Conformal Prediction Meets Test-Time Uncertainty in Vision Foundation Models

Huaze Liu
Harvey Mudd College
301 Platt Boulevard, Claremont, CA
hualiu@g.hmc.edu

Adyasha Mohanty
Harvey Mudd College
301 Platt Boulevard, Claremont, CA
admohanty@g.hmc.edu

Abstract

Foundation vision models have demonstrated remarkable capabilities in open-vocabulary semantic segmentation and object detection tasks. However, their deployment in safety-critical applications requires robust uncertainty quantification algorithms. We present an initial study of conformal prediction applied to CLIPSeg and Grounding DINO, two state-of-the-art open-vocabulary models. Our approach uses split conformal prediction to provide statistically valid prediction sets with user-specified coverage guarantees. We evaluate multiple uncertainty estimation baselines, including Monte Carlo dropout, test-time augmentation, and hybrid approaches across normal and adverse weather conditions on the KITTI dataset. Our results demonstrate that conformal prediction can successfully provide coverage guarantees for semantic segmentation, whereas object detection presents additional challenges requiring further improved conformity score designs.

1. Introduction

The rapid advancement of foundation vision models has revolutionized computer vision, enabling unified architectures trained on massive datasets to perform remarkably well across diverse tasks [8, 14]. Notably, models like CLIPSeg [10] and Grounding DINO [9] have demonstrated strong open-vocabulary generalization in semantic segmentation and object detection, respectively, through zero-shot and few-shot capabilities. However, deploying these powerful models in safety-critical domains requires rigorous uncertainty quantification with statistically valid reliability guarantees.

Traditional uncertainty quantification methods, including Bayesian neural networks [3], ensemble techniques [7], and test-time augmentation [18], provide approximate confidence estimates but often suffer from scalability limitations or lack formal guarantees. Calibration methods [4, 6] offer post-hoc corrections but do not address the underlying epistemic uncertainty or provide coverage guarantees.

Conformal prediction [15, 17] presents a compelling alternative by providing finite-sample coverage guarantees in a distribution-free manner and without requiring retraining. Using conformity scores derived from a held-out calibration set [12], this framework has shown promise in classification [1], segmentation [16], and detection [2]. However, applications to open-vocabulary settings remain relatively nascent. These scenarios introduce unique challenges: semantic ambiguity, concept drift, and the open-ended nature of the label space [13]. While recent work has explored robustness [5] and calibration [11] in vision-language models, few have tackled uncertainty quantification in this setting.

In this work, we address this gap by adapting split conformal prediction to two open-vocabulary foundation models. We design novel conformity scores tailored to pixel-level and object-level outputs, benchmark them under both normal and adverse weather conditions, and compare their performance across multiple baselines. Our experiments on the KITTI dataset reveal that conformal prediction provides reliable coverage in segmentation tasks but struggles with detection, where localization and label matching further complicate uncertainty estimation, highlighting the need for more structured, task-aware conformity designs.

2. Methods

We develop a selective prediction mechanism that, given an input image, returns prediction sets containing the ground-truth label with user-specified probability $1-\alpha$ for each spatial location (semantic segmentation) or object instance (detection). We employ $\alpha \in \{0.20, 0.10, 0.05\}$, corresponding to target coverage levels of 80%, 90%, and 95%, respectively. We adopt the **split conformal prediction** (split-CP) paradigm [12], which leverages a held-out calibration subset to determine a scalar threshold q_α. This threshold is then applied uniformly to the test split, ensuring finite-sample validity guarantees. Crucially, our procedure is training-free: the foundation models remain frozen throughout the entire process, making our approach directly applicable to existing pre-trained models without requiring additional computational resources for retraining.

We evaluate our method on two state-of-the-art open-vocabulary foundation models without any fine-tuning. **CLIPSeg** [10] performs open-vocabulary semantic segmentation by generating per-prompt logit maps $L_c(x) \in \mathbb{R}^{H\times W}$, which are softmax-normalized across prompts to yield pixel-wise class probabilities. **Grounding DINO** [9] performs open-vocabulary object detection by predicting bounding boxes $b_i \in \mathbb{R}^4$, class logits $\ell_i \in \mathbb{R}^C$, and confidence scores $s_i \in [0,1]$. This model leverages a transformer-based architecture to unify object detection with phrase grounding in natural language. All evaluations are conducted using the original pre-trained models to assess out-of-the-box performance.

