness. Post-hoc conformal calibration further improves coverage, particularly when extended with subgroup-based calibration to better approximate conditional coverage.

Empirical results demonstrate that our method maintains strong agreement with human judgments, while producing uncertainty estimates that are well-calibrated. Uncertainty correlates with prediction error and adapts to input ambiguity, indicating that the model captures meaningful signals about its own reliability. These properties are essential for downstream applications, where evaluation metrics inform benchmarking or model selection.

Limitations and Ethical Considerations

While BEiTScore++ improves uncertainty estimation for image-caption evaluation, providing calibrated and sharp confidence intervals, several important limitations remain.

First, our experiments are restricted to English captions, so the behavior of the method in multilingual settings is still unknown. Extending BEiTScore++ to other languages is an important direction for future work, especially given recent evidence that CLIP-based metrics can be adapted for multilingual caption evaluation (Gomes et al., 2025b). It will be important to study whether the proposed uncertainty estimates remain reliable when captions contain translation artifacts, language-specific constructions, or morphological complexity that differs from English.

Second, although our method improves evaluation interpretability, it does not remove the need for human evaluation. Automated metrics can support and complement human judgment, but they should not be treated as a substitute, particularly in applications where erroneous assessments may have downstream consequences. More generally, uncertainty intervals may help users identify low-confidence cases, but they do not guarantee fairness or correctness in all settings.

Third, the proposed framework is evaluated on benchmark datasets that may not fully capture the diversity of real-world image-caption pairs. As with many evaluation methods, its performance may be affected by dataset bias, domain shift, or limited coverage of rare concepts and styles. Future work should investigate how BEiTScore++ behaves under broader distributional shifts, and whether additional calibration strategies are needed to maintain robustness in more challenging settings.

Finally, while BEiTScore++ is designed to be efficient at inference time, the broader research area still depends on substantial computational resources for training and evaluation. In that sense, developing uncertainty-aware metrics that remain accurate, efficient, and broadly applicable across languages and domains remains as a particularly important research direction.

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*.

Amith Ananthram, Elias Stengel-Eskin, Lorena A Bradford, Julia Demarest, Adam Purvis, Keith Krut, Robert Stein, Rina Elster Pantalony, Mohit Bansal, and Kathleen McKeown. 2025. PoSh: Using scene graphs to guide LLMs-as-a-judge for detailed image descriptions. *arXiv preprint arXiv:2510.19060*.

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*.

David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. 2023. CLAIR: Evaluating Image Captions with Large Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Xiaofu Chen, Israfel Salazar, and Yova Kementchedjhieva. 2025. SPECS: Specificity-Enhanced CLIPScore for Long Image Caption Evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. 2021. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Proceedings of the Annual Meeting on Neural Information Processing Systems*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*.

Roopal Garg, Andrea Burns, Burcu Karagol-Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. 2024. ImageInWords: Unlocking hyper-detailed image descriptions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of the International Conference on Computational Linguistics*.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Conference on Empirical Methods in Natural Language Processing*.

Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*.

Goncalo Emanuel Cavaco Gomes, Bruno Martins, and Chrysoula Zerva. 2025a. A conformal risk control framework for granular word assessment and uncertainty calibration of CLIPScore quality estimates. In *Findings of the Association for Computational Linguistics*.

Goncalo Emanuel Cavaco Gomes, Bruno Martins, and Chrysoula Zerva. 2026. BEiTScore: Reference-free image captioning evaluation with an efficient cross-encoder model. *arXiv preprint arXiv:2605.21728*.

Goncalo Emanuel Cavaco Gomes, Chrysoula Zerva, and Bruno Martins. 2025b. Evaluation of multilingual image captioning: How far can we get with CLIP models? In *Findings of the Association for Computational Linguistics*.

Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and 1 others. 2024. Evalmuse-40k: A Reliable and Fine-Grained Benchmark with Comprehensive Human Annotations for Text-to-Image Generation Model Evaluation. *arXiv preprint arXiv:2412.18150*.

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*.

Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Proceedings of the Annual Meeting on Neural Information Processing Systems*.

Hyunjong Kim, Sangyeop Kim, Jongheon Jeong, Yeongjae Cho, and Sungzoon Cho. 2025. EXPERT: An Explainable Image Captioning Evaluation Metric with Structured Explanations. In *Findings of the Annual Meeting of the Association for Computational Linguistics*.

Roger Koenker. 2006. Quantile regresssion. *Encyclopedia of Environmetrics*.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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 *Proceedings of the European Conference on Computer Vision*.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating Text-to-Visual Generation with Image-to-Text Generation. In *Proceedings of the European Conference on Computer Vision*. Springer.

Yue Lu, Chao Guo, Xingyuan Dai, and Fei-Yue Wang. 2022. ArtCap: A dataset for image captioning of fine art paintings. *IEEE Transactions on Computational Social Systems*.

David John Cameron Mackay. 1992. *Bayesian Methods for Adaptive Models*. California Institute of Technology.

