Conformal Prediction for Zero-Shot Models

Supplementary Material

A. Extended related works

In the following, we provide extended remarks about prior literature on conformal prediction. Concretely, we delve into its historical transductive nature and two different families of methods for further improving the conformal inference in vision: conformal training and methods that assume access to additional data splits.

Transduction in conformal prediction. Initially, this reliability framework was transductive [1, 48, 53]. The original setting, usually called *full conformal prediction*, assumes access to *i.i.d.* labeled training and unlabeled test samples for producing the conformal sets. These methods require evaluating all the label space, $\mathcal{Y} = \{1, 2, ..., K\}$, to fit label-specific models, including all values of y that are sufficiently consistent for a given significance level. Such computational overload prohibits its use in modern deep-learning models. In contrast, *split conformal prediction* [46, 66] assumes a more resource-efficient, practical setting. Given a trained, black-box model that outputs logit predictions, it assumes access to a fresh labeled calibration set exchangeable [66] to testing data. Even though it was initially known as *inductive conformal prediction* [46, 64], split conformal prediction does not necessarily preclude transduction mechanisms [19]. Our transductive approach differs from these initial conceptions on full conformal prediction but exploits transduction while adapting a black-box model following the split conformal prediction setting.

Conformal training in vision. A family of works, known as the *conformal training*, focuses on preparing the base model to enforce small predictive sets [9, 59] or improving sample adaptiveness by regularizing learning to produce homogenous non-conformity scores [15]. Conformal training methods [9, 15, 59] are not applicable in the era of foundation models, where models are trained on limited occasions using general learning objectives and accessed for efficient adaptation. Thus, they do not apply to the explored black-box setting and fall out of the scope of this paper.

Improving conformal sets with additional data splits. Within the black-box split conformal prediction, recent literature explores novel aspects of well-known non-conformity scores, such as APS or RAPS. For example, [59] explores the impact of temperature scaling, and [13] studies its behavior on tasks with many classes and, more particularly, the class-conditional coverage these scores offer. Also, the authors propose different strategies for improving these approaches. First, Conf-OT [59] proposes to train the optimum temperature scaling by minimizing the efficiency gap in a costume loss function. Also, Clustered Conformal Prediction [13] integrates a clustering step to find subgroups of classes with similar quantized behavior for a given non-conformity score using K-means. The groups are then employed to perform conformal prediction at the cluster level. Such measures have shown improvement over the base adaptive methods, i.e., APS and RAPS. However, this comes at the cost of incorporating additional data splits to adjust the proposed methods without potentially breaking the exchangeability to test data. In this work, we stick to the standard experimental setting by accessing one calibration set uniquely. We argue that this setting is more realistic, especially in critical scenarios such as detecting rare, low-prevalence diseases [17, 30, 54].

B. Non-conformity scores

The following formally introduces the non-conformity scores employed in this paper. The measures are designed so that smaller conformal scores correspond to larger model confidence levels. These are employed to search its quantile that satisfies the 1-α coverage in the calibration set, following Eq. (3), and producing conformal sets in test samples as detailed in Eq. (4) in the main manuscript.

Least Ambiguos Classifier. The intuitive idea of LAC [39] is constructing sets by thresholding the output probabilities with a confidence level. Thus, the non-conformity score can be constructed as:

$$\mathcal{S}_{\text{LAC}}(\mathbf{x}, y) = 1 - x_{k=y}. \tag{11}$$

LAC produces the minimum set sizes in case the input probabilities are correct. However, it lacks adaptiveness, e.g., in under-represented categories.

Adaptive Prediction Sets. Aiming to improve adaptiveness, APS [51] construct the confidence sets based on accumulating ordered class confidences, such that:

$$\mathcal{S}_{\text{APS}}(\mathbf{x}, y) = \rho_x(y) + x_{k=y} \cdot u, \tag{12}$$

where $\rho_x(y)$ is the total probability mass of the set of labels more likely than the input label y, i.e., $\rho_x(y) = \sum_{k' \in \mathcal{Y}'(\mathbf{x},y)} x_{k=k'}$, with $\mathcal{Y}'(\mathbf{x}, y) = \{k | x_k > x_{k=y}\}$. Also, note that $u \in [0, 1]$ is a random variable to break ties and archive exact 1-α marginal coverage.

Regularized Adaptive Prediction Sets. Even though APS produces better coverage than LAC across different data subgroups, it comes at the cost of producing large set sizes. To tackle this issue, RAPS [2] incorporates penalties when adding categories into the accumulative confidence procedure, thus taming the score distribution tail.

Dataset	Classes	Splits			b/u	Task description	Text templates
		Train	Val	Test			
ImageNet [11]	1,000	1.28M	-	50,000	b	Natural objects recognition.	"Itap of a [CLS].","A bad photo of [CLS].", "A origami of [CLS].","A photo of the large [CLS].", "A [CLS] on a video game.","Art of the [CLS].", "A photo of the small [CLS].","A photo of a [CLS]."
ImageNet-A [23]	200	-	-	7,500	u	Natural objects recognition.	
ImageNet-V2 [50]	1,000	-	-	10,000	b	Natural objects recognition.	
ImageNet-R [24]	200	-	-	30,000	u	Natural objects recognition.	
ImageNet-Sketch [68]	1,000	-	-	50,889	b	Sketch-style images.	
SUN397 [70]	397	15,880	3,970	19,850	b	Scenes classification.	"A photo of a [CLS]."
FGVCAircraft [37]	100	3,334	3,333	3,333	u	Aircraft classification.	"A photo of [CLS], a type of aircraft."
EuroSAT [22]	10	13,500	5,400	8,100	u	Satellite image classification.	"A centered satellite photo of [CLS]."
StanfordCars [27]	196	6,509	1,635	8,041	u	Cars classification.	"A photo of a [CLS]."
Food101 [4]	101	50,500	20,200	30,300	b	Foods classification.	"A photo of a [CLS], a type of food."
OxfordPets [47]	37	2,944	736	3,669	u	Pets classification.	"A photo of a [CLS], a type of a pet."
Flowers102 [43]	102	4,093	1,633	2,463	u	Flowers classification.	"A photo of a [CLS], a type of flower."
Caltech101 [16]	100	4,128	1,649	2,465	u	Natural objects classification.	"A photo of a [CLS]."
DTD [8]	47	2,820	1,128	1,692	b	Textures classification.	"[CLS] texture."
UCF101 [58]	101	7,639	1,898	3,783	u	Action recognition.	"A photo of a person doing [CLS]."

Table 4. **Datasets overview**. Detailed description of the 15 datasets used to evaluate the conformal inference of zero-shot vision-language models. Also, the handcrafted textual templates for setting the zero-shot text-driven classifier for each dataset are indicated. These are the same ones used in relevant prior literature on this topic [56, 75]. "b"/"u" denotes balanced or unbalanced test partitions, respectively.

RAPS can be formally introduced as:

$$\mathcal{S}_{\text{RAPS}}(\mathbf{x}, y) = \mathcal{S}_{\text{APS}}(\mathbf{x}, y) + \lambda \cdot (o(\mathbf{x}, y) - k_{\text{reg}})^{+}, \quad (13)$$

where $\lambda, k_{\text{reg}} \geq 0$ are hyper-parameters that control the strength of the constraint, $o_x(y)$ is the rank of the label y in the sorted predictions, $o(\mathbf{x}, y) = |\mathcal{Y}'(\mathbf{x}, y)| + 1$, and $(\cdot)^{+}$ represented the ReLU function. Regarding the hyper-parameters, λ is the additional confidence cost of incorporating each new category to the set, and k_{reg} is the category index in which the penalty starts.

C. Additional datasets details

Datasets. As stated in the main manuscript, we perform a large-scale benchmark on the conformal prediction of CLIP models across typical datasets employed for transferability [18, 49, 75]. Concretely, we employ 15 datasets, which compile thousands of general concepts, fine-graned categories, textures, and actions. Also, it is worth noting that these benchmarks include 9/15 unbalanced tasks, which are commonly more challenging to address in vision classification literature. A summary of the employed datasets and task descriptions is depicted in Tab. 4. Regarding the dataset splits, we employed the ones proposed in seminal works for few-shot adaptation of CLIP [18, 75].

CLIP's text templates. For creating text prototypes for zero-shot and transfer learning using CLIP, the target category names ("`[CLS]`.") are forwarded through the text encoder. These category names are usually combined with hand-crafted text templates, which provide additional context on the task at hand, e.g., "`A centered satellite photo of [CLS]`." We followed prior works in this aspect [56, 75], using common text templates for each task, which are depicted in Tab. 4.

D. Evaluation of conformal inference

The evaluation metrics indicated in the experimental section are formally introduced in this section.

Top-1 accuracy. To evaluate the discriminative performance of CLIP models, we employ accuracy, a widely extended metric in few-shot literature [18], and conformal prediction in vision [2].

The following details the metrics employed to evaluate the conformal inference methods. For that purpose, let us assume an arbitrary data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{I}$, an error rate of coverage level α, and the function creating the conformal sets from a non-conformity score $C(\mathbf{x})$, which operates as in Eq. (4), after finding the non-conformity score threshold from a calibration subset. We refer the reader to Sec. 3.2 in the main manuscript for other definitions.

Coverage. The empirical coverage on the test domain is employed to measure the degree of satisfaction of the marginal coverage guarantees:

$$\text{Cov}(\mathcal{D}) = \frac{1}{I} \sum_{i \in \mathcal{D}} \delta[(y_i \subset C(\mathbf{x}_i)], \quad (14)$$

where δ denotes a delta function, that is, 1 if its argument is true, and 0 otherwise.

Set size. The average set size, also known as inefficiency [59], is a widely employed metric [2, 13, 69] for assessing the utility of conformal prediction methods for multi-class classification problems. An optimum conformal method in terms of efficiency should provide a lower set size:

$$\text{Size}(\mathcal{D}) = \frac{1}{I} \sum_{i \in \mathcal{D}} |C(\mathbf{x}_i)|. \quad (15)$$

Class-conditional coverage violation (CCV). A recently proposed metric in [13] to evaluate the adaptiveness of a

conformal prediction method based on measuring the empirical coverage gap observed in each category in the target task. An optimal conformal method in terms of adaptiveness is expected to provide a small average gap:

$$\text{CCV}(\mathcal{D}) = 100 \times \frac{1}{|\mathcal{Y}|} \sum_{k \in \mathcal{Y}} \big|\text{Cov}(\mathcal{D}_k) - (1-\alpha)\big|, \quad (16)$$

where $|\cdot|$ over the scalar in the summation term represents the absolute value, and $\mathcal{D}_k$ indicates the subset of samples labeled as the category k, such that $\mathcal{D}_k = \{(\mathbf{x}_i, y_i)\}_{i \in \mathcal{B}_k}$, with $\mathcal{B}_k = \{i \mid y_i = k\}$. Note that the metric is multiplied by 100 to provide a percentage scale.

E. Details on inductive adaptation

Fig. 1(b) refers to an experiment using the calibration set for adapting the zero-shot logits to the new tasks, so-called Adapt+SCP. Concretely, we train new class prototypes on the logit space, $\mathbf{W} \in \mathbb{R}^{K \times K}$. These weights are initialized in the simplex corners and are l2-normalized during training such that $\mathbf{w}_k = \frac{\mathbf{w}_k}{||\mathbf{w}_k||_2}$. Given the labeled calibration set, $\mathcal{D}_{\text{cal}} = \{(\boldsymbol{l}_i, y_i)\}_{i=1}^{N}$, we first l2-normalize the received logit representations, $\boldsymbol{l} = \frac{\boldsymbol{l}}{||\boldsymbol{l}||_2}$. Then, the new scores based on the learned class prototypes are obtained as follows:

$$l'_k = -\frac{\tau^{\text{LP}}}{2} ||\boldsymbol{l} - \mathbf{w}_k||, \quad (17)$$

where τ^{LP} is a temperature scaling parameter, which is searched greedily based on calibration data for each dataset, and $\boldsymbol{l}' = (l'_k)_{1 \leq k \leq K}$ are the new logits for a given sample. Finally, for a given sample, the new output probabilities are the softmax of the new logits: $\mathbf{p} = \sigma_k(\boldsymbol{l}')$.

The inductive adaptation consists of learning the new class weights based on a few shots available on calibration data by minimizing cross-entropy:

$$\min_{\mathbf{W}} \; -\frac{1}{NK} \sum_{i=1}^{I} \sum_{k=1}^{K} y_{ik} \log p_{ik}, \quad (18)$$

where $\{\{y_{ik}\}_{i=1}^{I}\}_{k=1}^{K}$ are the one-hot-encoded labels. Training is performed via mini-batch gradient descent, using large batches of $2,048$ samples and ADAM as an optimizer, with a learning rate of 0.1, during 500 iterations, using a cosine decay scheduler for the learning rate.