2.1. Uncertainty Estimation Baselines

We evaluate four test-time inference strategies to study their impact on conformal prediction without modifying model weights: **(1) Vanilla** uses a single forward pass as the deterministic baseline; **(2) Monte Carlo (MC)-Dropout** [3] enables dropout at test time with $N = 10$ stochastic passes to estimate epistemic uncertainty; **(3) Test-Time Augmentation (TTA)** applies $M = 8$ geometric augmentations (rotations, flip, identity) to model aleatoric uncertainty; and **(4) Hybrid** combines 5 original and 5 flipped passes for a balance of diversity and cost. For CLIPSeg, pixel-wise probabilities are averaged across passes; for Grounding DINO, predictions are merged via per-class NMS.

2.2. Conformity Score Design

We design scores that monotonically increase with prediction uncertainty, enabling consistent acceptance rules where predictions are retained only when their conformity score is at most q_α.

2.2.1. Segmentation Conformity Scores

For pixel-wise predictions, let $p(x) \in \Delta^{C-1}$ denote the class probability distribution at spatial location x. We evaluate entropy score as an uncertainty measure which is defined as:

$$\phi_{\text{ent}}(x) = -\sum_{c=1}^{C} p_c(x) \log p_c(x) \tag{1}$$

This measure captures overall prediction uncertainty, with larger values indicating greater ambiguity across all classes.

2.2.2. Detection Conformity Scores

Object detection presents additional complexity due to the joint requirements of accurate localization and classification. For each ground-truth object g with class label y_g, we define a **best-match confidence** by searching across all predicted bounding boxes $(b_i, s_i, \hat{y}_i)$:

$$\psi(g) = \max_{i:\text{IoU}(b_i,g)\geq\tau\wedge(\hat{y}_i=y_g \text{ if label matching enabled})} s_i \tag{2}$$

where $\tau = 0.5$ represents the IoU threshold for spatial overlap. When no prediction satisfies the matching criteria, we set $\psi(g) = 0$. Higher ψ values indicate more confident matches between predictions and ground truth. To maintain consistency with our acceptance rule, we apply the monotone transformation $\phi_{\text{det}}(g) = 1 - \psi(g)$, ensuring that uncertain predictions yield high conformity scores.

2.3. Split Conformal Prediction Procedure

Our implementation follows the standard split conformal prediction protocol [17]. Given conformity scores ϕ_i^{cal} calculated in the calibration split, representing pixels for segmentation or ground-truth objects for detection, we determine the decision threshold as:

$$q_\alpha = \text{Quantile}_{1-\alpha}^{\text{higher}}(\phi_i^{\text{cal}}) \tag{3}$$

The higher quantile rule ensures finite-sample coverage guarantees by selecting the smallest threshold that satisfies the coverage requirement.

During test-time inference, our selective mechanism operates as follows: for segmentation, we return the singleton prediction $\arg\max_c p_c(x)$ at pixel x if and only if $\phi(x) \leq q_\alpha$; for detection, we retain object g if and only if $\phi_{\text{det}}(g) \leq q_\alpha$. When the conformity score exceeds the threshold, the mechanism abstains from making a prediction. Under the standard exchangeability assumption between calibration and test distributions, this procedure guarantees that the probability of the ground-truth label being included in the returned prediction set is at least $1-\alpha$, providing formal statistical validity for our selective predictions.

3. Experiments and Results

We evaluate both *selective coverage* and *reliability under normal weather and adverse conditions* for segmentation and detection. Coverage refers to the fraction of test pixels (segmentation) or ground-truth objects (detection) where a non-empty prediction set is returned. In detection, a ground-truth object is considered covered if at least one predicted box has IoU $\geq \tau$ (default τ=0.5) and, unless noted otherwise, a matching class label. When exchangeability holds, observed coverage should approximate the nominal level $1-\alpha$. Additionally, we compute the Area Under the Risk–Coverage Curve (AURC), which summarizes the tradeoff between abstention and reliability by plotting error rate (risk) versus coverage under varying thresholds.