Kazuki Matsuda, Yuiga Wada, and Komei Sugiura. 2024. DENEB: A Hallucination-Robust Automatic Evaluation Metric for Image Captioning. In *Proceedings of the Asian Conference on Computer Vision*.

Taman Narayan, Serena Wang, Kevin Canini, and Maya Gupta. 2021. Regularization strategies for quantile regression. *arXiv preprint arXiv:2102.05135*.

Lothar D Narins, Andrew Scott, Aakash Gautam, Anagha Kulkarni, Mar Castanon, Benjamin Kao, Shasta Ihorn, Yue-Ting Siu, James M Mason, Alexander Blum, and 1 others. 2023. Validated Image Caption Rating Dataset. *Proceedings of the Annual Meeting on in Neural Information Processing Systems*.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, and 1 others. 2024. DOCCI: Descriptions of connected and contrasting images. In *Proceedings of the European Conference on Computer Vision*.

Tim Pearce, Felix Leibfried, and Alexandra Brintrup. 2020. Uncertainty in neural networks: Approximately Bayesian ensembling. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting Vision and Language with Localized Narratives. In *Proceedings of the European Conference on Computer Vision*.

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*.

Sara Sarto, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training. *International Journal of Computer Vision*.

Dustin Tran, Mike Dusenberry, Mark Van Der Wilk, and Danijar Hafner. 2019. Bayesian layers: A module for neural network uncertainty. *Proceedings of the Annual Meeting on Neural Information Processing Systems*.

Vladimir Vovk. 2012. Conditional validity of inductive conformal predictors. In *In Proceedings of the Asian Conference on Machine Learning*.

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*.

Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*.

Chrysoula Zerva and André F. T. Martins. 2024. Conformalizing machine translation evaluation. *Transactions of the Association for Computational Linguistics*.

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*.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-CLIP: Unlocking the Long-Text Capability of CLIP. In *Proceedings of the European Conference on Computer Vision*.

A Additional Results on Correlation with Human Judgments

In complement to the results presented in Table 1, this appendix provides a more detailed evaluation of the correlation between the median predicted scores and human judgments, across a wide range of benchmarks. We compare the proposed uncertainty-aware BEiTScore++ metric against several representative reference-free metrics from the literature, including both encoder-based and LLM-based approaches. This analysis is intended to verify that introducing uncertainty estimation does not compromise the model's ability to produce accurate quality assessments.

Table 2 reports Kendall's τ_c on VICR, Polaris, Nebula, EX-8K, and COM. Overall, the results confirm that the uncertainty-aware BEiTScore retains state-of-the-art performance in terms of agreement with human judgments. In particular, the large version of the proposed model achieves the best or second-best performance on most benchmarks, outperforming many existing alternatives while providing the additional information about the uncertainty of the model in its predictions.

B Datasets

The following datasets were used in the training, calibration, and evaluation of BEiTScore++, specifically in experiments that focused on assessing the Uncertainty Pearson Score (UPS) and the correlation with human judgments. We separate the datasets into those used for training and calibration, and those used only for evaluation.

Datasets used for training and calibration.

- **VICR** (Narins et al., 2023): The Validated Image Caption Rating (VICR) dataset features human ratings collected through a gamified approach for image-caption pairs involving a total of 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. The dataset is organized into 10,175 training pairs, 2,310 validation pairs, and 3,161 test pairs, with an average of 4.4 annotations per instance. In our experiments, we used the test split of the VICR dataset, which includes 3,161 unique image-caption pairs. For calibration, we used the validation split, which comprises 2,310 unique image-caption pairs.
- **Polaris** (Wada et al., 2024): The Polaris dataset comprises human judgments on image-caption pairs involving 13,691 distinct images, collected from evaluators. The dataset is organized into 26,340 training pairs, 8,738 validation pairs, and 8,726 test pairs, with an average of 3.0 annotations per instance. It surpasses existing datasets in scale and diversity. 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

