Hybrid Adaptive Prediction Sets for Conformal Prediction

Soundouss Messoudi [1] **Abdelhak Imoussaten** [2]

Abstract

Probability-based Conformal Prediction (CP) often fails under distribution shift due to deep neural network overconfidence, producing unreliable prediction sets. To address this, we introduce Hybrid Adaptive Prediction Sets (HAPS), a non-conformity framework combining the discriminative precision of supervised classifiers with the zero-shot geometric robustness of frozen Foundation Models (FMs). By multiplicatively fusing parameterized probabilities and semantic latent similarities, HAPS enforces a "geometric veto" to systematically suppress overconfidence. We propose three scalable variants (kNN-HAPS, Fix-HAPS, and Var-HAPS) to balance topological precision and computational cost. Evaluations on ImageNet-1K show HAPS drastically reduces average prediction set sizes while maintaining exact marginal coverage. Furthermore, HAPS significantly outperforms standard baselines under severe shifts (ImageNet-A/R), producing better coverage and efficiency performances.

1. Introduction

Deep neural networks frequently suffer from severe overconfidence, rendering their raw probabilistic outputs unreliable in high-stakes domains (Guo et al., 2017). Conformal Prediction (CP) addresses this by outputting rigorous prediction sets that guarantee marginal coverage of the true label (Vovk et al., 2005; Angelopoulos & Bates, 2023). Standard probability-based CP approaches offer highly efficient prediction sets for in-distribution data but extrapolate with extreme overconfidence under distribution shift (Papernot & McDaniel, 2018). Other emerging methods leverage the generalized latent spaces of FMs (FMs) (Silva-Rodríguez et al., 2025; Fillioux et al., 2024). While completely frozen, zero-shot FMs provide exceptional geometric robustness out-of-distribution but lack the explicitly optimized decision boundaries required for highly efficient prediction sets.

To bypass this trade-off without sacrificing zero-shot robustness, we propose Hybrid Adaptive Prediction Sets (HAPS). HAPS introduces a novel non-conformity framework that mathematically synergizes the specialized discriminative precision of a standard supervised classifier with the untampered geometric properties of a completely frozen multimodal FM. By employing a multiplicative fusion strategy, HAPS requires mutual agreement between the classifier's probability and the FM's semantic latent space. This acts as a "geometric veto," systematically suppressing overconfidence and preventing the construction of indiscriminately large or dangerously empty prediction sets under shift.

The primary contributions of this work are as follows:

- **Hybrid Conformity Scores:** We introduce a unified framework that fuses classifier probabilities with FM cosine similarities, easily integrating into established CP algorithms like LAC, APS and RAPS.
- **Scalable Latent Representations:** To balance topological precision with latency and memory constraints, we propose three score variants: approximate similarity (kNN-HAPS), fixed compressed centroids (Fix-HAPS), and adaptive multi-centroids (Var-HAPS).
- **Robust Empirical Validation:** Experiments on ImageNet-1K and its shifted variants (ImageNet-A/R) demonstrate that HAPS dynamically optimizes the trade-off between standard probability and semantic similarity, yielding highly efficient sets and improved coverage under distribution shift.

2. Related Work

Probability-Based Conformal Prediction. Conformal Prediction (CP) constructs prediction sets $\hat{C}_{1-\alpha}(x)$ for a new test instance x satisfying the marginal coverage property $\mathbb{P}(y \in \hat{C}_{1-\alpha}(x)) \geq 1 - \alpha$, assuming data exchangeability (Papadopoulos et al., 2002). Foundational methods like the Least Ambiguous Classifier (LAC) (Sadinle et al., 2019), Adaptive Prediction Sets (APS) (Romano et al., 2020), and

[1]Heudiasyc - UMR CNRS 7253, Université de Technologie de Compiègne, Compiègne, France. [2]SyCoIA, IMT Mines Alès, Alès, France.. Correspondence to: Soundouss Messoudi <soundouss.messoudi@hds.utc.fr>, Abdelhak Imoussaten <abdelhak.Imoussaten@mines-ales.fr>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

Regularized Adaptive Prediction Sets (RAPS) (Angelopoulos et al., 2020) construct these sets by thresholding or accumulating the softmax probabilities of candidate classes. However, because these methods rely strictly on parameterized classifier outputs, they frequently extrapolate with extreme overconfidence (Papernot & McDaniel, 2018), particularly under domain shifts, out-of-distribution (OOD) settings, or adversarial perturbations. This reliance leads to prediction sets that are either dangerously empty (in LAC) or indiscriminately large as they accumulate the heavy tails of uncalibrated probability mass (in APS).