We compute a global threshold q_α for each inference mode (Vanilla, MC-Dropout, TTA, Hybrid) using the calibration set, then evaluate fixed thresholds on the test split. MC and TTA outputs are aggregated as described in §2.1. Quantiles are computed using the "higher" rule to ensure coverage. Adverse weather tests follow the same protocol to assess generalization.

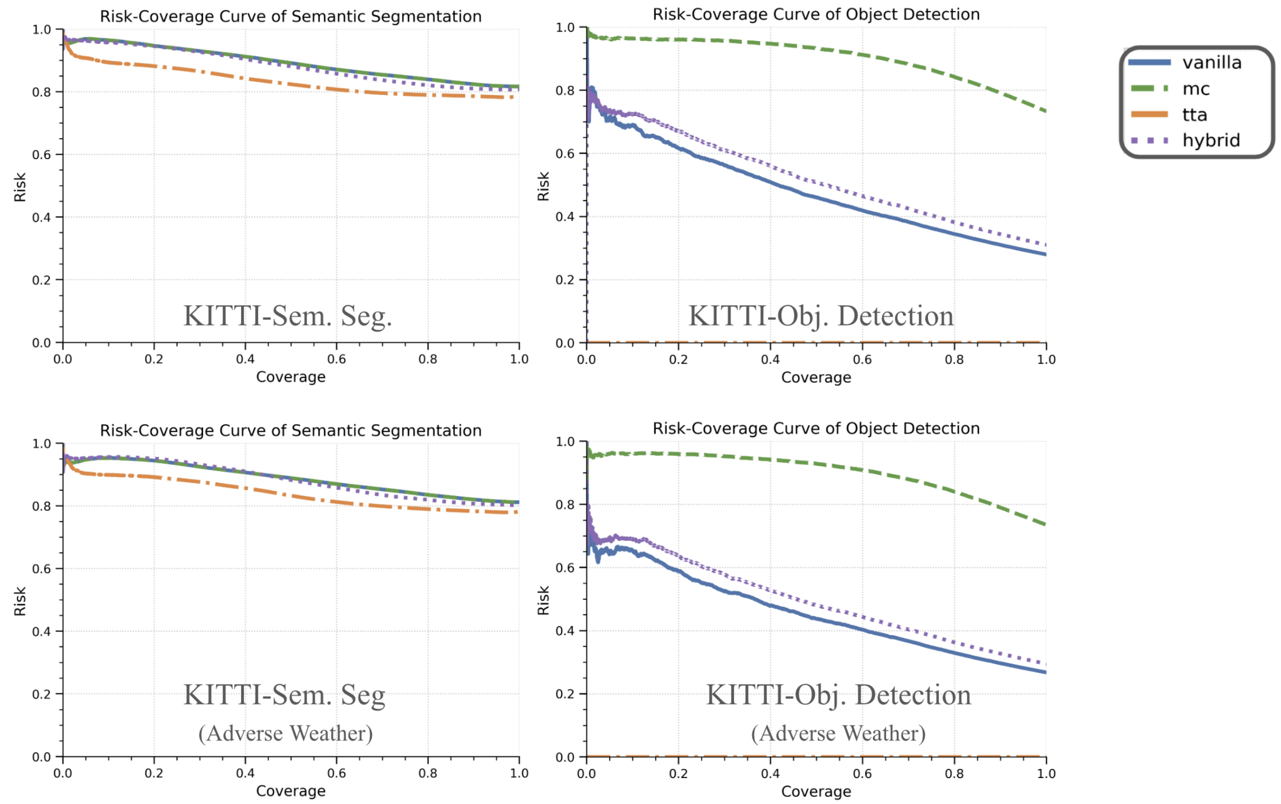

Figure 1. Risk-Coverage Curves for semantic segmentation (CLIPSeg) and object detection (Grounding DINO) across normal (top) and adverse weather (bottom) KITTI conditions. Lower risk at higher coverage indicates better calibration. For segmentation, TTA achieves consistently lower risk, while in object detection, MC-Dropout significantly underperforms, and Hybrid offers improved tradeoffs under adverse conditions