Model	Size (M)	GFLOPs	VICRτ_c	Polarisτ_c	Nebulaτ_c	EX-8Kτ_c	COMτ_c
LLM-based Methods							
FLEUR (Lee et al., 2024)	13350.84	9317.47	74.7	58.4	52.7	53.0	<u>60.8</u>
EXPERT (Kim et al., 2025)	13350.84	9317.47	77.4	60.9	54.9	56.7	**61.8**
VQAScore (CLIP-FlanT5-XXL) (Lin et al., 2024)	11459.59	3742.10	69.8	51.6	49.4	41.0	53.4
Encoder-based Methods							
CLIP-S/ViT-B (Hessel et al., 2021)	151.28	7.28	70.8	52.9	47.8	52.3	51.8
CLIP-S/ViT-L (Hessel et al., 2021)	427.62	84.4	72.6	53.2	48.8	54.7	55.4
LongCLIP-S/ViT-B (Zhang et al., 2024)	149.84	26.24	70.2	54.8	48.9	53.0	51.3
LongCLIP-S/ViT-L (Zhang et al., 2024)	427.75	98.90	69.0	53.9	48.6	53.9	50.5
PAC-S/ViT-B (Sarto et al., 2023)	151.28	4.88	71.4	52.3	47.2	54.3	53.5
PAC-S/ViT-L (Sarto et al., 2023)	427.62	56.26	73.2	53.1	47.9	55.7	55.4
PAC-S++/ViT-B (Sarto et al., 2025)	155.42	4.88	72.7	52.4	46.9	54.5	55.7
PAC-S++/ViT-L (Sarto et al., 2025)	430.58	56.26	75.2	53.6	48.9	57.4	59.4
SPECS (Chen et al., 2025)	151.49	9.20	33.4	39.7	34.4	30.4	24.1
BEiTScore-B/32	222.28	83.85	<u>78.2</u>	61.5	55.3	59.4	58.9
BEiTScore-L/14	674.69	296.73	**78.8**	<u>63.2</u>	<u>56.7</u>	<u>60.2</u>	59.5
BEiTScore++-B/32	222.28	83.85	77.6	61.8	55.6	59.7	55.3
BEiTScore++-L/14	674.69	296.73	<u>78.2</u>	**63.4**	**57.6**	**61.2**	57.6

Table 2: Comparison against state-of-the-art reference-free metrics across multiple human correlation benchmarks. Methods are categorized as encoder-based or LLM-based. Performance is measured using Kendall's τ_c correlation coefficient with human judgments on all benchmarks. The highest score is highlighted using **bold**, and the second best is highlighted using an <u>underline</u>.

calibration, we used the validation split, which comprises 8,738 unique image-caption pairs.

- **Nebula** (Matsuda et al., 2024): The Nebula dataset comprises 32,978 human annotations on image-caption pairs, spanning 32,978 unique images. It is organized into 26,382 training pairs, 3,298 validation pairs, and 3,298 test pairs. The dataset was introduced for training hallucination-robust image captioning evaluation metrics and was constructed by extending the Polaris dataset to increase visual diversity. In our experiments, we used the test split of Nebula, which contains 3,298 image-caption pairs. For calibration, we used the validation split, which also comprises 3,298 image-caption pairs.
- **EvalMuse-40K** (Han et al., 2024): The EvalMuse-40K dataset comprises 32,717 human annotations on image-caption pairs, spanning 32,717 unique images. It is organized into 29,445 training pairs and 3,272 validation pairs. EvalMuse-40K was introduced as a fine-grained benchmark for text-to-image generation evaluation, with comprehensive human annotations designed to better assess image-text alignment. Our experiments used the training pairs for model training, and the validation pairs for calibration.

Datasets used only for evaluation.

- **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 (2,007 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.

Table 3 summarizes the key statistics of the datasets used for training, calibration, and evaluation.

The following datasets were used in our experiments to evaluate the proposed uncertainty estimator under different conditions, including controlled input ambiguity and model knowledge gaps. Specifically, we rely on them to assess calibration, sharpness, and correlation with human judgments, as well as to study how uncertainty changes under out-of-domain settings and varying caption lengths. IIW, DOCCI, Localized Narratives, and Docent are all datasets with long, detailed captions, making them particularly suitable for analyzing the behavior of the model in the long-caption setting.

IIW (Garg et al., 2024): ImageInWords (IIW) is a human-in-the-loop dataset designed to support

Metric	VICR	Polaris	Nebula	Flickr8K-Expert	Composite	EvalMuse-40K
Training	10175	26340	26382	-	-	29445
Validation	2310	8738	3298	-	-	3272
Testing	3161	8726	3298	5664	13146	-
Unique Images	9990	13691	32978	1000	3995	32717
Caption Len. $\mu \pm \sigma$ (Q1/Q2/Q3)	10.9 ± 3.0 (9.0/10.0/12.0)	9.3 ± 5.5 (8.0/10.0/11.0)	8.8 ± 2.5 (8.0/9.0/10.0)	11.9 ± 4.2 (9.0/11.0/14.0)	13.1 ± 9.3 (9.0/11.0/14.0)	11.0 ± 7.0 (6.0/9.0/14.0)
Annot./Inst. $\mu \pm \sigma$ (Q1/Q2/Q3)	4.4 ± 0.8 (4.0/4.0/5.0)	3.0 ± 2.8 (1.0/1.0/6.0)	1.0 ± 0.0 (1.0/1.0/1.0)	3.0 ± 0.0 (3.0/3.0/3.0)	1.0 ± 0.0 (1.0/1.0/1.0)	1.0 ± 0.0 (1.0/1.0/1.0)
Human Score $\mu \pm \sigma$ (Q1/Q2/Q3)	2.67 ± 1.41 (1.20/2.40/4.19)	0.25 ± 0.34 (0.00/0.00/0.62)	0.52 ± 0.36 (0.00/0.60/0.78)	1.58 ± 0.72 (1.00/1.33/2.00)	3.00 ± 1.45 (1.50/3.00/4.50)	0.46 ± 0.25 (0.25/0.42/0.67)
Rating Scale	[1.0, 5.0]	[0.0, 1.0]	[0.0, 1.0]	[1.0, 4.0]	[1.0, 5.0]	[0.0, 1.0]
Score Distribution						