Distance and Representation-Based Conformity. To mitigate softmax overconfidence, distance-based Non-Conformity Measures (NCMs) evaluate conformity through spatial proximity. While foundational approaches like CP-kNN rely on the relative distance of an instance to candidate versus competing classes (Papadopoulos et al., 2002), the high inference cost over large datasets has driven a shift toward centroid-based and latent-space methods. For instance, (Fisch et al., 2021) conformalized Prototypical Networks (Snell et al., 2017) via the distance between query embeddings and empirical class centroids. Capitalizing on deep representation geometries, the Deep k-Nearest Neighbors (DkNN) algorithm aggregates label disagreements across multiple hidden layers (Papernot & McDaniel, 2018), while Neighborhood Conformal Prediction (NCP) (Ghosh et al., 2023) derives adaptive sets by weighting conformity scores within the classifier's penultimate representation space.

FMs for Conformal Prediction. Recently, the integration of FMs (FMs), such as CLIP (Radford et al., 2021), has replaced traditional softmax approximations with continuous cosine similarity metrics computed within a shared normalized embedding space (Silva-Rodríguez et al., 2025) or by adding a linear probing to the frozen FM (Fillioux et al., 2024). Thus, these conformal approaches take advantage of the fact that FMs possess well-documented zero-shot robustness to distribution shifts.

Building on these complementary strengths, our work proposes a hybrid non-conformity measure that mathematically fuses the classifier's output with the robust geometric properties of a frozen multimodal FM in order to construct efficient and robust conformal predictors.

3. HAPS: Hybrid Adaptive Prediction Sets

3.1. Hybrid Conformity Scoring

To generate the raw hybrid score $\tilde{s}_{\text{HAPS}}(x, y)$, we fuse the classifier's softmax probabilities $\hat{P}(y|x)$ with the FM's cosine similarities $Sim_{FM}(x, y)$ (clipped to $[0, 1]$) using a mixing parameter $\beta \in [0, 1]$. The multiplicative strategy functions as a logical *AND* gate, requiring mutual agreement between the classifier and the FM's latent space:

$$\tilde{s}_{\text{HAPS}}(x, y) = \left(\hat{P}(y|x)\right)^{1-\beta} \times (Sim_{FM}(x, y))^{\beta} \quad (1)$$

If the FM determines the image is geometrically disjoint from class y ($Sim_{FM} = 0$), the product collapses. This acts as a "geometric veto," suppressing overconfidence if the classifier assigns high probability to an implausible class (and vice versa).

3.2. Conformal Calibration Adaptation

The integration of the hybrid score depends on the mathematical requirements of the underlying CP algorithm.

- **Least Ambiguous Classifier (LAC):** Because LAC relies on independent marginals, we directly substitute the classifier's probability with our raw hybrid score to form the NCM: $S_{LAC}(x, y) = 1 - \tilde{s}_{\text{HAPS}}(x, y)$.
- **Adaptive Prediction Sets (APS) and Regularized APS (RAPS):** Because APS and RAPS accumulate probability mass, the scores across all candidate classes $\mathcal{Y}$ must form a valid distribution summing to 1. We apply L_1 normalization to satisfy this constraint:

$$s^{norm}_{\text{HAPS}}(x, y) = \frac{\tilde{s}_{\text{HAPS}}(x, y)}{\sum_{y' \in \mathcal{Y}} \tilde{s}_{\text{HAPS}}(x, y') + \epsilon_{norm}} \quad (2)$$

  where $\epsilon_{norm} \approx 10^{-10}$. These normalized scores are then seamlessly accumulated (and penalized for RAPS) to construct the conformal prediction sets.

3.3. Latent Space Representations and HAPS Variants

We utilize a frozen FM Φ to project images into a latent space: $z_i = \Phi(x_i)$. The formulation of the semantic similarity $Sim_{FM}(x, y)$ between a test image x and a candidate class y depends on the chosen HAPS variant.

1. kNN-HAPS (Approximate k Nearest Neighbors): For environments without strict latency or memory bottlenecks, kNN-HAPS uses a similarity computed against the Approximate kNN with Hierarchical Navigable Small World (HNSW) graphs (Malkov & Yashunin, 2018) in the raw high-dimensional space. While geometrically precise, it is relatively slow and requires massive memory for graph-based indices.