Figure 1 illustrates the risk-coverage trade-offs for CLIPSeg and Grounding DINO across normal and adverse weather scenarios. For semantic segmentation with CLIPSeg, the results strongly favor TTA as the most effective uncertainty-aware inference strategy. It consistently yields the lowest AURC which shows high selective reliability. At $\alpha = 0.05$, TTA achieves a coverage of 95.14 % under normal conditions and 95.45 % under adverse weather, both within 0.5 % of the nominal target, and with the lowest corresponding entropy threshold. The Hybrid method also shows improvement over Vanilla and MC-Dropout, particularly under domain shift. Notably, it achieves the highest static class coverage at $\alpha = 0.05$, with dynamic class coverage of 82.15 %, reflecting a balanced handling of both large background regions and smaller foreground objects. In contrast, MC-Dropout while replicating Vanilla's behavior in segmentation due to shared dropout masks, fails to provide additional benefits in this context. Across all tested methods, segmentation consistently shows higher conformity than detection, with robust coverage even under adverse conditions.

Baseline	$q_{80\%}$	$q_{90\%}$	$q_{95\%}$
Vanilla	0.2104	0.2481	0.2847
MC-Dropout	0.2104	0.2481	0.2847
TTA	0.2091	0.2270	0.2446
Hybrid	0.2108	0.2407	0.2695
Baseline	Cov@80%	Cov@90%	Cov@95%
Vanilla	0.7995	0.8984	0.9474
MC-Dropout	0.7995	0.8984	0.9474
TTA	0.7902	0.9000	0.9514
Hybrid	0.7927	0.8951	0.9479

Table 1. CLIPSeg conformal prediction results on KITTI (semantic segmentation) using entropy as the conformity score. The table shows both the threshold values (q_α) and observed coverage for target coverage levels $1 - \alpha \in \{0.80, 0.90, 0.95\}$.

However, the static–dynamic split reveals persistent calibration gaps: at $\alpha = 0.10$, dynamic object coverage for Vanilla and MC-Dropout remains under 60 %, while TTA and Hybrid exceed 70 %, underscoring their improved handling of epistemic uncertainty in fine-grained object classes.

Baseline	$q_{80\%}$	$q_{90\%}$	$q_{95\%}$
Vanilla	0.2134	0.2514	0.2863
MC-Dropout	0.2134	0.2514	0.2863
TTA	0.2126	0.2309	0.2490
Hybrid	0.2148	0.2442	0.2715

Baseline	Cov@80%	Cov@90%	Cov@95%
Vanilla	0.7972	0.8965	0.9454
MC-Dropout	0.7972	0.8965	0.9454
TTA	0.7891	0.9002	0.9545
Hybrid	0.7944	0.8969	0.9474

Table 2. CLIPSeg conformal prediction results on KITTI Adverse Weather (semantic segmentation) using entropy as the conformity score. The table reports both the threshold values (q_α) and the observed coverage for target coverage levels $1 - \alpha \in \{0.80, 0.90, 0.95\}$.

Baseline	α	Static Cov.	Dynamic Cov.	AURC
Vanilla	0.20	0.7793	0.3625	0.8920
Vanilla	0.10	0.8980	0.5791	0.8920
Vanilla	0.05	0.9484	0.7453	0.8920
MC-Dropout	0.20	0.7793	0.3625	0.8920
MC-Dropout	0.10	0.8980	0.5791	0.8920
MC-Dropout	0.05	0.9484	0.7453	0.8920
TTA	0.20	0.7951	0.4539	0.8339
TTA	0.10	0.8979	0.7240	0.8339
TTA	0.05	0.9481	0.8395	0.8339
Hybrid	0.20	0.8430	0.4459	0.8823
Hybrid	0.10	0.9220	0.7067	0.8823
Hybrid	0.05	0.9559	0.8215	0.8823

Table 3. CLIPSeg conformal prediction summary on KITTI (normal weather). The table reports static and dynamic pixel coverage and AURC values for different miscoverage levels.