Table 3: Statistical characterization of the datasets. For each metric (i.e., Caption Length, Annotations per Instance, and Human Score), we report the mean μ and standard deviation σ, along with quartiles (Q1: 25th percentile, Q2: median, Q3: 75th percentile). Score distributions are represented using sparkline histograms computed over the average human scores using the observed annotation scale inferred directly from each dataset.

rich, detailed image descriptions. It was created to capture hyper-detailed visual semantics and to facilitate the evaluation of models that generate long-form captions with fine-grained content. In our experiments, we use IIW as an in-domain dataset with realistic natural imagery and detailed textual descriptions, making it suitable for studying model behavior under standard captioning conditions. We evaluate results on the complete set of 400 test instances provided by the dataset.

DOCCI (Onoe et al., 2024): Descriptions of Connected and Contrasting Images (DOCCI) is a dataset of long, human-authored descriptions for approximately 15k images. The annotations are designed to be comprehensive and compositional, with an emphasis on challenges such as spatial relations, counting, text rendering, and world knowledge. We use DOCCI as an in-domain benchmark for long-caption evaluation, since it provides natural images paired with detailed descriptions that stress the model's ability to assess rich semantic alignment. In our experiments, we use the 100 test instances available in the dataset.

Localized Narratives (Pont-Tuset et al., 2020): Localized Narratives is a large-scale multimodal dataset in which annotators describe images while simultaneously hovering over the regions they are referring to, producing word-level grounding information. The dataset covers hundreds of thousands of images from sources such as COCO, Flickr30k, ADE20K, and Open Images, and is widely used for dense vision-language understanding. In our experiments, we use it as an in-domain dataset because it contains realistic scenes and natural language descriptions that are well aligned with standard image captioning evaluation settings. Since the test split contains more than 100k instances, we randomly sample 3k examples for our analysis.

MSCOCO (Lin et al., 2014): The Microsoft COCO Captions dataset is one of the most widely used benchmarks for image captioning, containing multiple human-generated captions for each image. It provides natural photographs with diverse everyday scenes and serves as a strong in-domain reference for both caption generation and evaluation. We use MSCOCO as an in-domain dataset in our experiments, including a random subset matched in size to ArtCaps for the short-caption setting. Specifically, we randomly sample 18k MSCOCO instances to compare directly against ArtCaps.

Docent (Ananthram et al., 2025): Docent is an artwork-oriented dataset used to evaluate captioning models under domain shift. Unlike natural-image benchmarks, it focuses on artistic imagery, where semantic interpretation can be more abstract and stylistically driven. We use Docent as an out-of-domain dataset to test whether the proposed uncertainty estimator reflects model knowledge gaps when processing artwork rather than photographic content. The dataset contains 100 distinct image-caption pairs, and we use the provided reference captions for our analysis.

ArtCaps (Lu et al., 2022): ArtCaps is a fine-art painting captioning dataset containing paintings paired with multiple human-written descriptions. It was constructed specifically for image captioning in the artistic domain, where visual content is often complex and culturally contextual. We use ArtCaps as an out-of-domain benchmark in our experiments, since it provides a clear contrast to natural-image datasets and is well suited for probing epistemic uncertainty under domain shift. The dataset contains approximately 18k instances, which we compare against a randomly sampled subset of 18k MSCOCO instances in the short-caption setting.

C Input Ambiguity Beyond Blur

While Gaussian blur provides a controlled way to degrade visual information, it represents only one form of input ambiguity. In practice, images may be affected by a wide range of distortions that alter different aspects of the visual signal, such as noise, compression artifacts, structural occlusions, or geometric deformations. To assess whether our conclusions for BEiTScore++ generalize beyond blur, we extend our analysis to additional perturbations, including salt-and-pepper noise, JPEG compression, random patch cutout, contrast adjustments, and elastic warping. Figure 6 presents a grid of examples illustrating these different distortions.

We apply each distortion to the images and measure the corresponding uncertainty estimates. The resulting distributions of confidence interval widths are presented in Figure 5. For most distortion types, we observe that introducing degradation leads to increased uncertainty compared to the no-distortion setting. This trend is consistent with the results obtained using Gaussian blur and further supports the hypothesis that the proposed estimator is sensitive to input ambiguity.

An exception is observed for contrast adjustments, where the interval widths remain unchanged. We find similar behavior when applying other color-based transformations, suggesting that the model is relatively robust to photometric changes that preserve the underlying geometries and shapes of content of the image. In contrast, distortions that disrupt structural or spatial information (e.g., noise, occlusions, and warping) lead to a clear increase in the model uncertainty.