F. Transductive adaptation baselines

As stated in the main manuscript, there is no clear candidate to include as a baseline for the proposed setting: training-free, transductive adaptation of VLMs using black-box logit predictions. Hence, we adapted two transductive training approaches for the task: a general transductive formulation based on mutual information, i.e., TIM [5], and the recently proposed TransCLIP [73], a GMM-based method specially designed for zero-shot VLMs.

Transductive information maximization [5]. TIM is a general framework based on mutual information for adjusting a set of class prototypes on unlabeled data. Formally, the employed input logits, class weights, and new probabilities are obtained as in the inductive setting detailed in Section Appendix E. More concretely, softmax probabilities are obtained following Eq. (17). In contrast, TIM operates unsupervised and targets an entropy minimization loss with the regularized label-marginal distribution. We follow the SGD-based version proposed by the authors and modify the Shannon entropy maximization term by a KL divergence, which allows us to employ the observed label-marginal on the calibration set, i.e., $\mathbf{m}$. Two versions are proposed, with different regularizations for the predicted label-marginal distribution, $\hat{\mathbf{m}} = \frac{1}{N+M} \sum_{i=1}^{N+M} \mathbf{p}_i$. These two versions are:

$\text{TIM}_{\text{KL}(\hat{\mathbf{m}}||\mathbf{u}_K)}$: the base version, closer to the original work in [5] employs a uniform target label-marginal distribution, such that:

$$\min_{\mathbf{W}} \; \frac{\alpha}{N} \sum_{i=1}^{I} \sum_{k=1}^{K} p_{ik} \log p_{ik} + \sum_{k=1}^{K} \mathbf{u}_K \log \frac{\mathbf{u}_K}{\hat{\mathbf{m}}}. \quad (19)$$

$\text{TIM}_{\text{KL}(\hat{\mathbf{m}}||\mathbf{m})}$: the modified version, leverages the marginal distributions observed in the calibration set, whose annotated labels are available:

$$\min_{\mathbf{W}} \; \frac{\alpha}{N} \sum_{i=1}^{I} \sum_{k=1}^{K} p_{ik} \log p_{ik} + \sum_{k=1}^{K} \mathbf{m} \log \frac{\mathbf{m}}{\hat{\mathbf{m}}}. \quad (20)$$

The training is performed by gradient descent, using large batches of $2,048$ samples during 100 iterations, ADAM as an optimizer, and a base learning rate of 0.001. The later hyper-parameters follow the advice provided in SGD-TIM [5]. Finally, it is worth mentioning that we found TIM highly sensitive to the choice of τ^{LP} in Eq. (17). To alleviate this issue, τ^{LP} is searched using a grid of $\tau^{\text{LP}} \in \{0.1, 1, 5, 10, 15, 30, 60, 100\}$ per dataset. We employed the accuracy on the calibration set after transductive adaptation as the maximization objective for such a search. Note that these labels are available in the split conformal inference scenario, and we did not observe any empirical degradation in marginal coverage. Also, α is fixed to $\alpha = 0.1$.

TransCLIP [73]. We perform minor modifications over the base zero-shot version proposed by the authors. First, we set the zero-shot CLIP prototypes as the corners of the logit simplex. Second, we find the method relatively sensitive to the text-guided KL divergence penalty λ, which we increased to $\lambda = 10$ upon a greedy search to avoid performance degradation. In terms of performance, our results obtained with this baseline are close to the figures reported by

the authors in TransCLIP. Concretely, for ViT-B/16, we obtained an average accuracy of 65.1 for 15 datasets. For the same tasks, the authors report 68.1 accuracy. This performance gap is explained by the difference in the input given to TransCLIP since logit scores can be seen as using projections of the original features, with the consequent information loss, especially for tasks with small categories, such as EuroSAT. Concretely, this dataset mainly contributes to such an average accuracy gap (-16.8).

As a final remark, we would want to highlight the presence of key hyper-parameters on the explored baselines. The performance for each specific task might be sensitive to the choice of these. In contrast, *Conf-OT does not introduce any critical hyper-parameter that could degrade the transfer learning performance.*

G. Additional results

This section introduces additional results and specific metrics that showcase Conf-OT's effectiveness compared to the prior art and validate its key elements and robustness across various scenarios.

G.1. Additional CLIP backbones

In Tab. 7, we report the performance of Conf-OT atop additional CLIP, i.e., ResNet-101, ViT-B/32, and ViT-L/14, and MetaCLIP, i.e., ViT-B/16, and ViT-H/14, backbones. The results are consistent with the main findings in Sec. 5.2 (Tab. 1) and indicate the generalization of the proposed setting across several black-box outputs. Concretely, efficiency improvements on set size are consistently maintained on CLIP ResNet-101 and CLIP ViT-B/32, with such a figure improved by nearly 20%. Also, Conf-OT is effective on larger backbones such as CLIP ViT-L/14 or MetaCLIP ViT-H/14, whose initial performance is also notably better. In this case, relative performance improvements of nearly 15% are observed.

G.2. Results per dataset

Performances are detailed for each backbone individually. Also, we split the five ImageNet shifts and 10 fine-grained tasks into different Tables to improve readability. First, Tab. 9 and 10 introduce the results for CLIP ResNet-50, and Tab. 11 and 12 do the same for CLIP ResNet-101. Tab. 13, 14, and 15, 16 introduce the results for both CLIP ViT-B backbones: ViT-B/32 and ViT-B/16, respectively. Finally, Tab. 17 and 18 present the results for ViT-L/14. Complementary, Tab. 19, 20, 21, and 22 depict the detailed results for MetaCLIP backbones across all datasets.

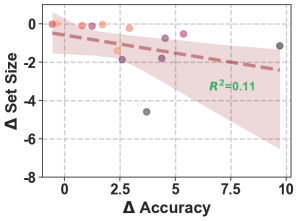
G.3. Additional results for transductive baselines

Tab. 8 introduces the performance obtained using TIM [5] and TransCLIP [73] baselines for additional datasets, i.e., CLIP ResNet-50, and more demanding error rates, i.e., $\alpha = 0.05$. These results complement the observations in Sec. 5.2 (Tab. 8). They showcase that Conf-OT consistently performs better across various scenarios, especially under demanding premises of low error rates ($\alpha = 0.05$). Regarding $\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}||\mathbf{u}_K)}$, we recall that this option provided better set efficiency compared to Conf-OT using APS in the narrower scenario introduced in Tab. 8. Nevertheless, these supplementary results show this trend is not generalizable across additional backbones and coverage levels. For example, for ResNet-50 in Tab. 8, $\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}||\mathbf{u}_K)}$ combined with APS produces set sizes nearly 15% larger than Conf-OT, for both $\alpha = 0.10$, and $\alpha = 0.05$.

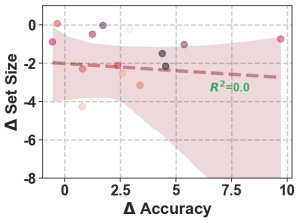
G.4. Further exploration on accuracy against set size improvement

In the main manuscript, we questioned whether the improvements provided by the proposed transductive approach were due solely to better discriminative performance (Sec. 5.3, Fig. 3). In this section, we provided additional observations on the discriminative and conformal inference improvements provided by Conf-OT.

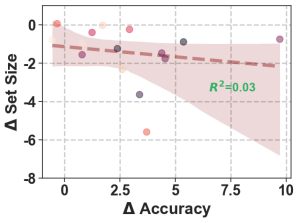
Relative improvements across datasets. We provide in Fig. 4 an extended version of Fig. 3(a). These visualizations indicate a limited correlation between the accuracy improvement and set size decreasing for all non-conformity scores employed, both adaptive and non-adaptive measures, i.e., LAC, APS, and RAPS.



(a) LAC [39]



(b) APS [51]

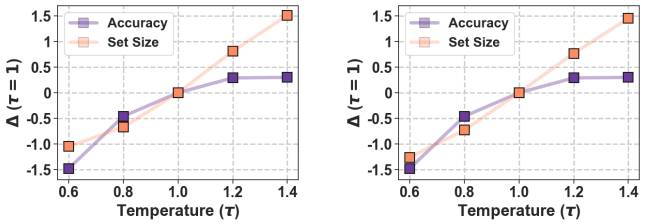


(c) RAPS [2]

Figure 4. **Correlation accross datasets of accuracy *vs.* set size change (Δ) increment using Conf-OT**, with popular non-conformal scores, i.e., LAC [39], APS [51], and RAPS [2]. Results using CLIP ViT-B/16 on 15 datasets with $\alpha = 0.10$. These results complement Fig. 3 in the main manuscript.

Effect of temperature scaling in adaptive non-conformity scores. In the main manuscript (Sec. 5.3),

we explored the positive effect of increasing confidence in prediction for adaptive methods using temperature scaling, as observed in [69]. We now recover this argument to explore how this affects accuracy, showcased in Fig. 5. Such results indicate that, if decreasing the temperature scaling value τ, the efficiency improvement comes at the cost of an accuracy detriment when using Conf-OT. This, again, suggests a disjoint behaviour between discriminative and conformal figures of merit.



(a) APS [51] (b) RAPS [2]

Figure 5. **Relation of accuracy against set size improvement on adaptive scores**, i.e., APS [51] (a) and RAPS [2] (b), resulting from modifying the distribution sharpness, via temperature scaling. Results using ViT-B/16 on 15 datasets with $\alpha = 0.10$. These results complement Fig. 3 in the main manuscript.

G.5. Conf-OT hyper-parameter studies

Fig. 6 presents the convergence of the Sinkhorn algorithm in Conf-OT regarding the number of iterations by measuring set size. These results demonstrate that such algorithms reach a satisfactory convergence after three iterations. This observation is consistent with typical values employed on prior works using Sinkhorn optimal transport in vision tasks [7]. Thus, we kept $T = 3$ across all experiments and configurations in our experiments.

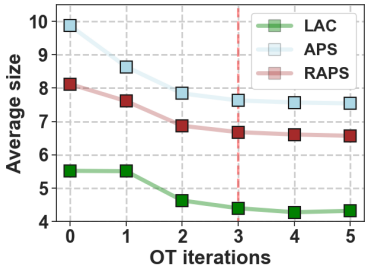


Figure 6. **Study on the number of iterations of Conf-OT**. Results using CLIP ViT-B/16 on 15 datasets. Dashed lines indicate the chosen value for this hyper-parameter, *i.e.* $T = 3$ repetitions.

G.6. Details on computational efficiency

The following section includes additional details on the hardware employed for our experiments and the computational efficiency of Conf-OT.

Hardware. All our experiments were conducted on a single GeForce RTX 3060. We extracted the vision features and class text-driven prototypes using GPU resources. However, we evaluated Conf-OT using solely CPU hardware. If available, Conf-OT requires less than 0.7 Gb of memory in a commodity GPU for the most demanding datasets, such as ImageNet. In this scenario, Sinkhorn optimal transport computation speed-up in a factor $\times 10$.

Runtime analysis in Conf-OT. Fig. 7 provides detailed runtimes for each of the stages in Conf-OT. Note that the baseline time refers to the feature extraction runtime when extracting feature embedding from the vision encoder (nearly 4 minutes). We recall that Conf-OT is equipped with three stages: *i*) transfer learning trough transductive optimal transport, *ii*) searching the $1 - \alpha$ quantile from a non-conformity score distribution in the calibration set, and *iii*) inference on the query samples, which consists of producing the output sets. On average, across 15 tasks, these three stages require less than 0.75 seconds for the full dataset. The results showcase that the latest step involves the main part of the runtime, which is also required if not following our transfer learning pipeline. On the other hand, threshold computing requires negligible computing times, and the optimal transport adaptation consumes nearly 1/3 of the whole runtime. These results showcase that Conf-OT is extremely efficient, compensating for the necessity of optimizing the optimal transport problem and finding the non-conformity score threshold for each query batch, e.g., if the inference is carried out in small testing batches of 8 images, the whole Conf-OT procedure should be computed for each batch. Also, it is worth mentioning that these processing times are orders of magnitude smaller than the feature extraction process. Also, *Conf-OT runtimes can be notably decreased using specialized hardware, such as GPUs, for the Sinkhorn algorithm*.