In contrast, object detection with Grounding DINO presents a more difficult challenge. The risk coverage curves remain flat, especially in TTA and Hybrid, which fail to cover dynamic objects at any α level. The Hybrid strategy marginally improves detection performance over Vanilla, reaching 69.47 % coverage at $\alpha = 0.10$ under normal weather. MC-Dropout shows the highest dynamic coverage under adverse conditions, at the cost of poor risk calibration, as reflected in the increased AURC values, which suggest that it lacks precision in abstention, often retaining low-quality boxes. Moreover, detection coverage remains systematically below the target: even the best performing MC-Dropout baseline falls by nearly 10 %. This result shows the difficulty of conformal prediction in detection tasks, where both localization and classification accuracy are crucial.

Dataset	Baseline	Cov@80%	Cov@90%	Cov@95%
Normal	Vanilla	0.6014	0.6717	0.7040
	MC-Dropout	0.6948	0.7752	0.8130
	TTA	0.6203	0.6947	0.7263
	Hybrid	0.6203	0.6947	0.7263
Adverse	Vanilla	0.5811	0.6530	0.6893
	MC-Dropout	0.6866	0.7683	0.8082
	TTA	0.6001	0.6775	0.7116
	Hybrid	0.6001	0.6775	0.7116

Table 4. Grounding DINO conformal prediction coverage on the KITTI dataset under normal and adverse weather conditions. The table reports observed coverage for target levels $1 - \alpha \in \{0.80, 0.90, 0.95\}$.

Dataset	Baseline	α	Dynamic Cov.	AURC
Normal	Vanilla	0.20	0.4334	0.4804
	Vanilla	0.10	0.4855	0.4804
	Vanilla	0.05	0.5179	0.4804
	MC-Dropout	0.20	0.4658	0.9043
	MC-Dropout	0.10	0.5440	0.9043
	MC-Dropout	0.05	0.6006	0.9043
	Hybrid	0.20	0.4540	0.5219
	Hybrid	0.10	0.5199	0.5219
	Hybrid	0.05	0.5521	0.5219
Adverse	Vanilla	0.20	0.4042	0.4549
	Vanilla	0.10	0.4675	0.4549
	Vanilla	0.05	0.4948	0.4549
	MC-Dropout	0.20	0.4586	0.9010
	MC-Dropout	0.10	0.5521	0.9010
	MC-Dropout	0.05	0.6022	0.9010
	Hybrid	0.20	0.4238	0.4948
	Hybrid	0.10	0.4984	0.4948
	Hybrid	0.05	0.5272	0.4948

Table 5. Conformal prediction results for Grounding DINO on the KITTI dataset under normal and adverse weather conditions. Static coverage is not applicable to detection. Results for TTA are excluded due to consistent failure, yielding zero valid predictions.

4. Discussion

Our results highlight the novelty of applying conformal prediction to open-vocabulary vision models. While CLIPSeg readily supports plug-and-play conformal inference due to its dense outputs, object detection with Grounding DINO exposes limitations of standard test-time strategies. This underscores the need for task-aware abstention and calibration methods tailored to detection. Future work should develop principled uncertainty measures leveraging multi-scale consistency or foundation model priors.

References

[1] Anastasios N Angelopoulos, Stephen Bates, Jitendra Malik, Benjamin Recht, John Duchi, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. 1

[2] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022. 1

[3] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016. 1, 2

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 1

[5] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1

[6] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pages 3787–3798, 2019. 1

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6414, 2017. 1

[8] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2

[10] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 1, 2

[11] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, pages 15682–15694, 2021. 1

[12] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002. 1

[13] Matthieu Parchet, Matthieu Cord, and Charles Ollion. Measuring semantic inconsistencies in text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7150–7159, 2023. 1

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual representations from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[15] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008. 1

[16] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022. 1

[17] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005. 1, 2

[18] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019. 1