D Analysis of Different Sources of Uncertainty Beyond BEITScore-L

In this appendix, we report the results obtained with the base-sized version of BEiTScore++ for the same set of experiments described in Section 5.4. Overall, the base model exhibits the same behavior of the large model, confirming that the observed trends are not specific to model scale. In particular, the uncertainty estimates increase consistently with stronger input corruptions, and the model also reports higher uncertainty on out-of-domain artwork datasets, in comparison to the in-domain natural image benchmarks.

The uncertainty estimates increase under stronger input ambiguity, remain stable under color-based perturbations, and rise on out-of-domain artwork data due to model knowledge gaps.

D.1 Input Ambiguity

We first analyze the effect of input ambiguity under Gaussian blur. Some visual examples of the instance images, with different blur levels applied to them, can be seen in Figure 7. As in the main paper, we use the Polaris test split for this analysis and progressively increase the blur strength applied to each image. To reduce the confounding effect of the highly skewed score distribution in Polaris, we again downsample instances with a human score of zero, obtaining a more balanced set of image-caption pairs. This allows us to more clearly examine how uncertainty changes as the visual information becomes progressively degraded.

Figure 8 shows the confidence interval widths produced by the base model under increasing blur levels. The results follow the same pattern observed for the large model: the interval width grows with blur intensity, indicating that the estimator responds appropriately to increasing input ambiguity. As before, the strongest blur levels produce a slight reduction in interval width, which we attribute to the model becoming confident that the image-caption pair is strongly misaligned.

We also evaluate the base model under the broader set of distortions described in Appendix C, including salt-and-pepper noise, JPEG compression, random patch cutout, elastic warping, and contrast adjustments. The results are consistent with the large model, and can be seen in Figure 9. Most distortions lead to larger confidence intervals as their severity increases, while contrast and other color-based transformations produce little to

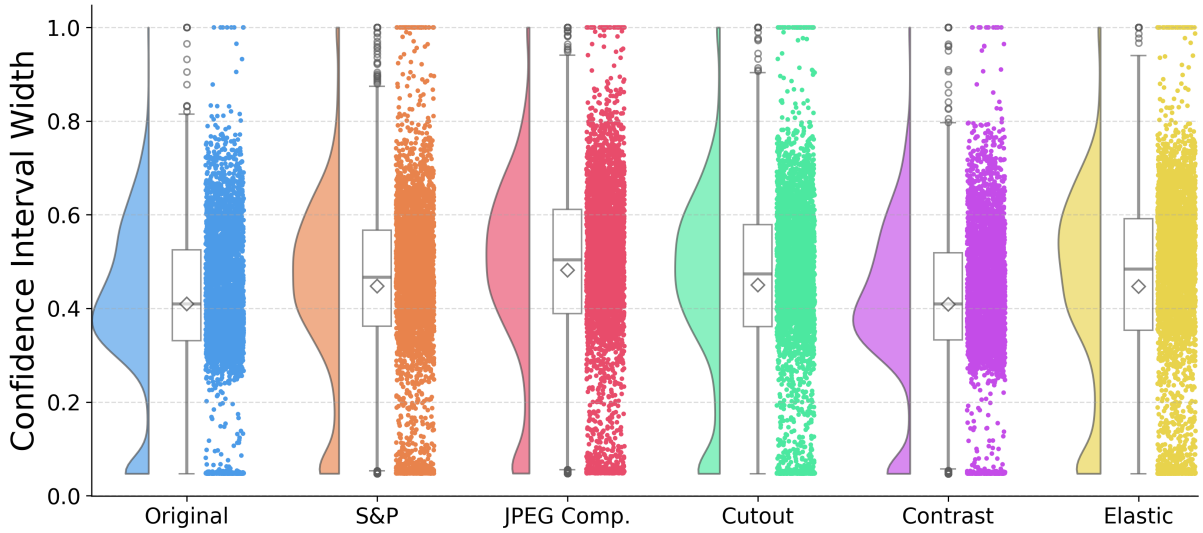

Figure 5: Box plots of confidence interval widths produced by our uncertainty estimator under different types of image distortions.

no change in uncertainty. This suggests that the base model is likewise robust to purely photometric perturbations and primarily sensitive to structural corruption of the image.

D.2 Model Knowledge Gaps

We next examine whether the base model reflects uncertainty arising from model knowledge gaps. As in the main experiments, we compare in-domain datasets containing natural images and realistic scenes, namely IIW, DOCCI, Localized Narratives, and MSCOCO, against out-of-domain artwork datasets, namely Docent and ArtCaps. We again also split the evaluation into long-caption and short-caption settings, to match the caption-length behavior expected by BEiTScore.

For long captions, we compare IIW, DOCCI, and Localized Narratives against Docent. For short captions, we compare ArtCaps against a randomly sampled subset of MSCOCO of equal size.

Figure 10 shows that the base model, like the large model, assigns substantially higher uncertainty to artwork datasets in both settings. This indicates that the uncertainty estimator captures epistemic uncertainty induced by domain shift, even at smaller model scale.