2. Fix-HAPS (Fixed Centroid Compression): If time or memory constraints are present, maintaining the full dataset is infeasible. Fix-HAPS compresses the latent space by extracting exactly one semantic centroid ($K = 1$) per class. The similarity becomes the cosine distance to this single global prototype. This drastically reduces complexity but may under-represent classes with high visual variance.

Table 1. Coverage, Efficiency, and Coverage Gap for LAC, APS, RAPS, and their HAPS variants across different target α levels. Evaluated using the PE-L embedding source.

	$\alpha = \mathbf{0.01}$			$\alpha = \mathbf{0.05}$			$\alpha = \mathbf{0.10}$		
Method	Cov. (%)	Eff. ↓	CovGap ↓	Cov. (%)	Eff. ↓	CovGap ↓	Cov. (%)	Eff. ↓	CovGap ↓
LAC (Baseline)	99.01	29.57	**1.73**	94.72	4.42	**5.78**	89.76	2.03	**8.86**
w/ kNN-HAPS	**99.00**	**12.98**	1.75	**94.86**	**2.98**	5.88	89.79	**1.66**	9.43
w/ Fix-HAPS	98.97	14.40	1.77	94.81	3.25	5.92	**89.91**	1.77	9.52
w/ Var-HAPS	99.03	13.83	**1.73**	94.76	3.12	5.89	89.87	1.73	9.44
APS (Baseline)	**99.01**	112.03	**1.71**	94.76	25.74	**5.24**	89.70	11.90	**7.41**
w/ kNN-HAPS	98.96	11.98	1.78	94.69	3.67	6.13	89.86	2.21	9.55
w/ Fix-HAPS	98.98	12.91	1.79	94.84	2.92	5.95	**89.94**	1.76	9.03
w/ Var-HAPS	98.96	**11.76**	1.78	**94.88**	**2.89**	5.90	89.77	**1.73**	8.98
RAPS (Baseline)	98.91	33.84	1.81	94.64	5.32	**5.68**	89.51	2.49	**8.48**
w/ kNN-HAPS	**99.00**	**11.25**	**1.75**	**94.96**	**2.76**	5.79	**90.06**	1.77	8.88
w/ Fix-HAPS	98.98	13.31	1.78	94.93	2.94	5.92	89.79	**1.72**	9.21
w/ Var-HAPS	98.95	11.71	1.79	94.84	2.85	5.92	90.14	**1.72**	8.95

3. Var-HAPS (Adaptive Centroid Compression): Since semantic categories exhibit varying degrees of polysemy, and to balance memory efficiency with topological accuracy, Var-HAPS adaptively determines the number of centroids K_j for each class j using the Silhouette coefficient (Rousseeuw, 1987). K-Means then extracts the centroids $C_j = \{c_{j,1}, \ldots, c_{j,K_j}\}$. During online inference, the semantic similarity $Sim_{FM}(x, y)$ between a test image x and a candidate class y is defined as the maximum cosine similarity to any of its centroids, clipped to $[0, 1]$:

$$Sim_{FM}(x, y) = \max\left(0, \max_{k \in \{1..K_y\}} \frac{\Phi(x) \cdot c_{y,k}}{\|\Phi(x)\| \|c_{y,k}\|}\right) \quad (3)$$

4. Experimental Setup

4.1. Datasets and Architectures

We use ImageNet-1K, strictly partitioning the 50K-image validation set to prevent data leakage: 20K images for conformal calibration ($\mathcal{D}_{cal}$), 5K for hyperparameter tuning ($\mathcal{D}_{tune}$), and 25K for testing ($\mathcal{D}_{test}$). To evaluate robustness under covariate shift, we test on ImageNet-R (renditions) (Hendrycks et al., 2021a) and ImageNet-A (adversarial) (Hendrycks et al., 2021b). The predictive pipeline uses a pretrained ResNet50 as the base classifier. For the semantic embeddings, we use the frozen Perception Encoder (Bolya et al., 2025) `PE-Core-L-14-336` (referred to as PE-L).