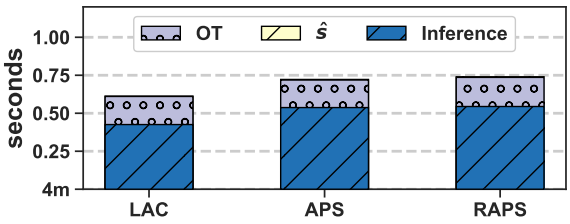


Figure 7. **Runtimes of Conf-OT**. The baseline (4 minutes) indicates feature extraction runtime. The required OT adaptation times are negligible ($\leq$ 10 ms) compared to the baseline and inference speed. "$\hat{s}$" indicates finding the conformity score from calibration data. Average results using CLIP ViT-B/16 on 15 datasets.

G.7. Details on data efficiency

The proposed Conf-OT is transductive. That means that access to both calibration and test data during inference is required to perform the transfer learning adaptation between source and target domains. In the following, we explore how sensitive Conf-OT is in two measures: *i*) calibration data requirements, and *ii*) size of the input query batch.

Method	Ratio		$\alpha = 0.10$		
	Calib - Test	Top-1↑	Cov.	Size↓	CCV↓
LAC	0.1 - 0.9	63.8	0.903	7.71	9.65
	0.2 - 0.8	63.8	0.899	5.56	9.80
	0.5 - 0.5	63.8	0.899	5.52	10.37
	0.8 - 0.2	63.8	0.899	5.56	11.70
Conf-OT+LAC	0.1 - 0.9	66.6	0.901	4.53	8.73
	0.2 - 0.8	66.7	0.899	4.39	8.86
	0.5 - 0.5	66.7	0.900	4.40	9.48
	0.8 - 0.2	66.7	0.899	4.41	11.12
APS	0.1 - 0.9	63.8	0.901	9.96	7.55
	0.2 - 0.8	63.8	0.900	9.94	7.65
	0.5 - 0.5	63.8	0.900	9.87	8.39
	0.8 - 0.2	63.8	0.900	9.88	10.32
Conf-OT+APS	0.1 - 0.9	66.6	0.902	7.7	6.44
	0.2 - 0.8	66.7	0.901	7.64	6.59
	0.5 - 0.5	66.7	0.899	7.64	7.44
	0.8 - 0.2	66.7	0.900	7.67	9.62
RAPS	0.1 - 0.9	63.8	0.900	8.12	7.67
	0.2 - 0.8	63.8	0.900	8.10	7.74
	0.5 - 0.5	63.8	0.900	8.12	8.50
	0.8 - 0.2	63.8	0.900	8.10	10.37
Conf-OT+RAPS	0.1 - 0.9	66.6	0.901	6.73	6.73
	0.2 - 0.8	66.7	0.900	6.70	6.64
	0.5 - 0.5	66.7	0.900	6.68	7.48
	0.8 - 0.2	66.7	0.899	6.70	9.69

Table 5. **Robustness to different data ratios**. Results using CLIP ViT-B/16 on 15 datasets averaged across 20 seeds.

Method	M		$\alpha = 0.10$		
		Top-1↑	Cov.	Size↓	CCV↓
LAC	-	63.8	0.899	5.52	10.37
w/ Conf-OT	Full	66.7	0.900	4.40	9.48
w/ Conf-OT	32	66.5	0.898	4.43	9.66
w/ Conf-OT	16	66.5	0.898	4.43	9.67
w/ Conf-OT	8	66.6	0.898	4.42	9.67
APS	-	63.8	0.900	9.87	8.39
w/ Conf-OT	Full	66.7	0.899	7.64	7.44
w/ Conf-OT	32	66.5	0.900	7.68	7.51
w/ Conf-OT	16	66.5	0.900	7.68	7.52
w/ Conf-OT	8	66.6	0.900	7.67	7.54
RAPS	-	63.8	0.900	8.12	8.50
w/ Conf-OT	Full	66.7	0.899	6.70	7.48
w/ Conf-OT	32	66.5	0.900	6.72	7.58
w/ Conf-OT	16	66.5	0.900	6.71	7.57
w/ Conf-OT	8	66.6	0.900	6.72	7.57

Table 6. **Robustness to small query batches**. Results using CLIP ViT-B/16 on 15 datasets averaged across 20 seeds. "M" indicates the size of the query batch for Conf-OT. Metrics are extracted by concatenating the predicted sets on the whole test subset.

Robustness to calibration data. Tab. 5 provides the results of base non-conformal scores over CLIP's zero-shot predictions and atop our Conf-OT method for different calibration/testing data ratios. Results demonstrate a constant performance regarding the efficiency improvements derived from Conf-OT across all these settings and non-conformity scores, even in challenging scenarios such as using solely 10% of data for calibration.

Robustness to small query sets. Tab. 6 contains the results when inference in Conf-OT is performed using extremely small mini-batches of images, e.g., 8, 16, or 32 images, sequentially. The figures of merit show consistent efficiency and class-conditional coverage improvements w.r.t. the base version of each non-conformal score, at the same level as the results observed when using the full batch (all testing samples) simultaneously, as in the main manuscript. As expected, the total runtimes for the whole dataset increase if small mini-batches are used since Conf-OT requires optimization for each batch.

	Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
			Cov.	Size↓	CCV↓	Cov.	Size↓	CCV↓
CLIP ResNet-101	LAC [39]	56.7	0.899	9.1	10.31	0.950	16.43	6.14
	w/ Conf-OT	**59.8**$_{+3.1}$	0.900	**7.22**$_{-1.9}$	**9.48**$_{-0.8}$	0.950	**12.99**$_{-3.4}$	**5.75**$_{-0.4}$
	APS [51]	56.7	0.900	14.97	8.67	0.950	24.76	5.55
	w/ Conf-OT	**59.8**$_{+3.1}$	0.900	**11.4**$_{-3.6}$	**7.74**$_{-0.9}$	0.950	**18.69**$_{-6.1}$	**5.15**$_{-0.4}$
	RAPS [2]	56.7	0.901	12.12	8.79	0.950	19.12	5.60
	w/ Conf-OT	**59.8**$_{+3.1}$	0.900	**9.93**$_{-2.2}$	**7.82**$_{-1.0}$	0.950	**15.27**$_{-3.9}$	**5.23**$_{-0.4}$
CLIP ViT-B/32	LAC [39]	58.7	0.900	8.20	10.40	0.950	15.21	6.12
	w/ Conf-OT	**61.3**$_{+2.6}$	0.899	**6.63**$_{-1.6}$	**9.39**$_{-1.0}$	0.950	**12.05**$_{-3.2}$	**5.76**$_{-0.4}$
	APS [51]	58.7	0.901	13.55	8.59	0.950	22.76	5.54
	w/ Conf-OT	**61.3**$_{+2.6}$	0.900	**10.69**$_{-2.9}$	**7.59**$_{-1.0}$	0.950	**17.49**$_{-5.3}$	**5.05**$_{-0.5}$
	RAPS [2]	58.7	0.901	11.00	8.70	0.950	18.09	5.61
	w/ Conf-OT	**61.3**$_{+2.6}$	0.899	**9.25**$_{-1.75}$	**7.66**$_{-1.0}$	0.950	**14.10**$_{-4.0}$	**5.13**$_{-0.5}$
MetaCLIP ViT-B/16	LAC [39]	69.8	0.900	3.79	10.04	0.950	7.08	5.92
	w/ Conf-OT	**71.8**$_{+2.0}$	0.900	**3.06**$_{-0.7}$	**9.47**$_{-0.6}$	0.950	**5.57**$_{-1.5}$	**5.63**$_{-0.3}$
	APS [51]	69.8	0.900	7.28	7.76	0.949	12.4	5.13
	w/ Conf-OT	**71.8**$_{+2.0}$	0.900	**5.58**$_{-1.7}$	**7.12**$_{-0.6}$	0.950	**9.80**$_{-2.6}$	**4.85**$_{-0.3}$
	RAPS [2]	69.8	0.900	6.03	7.83	0.950	9.21	5.16
	w/ Conf-OT	**71.8**$_{+2.0}$	0.900	**5.11**$_{-0.9}$	**7.13**$_{-0.7}$	0.950	**7.71**$_{-1.5}$	**4.90**$_{-0.3}$
CLIP ViT-L/14	LAC [39]	72.6	0.900	2.93	10.34	0.950	5.36	6.03
	w/ Conf-OT	**74.7**$_{+2.1}$	0.900	**2.5**$_{-0.4}$	**9.97**$_{-0.4}$	0.949	**4.49**$_{-0.9}$	**5.93**$_{-0.1}$
	APS [51]	72.6	0.900	6.40	7.80	0.949	11.27	5.21
	w/ Conf-OT	**74.7**$_{+2.1}$	0.900	**4.87**$_{-1.5}$	**6.98**$_{-0.8}$	0.949	**8.39**$_{-2.9}$	**4.84**$_{-0.4}$
	RAPS [2]	72.6	0.900	4.93	7.87	0.949	7.31	5.27
	w/ Conf-OT	**74.7**$_{+2.1}$	0.900	**4.13**$_{-0.8}$	**7.01**$_{-0.9}$	0.950	**6.22**$_{-1.1}$	**4.89**$_{-0.4}$
MetaCLIP ViT-H/14	LAC [39]	79.4	0.900	1.79	10.91	0.950	2.98	6.17
	w/ Conf-OT	**81.1**$_{+1.7}$	0.900	**1.59**$_{-0.2}$	**10.07**$_{-0.8}$	0.950	**2.53**$_{-0.4}$	**5.92**$_{-0.3}$
	APS [51]	79.4	0.900	4.45	7.48	0.950	7.70	4.97
	w/ Conf-OT	**81.1**$_{+1.7}$	0.899	**3.47**$_{-1.0}$	**6.60**$_{-0.9}$	0.949	**6.00**$_{-1.7}$	**4.55**$_{-0.4}$
	RAPS [2]	79.4	0.900	3.45	7.53	0.950	4.90	5.01
	w/ Conf-OT	**81.1**$_{+1.7}$	0.899	**2.93**$_{-0.5}$	**6.61**$_{-0.9}$	0.950	**4.30**$_{-0.6}$	**4.57**$_{-0.4}$

Table 7. **Performance for additional CLIP and MetaCLIP backbones**. Conf-OT performance above popular conformal inference methods such as LAC [39], APS [51], and RAPS [2]. Average performance across 15 datasets. We repeat each experiment 20 times. "↓" indicates smaller values are better. **Bold** numbers are superior results. These results complement Tab. 1 in the main manuscript.