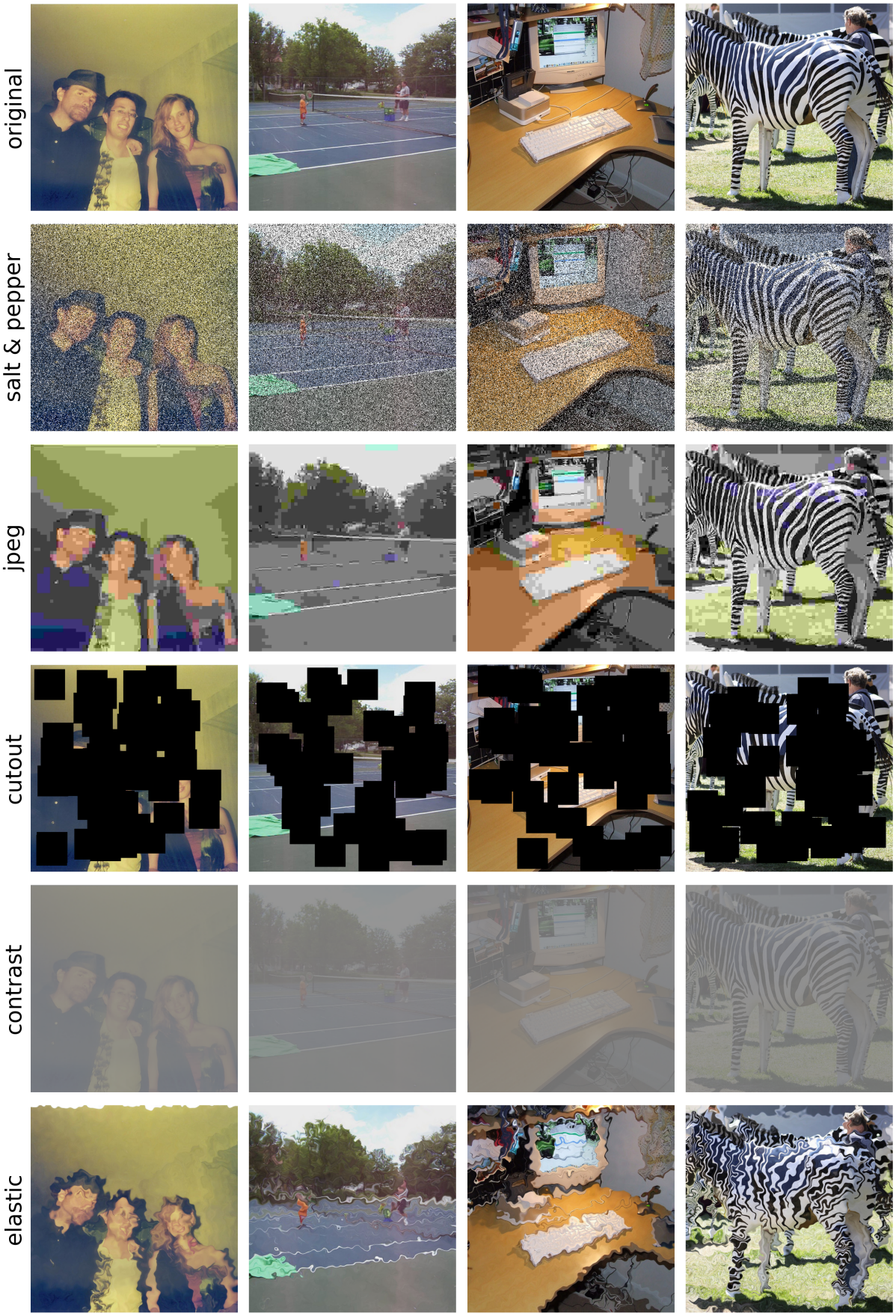

Figure 6: Examples of image distortions used to evaluate input ambiguity beyond Gaussian blur. Each row corresponds to a different type of perturbation.