4.2. Calibration and Evaluation Protocol

Prior to executing CP, we calibrate the base classifier's raw logits using temperature scaling (Guo et al., 2017). We calculate the performance of kNN-HAPS (using Qdrant HNSW with exploration factor 2000), Fix-HAPS ($K = 1$), and Var-HAPS (adaptive K). The mixing parameter $\beta \in \{0.0, 0.1, 0.25, 0.5, 0.75, 0.9, 1.0\}$ and RAPS penalty variables are strictly optimized on $\mathcal{D}_{tune}$. Sweeping β to the boundary values (0.0 and 1.0) acts as a fallback, ensuring HAPS only activates if it strictly ameliorates the probability-only or similarity-only baselines.

Performance is quantified across three metrics: Marginal Coverage, Efficiency (Average Set Size), and CovGap (Ding et al., 2023) to measure class-conditional coverage deviations. All reported metrics represent the median over 1000 bootstrap resamples of the test set.

5. Results

5.1. On ImageNet-1K

The experimental results detailed in Table 1 demonstrate that integrating the HAPS framework yields substantial improvements in predictive efficiency without compromising theoretical coverage guarantees. Across all evaluated confidence levels ($\alpha \in \{0.01, 0.05, 0.10\}$), both the baseline methods and their respective hybrid variants successfully maintain valid marginal coverage, consistently hitting the target $1 - \alpha$ thresholds. The primary advantage of the hybrid approach emerges in the drastic reduction of average prediction set sizes (Efficiency). This improvement is most pronounced at the highly conservative $\alpha = 0.01$ level. For instance, baseline APS struggles with an average set size of 112.03, which is aggressively reduced to under 13 across all HAPS variants. When comparing the score variants, while kNN-HAPS yields highly competitive set sizes, both Fix-HAPS (which aggressively compresses the space to exactly one centroid per class) and the adaptive Var-HAPS provide significantly greater time and memory efficiency

during online inference. Furthermore, Var-HAPS consistently achieves tighter sets than the more rigid Fix-HAPS compression, validating the use of adaptive multi-centroid clustering to capture intra-class variance. Crucially, these massive efficiency gains sustain only negligible fluctuations in the Coverage Gap, indicating that HAPS successfully suppresses overconfidence without degrading class-conditional coverage.

Due to space constraints, we present the remainder of the experimental results for RAPS.

To isolate the impact of multimodal FMs on the efficacy of our proposed framework, we substitute the frozen FM entirely, using representations extracted directly from the penultimate layer of the ResNet50 base classifier. Furthermore, we replace the PE-L model (`PE-Core-L-14-336`) with a smaller-scale PE-S model (`PE-Core-S-16-384`).

Table 2. Efficiency (↓) for RAPS and its HAPS variants across different embedding sources.

Method / Embedding Source	**Target Coverage** $(1-\alpha)$		
	0.99	**0.95**	**0.90**
RAPS (Baseline)	33.84	5.32	2.49
w/ kNN-HAPS			
ResNet-50	34.03	5.33	2.41
PE-S	21.56	4.41	2.34
PE-L	**11.25**	**2.76**	**1.77**
w/ Fix-HAPS			
ResNet-50	38.51	5.47	2.54
PE-S	25.08	4.89	2.32
PE-L	**13.31**	**2.94**	**1.72**
w/ Var-HAPS			
ResNet-50	35.91	5.45	2.41
PE-S	25.24	4.66	2.34
PE-L	**11.71**	**2.85**	**1.72**

Table 2 evaluates RAPS and its HAPS variants across these three distinct embedding sources. We report only predictive efficiency, as empirical marginal coverage and the conditional coverage gap (CovGap) remain largely consistent across the evaluated methods. The results empirically validate our core hypothesis regarding the necessity of representation disentanglement. When HAPS uses the ResNet50 embeddings, it yields negligible improvements and occasionally inflates the average set size compared to the baseline RAPS (e.g., expanding from 33.84 to 38.51 for Fix-HAPS at $1-\alpha = 0.99$). Because these supervised embeddings share the exact same optimization vulnerabilities as the softmax logits, they fail to provide a reliable, independent "geometric veto", confirming that task-specific latent spaces are insufficient for suppressing overconfidence.

Conversely, substituting the supervised representations with those from frozen FMs reveals a clear architectural scaling trend. Integrating the smaller PE-S model immediately yields moderate efficiency gains, successfully dropping the 99% target set size from the mid-30s down to the 21–25 range. Upgrading to the highly robust and larger-scale PE-L yields the tightest prediction sets across all evaluated variants (reaching as low as 11.25 for kNN-HAPS). This monotonic improvement demonstrates that the performance of the HAPS framework is directly bounded by the semantic richness of the external latent space: as the zero-shot geometric stability of the chosen FM scales, the resulting conformal prediction sets become significantly more efficient.

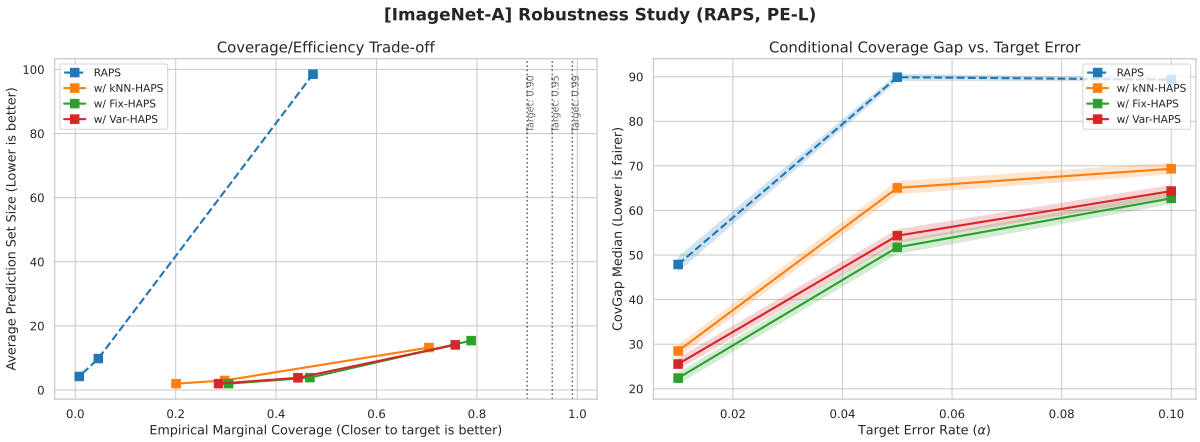
5.2. On ImageNet-A/R: Distribution Shift

To evaluate the robustness of our proposed framework under distribution shift, we extend our analysis to ImageNet-A (adversarial) and ImageNet-R (renditions). We benchmark the standard baseline RAPS against our HAPS variants using representations from the frozen multimodal FM (PE-L).

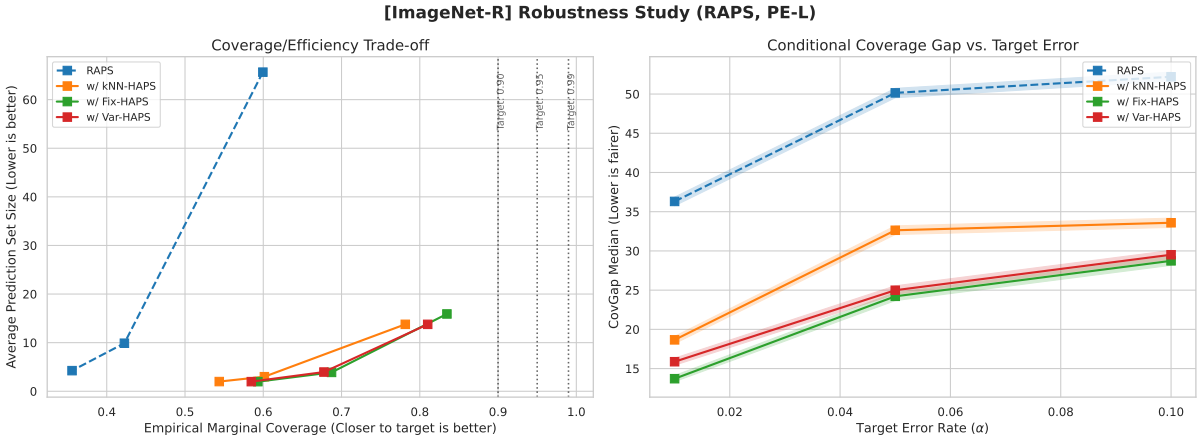
As illustrated in the figures for both ImageNet-A (1a) and ImageNet-R (1b), the fundamental exchangeability assumption of split conformal prediction is broken under these severe distribution shifts. Consequently, all evaluated methods ultimately fail to achieve the theoretical targeted confidence levels. As the target coverage levels $(1-\alpha \in \{0.90, 0.95, 0.99\})$ become more strict, the baseline RAPS algorithm produces large prediction sets (shooting vertically on the y-axis) while failing to approach the empirical target coverage (stalling near 0.5 on ImageNet-A and just below 0.6 on ImageNet-R).

In contrast, all HAPS variants exhibit significantly greater robustness. While the theoretical guarantee of marginal coverage remains voided by the distribution shift, leveraging the open-world semantic stability of the FM pushes the empirical coverage much closer to the intended theoretical targets (reaching approximately 80% coverage on ImageNet-A and nearly 85% on ImageNet-R). This closer validity is achieved while maintaining remarkably efficient average prediction set sizes (remaining well below 20). Furthermore, this robustness extends to the Conditional Coverage Gap (CovGap) where the baseline method suffers from extreme coverage disparities across classes under severe shift. The PE-L based HAPS variants, however, consistently compress this gap across all target error rates (α).

This confirms our core hypothesis: by anchoring task-specific probabilities with the zero-shot geometric stability of an FM, HAPS is more robust to distribution shift whereas standard supervised classifiers completely break down.



(a) Robustness evaluation on ImageNet-A (Adversarial).



(b) Robustness evaluation on ImageNet-R (Rendition).

Figure 1. Coverage vs. Efficiency curves and Conditional Coverage Gap under severe distribution shift for RAPS.

6. Conclusion

In this work, we presented Hybrid Adaptive Prediction Sets (HAPS), a robust non-conformity framework designed to mitigate the extreme overconfidence inherent in standard probability-based conformal predictors. By mathematically fusing the softmax outputs of a supervised classifier with the semantic similarities of a frozen FM, HAPS acts as a geometric veto that reliably rejects implausible predictions.

Our empirical evaluations on ImageNet-1K demonstrate that HAPS seamlessly integrates with established CP algorithms (LAC, APS, and RAPS) to yield substantial improvements in predictive efficiency, drastically reducing average set sizes without degrading marginal or class-conditional coverage. Ablation studies across different embedding spaces confirmed that these efficiency gains are intrinsically linked to the zero-shot stability of the external FM. Relying on the supervised classifier's own latent space fails to suppress overconfidence, whereas scaling to larger, robust FMs monotonically improves the tightness of the prediction sets.

Crucially, under severe distribution shifts (ImageNet-A and ImageNet-R), where the exchangeability assumption is violated and standard baselines collapse into invalid large prediction sets, HAPS exhibits exceptional robustness. It maintains remarkably efficient set sizes and drives empirical coverage much closer to theoretical targets. Ultimately, by anchoring supervised outputs to the untampered semantic geometry of FMs, HAPS provides a highly effective and scalable methodology for deploying reliable uncertainty quantification in dynamic, open-world environments.

While HAPS successfully leverages the zero-shot geometric robustness of FMs to achieve strong *empirical* resilience under severe distribution shifts (e.g., ImageNet-A/R), it remains constrained by the standard split conformal prediction assumption of data exchangeability. Because formal coverage guarantees are lost once this assumption is broken, a critical avenue for future work involves extending HAPS

to non-exchangeable settings. Specifically, integrating the framework with methodologies explicitly designed for distribution shifts (Tibshirani et al., 2019; Barber et al., 2023) would provide a rigorous pathway to bridge the gap between empirical performance and strict theoretical validity.

Acknowledgement

This work has received funding from the European Union, via the oc4-2025-TES-02 issued and implemented by the ENFIELD project, under the grant agreement No 101120657.

References

Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Bolya, D., Huang, P.-Y., Sun, P., Cho, J. H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

Ding, T., Angelopoulos, A., Bates, S., Jordan, M., and Tibshirani, R. J. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Fillioux, L., Silva-Rodriguez, J., Ayed, I. B., Cournede, P.-H., Vakalopoulou, M., Christodoulidis, S., and Dolz, J. Are foundation models for computer vision good conformal predictors? *arXiv preprint arXiv:2412.06082*, 2024.

Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, pp. 3329–3339. PMLR, 2021.

Ghosh, S., Belkhouja, T., Yan, Y., and Doppa, J. R. Improving uncertainty quantification of deep classifiers via neighborhood conformal prediction: Novel algorithm and theoretical analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7722–7730, 2023.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., 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*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Malkov, Y. A. and Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *European conference on machine learning*, pp. 345–356. Springer, 2002.

Papernot, N. and McDaniel, P. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Romano, Y., Sesia, M., and Candes, E. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Silva-Rodríguez, J., Ben Ayed, I., and Dolz, J. Conformal prediction for zero-shot models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19931–19941, 2025.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer, 2005.