	Method		$\alpha = 0.10$			$\alpha = 0.05$		
		Top-1↑	Cov.	Size↓	CCV↓	Cov.	Size↓	CCV↓
CLIP ResNet-50	LAC [39]	54.7	0.900	10.77	9.82	0.950	19.22	5.91
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [5]	55.7$_{+1.0}$	0.899	14.45$_{+3.7}$	9.69$_{-0.1}$	0.950	26.31$_{+7.1}$	5.70$_{-0.2}$
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [5]	56.4$_{+1.7}$	0.899	13.63$_{+2.9}$	12.02$_{+2.2}$	0.950	24.88$_{+5.7}$	6.03$_{+0.1}$
	TransCLIP [73]	55.7$_{+1.0}$	**0.861**	**7.83**$_{-2.9}$	12.20$_{+2.4}$	**0.881**	**8.81**$_{-10.4}$	11.03$_{+5.1}$
	Conf-OT	**57.3**$_{+2.6}$	0.900	8.61$_{-2.1}$	**9.15**$_{-0.6}$	0.951	15.53$_{-3.6}$	**5.61**$_{-0.3}$
	APS [51]	54.7	0.900	16.35	8.36	0.950	26.50	5.34
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [5]	55.7$_{+1.0}$	0.900	13.33$_{-3.0}$	8.64$_{+0.3}$	0.950	24.73$_{-1.8}$	5.37$_{+0.0}$
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [5]	56.4$_{+1.7}$	0.900	13.26$_{-3.1}$	8.69$_{+0.3}$	0.950	24.09$_{-2.4}$	5.47$_{+0.1}$
	TransCLIP [73]	55.7$_{+1.0}$	**0.873**	39.05$_{+22.7}$	11.15$_{+2.8}$	**0.899**	46.78$_{+20.3}$	9.41$_{+4.1}$
	Conf-OT	**57.3**$_{+2.6}$	0.900	**12.94**$_{-3.4}$	**7.64**$_{-0.7}$	0.950	**20.96**$_{-5.5}$	**5.03**$_{-0.3}$
	RAPS [2]	54.7	0.900	13.37	8.46	0.950	22.06	5.44
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [5]	55.7$_{+1.0}$	0.900	13.18$_{-0.2}$	8.64$_{+0.2}$	0.950	24.54$_{+2.5}$	5.38$_{-0.1}$
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [5]	56.4$_{+1.7}$	0.900	12.99$_{-0.4}$	8.71$_{+0.3}$	0.950	23.59$_{+1.5}$	5.51$_{+0.1}$
	TransCLIP [73]	55.7$_{+1.0}$	0.900	13.68$_{+0.3}$	9.94$_{+1.5}$	0.949	28.16$_{+6.1}$	6.03$_{+0.6}$
	Conf-OT	**57.3**$_{+2.6}$	0.900	**11.17**$_{-2.2}$	**7.72**$_{-0.7}$	0.950	**17.24**$_{-4.8}$	**5.19**$_{-0.2}$
CLIP ViT-B/16	LAC [39]	63.8	0.899	5.52	10.37	0.950	10.24	6.14
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [5]	64.7$_{+0.9}$	0.899	8.30$_{+2.8}$	10.41$_{+0.0}$	0.950	15.89$_{+5.7}$	6.03$_{-0.1}$
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [5]	65.0$_{+1.2}$	0.898	7.73$_{+2.2}$	10.89$_{+0.5}$	0.950	14.68$_{+4.4}$	6.40$_{+0.3}$
	TransCLIP [73]	65.1$_{+1.3}$	**0.892**	5.76$_{+0.2}$	11.02$_{+0.7}$	**0.921**	**7.02**$_{-3.2}$	8.31$_{+2.2}$
	Conf-OT	**66.7**$_{+2.9}$	0.900	**4.40**$_{-1.1}$	**9.48**$_{-0.8}$	0.949	7.99$_{-2.2}$	**5.80**$_{-0.3}$
	APS [51]	63.8	0.900	9.87	8.39	0.950	16.92	5.51
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [5]	64.7$_{+0.9}$	0.900	**7.24**$_{-2.6}$	9.32$_{+0.9}$	0.950	14.03$_{-2.9}$	5.88$_{+0.4}$
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [5]	65.0$_{+1.2}$	0.900	7.82$_{-2.1}$	9.38$_{+1.0}$	0.950	14.34$_{-2.6}$	6.03$_{+0.5}$
	TransCLIP [73]	65.1$_{+1.3}$	**0.892**	8.27$_{-1.6}$	11.50$_{+3.1}$	**0.931**	38.86$_{+21.9}$	7.47$_{+2.0}$
	Conf-OT	**66.7**$_{+2.9}$	0.899	7.64$_{-2.2}$	**7.44**$_{-0.9}$	0.949	**12.58**$_{-4.3}$	**5.09**$_{-0.4}$
	RAPS [2]	63.8	0.900	8.12	8.50	0.950	12.66	5.52
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [5]	64.7$_{+0.9}$	0.900	7.18$_{-0.9}$	9.32$_{+0.8}$	0.950	13.92$_{+1.3}$	5.99$_{+0.5}$
	$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [5]	65.0$_{+1.2}$	0.900	7.68$_{-0.4}$	9.42$_{+0.9}$	0.950	14.11$_{+1.5}$	6.04$_{+0.5}$
	TransCLIP [73]	65.1$_{+1.3}$	0.899	7.17$_{-1.0}$	10.20$_{+1.7}$	0.949	15.26$_{+2.6}$	6.31$_{+0.8}$
	Conf-OT	**66.7**$_{+2.9}$	0.900	**6.68**$_{-1.4}$	**7.48**$_{-1.0}$	0.949	**10.11**$_{-2.5}$	**5.16**$_{-0.3}$

Table 8. **Comparison of transductive baselines for additional CLIP backbones**. Average performance across 15 datasets. We repeat each experiment 20 times. "↓" indicates smaller values are better. **Bold** numbers are superior results. These results complement Tab. 2 in the main manuscript. **Grayish** marginal coverage ("Cov.") indicates an unsatisfied error rate, usually seen in TransCLIP.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
SUN397							
LAC [39]	58.7	0.899	4.28	8.90	0.950	8.42	5.44
w/ Conf-OT	62.0	0.899	3.62	7.64	0.951	6.77	4.83
APS [51]	58.7	0.901	8.98	7.59	0.949	16.16	5.01
w/ Conf-OT	62.0	0.900	7.05	6.60	0.950	12.20	4.46
RAPS [2]	58.7	0.902	8.16	7.59	0.949	13.63	5.06
w/ Conf-OT	62.0	0.900	6.57	6.62	0.951	10.70	4.53
FGVCAircraft							
LAC [39]	16.9	0.899	27.27	12.09	0.949	37.47	7.65
w/ Conf-OT	19.4	0.900	23.13	9.87	0.953	31.94	6.02
APS [51]	16.9	0.898	28.26	11.95	0.950	38.14	7.44
w/ Conf-OT	19.4	0.897	24.18	9.40	0.951	31.53	5.81
RAPS [2]	16.9	0.898	27.77	11.78	0.949	39.08	7.58
w/ Conf-OT	19.4	0.900	23.81	9.37	0.950	31.24	6.05
EuroSAT							
LAC [39]	36.1	0.900	4.74	8.75	0.952	5.84	4.43
w/ Conf-OT	43.4	0.899	4.03	7.97	0.951	5.38	4.24
APS [51]	36.1	0.901	5.03	6.87	0.952	5.97	3.54
w/ Conf-OT	43.4	0.901	4.61	6.26	0.952	5.76	3.59
RAPS [2]	36.1	0.900	5.02	6.95	0.953	5.98	3.57
w/ Conf-OT	43.4	0.900	4.59	6.29	0.951	5.74	3.61
StanfordCars							
LAC [39]	55.8	0.900	3.85	10.05	0.953	6.26	6.05
w/ Conf-OT	59.0	0.902	3.36	7.99	0.953	5.31	5.05
APS [51]	55.8	0.899	5.37	7.95	0.951	7.75	5.22
w/ Conf-OT	59.0	0.901	4.66	6.47	0.951	6.76	4.55
RAPS [2]	55.8	0.900	5.24	7.97	0.951	7.46	5.28
w/ Conf-OT	59.0	0.900	4.58	6.48	0.953	6.57	4.54
Food101							
LAC [39]	77.3	0.899	1.75	5.08	0.949	3.07	2.90
w/ Conf-OT	77.2	0.899	1.78	4.81	0.950	3.25	2.74
APS [51]	77.3	0.899	3.20	2.77	0.950	4.96	2.07
w/ Conf-OT	77.2	0.900	3.28	2.44	0.951	5.11	1.85
RAPS [2]	77.3	0.899	3.09	2.81	0.950	4.63	2.08
w/ Conf-OT	77.2	0.900	3.17	2.44	0.951	4.84	1.91
OxfordPets							
LAC [39]	85.7	0.901	1.11	8.00	0.950	1.38	4.37
w/ Conf-OT	86.3	0.897	1.09	6.73	0.949	1.35	3.73
APS [51]	85.7	0.904	1.56	3.86	0.952	1.94	2.71
w/ Conf-OT	86.3	0.905	1.58	4.11	0.951	1.95	2.82
RAPS [2]	85.7	0.904	1.55	3.87	0.952	1.92	2.76
w/ Conf-OT	86.3	0.905	1.57	4.09	0.951	1.93	2.82
Flowers102							
LAC [39]	66.2	0.902	7.47	15.16	0.952	16.06	8.07
w/ Conf-OT	67.2	0.904	5.23	15.09	0.952	9.94	8.32
APS [51]	66.2	0.903	8.55	14.40	0.951	17.31	7.79
w/ Conf-OT	67.2	0.901	6.07	13.02	0.951	10.59	7.87
RAPS [2]	66.2	0.903	8.23	14.75	0.951	17.98	8.06
w/ Conf-OT	67.2	0.901	5.91	13.25	0.952	10.63	8.21
Caltech101							
LAC [39]	85.7	0.903	1.12	11.89	0.950	1.38	7.91
w/ Conf-OT	87.4	0.899	1.06	12.35	0.950	1.35	8.26
APS [51]	85.7	0.891	3.12	9.43	0.943	4.56	6.87
w/ Conf-OT	87.4	0.891	1.92	8.63	0.946	2.66	6.44
RAPS [2]	85.7	0.892	2.91	9.42	0.944	4.03	7.00
w/ Conf-OT	87.4	0.892	1.88	8.54	0.945	2.52	6.41
DTD							
LAC [39]	42.9	0.897	11.95	10.44	0.948	18.66	6.40
w/ Conf-OT	45.5	0.898	9.29	8.53	0.948	14.56	5.59
APS [51]	42.9	0.900	13.20	9.48	0.950	19.53	6.08
w/ Conf-OT	45.5	0.902	10.19	8.40	0.950	15.17	5.50
RAPS [2]	42.9	0.899	12.91	9.61	0.951	19.34	6.32
w/ Conf-OT	45.5	0.901	10.09	8.44	0.951	15.15	5.59
UCF101							
LAC [39]	61.8	0.902	4.34	11.55	0.951	8.08	6.94
w/ Conf-OT	65.6	0.905	3.64	10.96	0.953	6.54	6.49
APS [51]	61.8	0.904	7.19	9.74	0.952	11.26	6.13
w/ Conf-OT	65.6	0.903	6.00	8.23	0.951	9.12	5.37
RAPS [2]	61.8	0.903	6.87	9.79	0.952	10.53	6.38
w/ Conf-OT	65.6	0.902	5.85	8.38	0.952	8.65	5.48

Table 9. Results on fine-grained tasks with CLIP ResNet-50.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
ImageNet							
LAC [39]	60.3	0.900	5.18	8.25	0.950	10.74	5.03
w/ Conf-OT	61.6	0.901	4.81	7.63	0.950	10.18	4.78
APS [51]	60.3	0.900	15.06	6.82	0.950	29.62	4.66
w/ Conf-OT	61.6	0.900	12.38	6.32	0.950	23.58	4.30
RAPS [2]	60.3	0.899	11.23	6.89	0.950	18.24	4.71
w/ Conf-OT	61.6	0.900	10.05	6.34	0.951	16.74	4.40
ImageNet-A							
LAC [39]	23.6	0.899	31.18	8.87	0.950	52.70	6.28
w/ Conf-OT	27.7	0.897	25.90	11.04	0.948	42.92	7.40
APS [51]	23.6	0.898	41.63	8.31	0.950	62.62	6.06
w/ Conf-OT	27.7	0.898	34.30	10.25	0.949	51.99	7.12
RAPS [2]	23.6	0.900	31.53	8.77	0.950	47.73	6.05
w/ Conf-OT	27.7	0.898	27.56	10.79	0.947	39.87	7.94
ImageNet-V2							
LAC [39]	53.5	0.900	9.21	12.60	0.951	20.06	7.84
w/ Conf-OT	54.7	0.903	8.55	12.53	0.952	20.08	7.73
APS [51]	53.5	0.902	25.22	12.39	0.951	47.96	7.84
w/ Conf-OT	54.7	0.900	19.49	12.19	0.951	37.71	7.78
RAPS [2]	53.5	0.901	17.35	12.42	0.953	27.55	7.75
w/ Conf-OT	54.7	0.902	15.34	12.18	0.952	25.17	7.75
ImageNet-R							
LAC [39]	60.1	0.901	7.38	5.77	0.950	15.31	3.41
w/ Conf-OT	62.9	0.901	6.25	5.71	0.951	13.41	3.65
APS [51]	60.1	0.901	12.44	4.36	0.950	20.51	2.92
w/ Conf-OT	62.9	0.901	11.05	4.60	0.951	18.52	3.04
RAPS [2]	60.1	0.899	11.26	4.50	0.950	17.87	3.06
w/ Conf-OT	62.9	0.902	10.15	4.71	0.950	16.19	3.33
ImageNet-Sketch							
LAC [39]	35.3	0.900	40.68	9.93	0.950	82.82	5.96
w/ Conf-OT	39.3	0.900	27.42	8.46	0.951	59.96	5.30
APS [51]	35.3	0.901	66.42	9.45	0.950	109.28	5.79
w/ Conf-OT	39.3	0.901	47.41	7.69	0.951	81.79	4.97
RAPS [2]	35.3	0.900	47.36	9.78	0.950	94.91	5.91
w/ Conf-OT	39.3	0.901	36.46	7.91	0.950	62.64	5.31

Table 10. Results on ImageNet shifts using CLIP ResNet-50.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
SUN397							
LAC [39]	58.8	0.901	4.43	9.22	0.951	8.80	5.57
w/ Conf-OT	62.2	0.900	3.46	7.55	0.951	6.67	4.95
APS [51]	58.8	0.901	9.93	8.04	0.950	18.80	5.14
w/ Conf-OT	62.2	0.900	7.04	6.56	0.950	12.39	4.44
RAPS [2]	58.8	0.901	8.88	8.09	0.950	15.01	5.22
w/ Conf-OT	62.2	0.900	6.55	6.55	0.951	10.92	4.48
FGVCAircraft							
LAC [39]	17.8	0.898	26.88	13.15	0.949	42.14	8.45
w/ Conf-OT	20.1	0.896	22.14	11.28	0.950	31.98	6.84
APS [51]	17.8	0.899	28.34	13.50	0.949	43.41	8.54
w/ Conf-OT	20.1	0.899	22.76	10.75	0.949	31.90	6.99
RAPS [2]	17.8	0.900	26.87	13.25	0.949	42.22	8.25
w/ Conf-OT	20.1	0.896	22.22	10.75	0.950	32.86	6.64
EuroSAT							
LAC [39]	32.9	0.899	6.99	10.24	0.950	7.71	5.46
w/ Conf-OT	38.4	0.899	5.55	6.82	0.951	6.60	3.62
APS [51]	32.9	0.900	6.93	10.03	0.949	7.84	5.69
w/ Conf-OT	38.4	0.899	5.55	6.96	0.951	6.69	3.97
RAPS [2]	32.9	0.899	6.92	10.15	0.949	7.89	5.74
w/ Conf-OT	38.4	0.900	5.55	7.02	0.951	6.71	3.98
StanfordCars							
LAC [39]	63.2	0.900	2.64	9.54	0.951	3.94	5.87
w/ Conf-OT	65.6	0.900	2.48	8.61	0.950	3.65	5.24
APS [51]	63.2	0.900	3.85	8.22	0.950	5.41	5.41
w/ Conf-OT	65.6	0.901	3.53	6.64	0.951	4.97	4.66
RAPS [2]	63.2	0.899	3.77	8.17	0.950	5.22	5.42
w/ Conf-OT	65.6	0.901	3.48	6.66	0.951	4.82	4.70
Food101							
LAC [39]	80.6	0.899	1.45	4.99	0.949	2.31	2.89
w/ Conf-OT	80.6	0.899	1.46	4.69	0.950	2.42	2.77
APS [51]	80.6	0.900	2.72	2.73	0.949	4.13	1.96
w/ Conf-OT	80.6	0.900	2.79	2.46	0.950	4.25	1.77
RAPS [2]	80.6	0.900	2.64	2.76	0.950	3.89	2.01
w/ Conf-OT	80.6	0.900	2.71	2.46	0.950	3.97	1.81
OxfordPets							
LAC [39]	86.9	0.900	1.09	9.32	0.949	1.42	5.50
w/ Conf-OT	87.5	0.899	1.06	7.18	0.952	1.27	3.73
APS [51]	86.9	0.907	1.51	4.42	0.950	1.81	2.98
w/ Conf-OT	87.5	0.905	1.48	3.73	0.952	1.80	2.97
RAPS [2]	86.9	0.907	1.51	4.51	0.950	1.80	2.99
w/ Conf-OT	87.5	0.905	1.47	3.76	0.952	1.79	2.93
Flowers102							
LAC [39]	64.5	0.901	5.55	15.33	0.950	15.16	8.18
w/ Conf-OT	69.8	0.903	4.00	15.86	0.951	7.85	8.95
APS [51]	64.5	0.902	7.31	13.64	0.951	15.01	8.00
w/ Conf-OT	69.8	0.904	5.19	12.92	0.952	8.23	8.21
RAPS [2]	64.5	0.902	7.09	13.84	0.951	15.85	8.10
w/ Conf-OT	69.8	0.906	5.08	13.13	0.951	8.28	8.47
Caltech101							
LAC [39]	89.9	0.895	0.99	13.08	0.948	1.16	8.21
w/ Conf-OT	91.9	0.897	0.94	13.39	0.950	1.11	8.52
APS [51]	89.9	0.895	2.71	9.01	0.946	4.13	6.80
w/ Conf-OT	91.9	0.891	1.56	8.68	0.946	2.10	6.49
RAPS [2]	89.9	0.895	2.52	9.01	0.946	3.61	6.76
w/ Conf-OT	91.9	0.893	1.53	8.74	0.946	2.01	6.48
DTD							
LAC [39]	36.9	0.900	13.76	10.50	0.953	18.31	5.81
w/ Conf-OT	44.4	0.899	11.29	9.94	0.952	15.79	5.54
APS [51]	36.9	0.900	15.05	9.42	0.953	19.16	5.66
w/ Conf-OT	44.4	0.898	12.07	8.99	0.953	16.73	5.14
RAPS [2]	36.9	0.902	14.64	9.75	0.951	18.57	5.63
w/ Conf-OT	44.4	0.901	11.93	9.14	0.953	16.48	5.15
UCF101							
LAC [39]	61.0	0.899	3.98	12.22	0.949	7.02	7.31
w/ Conf-OT	65.9	0.904	3.12	11.01	0.950	5.47	6.45
APS [51]	61.0	0.901	7.26	9.05	0.952	11.07	5.79
w/ Conf-OT	65.9	0.902	5.65	7.42	0.952	8.92	4.92
RAPS [2]	61.0	0.902	6.77	9.31	0.952	10.16	5.84
w/ Conf-OT	65.9	0.901	5.45	7.50	0.951	8.15	5.03

Table 11. Results on fine-grained tasks with CLIP ResNet-101.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
ImageNet							
LAC [39]	62.6	0.900	4.32	8.56	0.950	8.98	5.13
w/ Conf-OT	64.1	0.900	3.98	7.85	0.950	8.28	4.72
APS [51]	62.6	0.900	14.97	6.74	0.950	29.28	4.51
w/ Conf-OT	64.1	0.901	11.61	6.35	0.950	22.40	4.29
RAPS [2]	62.6	0.901	10.87	6.84	0.950	17.41	4.63
w/ Conf-OT	64.1	0.900	9.37	6.37	0.950	15.46	4.40
ImageNet-A							
LAC [39]	29.9	0.898	25.09	9.87	0.949	41.49	6.22
w/ Conf-OT	34.3	0.897	20.50	11.07	0.947	35.84	8.25
APS [51]	29.9	0.901	35.24	9.08	0.948	53.35	6.11
w/ Conf-OT	34.3	0.897	28.38	10.67	0.947	45.65	7.73
RAPS [2]	29.9	0.899	27.02	9.49	0.949	39.70	6.41
w/ Conf-OT	34.3	0.897	23.95	10.85	0.948	35.22	8.17
ImageNet-V2							
LAC [39]	56.3	0.903	7.74	12.63	0.951	18.72	7.97
w/ Conf-OT	57.0	0.901	7.04	12.53	0.951	17.55	7.85
APS [51]	56.3	0.901	24.30	12.52	0.951	48.33	7.89
w/ Conf-OT	57.0	0.901	18.66	12.34	0.949	36.51	7.93
RAPS [2]	56.3	0.902	16.85	12.49	0.951	26.83	7.93
w/ Conf-OT	57.0	0.901	14.48	12.43	0.949	24.24	7.97
ImageNet-R							
LAC [39]	67.8	0.899	4.14	5.86	0.950	9.19	3.50
w/ Conf-OT	69.8	0.900	3.65	5.89	0.950	8.48	3.61
APS [51]	67.8	0.900	9.10	4.23	0.949	15.24	2.99
w/ Conf-OT	69.8	0.899	8.48	4.21	0.950	14.04	2.91
RAPS [2]	67.8	0.901	8.18	4.31	0.949	12.77	3.06
w/ Conf-OT	69.8	0.900	7.70	4.27	0.950	12.09	3.08
ImageNet-Sketch							
LAC [39]	40.6	0.899	27.46	10.15	0.949	60.14	6.11
w/ Conf-OT	44.9	0.900	17.56	8.48	0.950	41.90	5.27
APS [51]	40.6	0.900	55.28	9.49	0.950	94.43	5.78
w/ Conf-OT	44.9	0.900	36.20	7.49	0.950	63.07	4.91
RAPS [2]	40.6	0.900	37.22	9.83	0.950	65.93	6.06
w/ Conf-OT	44.9	0.901	27.49	7.63	0.950	46.02	5.20

Table 12. Results on ImageNet shifts with CLIP ResNet-101.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
SUN397							
LAC [39]	61.9	0.900	3.64	9.10	0.949	6.55	5.54
w/ Conf-OT	64.8	0.898	3.02	7.66	0.951	5.36	4.77
APS [51]	61.9	0.900	7.62	7.30	0.950	13.52	4.85
w/ Conf-OT	64.8	0.901	5.92	6.27	0.949	9.75	4.28
RAPS [2]	61.9	0.900	6.85	7.36	0.949	11.12	4.87
w/ Conf-OT	64.8	0.901	5.55	6.32	0.949	8.58	4.30
FGVCAircraft							
LAC [39]	18.8	0.897	23.31	13.24	0.947	36.04	8.52
w/ Conf-OT	21.9	0.896	19.98	10.87	0.950	27.64	6.61
APS [51]	18.8	0.898	24.28	13.17	0.948	36.27	8.53
w/ Conf-OT	21.9	0.894	20.75	10.43	0.948	28.54	6.53
RAPS [2]	18.8	0.899	23.40	13.07	0.946	39.45	8.27
w/ Conf-OT	21.9	0.895	20.50	10.37	0.951	28.46	6.50
EuroSAT							
LAC [39]	45.2	0.899	4.97	12.26	0.949	6.02	6.47
w/ Conf-OT	52.0	0.900	3.40	7.056	0.950	4.54	4.66
APS [51]	45.2	0.900	5.08	11.02	0.949	6.18	6.27
w/ Conf-OT	52.0	0.900	3.93	6.57	0.951	5.02	3.81
RAPS [2]	45.2	0.900	5.07	11.15	0.949	6.22	6.37
w/ Conf-OT	52.0	0.899	3.91	6.64	0.950	5.01	3.84
StanfordCars							
LAC [39]	59.9	0.897	3.27	10.29	0.949	4.98	6.11
w/ Conf-OT	61.7	0.898	2.96	8.79	0.949	4.32	5.33
APS [51]	59.9	0.900	4.45	8.03	0.951	6.31	5.27
w/ Conf-OT	61.7	0.901	4.07	6.80	0.951	5.79	4.52
RAPS [2]	59.9	0.900	4.40	8.04	0.951	6.14	5.35
w/ Conf-OT	61.7	0.900	4.00	6.81	0.950	5.60	4.58
Food101							
LAC [39]	80.4	0.899	1.47	5.32	0.950	2.37	2.93
w/ Conf-OT	79.8	0.898	1.52	4.84	0.950	2.58	2.72
APS [51]	80.4	0.900	2.62	2.78	0.950	4.00	1.95
w/ Conf-OT	79.8	0.899	2.76	2.64	0.950	4.26	1.89
RAPS [2]	80.4	0.900	2.54	2.80	0.950	3.73	1.97
w/ Conf-OT	79.8	0.899	2.68	2.68	0.950	3.97	1.93
OxfordPets							
LAC [39]	87.4	0.902	1.06	8.94	0.951	1.33	4.91
w/ Conf-OT	88.1	0.899	1.04	6.93	0.951	1.26	3.70
APS [51]	87.4	0.91	1.47	4.11	0.956	1.80	2.79
w/ Conf-OT	88.1	0.904	1.45	3.71	0.952	1.77	2.71
RAPS [2]	87.4	0.911	1.46	4.15	0.955	1.78	2.78
w/ Conf-OT	88.1	0.905	1.44	3.64	0.952	1.76	2.74
Flowers102							
LAC [39]	66.5	0.900	5.43	15.75	0.951	16.5	8.43
w/ Conf-OT	69.8	0.899	3.53	15.65	0.953	8.74	8.86
APS [51]	66.5	0.901	6.43	13.89	0.950	15.21	8.52
w/ Conf-OT	69.8	0.899	4.87	12.85	0.952	8.70	8.25
RAPS [2]	66.5	0.900	6.26	14.06	0.952	19.37	8.39
w/ Conf-OT	69.8	0.896	4.71	13.08	0.952	8.93	8.40
Caltech101							
LAC [39]	91.2	0.897	0.97	13.59	0.948	1.10	8.11
w/ Conf-OT	92.2	0.896	0.94	14.44	0.949	1.08	8.97
APS [51]	91.2	0.899	2.07	8.93	0.946	3.05	6.79
w/ Conf-OT	92.2	0.898	1.39	8.67	0.945	1.81	6.39
RAPS [2]	91.2	0.899	1.96	9.07	0.946	2.72	6.79
w/ Conf-OT	92.2	0.898	1.37	8.65	0.945	1.74	6.39
DTD							
LAC [39]	41.9	0.904	11.49	10.11	0.950	16.58	5.94
w/ Conf-OT	46.4	0.894	8.46	8.05	0.947	13.44	5.30
APS [51]	41.9	0.904	12.82	9.38	0.951	16.96	5.49
w/ Conf-OT	46.4	0.900	9.97	7.73	0.949	14.28	5.07
RAPS [2]	41.9	0.903	12.49	9.45	0.953	17.05	5.82
w/ Conf-OT	46.4	0.898	9.89	7.80	0.948	14.06	5.20
UCF101							
LAC [39]	63.6	0.902	3.74	12.08	0.951	6.19	6.91
w/ Conf-OT	67.3	0.902	2.83	11.04	0.950	5.26	6.83
APS [51]	63.6	0.903	6.08	9.624	0.951	9.55	6.20
w/ Conf-OT	67.3	0.902	4.84	8.02	0.952	7.24	5.37
RAPS [2]	63.6	0.904	5.78	9.67	0.952	8.65	6.18
w/ Conf-OT	67.3	0.902	4.68	7.98	0.953	7.01	5.46