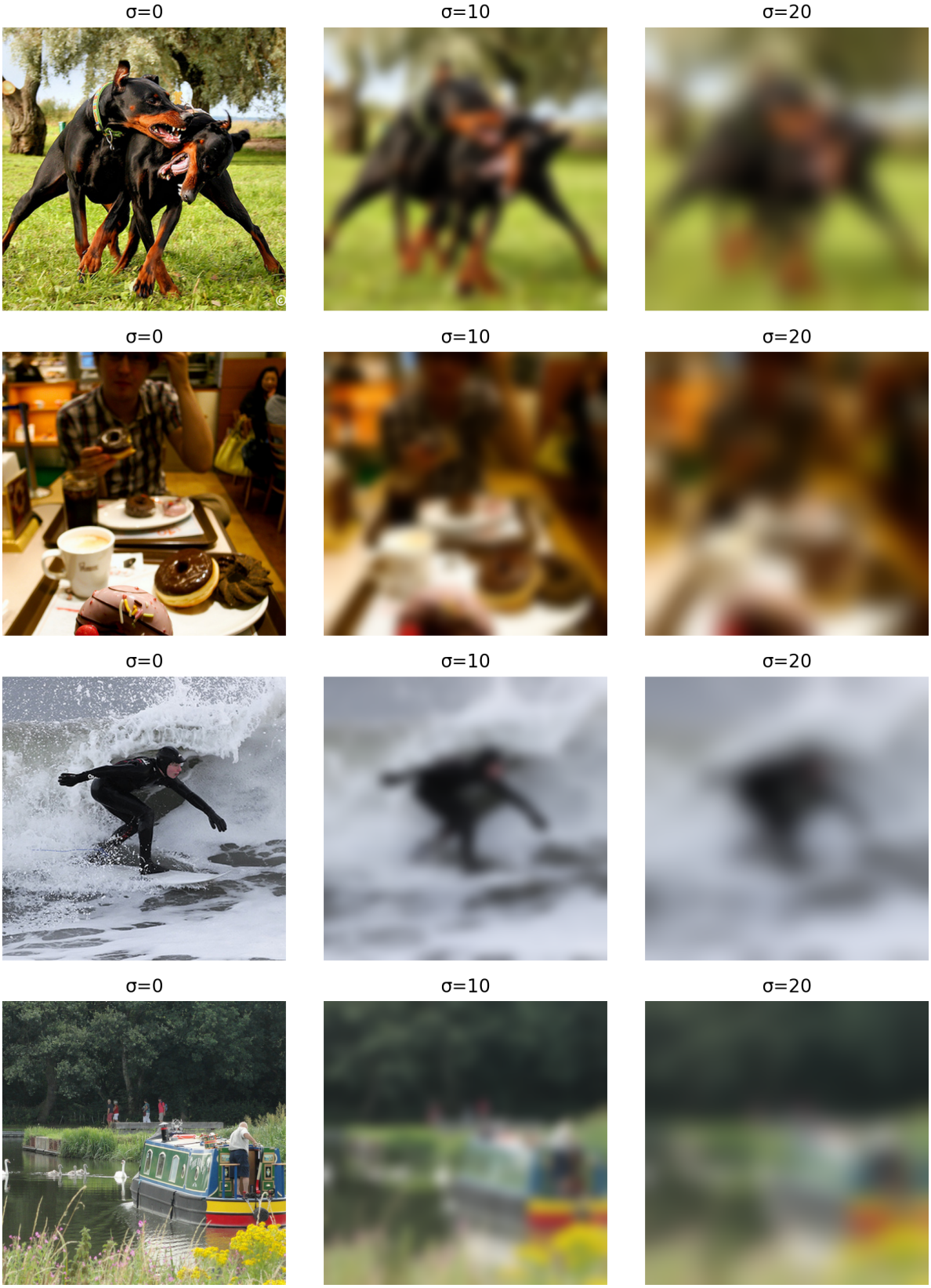

Figure 7: Examples of different levels of blur applied to the images of the Polaris test split during the evaluation of input ambiguity.

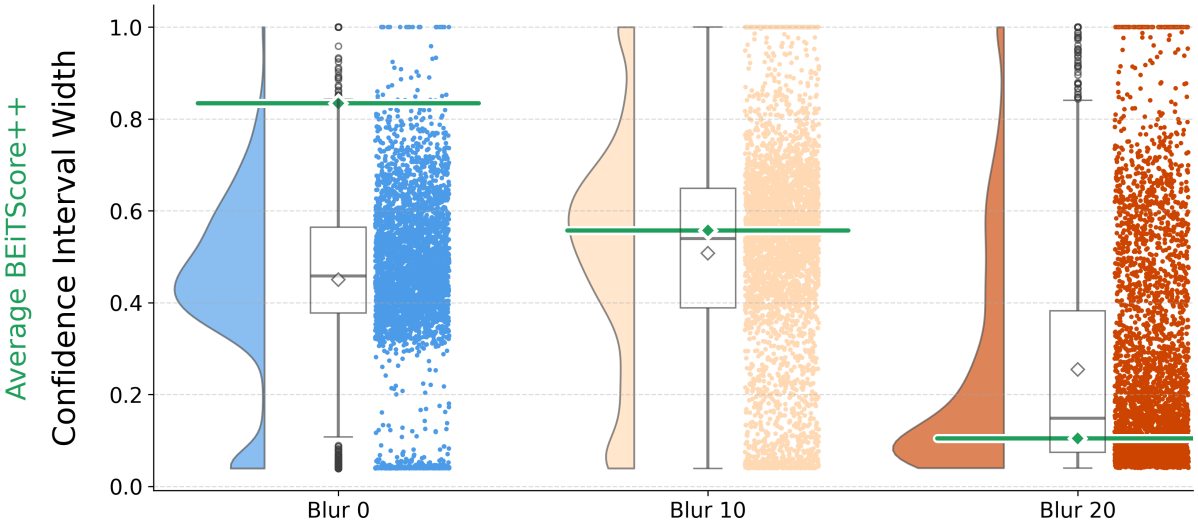

Figure 8: Distribution of 90% confidence interval widths generated by the base variant of BEiTScore++ as Gaussian blur increases on images from the Polaris test split. The horizontal line indicates the average median score produced by BEiTScore++ at each blur level.

