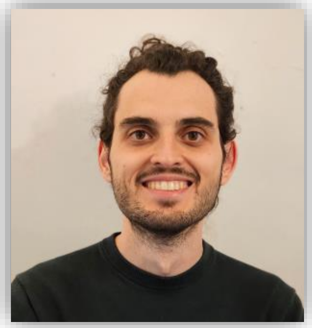
# Conformal Prediction for Zero-Shot Models


Julio
Silva-Rodríguez

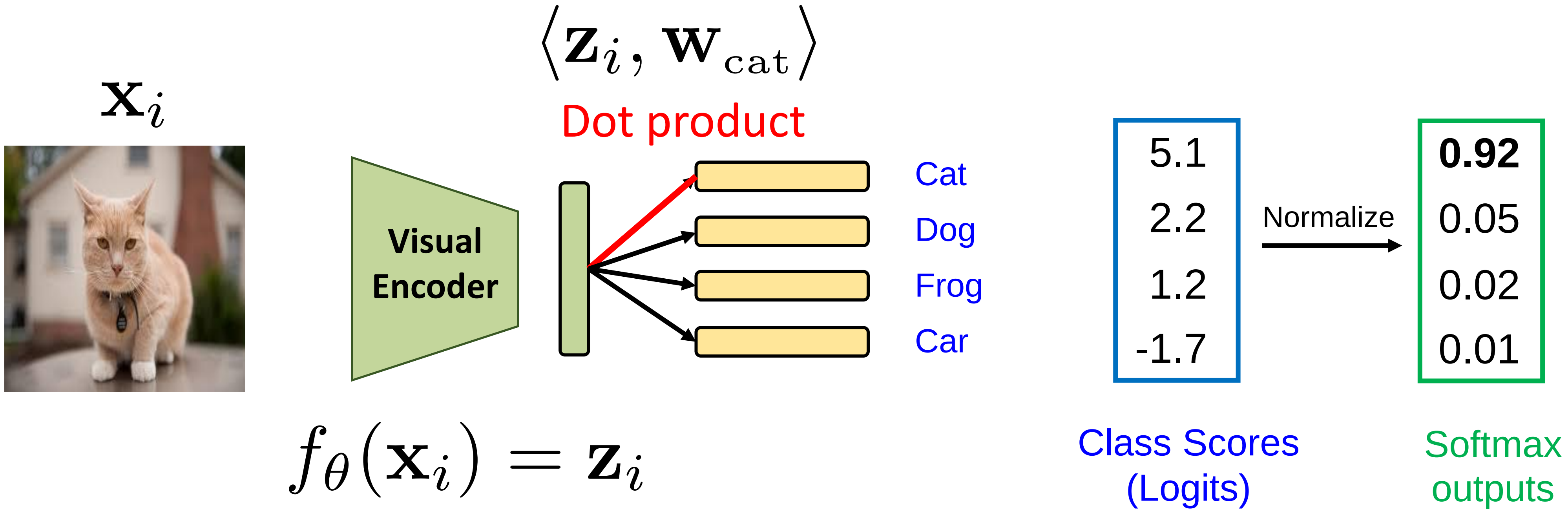


Ismail
Ben Ayed

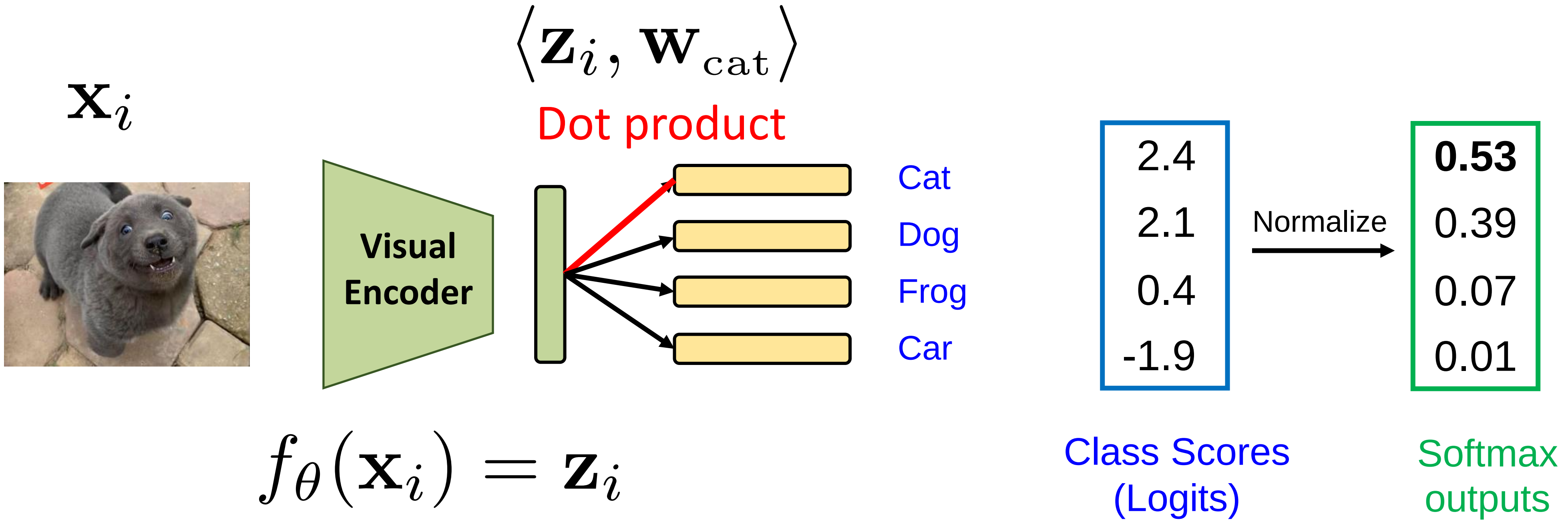
Jose Dolz

ÉTS Montréal

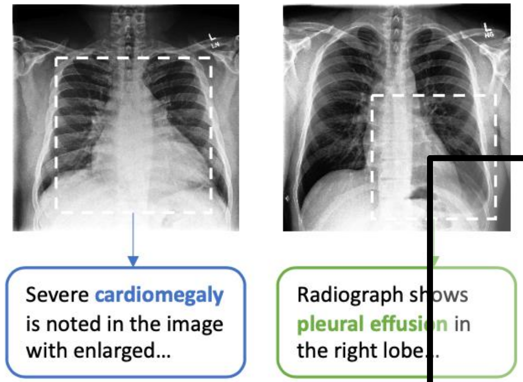
# Reliability and uncertainty on vision classifiers

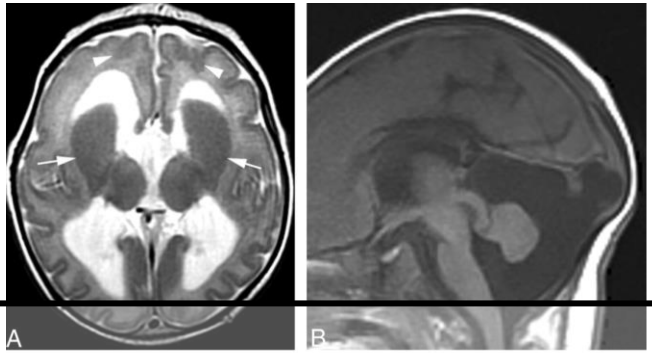


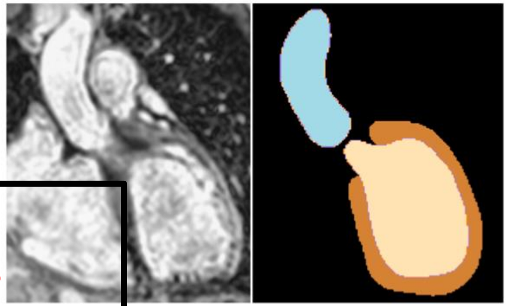
# Reliability and uncertainty on vision classifiers



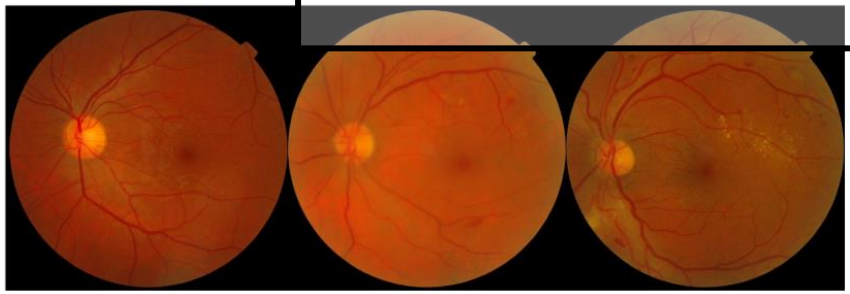
# Reliability and uncertainty on vision classifiers

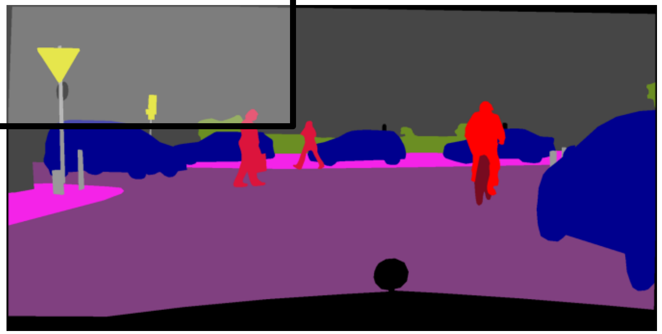


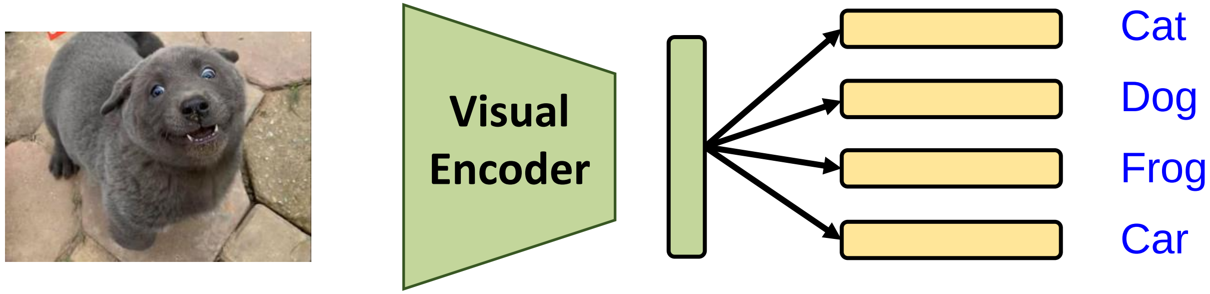
Vision classifiers are being deployed at high-stake applications!

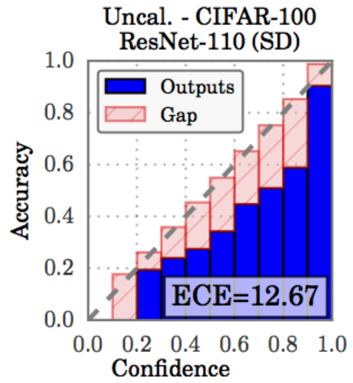




# Reliability and uncertainty on vision classifiers

- **Model calibration**



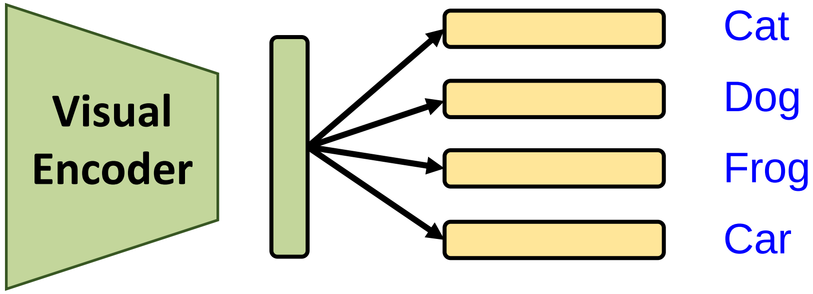

$$e^{l/T} / \sum_i e^{l_i/T}$$

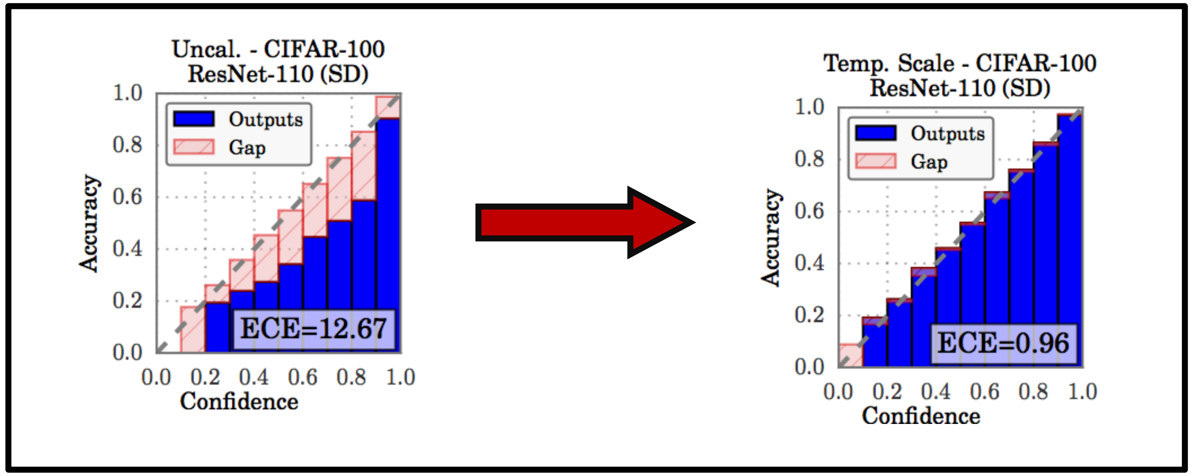
| Class Scores (Logits) | | Softmax outputs |
|---|---|---|
| 2.4 | | **0.53** |
| 2.1 | Normalize → | 0.39 |
| 0.4 | | 0.07 |
| -1.9 | | 0.01 |