Table 13. Results on fine-grained tasks with CLIP ViT-B/32.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
ImageNet							
LAC [39]	63.8	0.901	4.13	8.40	0.950	8.61	5.00
w/ Conf-OT	64.7	0.901	3.84	7.57	0.950	8.18	4.71
APS [51]	63.8	0.900	13.94	6.89	0.950	27.61	4.52
w/ Conf-OT	64.7	0.900	10.90	6.14	0.950	20.93	4.26
RAPS [2]	63.8	0.901	10.15	7.00	0.950	16.24	4.56
w/ Conf-OT	64.7	0.900	8.84	6.21	0.950	14.61	4.29
ImageNet-A							
LAC [39]	31.758	0.897	24.14	8.93	0.949	40.57	5.80
w/ Conf-OT	35.128	0.899	21.44	11.14	0.949	35.75	7.24
APS [51]	31.758	0.899	33.47	8.15	0.949	51.72	5.36
w/ Conf-OT	35.128	0.899	28.22	10.10	0.950	45.22	7.09
RAPS [2]	31.758	0.899	26.80	8.57	0.950	39.78	6.16
w/ Conf-OT	35.128	0.899	24.02	10.5	0.950	35.21	7.40
ImageNet-V2							
LAC [39]	56.4	0.901	7.28	12.58	0.951	16.66	7.85
w/ Conf-OT	57.5	0.900	6.92	12.47	0.950	16.21	7.94
APS [51]	56.4	0.900	23.97	12.43	0.949	47.19	7.92
w/ Conf-OT	57.5	0.901	18.38	12.22	0.950	35.61	7.85
RAPS [2]	56.4	0.900	15.99	12.44	0.950	25.06	7.93
w/ Conf-OT	57.5	0.900	13.88	12.26	0.950	22.94	7.87
ImageNet-R							
LAC [39]	69.2	0.899	3.69	5.66	0.949	8.38	3.36
w/ Conf-OT	71.1	0.900	3.30	5.98	0.950	7.10	3.59
APS [51]	69.2	0.899	8.77	4.03	0.950	14.61	2.85
w/ Conf-OT	71.1	0.898	8.07	4.07	0.950	13.05	2.82
RAPS [2]	69.2	0.899	7.92	4.12	0.950	12.26	2.89
w/ Conf-OT	71.1	0.899	7.30	4.12	0.950	11.22	2.92
ImageNet-Sketch							
LAC [39]	42.1	0.900	24.35	9.68	0.950	56.32	5.88
w/ Conf-OT	46.0	0.899	16.32	8.41	0.950	39.31	5.22
APS [51]	42.1	0.900	50.18	9.17	0.950	87.46	5.78
w/ Conf-OT	46.0	0.899	34.77	7.66	0.950	60.37	4.94
RAPS [2]	42.1	0.900	33.90	9.51	0.950	61.85	5.83
w/ Conf-OT	46.0	0.900	25.95	7.82	0.950	42.45	5.16

Table 14. Results on ImageNet shifts with CLIP ViT-B/32.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
SUN397							
LAC [39]	62.5	0.900	3.47	9.60	0.951	6.52	5.68
w/ Conf-OT	67.0	0.899	2.72	7.81	0.951	4.83	4.86
APS [51]	62.5	0.901	7.42	7.74	0.951	13.70	5.03
w/ Conf-OT	67.0	0.900	5.26	6.42	0.951	9.05	4.33
RAPS [2]	62.5	0.902	6.70	7.90	0.951	10.89	5.14
w/ Conf-OT	67.0	0.900	4.93	6.44	0.950	7.95	4.43
FGVCAircraft							
LAC [39]	24.4	0.894	17.94	13.37	0.948	27.90	7.92
w/ Conf-OT	27.8	0.896	13.81	10.03	0.944	19.89	6.62
APS [51]	24.4	0.895	17.66	13.11	0.948	28.44	8.05
w/ Conf-OT	27.8	0.893	14.51	10.18	0.945	20.93	6.36
RAPS [2]	24.4	0.896	17.96	13.38	0.947	28.97	7.98
w/ Conf-OT	27.8	0.895	14.32	10.24	0.946	20.80	6.30
EuroSAT							
LAC [39]	48.2	0.899	4.13	7.48	0.950	4.90	4.00
w/ Conf-OT	58.0	0.898	2.98	5.33	0.952	4.32	2.69
APS [51]	48.2	0.900	4.22	6.80	0.948	5.05	4.23
w/ Conf-OT	58.0	0.900	3.48	4.99	0.951	4.54	3.48
RAPS [2]	48.2	0.899	4.21	6.89	0.948	5.05	4.27
w/ Conf-OT	58.0	0.900	3.46	4.96	0.951	4.52	3.52
StanfordCars							
LAC [39]	65.5	0.899	2.36	10.73	0.951	3.34	6.50
w/ Conf-OT	68.4	0.900	2.13	9.33	0.950	3.03	5.73
APS [51]	65.5	0.901	3.26	8.00	0.952	4.43	5.24
w/ Conf-OT	68.4	0.898	3.00	6.81	0.948	4.07	4.60
RAPS [2]	65.5	0.901	3.21	8.08	0.952	4.29	5.28
w/ Conf-OT	68.4	0.900	2.97	6.83	0.947	3.97	4.67
Food101							
LAC [39]	85.8	0.899	1.14	5.28	0.950	1.56	2.70
w/ Conf-OT	85.5	0.898	1.15	4.84	0.949	1.64	2.69
APS [51]	85.8	0.900	1.99	2.60	0.949	2.87	1.87
w/ Conf-OT	85.5	0.899	2.05	2.44	0.949	2.99	1.84
RAPS [2]	85.8	0.900	1.94	2.62	0.950	2.70	1.85
w/ Conf-OT	85.5	0.900	2.00	2.42	0.949	2.83	1.84
OxfordPets							
LAC [39]	88.9	0.903	1.02	9.97	0.949	1.20	5.50
w/ Conf-OT	90.6	0.899	0.98	7.90	0.949	1.11	4.36
APS [51]	88.9	0.905	1.37	3.75	0.950	1.65	2.74
w/ Conf-OT	90.6	0.902	1.34	3.94	0.952	1.65	3.07
RAPS [2]	88.9	0.904	1.37	3.74	0.950	1.64	2.75
w/ Conf-OT	90.6	0.902	1.34	3.92	0.951	1.64	3.04
Flowers102							
LAC [39]	71.0	0.899	4.75	16.64	0.951	10.81	9.03
w/ Conf-OT	75.4	0.903	2.95	16.68	0.950	5.99	9.40
APS [51]	71.0	0.899	5.60	14.45	0.949	11.18	9.15
w/ Conf-OT	75.4	0.901	4.10	12.59	0.951	6.50	8.70
RAPS [2]	71.0	0.898	5.49	14.62	0.950	11.49	9.03
w/ Conf-OT	75.4	0.900	4.02	12.69	0.950	6.49	8.81
Caltech101							
LAC [39]	93.1	0.893	0.95	12.73	0.950	1.06	8.31
w/ Conf-OT	92.5	0.900	0.93	13.82	0.946	1.06	9.04
APS [51]	93.1	0.895	2.16	9.48	0.949	3.30	7.20
w/ Conf-OT	92.5	0.898	1.27	8.50	0.945	1.56	6.30
RAPS [2]	93.1	0.895	2.05	9.48	0.949	2.93	7.20
w/ Conf-OT	92.5	0.898	1.26	8.46	0.945	1.52	6.31
DTD							
LAC [39]	43.6	0.898	10.88	11.47	0.950	17.09	6.53
w/ Conf-OT	46.2	0.904	9.01	9.20	0.952	12.98	5.35
APS [51]	43.6	0.902	12.71	10.10	0.950	18.76	6.52
w/ Conf-OT	46.2	0.901	10.19	8.32	0.952	14.03	5.35
RAPS [2]	43.6	0.904	12.47	10.20	0.950	17.52	6.31
w/ Conf-OT	46.2	0.904	10.14	8.38	0.953	13.88	5.17
UCF101							
LAC [39]	67.6	0.902	2.86	11.69	0.953	5.09	6.69
w/ Conf-OT	72.9	0.902	2.33	11.51	0.950	4.13	6.79
APS [51]	67.6	0.902	5.13	8.88	0.951	7.78	5.85
w/ Conf-OT	72.9	0.898	4.10	7.60	0.950	6.17	5.30
RAPS [2]	67.6	0.902	4.88	8.88	0.950	7.18	5.98
w/ Conf-OT	72.9	0.898	3.99	7.70	0.951	5.92	5.35

Table 15. Results on fine-grained tasks with CLIP ViT-B/16.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
ImageNet							
LAC [39]	68.7	0.901	2.81	8.56	0.950	5.51	5.20
w/ Conf-OT	69.5	0.901	2.70	7.71	0.950	5.22	4.77
APS [51]	68.7	0.901	10.07	6.84	0.950	20.65	4.54
w/ Conf-OT	69.5	0.901	7.77	6.32	0.950	15.15	4.32
RAPS [2]	68.7	0.901	7.34	6.92	0.950	11.82	4.65
w/ Conf-OT	69.5	0.901	6.38	6.32	0.950	10.43	4.39
ImageNet-A							
LAC [39]	50.7	0.897	9.25	9.60	0.949	18.17	6.48
w/ Conf-OT	53.0	0.897	7.84	10.81	0.946	16.77	7.87
APS [51]	50.7	0.898	16.19	8.71	0.950	27.55	5.88
w/ Conf-OT	53.0	0.899	14.07	10.01	0.947	24.18	7.11
RAPS [2]	50.7	0.896	13.47	8.95	0.949	20.00	6.09
w/ Conf-OT	53.0	0.898	12.23	10.06	0.947	18.89	7.71
ImageNet-V2							
LAC [39]	62.2	0.898	4.54	12.81	0.949	10.2	8.00
w/ Conf-OT	63.0	0.901	4.44	12.58	0.951	10.11	7.83
APS [51]	62.2	0.898	17.06	12.49	0.950	33.17	7.91
w/ Conf-OT	63.0	0.899	12.79	12.26	0.949	24.25	7.92
RAPS [51]	62.2	0.899	11.47	12.40	0.951	17.99	7.87
w/ Conf-OT	63.0	0.897	9.91	12.36	0.948	15.74	7.97
ImageNet-R							
LAC [39]	77.5	0.901	1.92	5.86	0.950	4.02	3.68
w/ Conf-OT	78.7	0.900	1.80	6.12	0.950	3.62	3.82
APS [51]	77.5	0.903	5.82	3.90	0.952	9.55	2.77
w/ Conf-OT	78.7	0.901	5.33	3.72	0.951	8.61	2.70
RAPS [2]	77.5	0.903	5.23	3.93	0.951	7.87	2.85
w/ Conf-OT	78.7	0.901	4.83	3.77	0.952	7.32	2.73
ImageNet-Sketch							
LAC [39]	48.2	0.900	14.84	9.82	0.950	36.23	5.82
w/ Conf-OT	51.9	0.900	10.25	8.61	0.950	25.18	5.23
APS [51]	48.2	0.901	37.35	9.02	0.950	65.74	5.65
w/ Conf-OT	51.9	0.900	25.31	7.55	0.950	45.06	4.97
RAPS [2]	48.2	0.900	24.01	9.43	0.950	39.52	5.79
w/ Conf-OT	51.9	0.900	18.42	7.73	0.950	29.75	5.18

Table 16. Results on ImageNet shifts with CLIP ViT-B/16.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
SUN397							
LAC [39]	67.6	0.900	2.85	9.78	0.950	5.2	5.51
w/ Conf-OT	71.0	0.901	2.24	7.70	0.950	3.87	4.69
APS [51]	67.6	0.900	6.99	7.70	0.951	14.12	4.95
w/ Conf-OT	71.0	0.901	4.46	6.11	0.951	8.14	4.20
RAPS [2]	67.6	0.901	5.96	7.77	0.951	10.12	5.05
w/ Conf-OT	71.0	0.902	4.15	6.12	0.951	6.94	4.25
FGVCAircraft							
LAC [39]	32.5	0.895	7.27	11.15	0.947	10.47	6.67
w/ Conf-OT	36.2	0.893	6.25	10.32	0.946	9.13	6.52
APS [51]	32.5	0.896	8.38	10.27	0.948	11.34	6.77
w/ Conf-OT	36.2	0.898	7.28	8.824	0.949	10.04	6.00
RAPS [2]	32.5	0.896	8.25	10.35	0.948	11.00	6.82
w/ Conf-OT	36.2	0.898	7.21	8.85	0.950	9.95	6.04
EuroSAT							
LAC [39]	60.2	0.900	2.47	7.74	0.950	3.21	4.32
w/ Conf-OT	65.4	0.900	2.16	7.33	0.949	3.04	3.95
APS [51]	60.2	0.897	2.87	6.02	0.948	3.60	3.63
w/ Conf-OT	65.4	0.899	2.57	4.22	0.949	3.36	2.67
RAPS [2]	60.2	0.897	2.86	5.97	0.948	3.59	3.62
w/ Conf-OT	65.4	0.899	2.56	4.21	0.950	3.34	2.70
StanfordCars							
LAC [39]	76.9	0.904	1.52	11.04	0.952	2.04	6.41
w/ Conf-OT	79.6	0.901	1.37	9.24	0.954	1.79	5.27
APS [51]	76.9	0.903	2.06	7.51	0.950	2.62	4.99
w/ Conf-OT	79.6	0.903	1.90	6.45	0.950	2.43	4.29
RAPS [2]	76.9	0.903	2.04	7.50	0.950	2.57	4.99
w/ Conf-OT	79.6	0.903	1.88	6.49	0.950	2.40	4.28
Food101							
LAC [39]	90.8	0.899	0.97	5.48	0.949	1.15	2.79
w/ Conf-OT	90.4	0.900	0.98	5.38	0.950	1.20	2.78
APS [51]	90.8	0.899	1.50	2.40	0.949	2.02	1.64
w/ Conf-OT	90.4	0.898	1.57	2.30	0.949	2.16	1.71
RAPS [2]	90.8	0.899	1.47	2.41	0.949	1.92	1.64
w/ Conf-OT	90.4	0.898	1.54	2.31	0.949	2.05	1.72
OxfordPets							
LAC [39]	93.4	0.898	0.93	10.51	0.949	1.03	5.75
w/ Conf-OT	93.4	0.895	0.93	9.82	0.949	1.03	4.99
APS [51]	93.4	0.905	1.16	3.74	0.952	1.36	2.72
w/ Conf-OT	93.4	0.907	1.16	3.88	0.952	1.37	2.81
RAPS [2]	93.4	0.905	1.16	3.72	0.952	1.35	2.72
w/ Conf-OT	93.4	0.907	1.16	3.88	0.953	1.36	2.81
Flowers102							
LAC [39]	79.4	0.900	1.83	15.75	0.949	3.60	8.62
w/ Conf-OT	83.1	0.898	1.40	15.45	0.950	2.31	9.08
APS [51]	79.4	0.897	2.69	12.16	0.948	4.15	8.06
w/ Conf-OT	83.1	0.900	2.18	10.32	0.947	3.07	7.08
RAPS [2]	79.4	0.897	2.63	12.27	0.949	4.02	8.09
w/ Conf-OT	83.1	0.900	2.15	10.39	0.946	3.00	7.26
Caltech101							
LAC [39]	95.0	0.900	0.93	12.35	0.947	1.01	7.43
w/ Conf-OT	96.7	0.897	0.90	15.85	0.948	0.97	9.60
APS [51]	95.0	0.898	1.66	8.66	0.946	2.35	6.65
w/ Conf-OT	96.7	0.899	1.06	8.34	0.947	1.23	6.16
RAPS [2]	95.0	0.898	1.59	8.62	0.946	2.12	6.65
w/ Conf-OT	96.7	0.899	1.06	8.32	0.947	1.22	6.15
DTD							
LAC [39]	53.2	0.902	7.24	10.43	0.951	11.94	6.13
w/ Conf-OT	57.5	0.900	5.79	9.91	0.950	9.87	5.55
APS [51]	53.2	0.901	8.60	9.16	0.947	13.01	5.80
w/ Conf-OT	57.5	0.894	7.18	8.28	0.948	11.17	5.50
RAPS [2]	53.2	0.902	8.31	9.53	0.950	12.61	5.95
w/ Conf-OT	57.5	0.897	7.06	8.26	0.952	10.76	5.62
UCF101							
LAC [39]	74.8	0.901	1.80	12.47	0.952	3.05	7.00
w/ Conf-OT	78.2	0.903	1.65	11.57	0.951	2.62	6.86
APS [51]	74.8	0.905	3.34	9.33	0.952	5.12	6.22
w/ Conf-OT	78.2	0.905	2.81	7.38	0.951	4.10	5.21
RAPS [2]	74.8	0.904	3.16	9.38	0.951	4.55	6.19
w/ Conf-OT	78.2	0.903	2.72	7.36	0.950	3.85	5.24

Table 17. Results on fine-grained tasks with CLIP ViT-L/14.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
ImageNet							
LAC [39]	75.8	0.900	1.86	8.70	0.950	3.39	5.26
w/ Conf-OT	75.7	0.901	1.88	7.78	0.950	3.48	4.81
APS [51]	75.8	0.901	6.96	6.82	0.949	15.22	4.63
w/ Conf-OT	75.7	0.900	5.53	6.25	0.949	11.63	4.41
RAPS [2]	75.8	0.900	4.86	6.86	0.949	7.75	4.72
w/ Conf-OT	75.7	0.900	4.39	6.30	0.949	7.24	4.42
ImageNet-A							
LAC [39]	70.6	0.898	3.11	10.43	0.949	6.44	7.00
w/ Conf-OT	72.9	0.897	2.73	10.72	0.948	5.84	7.45
APS [51]	70.5	0.898	7.46	8.15	0.947	12.67	5.74
w/ Conf-OT	72.9	0.898	6.08	8.98	0.947	11.01	6.97
RAPS [2]	70.6	0.897	6.31	8.38	0.949	9.54	6.08
w/ Conf-OT	72.9	0.898	5.38	9.07	0.948	8.69	7.07
ImageNet-V2							
LAC [39]	70.2	0.899	3.05	12.93	0.951	6.78	7.95
w/ Conf-OT	69.6	0.900	2.91	12.61	0.949	7.02	7.97
APS [51]	70.2	0.899	13.75	12.43	0.948	28.69	8.01
w/ Conf-OT	69.6	0.899	9.56	12.25	0.948	19.62	8.00
RAPS [2]	70.2	0.900	8.39	12.42	0.949	12.70	7.94
w/ Conf-OT	69.6	0.898	7.10	12.24	0.950	11.53	7.95
ImageNet-R							
LAC [39]	87.7	0.899	1.06	6.41	0.950	1.49	3.79
w/ Conf-OT	88.0	0.899	1.05	7.11	0.949	1.47	4.17
APS [51]	87.7	0.899	3.08	3.63	0.950	4.99	2.68
w/ Conf-OT	88.0	0.898	3.00	3.58	0.950	4.81	2.59
RAPS [2]	87.7	0.899	2.78	3.65	0.950	4.05	2.71
w/ Conf-OT	88.0	0.899	2.73	3.59	0.951	3.96	2.68
ImageNet-Sketch							
LAC [39]	59.6	0.900	6.99	9.94	0.950	19.63	5.89
w/ Conf-OT	61.8	0.899	5.27	8.78	0.949	13.67	5.29
APS [51]	59.6	0.900	25.44	9.06	0.950	47.83	5.70
w/ Conf-OT	61.8	0.900	16.68	7.57	0.950	31.70	5.02
RAPS [2]	59.6	0.900	14.13	9.22	0.949	21.80	5.92
w/ Conf-OT	61.8	0.899	10.86	7.74	0.950	17.03	5.21

Table 18. Results on ImageNet shifts with CLIP ViT-L/14.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
SUN397							
LAC [39]	69.9	0.899	2.26	8.48	0.950	3.82	5.10
w/ Conf-OT	71.7	0.900	2.11	7.40	0.950	3.53	4.62
APS [51]	69.9	0.899	4.23	6.66	0.949	7.19	4.51
w/ Conf-OT	71.7	0.899	3.70	6.01	0.949	6.05	4.18
RAPS [2]	69.9	0.899	3.97	6.74	0.949	6.26	4.52
w/ Conf-OT	71.7	0.899	3.53	6.06	0.948	5.44	4.22
FGVCAircraft							
LAC [39]	30.1	0.898	13.70	14.35	0.949	24.08	8.38
w/ Conf-OT	34.4	0.897	9.01	10.46	0.949	14.09	6.73
APS [51]	30.1	0.897	15.34	13.74	0.949	24.59	8.35
w/ Conf-OT	34.4	0.897	10.70	10.74	0.949	15.35	6.70
RAPS [2]	30.1	0.896	14.58	13.79	0.952	24.15	8.26
w/ Conf-OT	34.4	0.898	10.49	10.76	0.949	14.93	6.76
EuroSAT							
LAC [39]	49.5	0.901	3.80	7.29	0.952	4.83	3.81
w/ Conf-OT	56.3	0.900	2.82	5.06	0.950	3.81	2.58
APS [51]	49.5	0.903	4.15	6.51	0.951	5.03	3.82
w/ Conf-OT	56.3	0.901	3.47	4.95	0.950	4.40	2.65
RAPS [2]	49.5	0.903	4.14	6.59	0.951	5.01	3.91
w/ Conf-OT	56.3	0.900	3.45	4.96	0.950	4.38	2.62
StanfordCars							
LAC [39]	83.1	0.898	1.20	9.25	0.949	1.55	5.49
w/ Conf-OT	84.5	0.899	1.15	9.10	0.950	1.46	5.25
APS [51]	83.1	0.902	1.73	6.34	0.950	2.26	4.24
w/ Conf-OT	84.5	0.899	1.67	6.07	0.951	2.20	4.06
RAPS [2]	83.1	0.902	1.72	6.35	0.950	2.22	4.24
w/ Conf-OT	84.5	0.899	1.66	6.04	0.950	2.16	4.08
Food101							
LAC [39]	86.0	0.899	1.13	4.98	0.949	1.62	2.60
w/ Conf-OT	85.6	0.899	1.14	4.65	0.950	1.65	2.51
APS [51]	86.0	0.899	1.98	2.39	0.949	2.84	1.73
w/ Conf-OT	85.6	0.900	2.04	2.42	0.950	2.97	1.77
RAPS [2]	86.0	0.899	1.94	2.40	0.949	2.69	1.73
w/ Conf-OT	85.6	0.900	1.99	2.40	0.950	2.80	1.80
OxfordPets							
LAC [39]	91.5	0.898	0.96	9.32	0.948	1.08	5.10
w/ Conf-OT	92.5	0.901	0.95	8.29	0.952	1.06	4.41
APS [51]	91.5	0.906	1.26	3.77	0.952	1.49	2.66
w/ Conf-OT	92.5	0.907	1.24	3.73	0.953	1.49	2.77
RAPS [2]	91.5	0.907	1.26	3.80	0.952	1.48	2.63
w/ Conf-OT	92.5	0.906	1.24	3.71	0.953	1.47	2.79
Flowers102							
LAC [39]	75.2	0.900	3.34	15.78	0.951	7.37	8.29
w/ Conf-OT	78.5	0.901	2.16	15.87	0.950	4.04	8.98
APS [51]	75.2	0.900	4.08	12.97	0.949	7.57	8.15
w/ Conf-OT	78.5	0.900	3.25	11.55	0.948	4.79	8.28
RAPS [2]	75.2	0.899	4.01	13.21	0.949	7.47	8.34
w/ Conf-OT	78.5	0.901	3.17	11.59	0.947	4.63	8.28
Caltech101							
LAC [39]	96.3	0.898	0.90	14.32	0.948	0.97	9.06
w/ Conf-OT	96.2	0.900	0.90	16.16	0.949	0.96	9.15
APS [51]	96.3	0.899	1.17	8.56	0.946	1.53	6.53
w/ Conf-OT	96.2	0.901	1.06	8.02	0.946	1.23	6.24
RAPS [2]	96.3	0.900	1.15	8.56	0.946	1.45	6.53
w/ Conf-OT	96.2	0.901	1.06	8.02	0.946	1.21	6.16
DTD							
LAC [39]	61.5	0.907	3.63	8.87	0.949	5.75	5.17
w/ Conf-OT	64.3	0.902	3.13	7.77	0.947	4.84	5.14
APS [51]	61.5	0.902	5.87	7.76	0.949	8.33	4.74
w/ Conf-OT	64.3	0.903	5.01	7.55	0.953	7.58	4.65
RAPS [2]	61.5	0.902	5.80	7.66	0.950	8.16	4.61
w/ Conf-OT	64.3	0.902	4.92	7.49	0.950	7.33	4.82
UCF101							
LAC [39]	72.0	0.904	2.15	11.85	0.952	3.68	6.61
w/ Conf-OT	75.3	0.902	1.84	11.00	0.950	2.92	6.36
APS [51]	72.0	0.899	3.71	7.92	0.949	5.52	5.45
w/ Conf-OT	75.3	0.899	3.25	6.89	0.949	4.67	4.96
RAPS [2]	72.0	0.899	3.57	7.96	0.950	5.06	5.42
w/ Conf-OT	75.3	0.899	3.14	6.79	0.950	4.45	4.95

Table 19. Results on fine-grained tasks with MetaCLIP ViT-B/16.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
ImageNet							
LAC [39]	72.2	0.901	2.30	8.24	0.950	4.45	5.19
w/ Conf-OT	72.7	0.901	2.24	7.58	0.951	4.36	4.77
APS [51]	72.2	0.899	8.26	6.52	0.950	16.70	4.46
w/ Conf-OT	72.7	0.900	6.80	6.16	0.950	13.61	4.23
RAPS [2]	72.2	0.899	6.20	6.62	0.950	9.93	4.52
w/ Conf-OT	72.7	0.899	5.58	6.22	0.950	9.21	4.27
ImageNet-A							
LAC [39]	49.6	0.901	9.98	9.47	0.949	19.34	6.49
w/ Conf-OT	53.1	0.895	8.48	11.38	0.949	18.33	7.09
APS [51]	49.6	0.899	15.99	8.22	0.948	26.41	6.08
w/ Conf-OT	53.1	0.898	14.22	9.66	0.948	24.67	6.91
RAPS [2]	49.6	0.900	13.73	8.49	0.947	20.57	6.17
w/ Conf-OT	53.1	0.899	12.57	9.76	0.948	19.61	7.09
ImageNet-V2							
LAC [39]	65.3	0.900	3.96	12.66	0.952	9.14	7.81
w/ Conf-OT	65.1	0.900	3.82	12.46	0.950	8.42	7.85
APS [51]	65.3	0.901	14.06	12.34	0.951	28.58	7.77
w/ Conf-OT	65.1	0.901	11.28	12.30	0.952	22.67	7.72
RAPS [2]	65.3	0.901	9.90	12.28	0.952	15.92	7.76
w/ Conf-OT	65.1	0.900	8.87	12.24	0.952	14.77	7.70
ImageNet-R							
LAC [39]	84.0	0.899	1.28	5.59	0.950	2.19	3.69
w/ Conf-OT	84.5	0.899	1.24	6.00	0.950	2.11	3.69
APS [51]	84.0	0.899	4.24	3.65	0.950	6.92	2.64
w/ Conf-OT	84.5	0.900	4.02	3.58	0.951	6.41	2.59
RAPS [2]	84.0	0.899	3.84	3.67	0.951	5.76	2.70
w/ Conf-OT	84.5	0.900	3.65	3.55	0.950	5.41	2.67
ImageNet-Sketch							
LAC [39]	59.8	0.900	6.24	10.22	0.950	16.26	6.00
w/ Conf-OT	62.2	0.900	4.85	8.82	0.950	11.98	5.38
APS [51]	59.8	0.901	23.31	9.11	0.950	41.08	5.76
w/ Conf-OT	62.2	0.900	15.98	7.18	0.950	28.88	5.01
RAPS [2]	59.8	0.901	14.60	9.29	0.950	22.08	6.02
w/ Conf-OT	62.2	0.900	11.27	7.34	0.950	17.83	5.21

Table 20. Results on ImageNet shifts with MetaCLIP ViT-B/16.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
SUN397							
LAC [39]	76.0	0.900	1.69	9.12	0.951	2.70	5.26
w/ Conf-OT	77.4	0.900	1.59	7.96	0.950	2.44	4.81
APS [51]	76.0	0.899	3.22	6.53	0.950	5.60	4.36
w/ Conf-OT	77.4	0.898	2.77	6.00	0.949	4.62	4.13
RAPS [2]	76.0	0.898	3.03	6.58	0.949	4.78	4.37
w/ Conf-OT	77.4	0.899	2.65	6.00	0.949	4.11	4.15
FGVCAircraft							
LAC [39]	49.9	0.898	3.30	11.91	0.950	4.68	7.06
w/ Conf-OT	55.7	0.901	2.70	9.44	0.947	3.47	5.84
APS [51]	49.9	0.899	4.29	11.02	0.950	5.83	6.65
w/ Conf-OT	55.7	0.898	3.43	7.51	0.949	4.52	5.02
RAPS [2]	49.9	0.900	4.20	11.07	0.948	5.63	6.72
w/ Conf-OT	55.7	0.898	3.39	7.46	0.949	4.42	4.99
EuroSAT							
LAC [39]	62.2	0.899	3.42	11.54	0.950	4.43	5.88
w/ Conf-OT	63.5	0.899	2.27	7.04	0.951	3.15	3.50
APS [51]	62.2	0.899	3.96	10.12	0.949	4.88	5.27
w/ Conf-OT	63.5	0.899	2.62	4.50	0.951	3.47	2.63
RAPS [2]	62.2	0.900	3.95	10.17	0.949	4.90	5.42
w/ Conf-OT	63.5	0.898	2.62	4.50	0.951	3.45	2.68
StanfordCars							
LAC [39]	89.5	0.901	1.00	11.65	0.951	1.17	5.93
w/ Conf-OT	91.0	0.902	0.98	10.87	0.950	1.13	5.82
APS [51]	89.5	0.901	1.30	6.04	0.949	1.58	4.15
w/ Conf-OT	91.0	0.901	1.30	5.90	0.949	1.57	3.97
RAPS [2]	89.5	0.901	1.03	6.04	0.949	1.57	4.11
w/ Conf-OT	91.0	0.901	1.29	5.91	0.949	1.55	3.98
Food101							
LAC [39]	92.2	0.900	0.95	5.58	0.950	1.10	2.80
w/ Conf-OT	91.8	0.900	0.95	5.57	0.950	1.11	2.91
APS [51]	92.2	0.899	1.41	2.19	0.949	1.86	1.58
w/ Conf-OT	91.8	0.899	1.44	2.22	0.949	1.91	1.66
RAPS [2]	92.2	0.899	1.39	2.16	0.950	1.77	1.57
w/ Conf-OT	91.8	0.899	1.42	2.23	0.949	1.81	1.67
OxfordPets							
LAC [39]	95.5	0.903	0.92	10.75	0.950	0.99	5.35
w/ Conf-OT	95.6	0.901	0.92	10.73	0.950	0.99	5.31
APS [51]	95.5	0.904	1.10	3.64	0.953	1.26	2.68
w/ Conf-OT	95.6	0.903	1.09	3.77	0.952	1.26	2.76
RAPS [2]	95.5	0.904	1.10	3.63	0.953	1.25	2.68
w/ Conf-OT	95.6	0.903	1.09	3.77	0.953	1.25	2.76
Flowers102							
LAC [39]	84.7	0.902	1.15	15.29	0.950	1.82	8.42
w/ Conf-OT	89.9	0.902	1.01	14.77	0.951	1.27	8.82
APS [51]	84.7	0.902	2.04	10.21	0.948	2.81	7.31
w/ Conf-OT	89.9	0.900	1.62	9.63	0.947	2.25	6.71
RAPS [2]	84.7	0.900	1.99	10.28	0.948	2.72	7.29
w/ Conf-OT	89.9	0.900	1.61	9.67	0.947	2.19	6.71
Caltech101							
LAC [39]	98.0	0.897	0.90	15.61	0.951	0.95	9.36
w/ Conf-OT	98.2	0.896	0.90	17.27	0.947	0.95	10.57
APS [51]	98.0	0.898	1.00	8.29	0.946	1.17	6.32
w/ Conf-OT	98.2	0.897	1.01	8.01	0.945	1.15	6.10
RAPS [2]	98.0	0.898	1.00	8.35	0.946	1.15	6.32
w/ Conf-OT	98.2	0.897	1.01	7.98	0.945	1.13	6.10
DTD							
LAC [39]	65.8	0.899	2.57	10.75	0.950	4.27	6.32
w/ Conf-OT	70.6	0.900	2.24	9.10	0.950	3.06	5.50
APS [51]	65.8	0.905	4.64	8.00	0.952	6.50	4.98
w/ Conf-OT	70.6	0.900	3.66	7.37	0.949	5.55	5.11
RAPS [2]	65.8	0.905	4.53	8.10	0.950	6.25	4.98
w/ Conf-OT	70.6	0.900	3.61	7.37	0.951	5.41	5.00
UCF101							
LAC [39]	81.9	0.899	1.24	12.27	0.949	1.64	6.76
w/ Conf-OT	85.3	0.898	1.12	11.21	0.949	1.52	6.56
APS [51]	81.9	0.899	2.06	7.46	0.951	2.89	4.90
w/ Conf-OT	85.3	0.897	2.00	6.83	0.952	2.83	4.60
RAPS [2]	81.9	0.899	2.00	7.47	0.952	2.71	4.94
w/ Conf-OT	85.3	0.897	1.95	6.79	0.951	2.66	4.60

Table 21. Results on fine-grained tasks with MetaCLIP ViT-H/14.

Method	Top-1↑	$\alpha = 0.10$			$\alpha = 0.05$		
		Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
ImageNet							
LAC [39]	80.7	0.901	1.42	8.84	0.950	2.33	5.38
w/ Conf-OT	80.6	0.900	1.42	7.99	0.950	2.36	4.89
APS [51]	80.7	0.899	5.60	6.64	0.949	12.37	4.58
w/ Conf-OT	80.6	0.898	4.24	6.22	0.950	8.87	4.30
RAPS [2]	80.7	0.899	3.90	6.68	0.949	6.15	4.66
w/ Conf-OT	80.6	0.898	3.39	6.25	0.950	5.47	4.34
ImageNet-A							
LAC [39]	75.5	0.898	2.24	10.17	0.950	5.58	6.30
w/ Conf-OT	77.0	0.897	2.10	10.09	0.950	4.94	6.85
APS [51]	75.5	0.900	6.70	7.70	0.951	11.69	5.65
w/ Conf-OT	77.0	0.898	5.63	8.10	0.949	10.48	5.91
RAPS [2]	75.5	0.901	5.71	7.86	0.951	8.77	5.83
w/ Conf-OT	77.0	0.898	4.98	8.18	0.951	8.18	6.14
ImageNet-V2							
LAC [39]	74.3	0.900	2.14	12.80	0.950	4.35	7.89
w/ Conf-OT	74.5	0.900	2.15	12.55	0.950	4.54	7.90
APS [51]	74.3	0.898	10.07	12.37	0.950	22.16	7.91
w/ Conf-OT	74.5	0.899	7.74	12.20	0.950	16.96	7.86
RAPS [2]	74.3	0.898	6.31	12.43	0.950	9.84	7.89
w/ Conf-OT	74.5	0.899	5.67	12.24	0.950	9.34	7.81
ImageNet-R							
LAC [39]	93.4	0.899	0.93	6.88	0.950	1.05	3.95
w/ Conf-OT	93.4	0.900	0.93	7.06	0.950	1.05	3.99
APS [51]	93.4	0.899	2.46	3.43	0.950	3.82	2.49
w/ Conf-OT	93.4	0.900	2.28	3.42	0.951	3.51	2.50
RAPS [2]	93.4	0.899	2.19	3.44	0.950	3.07	2.51
w/ Conf-OT	93.4	0.900	2.06	3.43	0.951	2.86	2.49
ImageNet-Sketch							
LAC [39]	70.3	0.900	3.00	10.53	0.951	7.69	5.98
w/ Conf-OT	71.6	0.900	2.58	9.36	0.950	5.93	5.45
APS [51]	70.3	0.900	16.95	8.49	0.950	31.08	5.74
w/ Conf-OT	71.6	0.900	11.29	7.26	0.950	21.11	5.01
RAPS [2]	70.3	0.900	9.18	8.71	0.950	13.01	5.86
w/ Conf-OT	71.6	0.900	7.15	7.35	0.950	10.71	5.19

Table 22. Results on ImageNet shifts with MetaCLIP ViT-H/14.