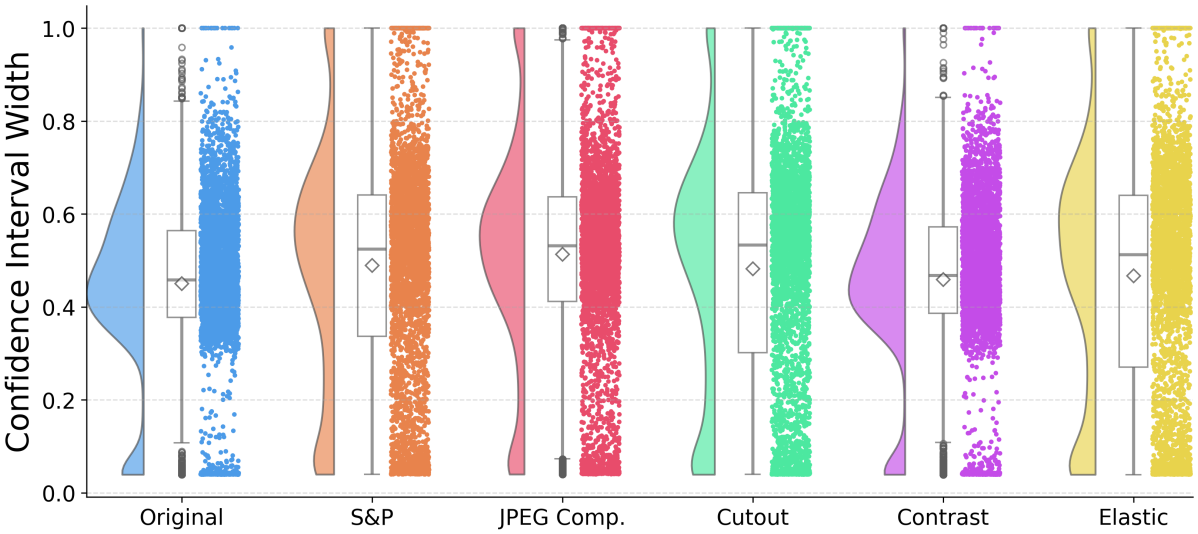

Figure 9: Box plots of confidence interval widths produced by the base-sized model under different types of distortions applied to images from the Polaris test split.

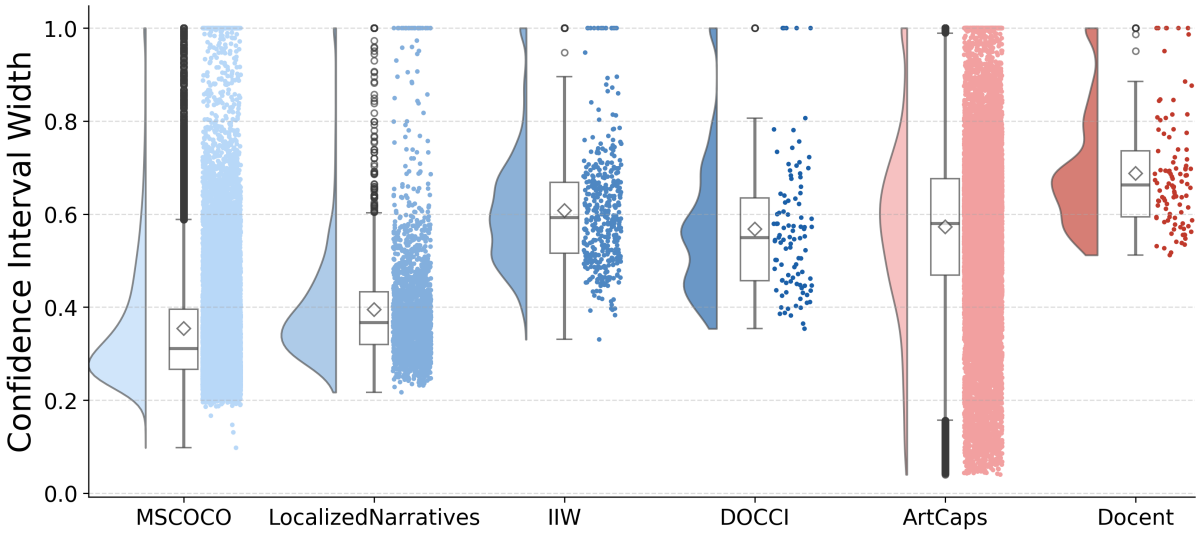

Figure 10: Distribution of 90% confidence interval widths produced by the base variant of BEiTScore++ for in-domain (natural images) and out-of-domain (artwork) datasets, across both short- and long-caption settings. The color gradient represents caption length, with pastel tones indicate short, direct captions, while more saturated tones correspond to datasets with more detailed descriptions.