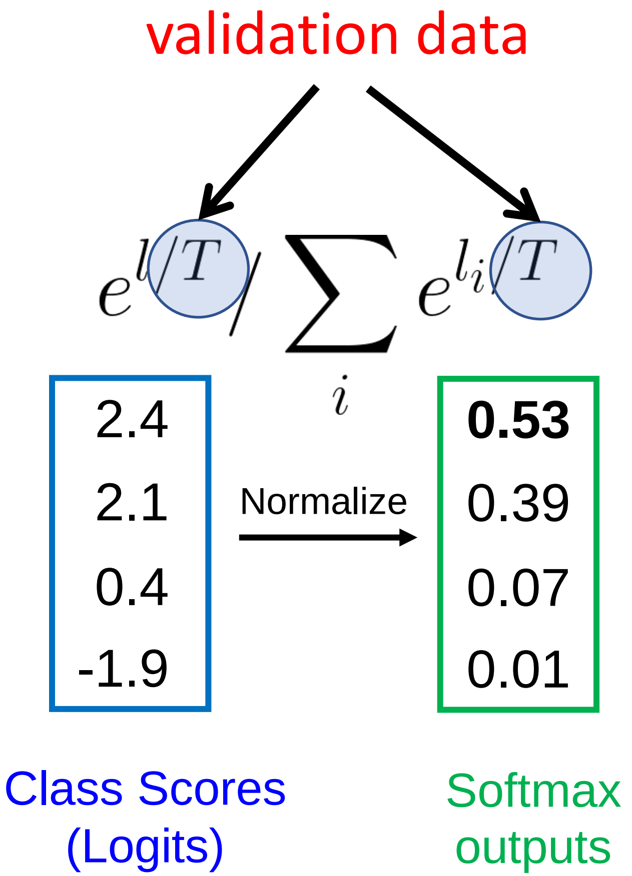
Plot from [Guo et al., On Calibration of Modern Neural Networks, ICML 2020]

# Reliability and uncertainty on vision classifiers

- **Model calibration**



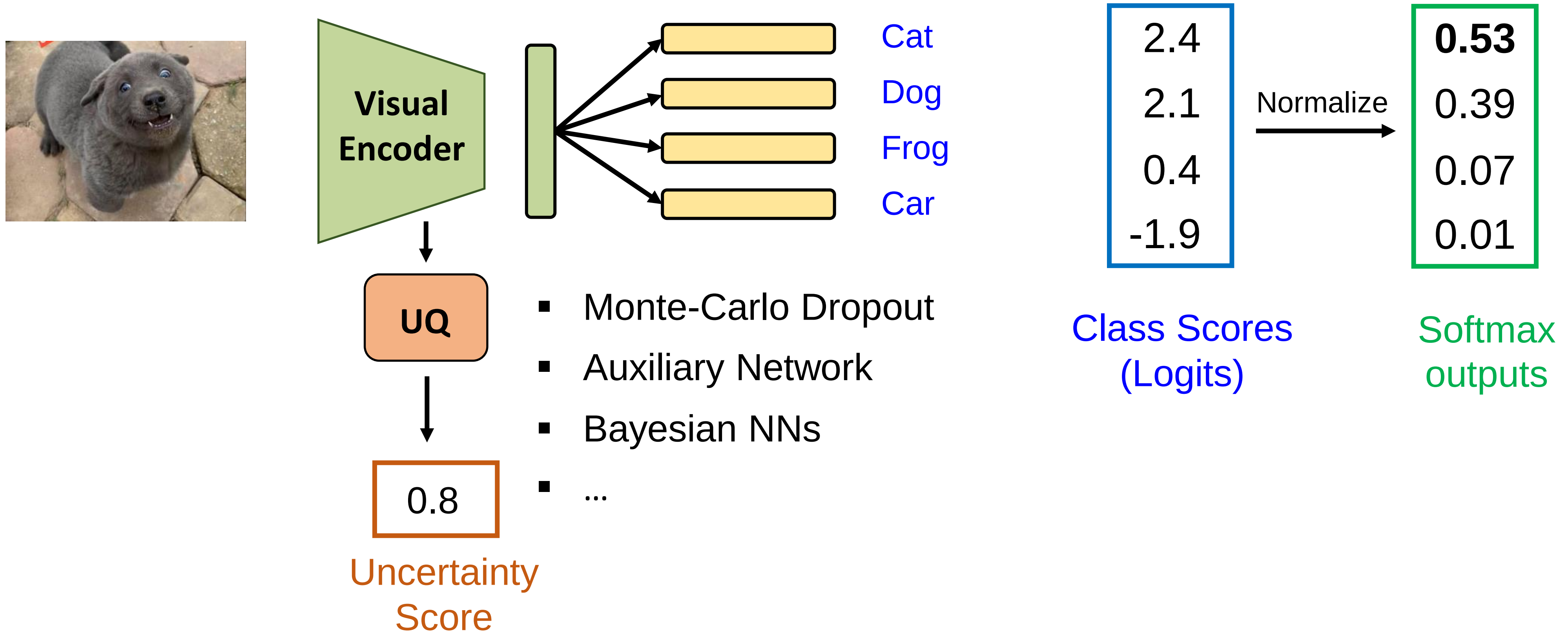




Plot from [Guo et al., On Calibration of Modern Neural Networks, ICML 2020]

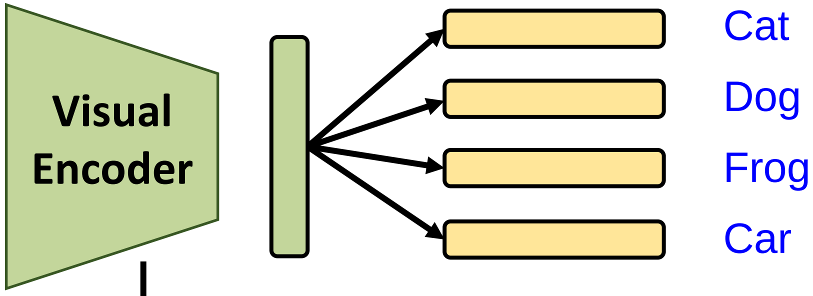
# Reliability and uncertainty on vision classifiers

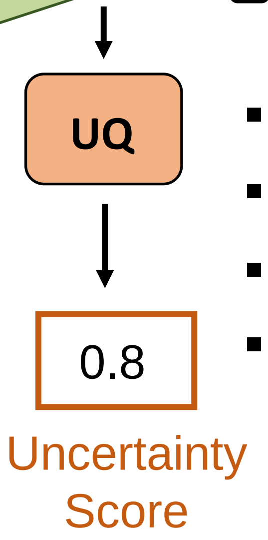
- **Uncertainty quantification**



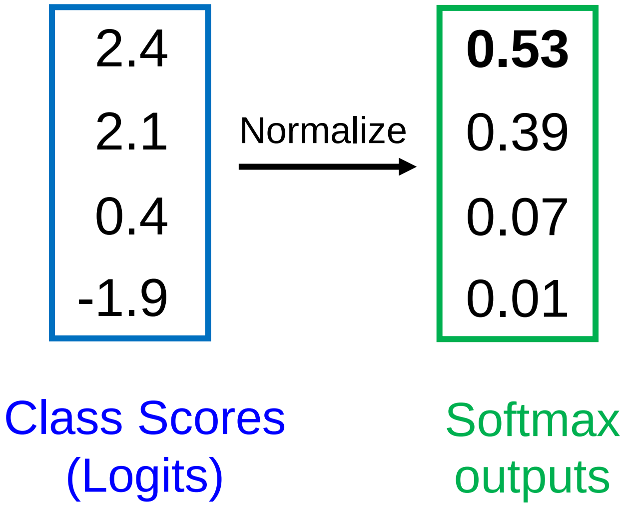
# Reliability and uncertainty on vision classifiers

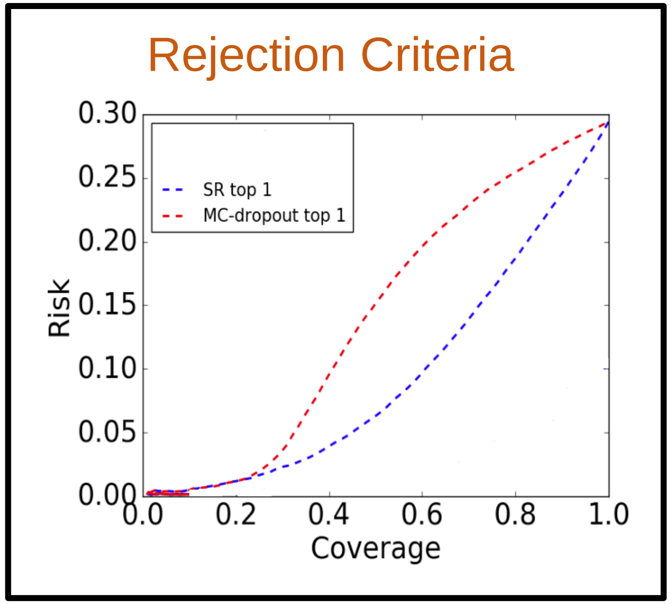
- **Uncertainty quantification**




- Monte-Carlo Dropout
- Auxiliary Network
- Bayesian NNs
- …




Plot modified from [Geifman et al., Selective Classification for Deep Neural Networks, NIPS 2017]

# Reliability and uncertainty on vision classifiers

- **Limitations, pitfalls.**

**1. Why to reject samples?**

Cat (p=0.53, u=0.8)
REJECT ❌

{Cat (p=0.53),
**Dog** (p=0.29)}

# Reliability and uncertainty on vision classifiers

- **Limitations, pitfalls.**

**1. Why to reject samples?**

Cat (p=0.53, u=0.8)
REJECT ❌

{Cat (p=0.53),
**Dog** (p=0.29)}

| Weight | Acc@1 | Acc@5 |
|---|---|---|
| AlexNet_Weights.IMAGENET1K_V1 | 56.522 | 79.066 |
| ConvNeXt_Base_Weights.IMAGENET1K_V1 | 84.062 | 96.87 |
| ConvNeXt_Large_Weights.IMAGENET1K_V1 | 84.414 | 96.976 |
| ConvNeXt_Small_Weights.IMAGENET1K_V1 | 83.616 | 96.65 |
| ConvNeXt_Tiny_Weights.IMAGENET1K_V1 | 82.52 | 96.146 |
| DenseNet121_Weights.IMAGENET1K_V1 | 74.434 | 91.972 |
| DenseNet161_Weights.IMAGENET1K_V1 | 77.138 | 93.56 |
| DenseNet169_Weights.IMAGENET1K_V1 | 75.6 | 92.806 |
| DenseNet201_Weights.IMAGENET1K_V1 | 76.896 | 93.37 |
| EfficientNet_B0_Weights.IMAGENET1K_V1 | 77.692 | 93.532 |
| EfficientNet_B1_Weights.IMAGENET1K_V1 | 78.642 | 94.186 |

https://pytorch.org/vision/stable/models.html

{**fox squirrel**}


{marmot, **fox squirrel**, mink, weasel, beaver}

From [Uncertainty Sets for Image Classifiers Using Conformal Prediction, ICLR 2021]

# Reliability and uncertainty on vision classifiers

- **Limitations, pitfalls.**

**1. Why to reject samples?**

Cat (p=0.53, u=0.8)
REJECT ❌

{Cat (p=0.53),
**Dog** (p=0.29)}

| Weight | Acc@1 | Acc@5 |
|---|---|---|
| AlexNet_Weights.IMAGENET1K_V1 | 56.522 | 79.066 |
| ConvNeXt_Base_Weights.IMAGENET1K_V1 | 84.062 | 96.87 |
| ConvNeXt_Large_Weights.IMAGENET1K_V1 | 84.414 | 96.976 |
| ConvNeXt_Small_Weights.IMAGENET1K_V1 | 83.616 | 96.65 |
| ConvNeXt_Tiny_Weights.IMAGENET1K_V1 | 82.52 | 96.146 |
| DenseNet121_Weights.IMAGENET1K_V1 | 74.434 | 91.972 |
| DenseNet161_Weights.IMAGENET1K_V1 | 77.138 | 93.56 |
| DenseNet169_Weights.IMAGENET1K_V1 | 75.6 | 92.806 |
| DenseNet201_Weights.IMAGENET1K_V1 | 76.896 | 93.37 |
| EfficientNet_B0_Weights.IMAGENET1K_V1 | 77.692 | 93.532 |
| EfficientNet_B1_Weights.IMAGENET1K_V1 | 78.642 | 94.186 |

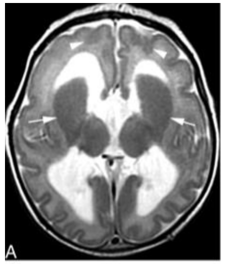
https://pytorch.org/vision/stable/models.html

{**fox squirrel**}

{marmot, **fox squirrel**, mink, weasel, beaver}

From [Uncertainty Sets for Image Classifiers Using Conformal Prediction, ICLR 2021]

**2. Lack of *guarantees.***



*"Set of predictions that covers the true diagnosis with a high probability (e.g., 95%)".*

# (Brief) Introduction to (split) Conformal Prediction

***Conformal prediction*** *(CP) is a machine learning freamework that provides* ***model agnostic****, and* ***distribution-fre****e,* ***finite-sample vailidy guarantees*** *for handling reliability, by producing* ***predictive sets.***

# (Brief) Introduction to (split) Conformal Prediction

***Conformal prediction** (CP) is a machine learning freamework that provides **model agnostic**, and **distribution-free**, **finite-sample vailidy guarantees** for handling reliability, by producing **predictive sets.***

- Random data points $(\mathbf{x}, y)$ from a data distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ .
- Label space $\mathcal{Y} = \{1, 2, ..., K\}$.
- Set-valued mapping function $\mathcal{C} : \mathcal{X} \rightarrow 2^K$, such that $C(\mathbf{x}) \subset \mathcal{Y}$ .
- Desired error level $\alpha \in (0, 1)$ .

# (Brief) Introduction to (split) Conformal Prediction

***Conformal prediction** (CP) is a machine learning freamework that provides **model agnostic**, and **distribution-free**, **finite-sample vailidy guarantees** for handling reliability, by producing **predictive sets.***

- Random data points $(\mathbf{x}, y)$ from a data distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ .
- Label space $\mathcal{Y} = \{1, 2, ..., K\}$.
- Set-valued mapping function $\mathcal{C} : \mathcal{X} \rightarrow 2^K$, such that $C(\mathbf{x}) \subset \mathcal{Y}$ .
- Desired error level $\alpha \in (0, 1)$ .

***Coverage property***

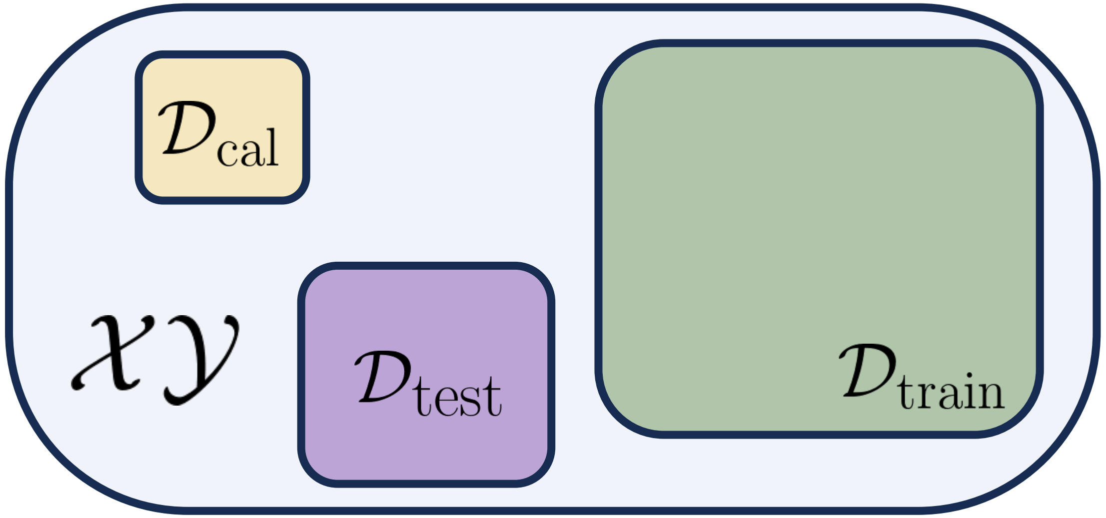
$$\mathcal{P}(Y \in C(\mathbf{x})) \geq 1 - \alpha$$

**(marginal over** $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ **)**

For more details, see [Vovk et al., Learning in a Random World, 2005]

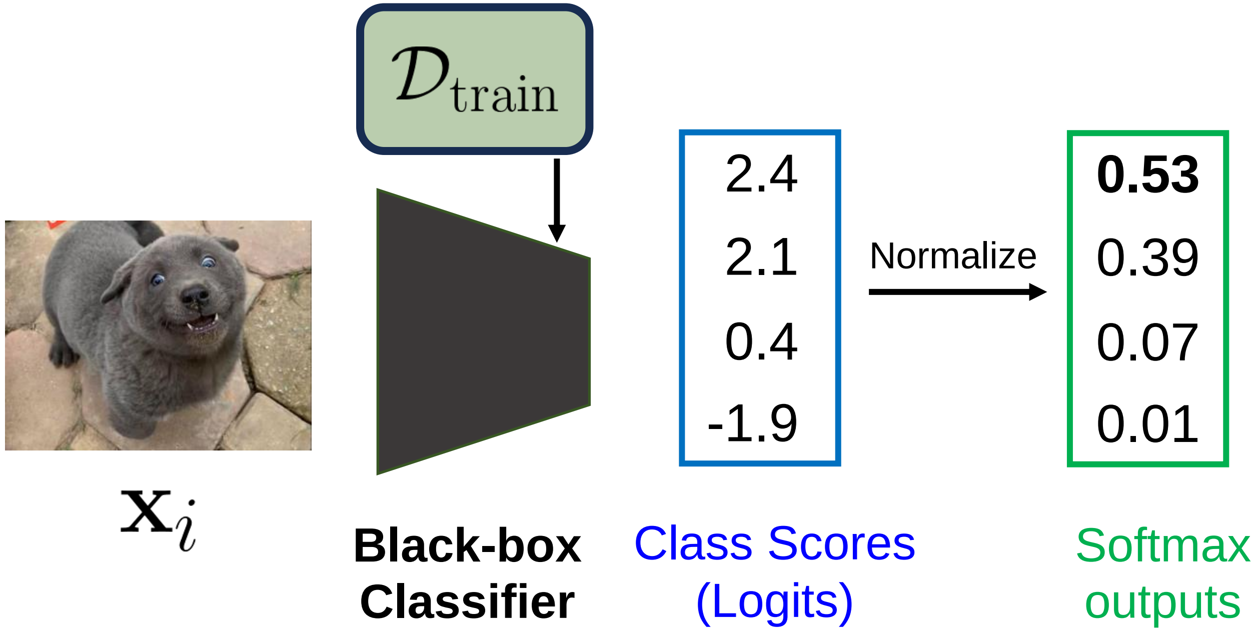
# (Brief) Introduction to (split) Conformal Prediction

- **Split conformal prediction (SCP).**

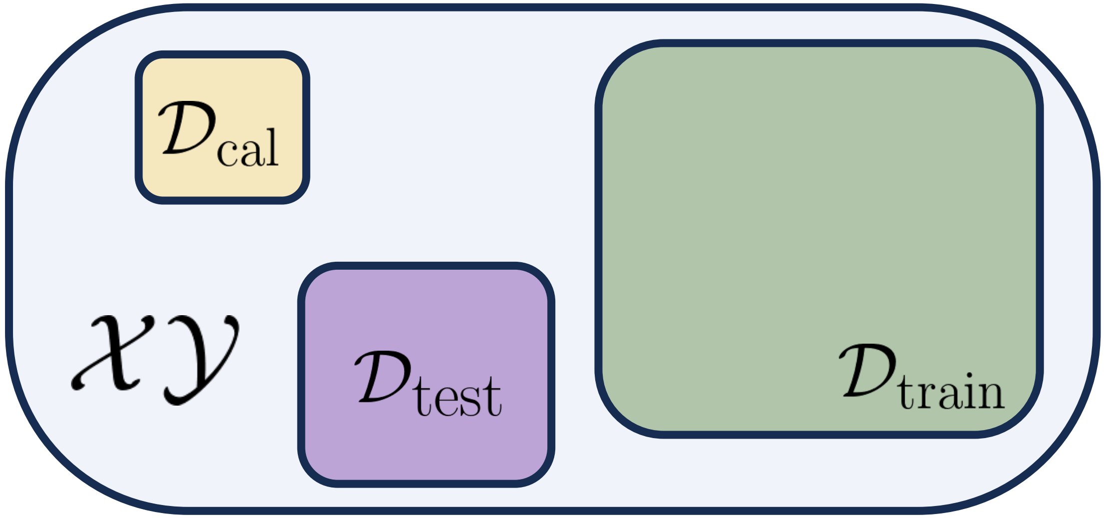


# (Brief) Introduction to (split) Conformal Prediction

- **Split conformal prediction (SCP).**

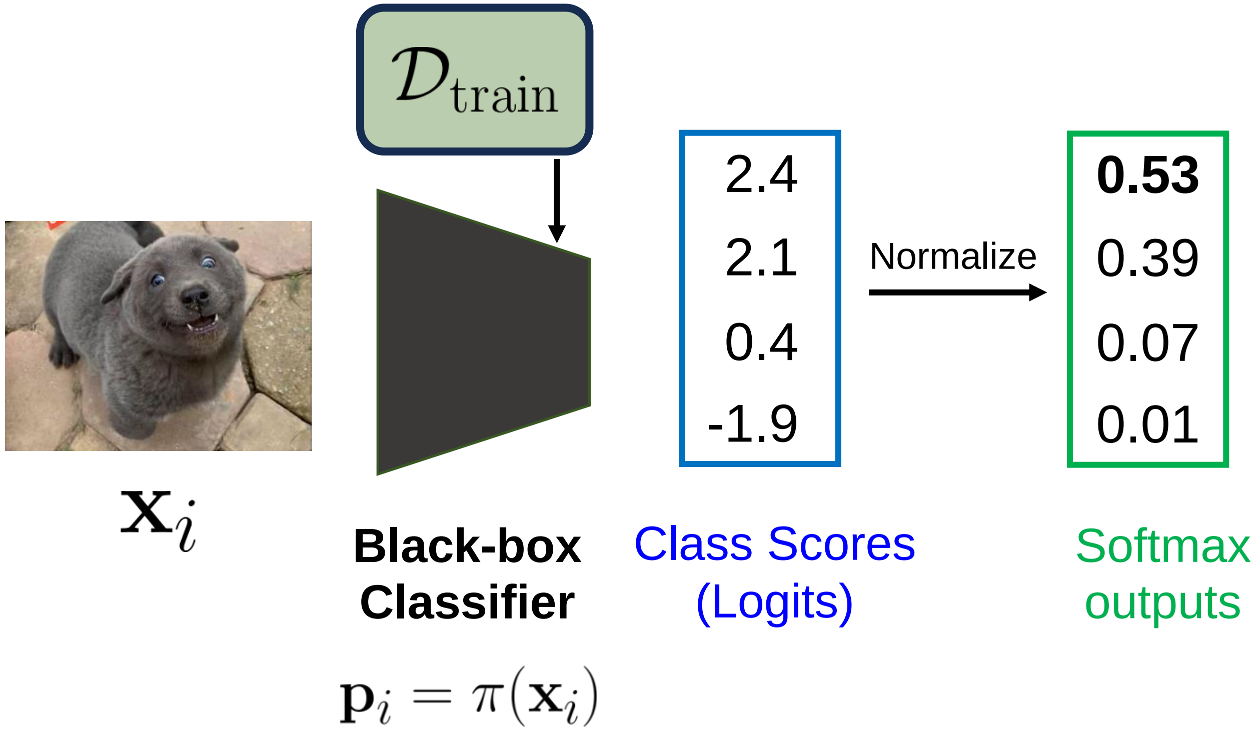


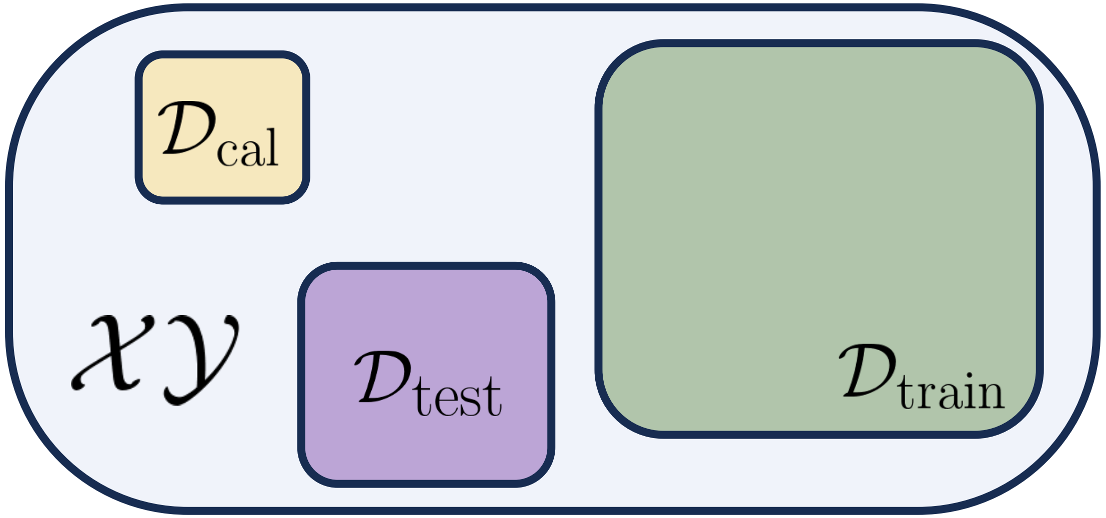

$\mathbf{p}_i = \pi(\mathbf{x}_i)$



# (Brief) Introduction to (split) Conformal Prediction

- **Split conformal prediction (SCP).**




$$\mathcal{D}_{\text{cal}} = \{(\pi(\mathbf{x}_i), y_i)\}_{i=1}^{N}$$

$$\mathcal{D}_{\text{test}} = \{(\pi(\mathbf{x}_i), )\}_{i=N+1}^{N+M}$$

# (Brief) Introduction to (split) Conformal Prediction

- **Split conformal prediction (SCP).**

**1. Define a <u>non-conformity score</u>.**

$$s(y) = \mathcal{S}(\mathbf{p}, y)$$

evaluated per label

| | | |
|---|---|---|
| **0.53** | | **0.47** |
| 0.39 | LAC → | 0.61 |
| 0.07 | | 0.93 |
| 0.01 | | 0.99 |

Procedure

# (Brief) Introduction to (split) Conformal Prediction

- **Split conformal prediction (SCP).**

**1. Define a <u>non-conformity score</u>.**

$$s(y) = \mathcal{S}(\mathbf{p}, y)$$

evaluated per label

| | LAC | |
|---|---|---|
| **0.53** | → | **0.47** |
| 0.39 | | 0.61 |
| 0.07 | | 0.93 |
| 0.01 | | 0.99 |

**2. Compute the cumulative score distribution from the <u>calibration set</u> for <u>true labels</u>.**

$$s_i = \mathcal{S}(\mathbf{p}_i, y_i)$$

0.61

# (Brief) Introduction to (split) Conformal Prediction

- **Split conformal prediction (SCP).**

**1. Define a <u>non-conformity score</u>.**

$$s(y) = \mathcal{S}(\mathbf{p}, y)$$

evaluated per label

| | | |
|---|---|---|
| **0.53** | LAC → | **0.47** |
| 0.39 | | 0.61 |
| 0.07 | | 0.93 |
| 0.01 | | 0.99 |

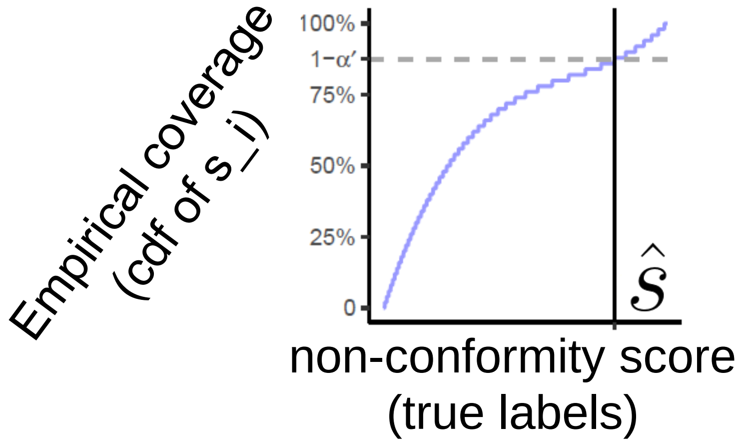
**2. Compute the cumulative score distribution from the <u>calibration set</u> for <u>true labels</u>.**

**3. Search the <u>1-alpha quantile</u> in such distribution.**

$$s_i = \mathcal{S}(\mathbf{p}_i, y_i)$$

0.61

$\mathcal{D}_{\text{cal}}$



Procedure

# (Brief) Introduction to (split) Conformal Prediction

- **Split conformal prediction (SCP).**

**1. Define a non-conformity score.**

$$s(y) = \mathcal{S}(\mathbf{p}, y)$$

evaluated per label

| | |
|---|---|
| **0.53** | **0.47** |
| 0.39 | 0.61 |
| 0.07 | 0.93 |
| 0.01 | 0.99 |

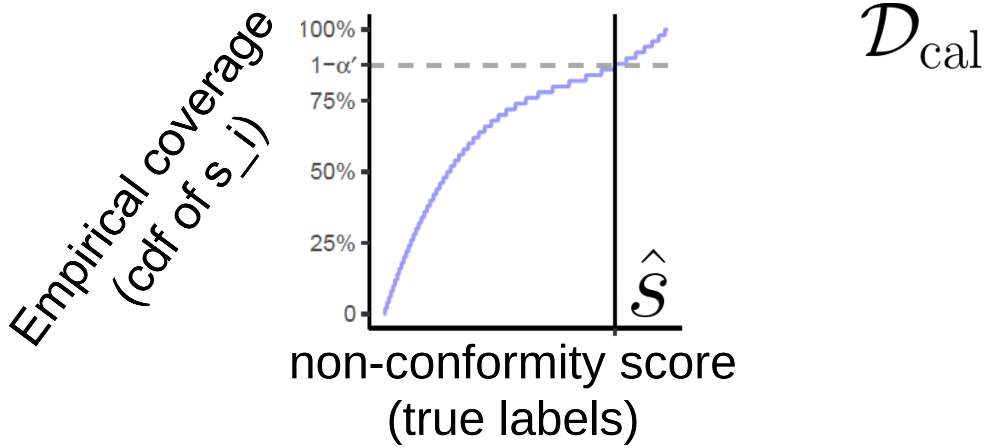
LAC →

**2. Compute the cumulative score distribution from the calibration set for true labels.**

**3. Search the 1-alpha quantile in such distribution.**

$$s_i = \mathcal{S}(\mathbf{p}_i, y_i)$$

0.61




**4. Produce output sets for new data points.**

$$\mathcal{C}(\mathbf{x}) = \{y \in \mathcal{Y} : \mathcal{S}(\mathbf{p}, y) \leq \hat{s}\}$$

Remarks

# (Brief) Introduction to (split) Conformal Prediction

- **Split conformal prediction (SCP).**

**Theoretical guarantees**

$$\mathcal{P}(Y \in C(\mathbf{x})) \geq 1 - \alpha$$

*Generaly, there exist theoretical finite-sample coverage guarantees under the assumption of **i.i.d** or, at least, **exchangable** data distributions for calibration and testing.*

$\mathcal{D}_{\text{cal}}$ $\mathcal{D}_{\text{test}}$

same marginals!

For more details, see [Vovk et al., Learning in a Random World, 2005]

Evaluate

# (Brief) Introduction to (split) Conformal Prediction

- **Split conformal prediction (SCP).**

**1. Efficiency**
(we want small sets)

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

**2. Empirical Coverage**
(keep the desired error)

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

**3. Adaptability**
(set size should adapt to give coverage to difficult subgroups)

$$\mathrm{CCV}(\mathcal{D}) = 100\times\frac{1}{|\mathcal{Y}|}\sum_{k\in\mathcal{Y}} \left|\mathrm{Cov}(\mathcal{D}_k) - (1-\alpha)\right|$$

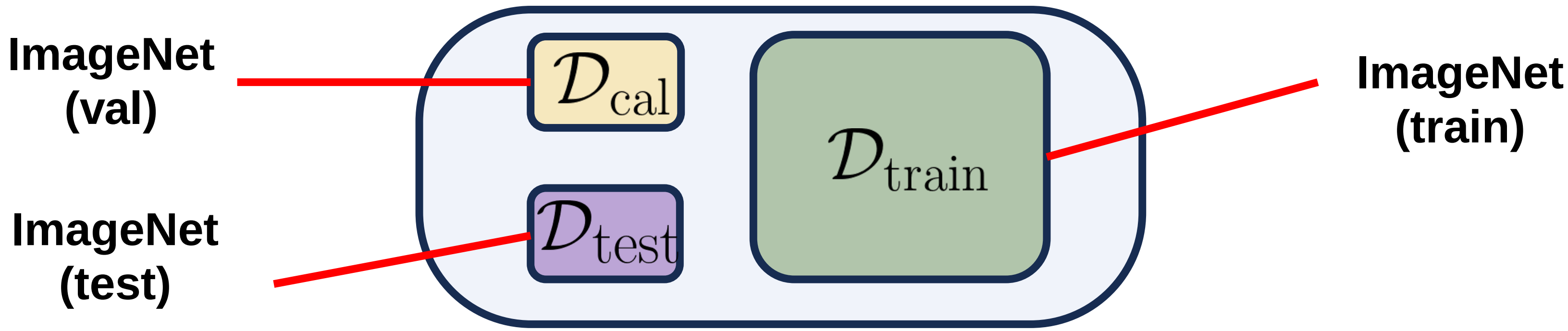
{**fox squirrel**}

{marmot, **fox squirrel**, mink, weasel, beaver}

From [Angelopoulos et al, Uncertainty Sets for Image Classifiers Using Conformal Prediction, ICLR 2021]

# Literature in Vision Classifiers

- **Explored in the standard supervised scenario.**




- **Different adaptive non-conformity scores have been proposed.**

$$\mathcal{S}_{\text{LAC}}(\mathbf{x}, y) = 1 - p_{k=y}$$

[Sadinle et al., Least ambiguous set-valued classifiers with bounded error levels, Jour. American Statistical Association 2019]

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

[Romano et al., Classification with valid and adaptive coverage., NeurIPS 2020]

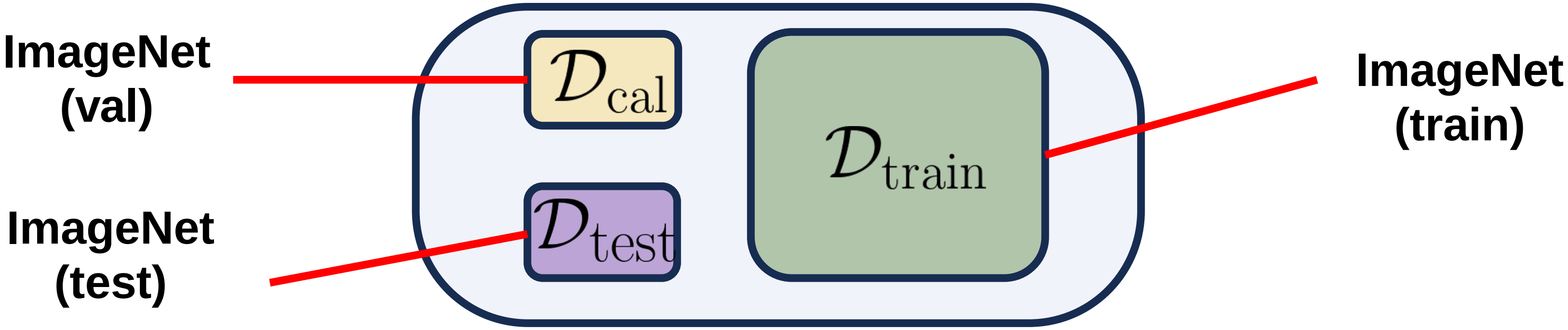
$$\mathcal{S}_{\text{RAPS}}(\mathbf{x}, y) = \mathcal{S}_{\text{APS}}(\mathbf{x}, y) + \lambda \cdot (o(\mathbf{x}, y) - k_{\text{reg}})^{+}$$

[Angelopoulos et al., Uncertainty Sets for Image Classifiers Using Conformal Prediction, ICLR 2021]

# Literature in Vision Classifiers

**Not yet explored for vision-language (CLIP) models**

- **Explored in the standard supervised scenario.**




- **Different adaptive non-conformity scores have been proposed.**

$$\mathcal{S}_{\text{LAC}}(\mathbf{x}, y) = 1 - p_{k=y}$$

[Sadinle et al., Least ambiguous set-valued classifiers with bounded error levels, Jour. American Statistical Association 2019]

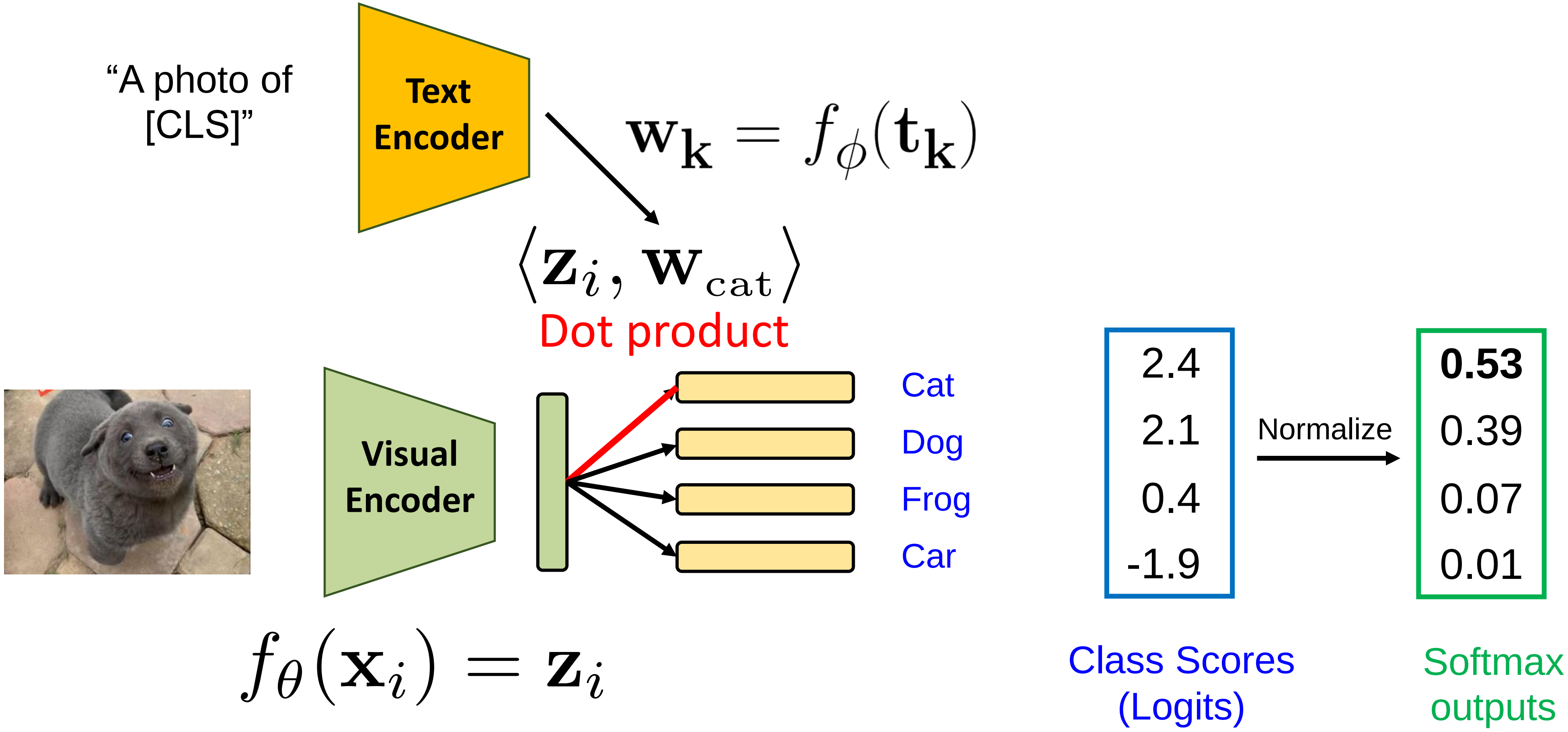
$$\mathcal{S}_{\text{APS}}(\mathbf{x}, y) = \rho_x(y) + p_{k=y} \cdot u$$

[Romano et al., Classification with valid and adaptive coverage., NeurIPS 2020]

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

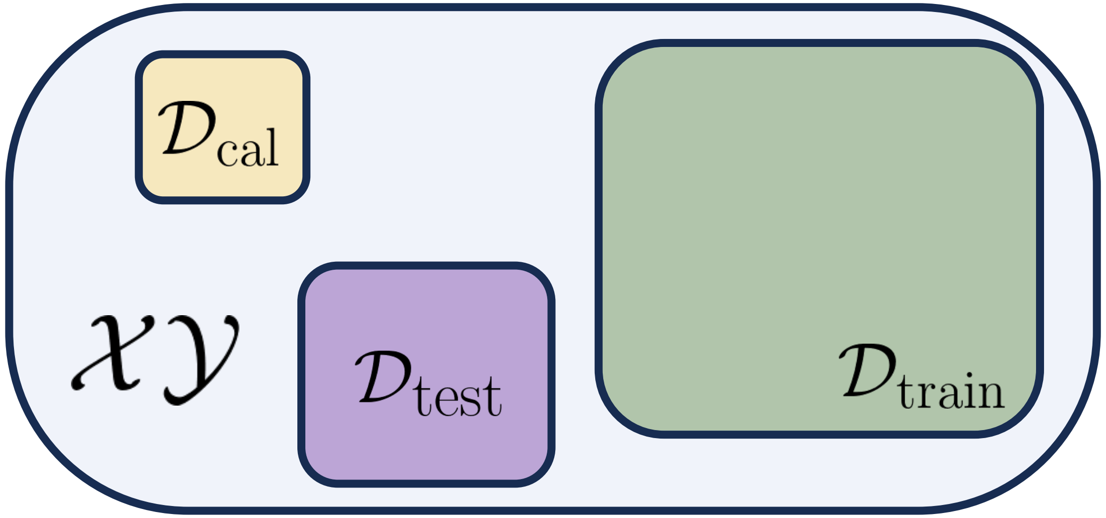
[Angelopoulos et al., Uncertainty Sets for Image Classifiers Using Conformal Prediction, ICLR 2021]

# Vision-Language (zero-shot) Models

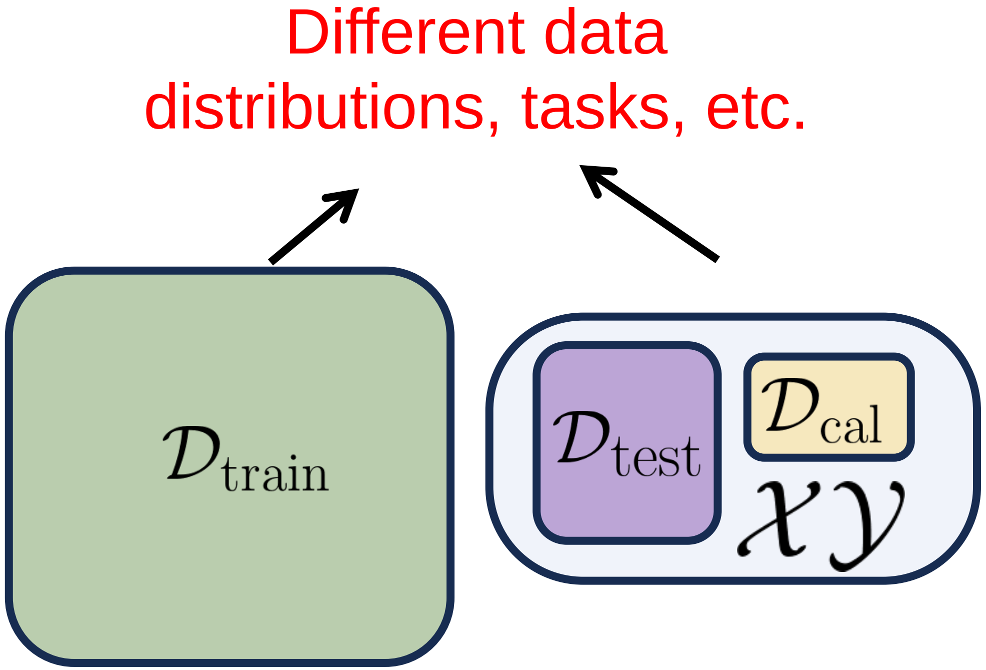


# Conformal Prediction for Zero-Shot Models

- **Transfer learning setting.**



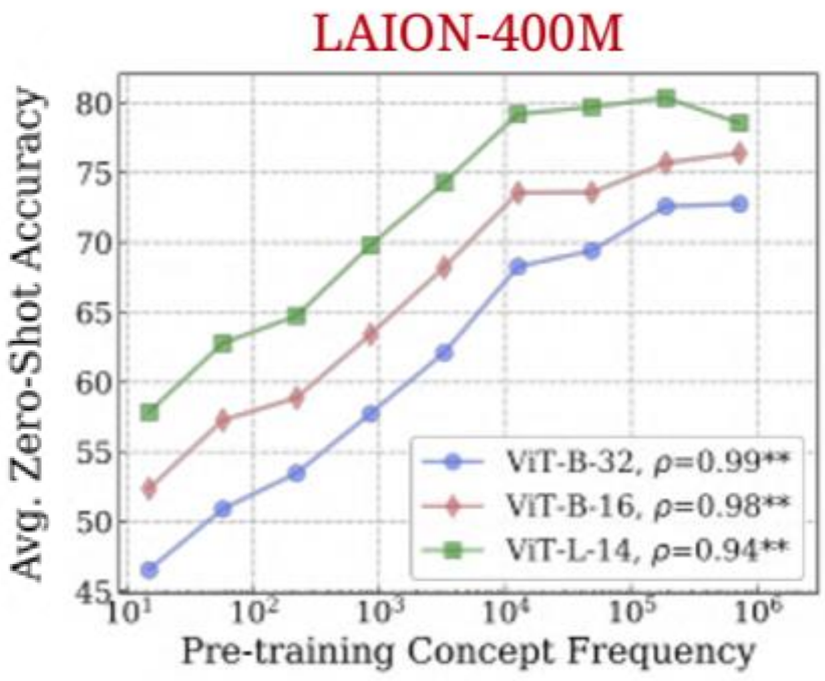

**Classical, supervised scenario**




**Foundation models**

# Conformal Prediction for Zero-Shot Models

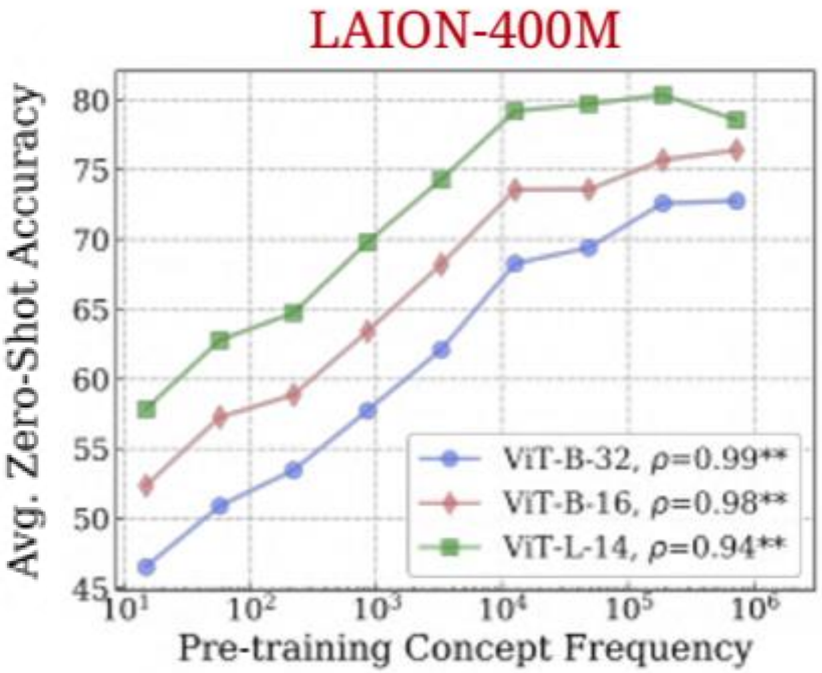
- **Transfer learning setting.**

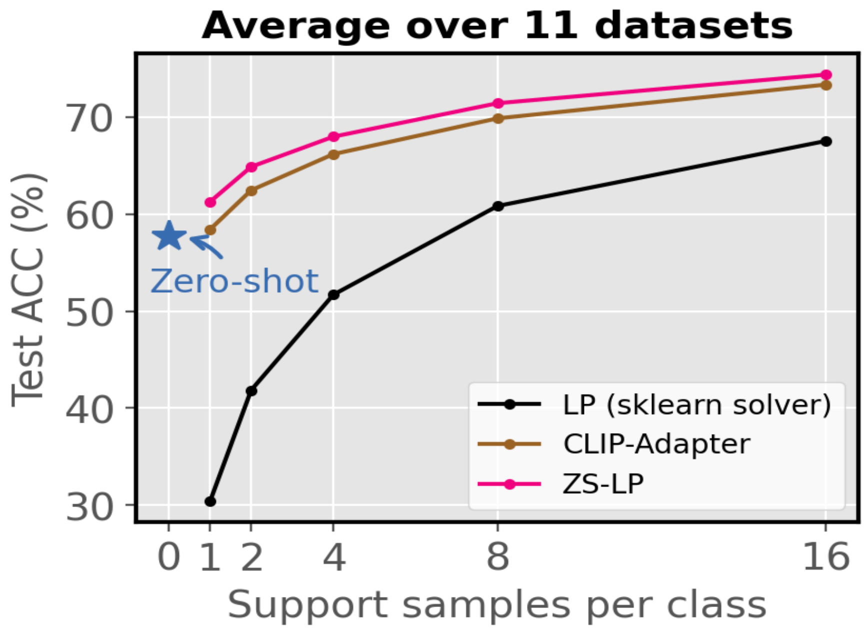



Plot 1 from [Udandaro et al., No “Zero-Shot” Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance, NeurIPS 2024]

# Conformal Prediction for Zero-Shot Models

- **Transfer learning setting.**




Tackled trough few-shot Linear Probing

Plot 1 from [Udandaro et al., No “Zero-Shot” Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance, NeurIPS 2024]
Plot 2 from [Silva-Rodríguez et al., A Closer Look at the Few-Shot Adaptation of Large Vision-Language Models, CVPR 2024]

# Conformal Prediction for Zero-Shot Models

- **Can we adapt and conformalize using the same data?**

# Conformal Prediction for Zero-Shot Models


- **Can we adapt and conformalize using the same data?**
  - **Training a Linear Probe on the logit space**

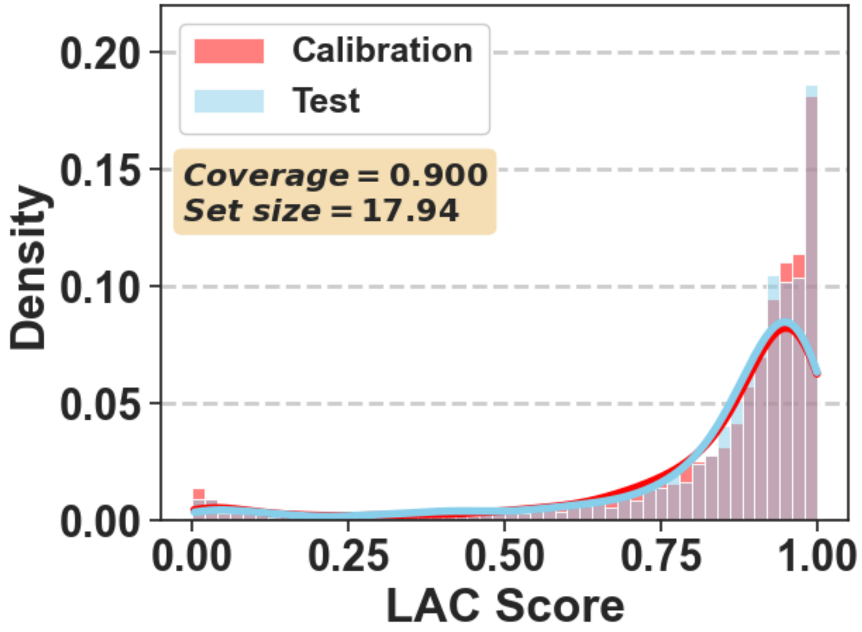
$$\mathcal{D}_{\text{cal}} = \{(\mathbf{l}_i, y_i)\}_{i=1}^{N} \qquad \mathcal{D}_{\text{test}} = \{(\mathbf{l}_i, )\}_{i=N+1}^{N+M}$$

- **New class prototypes** on the **logit projections** are defined $\mathbf{W} \in \mathbb{R}^{K \times K}$ .
- These obtain new class scores based on the **temperature-scaled Euclidean distance** $l'_k = -\frac{\tau^{\text{LP}}}{2}||\mathbf{l} - \mathbf{w}_k||$ .
- **Using calibration data**, optimize the class prototypes to **minimize cross-entropy loss.**

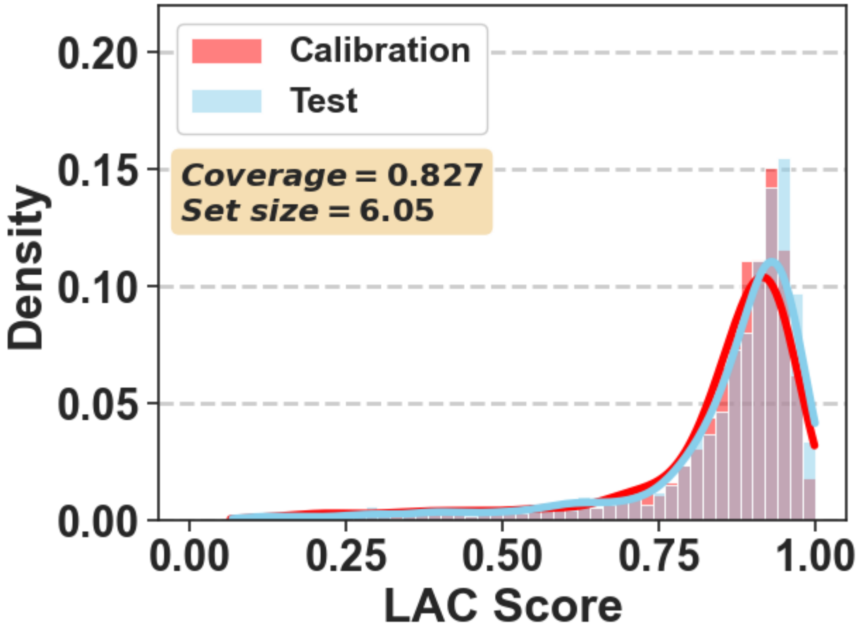
$$\min_{\mathbf{W}} \; -\frac{1}{NK}\sum_{i=1}^{I}\sum_{k=1}^{K} y_{ik} \log p_{ik},$$

# Conformal Prediction for Zero-Shot Models

- **Can we adapt and conformalize using the same data?**
  - **Conformal Prediction performance**


**Zero-shot**

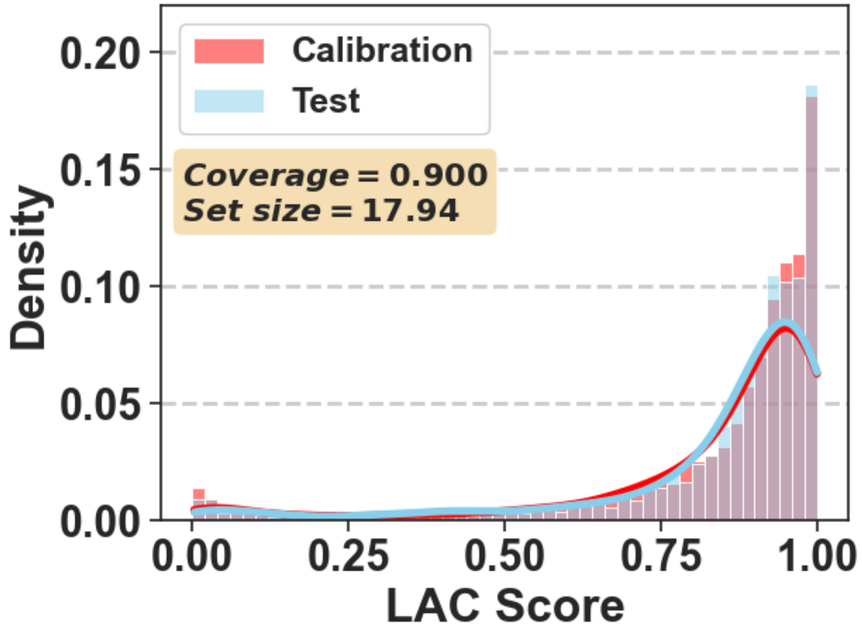



**Adapt + Conformalize in Calibration**

# Conformal Prediction for Zero-Shot Models

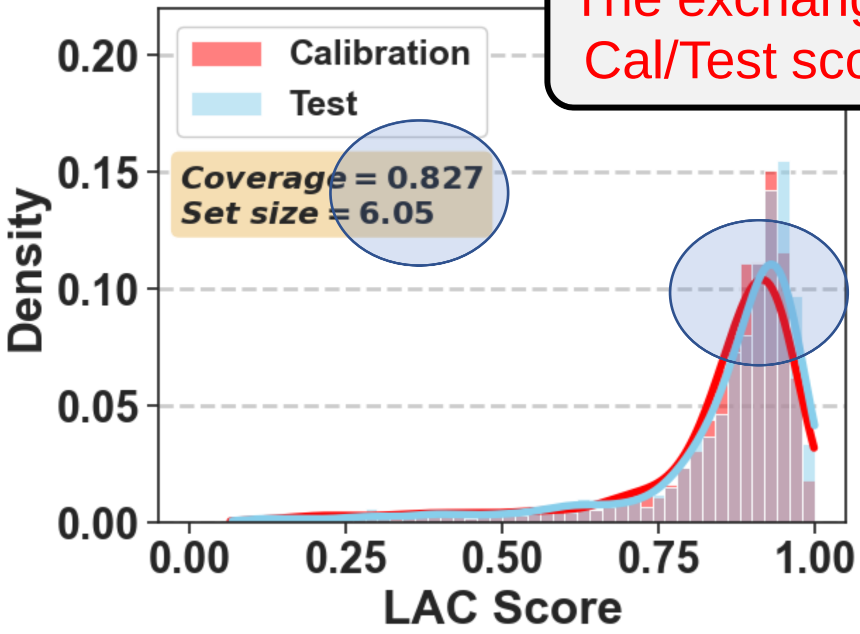
- **Can we adapt and conformalize using the same data?**


  - **Conformal Prediction performance**




**Zero-shot**

The exchangeability of the Cal/Test scores is broken




**Adapt + Conformalize in Calibration**

# Conformal Prediction for Zero-Shot Models

- **Transfer Learning for Conformal Prediction**


**1. Does <u>not directly rely on Cal labels</u>.**

Unsupervised

$$\mathcal{D}_{\text{cal}} = \{(\mathbf{l}_i), )\}_{i=1}^{N}$$

# Conformal Prediction for Zero-Shot Models


- **Transfer Learning for Conformal Prediction**

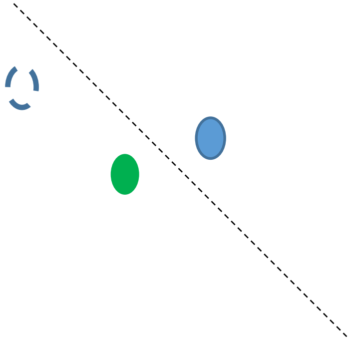
**1. Does <u>not directly rely on Cal labels</u>.**

Unsupervised

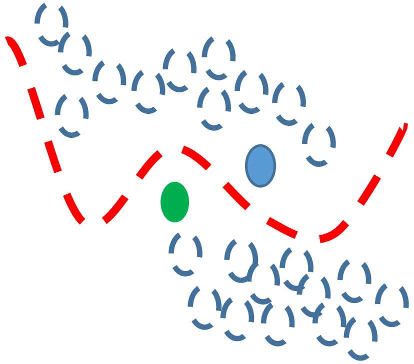
$$\mathcal{D}_{\text{cal}} = \{(\mathbf{l}_i), )\}_{i=1}^{N}$$

**2. <u>Jointly</u> modifies <u>Cal/Test score distributions</u>.**

Transductive



**Inductive**
One test sample
at a time



**Transductive**
Joint test-time
prediction

# Conformal Prediction for Zero-Shot Models


- **Transfer Learning for Conformal Prediction**

**1. Does <u>not directly rely on Cal labels</u>.**

Unsupervised

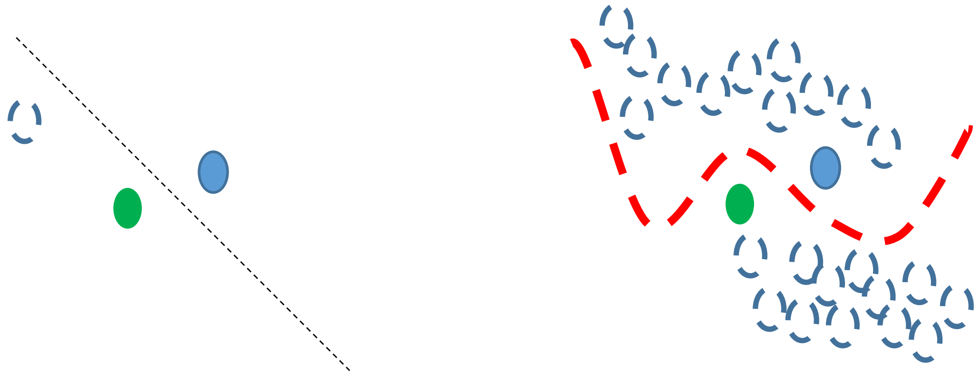
$$\mathcal{D}_{\text{cal}} = \{(\mathbf{l}_i),\, )\}_{i=1}^{N}$$

**2. <u>Jointly</u> modifies <u>Cal/Test score distributions</u>.**

Transductive

- Similarity matrix.

$$\mathbf{S} \in \mathbb{R}^{K \times (N+M)} = [l_{ki}]_{k=1,i=1}^{k=K,i=N+M}$$



**Inductive**
One test sample at a time

**Transductive**
Joint test-time prediction

# Conformal Prediction for Zero-Shot Models

- **Conformal Optimal Transport**

***Learning goal***: find the joint probability matrix (codes) which maximize the similarity assignment.

$$\max_{\mathbf{Q}\in\mathcal{Q}} \; tr(\mathbf{Q}^\top \mathbf{S})$$

where $\mathbf{Q} \in \mathbb{R}_+^{K\times(N+M)}$ is the assignment matrix, formed by individual codes for each sample, $\mathbf{q}_i$.

**Algorithm 1** Conf-OT conformal prediction.

**input:** calibration dataset $\mathcal{D}_{\text{cal}} = \{(l_i, y_i)\}_{i=1}^{N}$, query set $\mathcal{D}_{\text{test}} = \{(l_i)\}_{i=N+1}^{N+M}$, non-conformity score function $\mathcal{S}$, error level $\alpha$, entropic weight $\tau$, iterations $T$.

// **Block 1.** - Transductive transfer learning.

// **Step 1.1.** - Init. optimal transport problem.

$\mathbf{S} \in \mathbb{R}^{K\times(N+M)} = [l_{ki}]_{k=1,i=1}^{k=K,i=N+M}$ // Sim. matrix.

$\mathbf{m} = \frac{1}{N}\sum_1^N \mathbf{y}_i^{\text{ohe}}$ // Label-marginal.

$\mathbf{u}_{(\mathbf{N+M})} = \frac{1}{(N+M)}\mathbf{1}_{(N+M)}$ // Sample marginal.

// **Step 1.2.** - Compute renormalization vectors.

$\mathbf{Q}^{(0)} = (\exp(\mathbf{S}/\tau)/\sum(\exp(\mathbf{S}/\tau))$ // Init. codes.

$\mathbf{c}^{(0)} = \mathbf{1}_{(N+M)}$ // Init. renormalization vector.

**for** $t$ in $[1,\ldots,T]$ **do**

  $\mathbf{r}^{(t)} = \mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t-1)})$ // Eq. (9).

  $\mathbf{c}^{(t)} = \mathbf{u}_{(\mathbf{N+M})}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t)})$ // Eq. (10).

**end for**

// **Step 1.3.** - Compute codes.

$\mathbf{Q}^* = \text{Diag}(\mathbf{r}^{(T)})\mathbf{Q}^{(0)}\text{Diag}(\mathbf{c}^{(T)})$ // Transport codes.

$\mathbf{Q}^* = \mathbf{Q}^*\text{Diag}(1/\sum_k q_{ki}^*)$ // Normalize.

// **Block 2.** - Conformal prediction.

$\mathcal{D}_{\text{cal}} = \{({q_i^*}^\top, y_i)\}_{i=1}^{N}, \mathcal{D}_{\text{test}} = \{({q_i^*}^\top)\}_{i=N+1}^{N+M}$

// **Step 2.1.** - $1-\alpha$ non-conformity score quantile.

$\{s_i\}_{i=1}^N = \{\mathcal{S}({q_i^*}^\top, y_i)\}_{i=1}^N$ // Non-conformity scores.

$\hat{s} \leftarrow \{s_i\}_{i=1}^N, \alpha$ // Search threshold - Eq. (3).

// **Step 2.2.** - Create query sets.

**return:** $\{\mathcal{C}({q_i^*}^\top)\}_{i=N+1}^{M}$ // Eq. (4).

# Conformal Prediction for Zero-Shot Models

- **Conformal Optimal Transport**

***Learning goal***: find the joint probability matrix (codes) which maximize the similarity assignment.

$$\max_{\mathbf{Q}\in\mathcal{Q}} \; tr(\mathbf{Q}^\top\mathbf{S})$$

where $\mathbf{Q}\in\mathbb{R}_+^{K\times(N+M)}$ is the assignment matrix, formed by individual codes for each sample, $\mathbf{q}_i$.

More concretely, $\mathbf{Q}$ is restricted to be an element of the transportation polytope:

$$\mathcal{Q}=\{\mathbf{Q}\mid\mathbf{Q}\mathbf{1}_{(N+M)}=\mathbf{m},\mathbf{Q}^\top\mathbf{1}_K=\mathbf{u}_{(\mathbf{N+M})}\}$$

**Algorithm 1** Conf-OT conformal prediction.

**input:** calibration dataset $\mathcal{D}_{\text{cal}}=\{(l_i,y_i)\}_{i=1}^N$, query set $\mathcal{D}_{\text{test}}=\{(l_i)\}_{i=N+1}^{N+M}$, non-conformity score function $\mathcal{S}$, error level $\alpha$, entropic weight $\tau$, iterations $T$.
// **Block 1.** - Transductive transfer learning.
// **Step 1.1.** - Init. optimal transport problem.
$\mathbf{S}\in\mathbb{R}^{K\times(N+M)}=[l_{ki}]_{k=1,i=1}^{k=K,i=N+M}$ // Sim. matrix.
$\mathbf{m}=\frac{1}{N}\sum_1^N\mathbf{y}_i^{\text{obs}}$ // Label-marginal.
$\mathbf{u}_{(\mathbf{N+M})}=\frac{1}{(N+M)}\mathbf{1}_{(N+M)}$ // Sample marginal.
// **Step 1.2.** - Compute renormalization vectors.
$\mathbf{Q}^{(0)}=(\exp(\mathbf{S}/\tau)/\sum(\exp(\mathbf{S}/\tau))$ // Init. codes.
$\mathbf{c}^{(0)}=\mathbf{1}_{(N+M)}$ // Init. renormalization vector.
**for** $t$ in $[1,\ldots,T]$ **do**
    $\mathbf{r}^{(t)}=\mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t-1)})$ // Eq. (9).
    $\mathbf{c}^{(t)}=\mathbf{u}_{(\mathbf{N+M})}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t)})$ // Eq. (10).
**end for**
// **Step 1.3.** - Compute codes.
$\mathbf{Q}^*=\text{Diag}(\mathbf{r}^{(T)})\mathbf{Q}^{(0)}\text{Diag}(\mathbf{c}^{(T)})$ // Transport codes.
$\mathbf{Q}^*=\mathbf{Q}^*\text{Diag}(1/\sum_k q_{ki}^*)$ // Normalize.
// **Block 2.** - Conformal prediction.
$\mathcal{D}_{\text{cal}}=\{(q_i^{*\top},y_i)\}_{i=1}^N,\mathcal{D}_{\text{test}}=\{(q_i^{*\top})\}_{i=N+1}^{N+M}$
// **Step 2.1.** - $1-\alpha$ non-conformity score quantile.
$\{s_i\}_{i=1}^N=\{\mathcal{S}(q_i^{*\top},y_i)\}_{i=1}^N$ // Non-conformity scores.
$\hat{s}\leftarrow\{s_i\}_{i=1}^N,\alpha$ // Search threshold - Eq. (3).
// **Step 2.2.** - Create query sets.
**return:** $\{\mathcal{C}(q_i^{*\top})\}_{i=N+1}^M$ // Eq. (4).

# Conformal Prediction for Zero-Shot Models

- **Conformal Optimal Transport**

***<u>Learning goal</u>***: find the joint probability matrix (codes) which maximize the similarity assignment.

$$\max_{\mathbf{Q}\in\mathcal{Q}} \; tr(\mathbf{Q}^\top \mathbf{S})$$

where $\mathbf{Q} \in \mathbb{R}_{+}^{K\times(N+M)}$ is the assignment matrix, formed by individual codes for each sample, $\mathbf{q}_i$.

More concretely, $\mathbf{Q}$ is restricted to be an element of the transportation polytope:

$$\mathcal{Q} = \{\mathbf{Q} \mid \mathbf{Q}\mathbf{1}_{(N+M)} = \mathbf{m}, \mathbf{Q}^\top \mathbf{1}_K = \mathbf{u}_{(\mathbf{N+M})}\}$$

marginals!

**Algorithm 1** Conf-OT conformal prediction.

**input:** calibration dataset $\mathcal{D}_{\text{cal}} = \{(l_i, y_i)\}_{i=1}^{N}$, query set $\mathcal{D}_{\text{test}} = \{(l_i)\}_{i=N+1}^{N+M}$, non-conformity score function $\mathcal{S}$, error level $\alpha$, entropic weight $\tau$, iterations $T$.
// **Block 1.** - Transductive transfer learning.
// **Step 1.1.** - Init. optimal transport problem.
$\mathbf{S} \in \mathbb{R}^{K\times(N+M)} = [l_{ki}]_{k=1,i=1}^{k=K,i=N+M}$ // Sim. matrix.
$\mathbf{m} = \frac{1}{N}\sum_1^N \mathbf{y}_i^{\text{ohe}}$ // Label-marginal.
$\mathbf{u}_{(\mathbf{N+M})} = \frac{1}{(N+M)}\mathbf{1}_{(N+M)}$ // Sample marginal.
// **Step 1.2.** - Compute renormalization vectors.
$\mathbf{Q}^{(0)} = (\exp(\mathbf{S}/\tau)/\sum(\exp(\mathbf{S}/\tau))$ // Init. codes.
$\mathbf{c}^{(0)} = \mathbf{1}_{(N+M)}$ // Init. renormalization vector.
**for** $t$ in $[1,\ldots,T]$ **do**
  $\mathbf{r}^{(t)} = \mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t-1)})$ // Eq. (9).
  $\mathbf{c}^{(t)} = \mathbf{u}_{(\mathbf{N+M})}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t)})$ // Eq. (10).
**end for**
// **Step 1.3.** - Compute codes.
$\mathbf{Q}^* = \text{Diag}(\mathbf{r}^{(T)})\mathbf{Q}^{(0)}\text{Diag}(\mathbf{c}^{(T)})$ // Transport codes.
$\mathbf{Q}^* = \mathbf{Q}^*\text{Diag}(1/\sum_k q_{ki}^*)$ // Normalize.
// **Block 2.** - Conformal prediction.
$\mathcal{D}_{\text{cal}} = \{({q_i^*}^\top, y_i)\}_{i=1}^{N}, \mathcal{D}_{\text{test}} = \{({q_i^*}^\top)\}_{i=N+1}^{N+M}$
// **Step 2.1.** - $1-\alpha$ non-conformity score quantile.
$\{s_i\}_{i=1}^{N} = \{\mathcal{S}({q_i^*}^\top, y_i)\}_{i=1}^{N}$ // Non-conformity scores.
$\hat{s} \leftarrow \{s_i\}_{i=1}^{N}, \alpha$ // Search threshold - Eq. (3).
// **Step 2.2.** - Create query sets.
**return:** $\{\mathcal{C}({q_i^*}^\top)\}_{i=N+1}^{M}$ // Eq. (4).

# Conformal Prediction for Zero-Shot Models

- **Conformal Optimal Transport**

***Optimization*:** We solve the linear program trough the efficient **Sinkhorn algorithm**, which incorporates an **entropic-constraint**.

$$\max_{\mathbf{Q}\in\mathcal{Q}} \; tr(\mathbf{Q}^\top \mathbf{S}) + \varepsilon \mathcal{H}(\mathbf{Q})$$

**Algorithm 1** Conf-OT conformal prediction.

**input:** calibration dataset $\mathcal{D}_{\text{cal}} = \{(l_i, y_i)\}_{i=1}^{N}$, query set $\mathcal{D}_{\text{test}} = \{(l_i)\}_{i=N+1}^{N+M}$, non-conformity score function $\mathcal{S}$, error level $\alpha$, entropic weight $\tau$, iterations $T$.
// **Block 1.** - Transductive transfer learning.
// **Step 1.1.** - Init. optimal transport problem.
$\mathbf{S} \in \mathbb{R}^{K\times(N+M)} = [l_{ki}]_{k=1,i=1}^{k=K,i=N+M}$ // Sim. matrix.
$\mathbf{m} = \frac{1}{N}\sum_1^N \mathbf{y}_i^{\text{ohe}}$ // Label-marginal.
$\mathbf{u}_{(\mathbf{N+M})} = \frac{1}{(N+M)}\mathbf{1}_{(N+M)}$ // Sample marginal.
// **Step 1.2.** - Compute renormalization vectors.
$\mathbf{Q}^{(0)} = (\exp(\mathbf{S}/\tau)/\sum(\exp(\mathbf{S}/\tau))$ // Init. codes.
$\mathbf{c}^{(0)} = \mathbf{1}_{(N+M)}$ // Init. renormalization vector.
**for** $t$ in $[1,\dots,T]$ **do**
  $\mathbf{r}^{(t)} = \mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t-1)})$ // Eq. (9).
  $\mathbf{c}^{(t)} = \mathbf{u}_{(\mathbf{N+M})}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t)})$ // Eq. (10).
**end for**
// **Step 1.3.** - Compute codes.
$\mathbf{Q}^* = \text{Diag}(\mathbf{r}^{(T)})\mathbf{Q}^{(0)}\text{Diag}(\mathbf{c}^{(T)})$ // Transport codes.
$\mathbf{Q}^* = \mathbf{Q}^*\text{Diag}(1/\sum_k q_{ki}^*)$ // Normalize.
// **Block 2.** - Conformal prediction.
$\mathcal{D}_{\text{cal}} = \{(q_i^{*\top}, y_i)\}_{i=1}^{N}, \mathcal{D}_{\text{test}} = \{(q_i^{*\top})\}_{i=N+1}^{N+M}$
// **Step 2.1.** - $1-\alpha$ non-conformity score quantile.
$\{s_i\}_{i=1}^{N} = \{\mathcal{S}(q_i^{*\top}, y_i)\}_{i=1}^{N}$ // Non-conformity scores.
$\hat{s} \leftarrow \{s_i\}_{i=1}^{N}, \alpha$ // Search threshold - Eq. (3).
// **Step 2.2.** - Create query sets.
**return:** $\{\mathcal{C}(q_i^{*\top})\}_{i=N+1}^{M}$ // Eq. (4).

[Cuturi et al., Sinkhorn distances: Lightspeed computation of optimal transport, NeurIPS 2013]

# Conformal Prediction for Zero-Shot Models

- **Conformal Optimal Transport**

***Optimization***: We solve the linear program trough the efficient **Sinkhorn algorithm**, which incorporates an **entropic-constraint**.

$$\max_{\mathbf{Q}\in\mathcal{Q}} \; tr(\mathbf{Q}^\top \mathbf{S}) + \varepsilon \mathcal{H}(\mathbf{Q})$$

Now, the soft codes $\mathbf{Q}^*$ are the solution of the optimization problem, which can be efficiently optimized by computing marginal-renormalization vectors, such that:

$$\mathbf{Q}^* = \mathrm{Diag}(\mathbf{r}^{(t)})\mathbf{Q}^{(0)}\mathrm{Diag}(\mathbf{c}^{(t)})$$

**Algorithm 1** Conf-OT conformal prediction.

**input:** calibration dataset $\mathcal{D}_{\text{cal}} = \{(l_i, y_i)\}_{i=1}^N$, query set $\mathcal{D}_{\text{test}} = \{(l_i)\}_{i=N+1}^{N+M}$, non-conformity score function $\mathcal{S}$, error level $\alpha$, entropic weight $\tau$, iterations $T$.
// **Block 1.** - Transductive transfer learning.
// **Step 1.1.** - Init. optimal transport problem.
$\mathbf{S} \in \mathbb{R}^{K\times(N+M)} = [l_{ki}]_{k=1,i=1}^{k=K,i=N+M}$ // Sim. matrix.
$\mathbf{m} = \frac{1}{N}\sum_1^N \mathbf{y}_i^{\text{ohe}}$ // Label-marginal.
$\mathbf{u}_{(\mathbf{N+M})} = \frac{1}{(N+M)}\mathbf{1}_{(N+M)}$ // Sample marginal.
// **Step 1.2.** - Compute renormalization vectors.
$\mathbf{Q}^{(0)} = (\exp(\mathbf{S}/\tau)/\sum(\exp(\mathbf{S}/\tau))$ // Init. codes.
$\mathbf{c}^{(0)} = \mathbf{1}_{(N+M)}$ // Init. renormalization vector.
**for** $t$ in $[1, \ldots, T]$ **do**
  $\mathbf{r}^{(t)} = \mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t-1)})$ // Eq. (9).
  $\mathbf{c}^{(t)} = \mathbf{u}_{(\mathbf{N+M})}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t)})$ // Eq. (10).
**end for**
// **Step 1.3.** - Compute codes.
$\mathbf{Q}^* = \mathrm{Diag}(\mathbf{r}^{(T)})\mathbf{Q}^{(0)}\mathrm{Diag}(\mathbf{c}^{(T)})$ // Transport codes.
$\mathbf{Q}^* = \mathbf{Q}^*\mathrm{Diag}(1/\sum_k q^*_{ki})$ // Normalize.
// **Block 2.** - Conformal prediction.
$\mathcal{D}_{\text{cal}} = \{({q_i^*}^\top, y_i)\}_{i=1}^N$, $\mathcal{D}_{\text{test}} = \{({q_i^*}^\top)\}_{i=N+1}^{N+M}$
// **Step 2.1.** - $1-\alpha$ non-conformity score quantile.
$\{s_i\}_{i=1}^N = \{\mathcal{S}({q_i^*}^\top, y_i)\}_{i=1}^N$ // Non-conformity scores.
$\hat{s} \leftarrow \{s_i\}_{i=1}^N, \alpha$ // Search threshold - Eq. (3).
// **Step 2.2.** - Create query sets.
**return:** $\{\mathcal{C}({q_i^*}^\top)\}_{i=N+1}^M$ // Eq. (4).

[Cuturi et al., Sinkhorn distances: Lightspeed computation of optimal transport, NeurIPS 2013]

# Conformal Prediction for Zero-Shot Models

## ▪ Conformal Optimal Transport

***Optimization***: We solve the linear program trough the efficient **Sinkhorn algorithm**, which incorporates an **entropic-constraint**.

$$\max_{\mathbf{Q}\in\mathcal{Q}} \; tr(\mathbf{Q}^\top \mathbf{S}) + \varepsilon\mathcal{H}(\mathbf{Q})$$

Norm S as initial Q

Now, the soft codes $\mathbf{Q}^*$ are the solution of the optimization problem, which can be efficiently optimized by computing marginal-renormalization vectors, such that:

$$\mathbf{Q}^* = \mathrm{Diag}(\mathbf{r}^{(t)})\mathbf{Q}^{(0)}\mathrm{Diag}(\mathbf{c}^{(t)})$$

**Algorithm 1** Conf-OT conformal prediction.

**input:** calibration dataset $\mathcal{D}_{\text{cal}} = \{(l_i, y_i)\}_{i=1}^{N}$, query set $\mathcal{D}_{\text{test}} = \{(l_i)\}_{i=N+1}^{N+M}$, non-conformity score function $\mathcal{S}$, error level $\alpha$, entropic weight $\tau$, iterations $T$.
// Block 1. - Transductive transfer learning.
// Step 1.1. - Init. optimal transport problem.
$\mathbf{S} \in \mathbb{R}^{K\times(N+M)} = [l_{ki}]_{k=1,i=1}^{k=K,i=N+M}$ // Sim. matrix.
$\mathbf{m} = \frac{1}{N}\sum_1^N \mathbf{y}_i^{\text{ohe}}$ // Label-marginal.
$\mathbf{u}_{(\mathbf{N+M})} = \frac{1}{(N+M)}\mathbf{1}_{(N+M)}$ // Sample marginal.
// Step 1.2. - Compute renormalization vectors.
$\mathbf{Q}^{(0)} = (\exp(\mathbf{S}/\tau)/\sum(\exp(\mathbf{S}/\tau))$ // Init. codes.
$\mathbf{c}^{(0)} = \mathbf{1}_{(N+M)}$ // Init. renormalization vector.
**for** $t$ in $[1,\dots,T]$ **do**
  $\mathbf{r}^{(t)} = \mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t-1)})$ // Eq. (9).
  $\mathbf{c}^{(t)} = \mathbf{u}_{(\mathbf{N+M})}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t)})$ // Eq. (10).
**end for**
// Step 1.3. - Compute codes.
$\mathbf{Q}^* = \mathrm{Diag}(\mathbf{r}^{(T)})\mathbf{Q}^{(0)}\mathrm{Diag}(\mathbf{c}^{(T)})$ // Transport codes.
$\mathbf{Q}^* = \mathbf{Q}^*\mathrm{Diag}(1/\sum_k q_{ki}^*)$ // Normalize.
// Block 2. - Conformal prediction.
$\mathcal{D}_{\text{cal}} = \{(q_i^{*\top}, y_i)\}_{i=1}^{N}, \mathcal{D}_{\text{test}} = \{(q_i^{*\top})\}_{i=N+1}^{N+M}$
// Step 2.1. - $1-\alpha$ non-conformity score quantile.
$\{s_i\}_{i=1}^{N} = \{\mathcal{S}(q_i^{*\top}, y_i)\}_{i=1}^{N}$ // Non-conformity scores.
$\hat{s} \leftarrow \{s_i\}_{i=1}^{N}, \alpha$ // Search threshold - Eq. (3).
// Step 2.2. - Create query sets.
**return:** $\{\mathcal{C}(q_i^{*\top})\}_{i=N+1}^{M}$ // Eq. (4).

[Cuturi et al., Sinkhorn distances: Lightspeed computation of optimal transport, NeurIPS 2013]

# Conformal Prediction for Zero-Shot Models

- **Conformal Optimal Transport**

***Optimization***: We solve the linear program trough the efficient **Sinkhorn algorithm**, which incorporates an **entropic-constraint**.

$$\max_{\mathbf{Q}\in\mathcal{Q}} \; tr(\mathbf{Q}^\top \mathbf{S}) + \varepsilon \mathcal{H}(\mathbf{Q})$$

Incorporate prior label marginal

Now, the soft codes $\mathbf{Q}^*$ are the solution of the optimization problem, which can be efficiently optimized by computing marginal-renormalization vectors, such that:

$$\mathbf{Q}^* = \mathrm{Diag}(\mathbf{r}^{(t)})\mathbf{Q}^{(0)}\mathrm{Diag}(\mathbf{c}^{(t)})$$

**Algorithm 1** Conf-OT conformal prediction.

```
input: calibration dataset D_cal = {(l_i, y_i)}_{i=1}^N, query set D_test = {(l_i)}_{i=N+1}^{N+M}, non-conformity score function S, error level α, entropic weight τ, iter…
// Block 1. - Transductive transfer lear…
// Step 1.1. - Init. optimal transport pr…
S ∈ R^{K×(N+M)} = [l_ki]_{k=1,i=1}^{k=K,i=N+M} //…
m = 1/N Σ_1^N y_i^ohe // Label-marginal.
u_(N+M) = 1/(N+M) 1_(N+M) // Sample marginal.
// Step 1.2. - Compute renormalization vectors.
Q^(0) = (exp(S/τ) / Σ(exp(S/τ)) // Init. codes.
c^(0) = 1_(N+M) // Init. renormalization vector.
for t in [1, ..., T] do
    r^(t) = m/(Q^(0) c^(t-1)) // Eq. (9).
    c^(t) = u_(N+M)/(Q^(0) r^(t)) // Eq. (10).
end for
// Step 1.3. - Compute codes.
Q* = Diag(r^(T)) Q^(0) Diag(c^(T)) // Transport codes.
Q* = Q* Diag(1/Σ_k q*_ki) // Normalize.
// Block 2. - Conformal prediction.
D_cal = {(q_i*^T, y_i)}_{i=1}^N, D_test = {(q_i*^T)}_{i=N+1}^{N+M}
// Step 2.1. - 1 − α non-conformity score quantile.
{s_i}_{i=1}^N = {S(q_i*^T, y_i)}_{i=1}^N // Non-conformity scores.
ŝ ← {s_i}_{i=1}^N, α // Search threshold - Eq. (3).
// Step 2.2. - Create query sets.
return: {C(q_i*^T)}_{i=N+1}^M // Eq. (4).
```

Divide rows by initial label marginal

[Cuturi et al., Sinkhorn distances: Lightspeed computation of optimal transport, NeurIPS 2013]

# Conformal Prediction for Zero-Shot Models

- **Conformal Optimal Transport**

***Optimization***: We solve the linear program trough the efficient **Sinkhorn algorithm**, which incorporates an **entropic-constraint**.

$$\max_{\mathbf{Q}\in\mathcal{Q}} \; tr(\mathbf{Q}^\top \mathbf{S}) + \varepsilon \mathcal{H}(\mathbf{Q})$$

Incorporate prior sample marginal

Now, the soft codes $\mathbf{Q}^*$ are the solution of the optimization problem, which can be efficiently optimized by computing marginal-renormalization vectors, such that:

$$\mathbf{Q}^* = \mathrm{Diag}(\mathbf{r}^{(t)})\mathbf{Q}^{(0)}\mathrm{Diag}(\mathbf{c}^{(t)})$$

**Algorithm 1** Conf-OT conformal prediction.

```
input: calibration dataset D_cal = {(l_i, y_i)}_{i=1}^N, query
  set D_test = {(l_i)}_{i=N+1}^{N+M}, non-conformity score function
  S, error level α, entropic weight τ, it
// Block 1. - Transductive transfer lea
// Step 1.1. - Init. optimal transport p
S ∈ R^{K×(N+M)} = [l_ki]_{k=1,i=1}^{k=K,i=N+M}
m = 1/N Σ_1^N y_i^obs // Label-marginal.
u_(N+M) = 1/(N+M) 1_(N+M) // Sample marginal
// Step 1.2. - Compute renormalization vectors.
Q^(0) = (exp(S/τ)/Σ(exp(S/τ)) // Init. codes.
c^(0) = 1_(N+M) // Init. renormalization vector.
for t in [1,...,T] do
    r^(t) = m/(Q^(0) c^(t-1)) // Eq. (9).
    c^(t) = u_(N+M)/(Q^(0) r^(t)) // Eq. (10).
end for
// Step 1.3. - Compute codes.
Q* = Diag(r^(T)) Q^(0) Diag(c^(T)) // Transport codes.
Q* = Q* Diag(1/Σ_k q*_ki) // Normalize.
// Block 2. - Conformal prediction.
D_cal = {(q_i*^T, y_i)}_{i=1}^N, D_test = {(q_i*^T)}_{i=N+1}^{N+M}
// Step 2.1. - 1 − α non-conformity score quantile.
{s_i}_{i=1}^N = {S(q_i*^T, y_i)}_{i=1}^N // Non-conformity scores.
ŝ ← {s_i}_{i=1}^N, α // Search threshold - Eq. (3).
// Step 2.2. - Create query sets.
return: {C(q_i*^T)}_{i=N+1}^M // Eq. (4).
```

Divide columns by observed sample marginal

[Cuturi et al., Sinkhorn distances: Lightspeed computation of optimal transport, NeurIPS 2013]

# Conformal Prediction for Zero-Shot Models

## ▪ Conformal Optimal Transport

***Optimization***: We solve the linear program trough the efficient **Sinkhorn algorithm**, which incorporates an **entropic-constraint**.

$$\max_{\mathbf{Q}\in\mathcal{Q}} \; tr(\mathbf{Q}^\top \mathbf{S}) + \varepsilon\mathcal{H}(\mathbf{Q})$$

Now, the soft codes $\mathbf{Q}^*$ are the solution of the optimization problem, which can be efficiently optimized by computing marginal-renormalization vectors, such that:

$$\mathbf{Q}^* = \mathrm{Diag}(\mathbf{r}^{(t)})\mathbf{Q}^{(0)}\mathrm{Diag}(\mathbf{c}^{(t)})$$

**Algorithm 1** Conf-OT conformal prediction.

**input:** calibration dataset $\mathcal{D}_{\text{cal}} = \{(l_i, y_i)\}_{i=1}^N$, query set $\mathcal{D}_{\text{test}} = \{(l_i)\}_{i=N+1}^{N+M}$, non-conformity score function $\mathcal{S}$, error level $\alpha$, entropic weight $\tau$, iterations $T$.
// **Block 1.** - Transductive transfer learning.
// **Step 1.1.** - Init. optimal transport problem.
$\mathbf{S} \in \mathbb{R}^{K\times(N+M)} = [l_{ki}]_{k=1,i=1}^{k=K,i=N+M}$ // Sim. matrix.
$\mathbf{m} = \frac{1}{N}\sum_1^N \mathbf{y}_i^{\text{ohe}}$ // Label-marginal.
$\mathbf{u}_{(\mathbf{N+M})} = \frac{1}{(N+M)}\mathbf{1}_{(N+M)}$ // Sample marginal.
// **Step 1.2.** - Compute renormalization vectors.
$\mathbf{Q}^{(0)} = (\exp(\mathbf{S}/\tau)/\sum(\exp(\mathbf{S}/\tau))$ // Init. codes.
$\mathbf{c}^{(0)} = \mathbf{1}_{(N+M)}$ // Init. renormalization vector.
**for** $t$ in $[1,\ldots,T]$ **do**
    $\mathbf{r}^{(t)} = \mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t-1)})$ // Eq. (9).
    $\mathbf{c}^{(t)} = \mathbf{u}_{(\mathbf{N+M})}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t)})$ // Eq. (10).
**end for**
// **Step 1.3.** - Compute codes.
$\mathbf{Q}^* = \mathrm{Diag}(\mathbf{r}^{(T)})\mathbf{Q}^{(0)}\mathrm{Diag}(\mathbf{c}^{(T)})$ // Transport codes.
$\mathbf{Q}^* = \mathbf{Q}^*\mathrm{Diag}(1/\sum_k q_{ki}^*)$ // Normalize.
// **Block 2.** - Conformal prediction.
// **Step 2.2.** - Create query sets.
**return:** $\{\mathcal{C}(q_i^{*\top})\}_{i=N+1}^M$ // Eq. (4).

Apply renorm. vectors and sample-wise.

[Cuturi et al., Sinkhorn distances: Lightspeed computation of optimal transport, NeurIPS 2013]

# Conformal Prediction for Zero-Shot Models

- **Conformal Optimal Transport**

***Conformal prediction***: we follow the standard SCP setting using codes instead of the original probabilities.

$$\mathcal{D}_{\text{cal}} = \{(\mathbf{q}_i, y_i)\}_{i=1}^{N}$$

$$\mathcal{D}_{\text{test}} = \{(\mathbf{q}_i, )\}_{i=N+1}^{N+M}$$

**Algorithm 1** Conf-OT conformal prediction.

**input:** calibration dataset $\mathcal{D}_{\text{cal}} = \{(\mathbf{l}_i, y_i)\}_{i=1}^{N}$, query set $\mathcal{D}_{\text{test}} = \{(\mathbf{l}_i)\}_{i=N+1}^{N+M}$, non-conformity score function $\mathcal{S}$, error level $\alpha$, entropic weight $\tau$, iterations $T$.
// **Block 1.** - Transductive transfer learning.
// **Step 1.1.** - Init. optimal transport problem.
$\mathbf{S} \in \mathbb{R}^{K \times (N+M)} = [l_{ki}]_{k=1,i=1}^{k=K,i=N+M}$ // Sim. matrix.
$\mathbf{m} = \frac{1}{N}\sum_1^N \mathbf{y}_i^{\text{ohe}}$ // Label-marginal.
$\mathbf{u}_{(\mathbf{N+M})} = \frac{1}{(N+M)}\mathbf{1}_{(N+M)}$ // Sample marginal.
// **Step 1.2.** - Compute renormalization vectors.
$\mathbf{Q}^{(0)} = (\exp(\mathbf{S}/\tau)/\sum(\exp(\mathbf{S}/\tau))$ // Init. codes.
$\mathbf{c}^{(0)} = \mathbf{1}_{(N+M)}$ // Init. renormalization vector.
**for** $t$ in $[1, \dots, T]$ **do**
  $\mathbf{r}^{(t)} = \mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t-1)})$ // Eq. (9).
  $\mathbf{c}^{(t)} = \mathbf{u}_{(\mathbf{N+M})}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t)})$ // Eq. (10).
**end for**
// **Step 1.3.** - Compute codes.
$\mathbf{Q}^{*} = \text{Diag}(\mathbf{r}^{(T)})\mathbf{Q}^{(0)}\text{Diag}(\mathbf{c}^{(T)})$ // Transport codes.
$\mathbf{Q}^{*} = \mathbf{Q}^{*}\text{Diag}(1/\sum_k q_{ki}^{*})$ // Normalize.
// **Block 2.** - Conformal prediction.
$\mathcal{D}_{\text{cal}} = \{(q_i^{*\top}, y_i)\}_{i=1}^{N}$, $\mathcal{D}_{\text{test}} = \{(q_i^{*\top})\}_{i=N+1}^{N+M}$
// **Step 2.1.** - $1-\alpha$ non-conformity score quantile.
$\{s_i\}_{i=1}^{N} = \{\mathcal{S}(q_i^{*\top}, y_i)\}_{i=1}^{N}$ // Non-conformity scores.
$\hat{s} \leftarrow \{s_i\}_{i=1}^{N}, \alpha$ // Search threshold - Eq. (3).
// **Step 2.2.** - Create query sets.
**return:** $\{\mathcal{C}(q_i^{*\top})\}_{i=N+1}^{M}$ // Eq. (4).

# Conformal Prediction for Zero-Shot Models

- **Enhancing popular non-conformity scores**

| Method | | $\alpha = 0.10$ | | | $\alpha = 0.05$ | | |
|---|---|---|---|---|---|---|---|
| | Top-1↑ | Cov | Size↓ | CCV↓ | Cov. | Size↓ | CCV↓ |
| **CLIP ResNet-50** | | | | | | | |
| LAC [42] | 54.7 | 0.900 | 10.77 | 9.82 | 0.950 | 19.22 | 5.91 |
| w/ Conf-OT | **57.3**+2.6 | 0.900 | **8.61**-2.2 | **9.15**-0.7 | 0.951 | **15.53**-3.7 | **5.61**-0.3 |
| APS [54] | 54.7 | 0.900 | 16.35 | 8.36 | 0.950 | 26.50 | 5.34 |
| w/ 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 |
| w/ Conf-OT | **57.3**+2.6 | 0.900 | **11.17**-2.2 | **7.72**-0.7 | 0.950 | **17.24**-4.8 | **5.19**-0.3 |
| **CLIP ViT-B/16** | | | | | | | |
| LAC [42] | 63.8 | 0.899 | 5.52 | 10.37 | 0.950 | 10.24 | 6.14 |
| w/ Conf-OT | **66.7**+2.9 | 0.900 | **4.40**-1.1 | **9.48**-0.9 | 0.949 | **7.99**-2.3 | **5.80**-0.3 |
| APS [54] | 63.8 | 0.900 | 9.87 | 8.39 | 0.950 | 16.92 | 5.51 |
| w/ Conf-OT | **66.7**+2.9 | 0.899 | **7.64**-2.2 | **7.44**-1.0 | 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 |
| w/ Conf-OT | **66.7**+2.9 | 0.900 | **6.68**-1.4 | **7.48**-1.0 | 0.949 | **10.11**-2.6 | **5.16**-0.4 |

# Conformal Prediction for Zero-Shot Models

- **Enhancing popular non-conformity scores**

Valid coverage!

| Method | | $\alpha = 0.10$ | | | $\alpha = 0.05$ | | |
|---|---|---|---|---|---|---|---|
| | Top-1↑ | Cov | Size↓ | CCV↓ | Cov | Size↓ | CCV↓ |
| **CLIP ResNet-50** | | | | | | | |
| LAC [42] | 54.7 | 0.900 | 10.77 | 9.82 | 0.950 | 19.22 | 5.91 |
| w/ Conf-OT | **57.3**$_{+2.6}$ | 0.900 | **8.61**$_{-2.2}$ | **9.15**$_{-0.7}$ | 0.951 | **15.53**$_{-3.7}$ | **5.61**$_{-0.3}$ |
| APS [54] | 54.7 | 0.900 | 16.35 | 8.36 | 0.950 | 26.50 | 5.34 |
| w/ 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 |
| w/ Conf-OT | **57.3**$_{+2.6}$ | 0.900 | **11.17**$_{-2.2}$ | **7.72**$_{-0.7}$ | 0.950 | **17.24**$_{-4.8}$ | **5.19**$_{-0.3}$ |
| **CLIP ViT-B/16** | | | | | | | |
| LAC [42] | 63.8 | 0.899 | 5.52 | 10.37 | 0.950 | 10.24 | 6.14 |
| w/ Conf-OT | **66.7**$_{+2.9}$ | 0.900 | **4.40**$_{-1.1}$ | **9.48**$_{-0.9}$ | 0.949 | **7.99**$_{-2.3}$ | **5.80**$_{-0.3}$ |
| APS [54] | 63.8 | 0.900 | 9.87 | 8.39 | 0.950 | 16.92 | 5.51 |
| w/ Conf-OT | **66.7**$_{+2.9}$ | 0.899 | **7.64**$_{-2.2}$ | **7.44**$_{-1.0}$ | 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 |
| w/ Conf-OT | **66.7**$_{+2.9}$ | 0.900 | **6.68**$_{-1.4}$ | **7.48**$_{-1.0}$ | 0.949 | **10.11**$_{-2.6}$ | **5.16**$_{-0.4}$ |

15-20% better

# Conformal Prediction for Zero-Shot Models

- **Comparison to popular transductive methods**

| Method | | | | $\alpha = 0.10$ | | |
|---|---|---|---|---|---|---|
| | Top-1↑ | T | GPU | Cov. | Size↓ | CCV↓ |
| LAC [42] | 63.8 | **0.42** | - | 0.899 | 5.52 | 10.37 |
| $\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [6] | $64.7_{+0.9}$ | 11.05 | 8.24 | 0.899 | $8.30_{+2.8}$ | $10.41_{+0.0}$ |
| $\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [6] | $65.0_{+1.2}$ | 11.03 | 8.24 | 0.898 | $7.73_{+2.2}$ | $10.89_{+0.5}$ |
| TransCLIP [74] | $65.1_{+1.3}$ | 12.00 | 12.2 | **0.892** | $5.76_{+0.2}$ | $11.02_{+0.7}$ |
| