Conformal Prediction for Reliable Test-Time Uncertainty in Robotic Vision

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 show strong open-vocabulary performance in segmentation and detection, yet robotics applications demand reliable uncertainty estimation for safe deployment. We study split conformal prediction on CLIPSeg and Grounding DINO, demonstrating statistically valid coverage without retraining. Evaluations on the KITTI dataset under normal and adverse weather indicate that conformal prediction provides dependable coverage for semantic segmentation, while object detection remains more challenging, highlighting the need for task-aware conformity designs in robotic perception systems.

1. Introduction

Recent foundation vision models have advanced robotic perception by enabling open-vocabulary understanding via large-scale pretraining [8, 14]. CLIPSeg [10] and Grounding DINO [9] generalize to semantic segmentation and object detection through zero-shot and few-shot transfer. For safety-critical robotic systems, deployment requires uncertainty quantification with statistically valid guarantees.

Classical approaches—Bayesian neural networks [3], ensembles [7], and test-time augmentation [18]—yield approximate confidence but lack formal guarantees and may scale poorly. Calibration methods [4, 6] improve consistency yet do not capture epistemic uncertainty or provide coverage control.

Conformal prediction [15, 17] offers distribution-free, finite-sample coverage without retraining, using conformity scores from a held-out calibration set [12]. It shows promise in classification [1], segmentation [16], and detection [2], but open-vocabulary applications remain limited. These settings introduce semantic ambiguity, concept drift, and unbounded label spaces [13]. Prior work studies robustness [5] and calibration [11] in vision-language models, but uncertainty quantification is still underexplored.

We apply split conformal prediction to CLIPSeg and Grounding DINO for open-vocabulary robotic perception. We design pixel- and object-level conformity scores, evaluate on KITTI under normal and adverse weather, and compare inference strategies. Results show reliable segmentation coverage, while detection is harder, motivating task-aware conformity for robotic vision.

2. Methods

We develop a training-free selective prediction mechanism that, given an image, outputs prediction sets guaranteed to include the ground-truth label with user-specified probability $1 - \alpha$. We adopt the **split conformal prediction (split-CP)** framework [12], which computes a threshold q_α on a held-out calibration set and applies it to test data to ensure finite-sample validity. We use $\alpha \in \{0.20, 0.10, 0.05\}$, corresponding to target coverage of 80%, 90%, and 95%. The foundation models remain frozen, allowing conformal prediction to be applied directly without retraining—an essential property for robotic deployment.

We evaluate two open-vocabulary vision models: **CLIPSeg** [10] for semantic segmentation and **Grounding DINO** [9] for object detection. CLIPSeg generates per-prompt logit maps $L_c(x) \in \mathbb{R}^{H \times W}$, normalized to pixel-wise class probabilities, while Grounding DINO predicts bounding boxes $b_i \in \mathbb{R}^4$, class logits ℓ_i, and confidence scores $s_i \in [0, 1]$ using a transformer-based detection backbone. All evaluations are performed on the pre-trained checkpoints under normal and adverse KITTI weather conditions.

2.1. Uncertainty Estimation Baselines

We study several inference-time strategies without modifying weights: **(1) Vanilla**, a single deterministic forward pass; **(2) Test-Time Augmentation (TTA)**, applying geometric transformations ($M = 8$ rotations and flips) to estimate aleatoric uncertainty; and **(3) Hybrid**, combining original and flipped passes to balance diversity and cost. For CLIPSeg, pixel-wise probabilities are averaged across passes; for Grounding DINO, predictions are merged using per-class non-maximum suppression (NMS).

2.2. Conformity Score Design

We design conformity scores that increase monotonically with uncertainty, enabling a consistent accept/reject rule based on q_α.

Segmentation. For pixel-wise class probabilities $p(x) \in \Delta^{C-1}$, we use the entropy-based score:

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

which measures uncertainty across classes, with higher values indicating less confident predictions.

Detection. For each ground-truth object g with label y_g, we compute the best-match confidence:

$$\psi(g) = \max_{i:\, \text{IoU}(b_i, g) \geq \tau \wedge (\hat{y}_i = y_g)} s_i, \quad \tau = 0.5, \tag{2}$$

and define $\phi_{\text{det}}(g) = 1 - \psi(g)$ so that larger values correspond to higher uncertainty. When no prediction matches g, $\psi(g) = 0$.

2.3. Split Conformal Procedure

Given calibration scores $\{\phi_i^{\text{cal}}\}$ from the held-out split, we compute:

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

ensuring valid finite-sample coverage [17]. During inference, a prediction is accepted if its score $\phi(x) \leq q_\alpha$ and abstained otherwise. For segmentation, this retains pixels with sufficiently low entropy; for detection, it filters objects with low confidence. Under the exchangeability assumption between calibration and test sets, the probability that the ground truth lies within the returned prediction set is at least $1-\alpha$, providing statistically valid reliability for robotic perception.

3. Experiments and Results

We evaluate *selective coverage* and *reliability under normal and adverse weather* on KITTI for both segmentation and detection. Coverage measures the fraction of pixels (segmentation) or ground-truth objects (detection) with a non-empty prediction set. A prediction is considered covered if its conformity score is below the calibrated threshold q_α. When exchangeability holds, observed coverage should approximate the nominal level $1 - \alpha$. We further report the Area Under the Risk–Coverage Curve (AURC) to quantify the tradeoff between abstention and accuracy.

A single global threshold q_α is computed for each inference mode (Vanilla, TTA, Hybrid) using the calibration split and then applied to the test set. Quantiles use the "higher" rule to guarantee coverage, and the same protocol is repeated under adverse weather conditions.

Baseline	$q_{80\%}$	$q_{90\%}$	$q_{95\%}$
Vanilla	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
TTA	0.7902	0.9000	0.9514
Hybrid	0.7927	0.8951	0.9479

Table 1. CLIPSeg conformal prediction on KITTI (semantic segmentation) using entropy as the conformity score. Observed coverage matches nominal targets, with TTA achieving slightly lower AURC and better robustness under domain shift.

Baseline	Cov@80%	Cov@90%	Cov@95%
Vanilla	0.601	0.672	0.704
TTA	0.620	0.695	0.726
Hybrid	0.623	0.698	0.729

Table 2. Grounding DINO conformal coverage on KITTI under normal conditions. Detection coverage remains below segmentation but improves under Hybrid inference.

Overall risk–coverage trends align with the quantitative results, showing consistent calibration for segmentation and lower coverage for detection.

For semantic segmentation, conformal prediction maintains near-nominal coverage across all α levels. TTA provides the lowest AURC and the most stable thresholds under both normal and adverse weather, indicating reliable calibration for pixel-level uncertainty. Hybrid inference further improves coverage for small dynamic classes while remaining computationally efficient.

For object detection, Grounding DINO achieves lower overall coverage due to compounded localization and classification uncertainty. Hybrid inference slightly improves reliability compared to the deterministic baseline. Despite the remaining gap to nominal levels, conformal prediction provides interpretable coverage control, a useful property for safety-critical robotic perception systems.

4. Discussion

This study demonstrates the use of conformal prediction for open-vocabulary foundation models in robotic perception. CLIPSeg provides reliable pixel-level coverage, while detection with Grounding DINO remains less calibrated due to joint localization–classification uncertainty. These findings suggest that task-aware conformity and structured uncertainty measures are essential for extending conformal prediction to complex robotic vision tasks.

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

[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

[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

[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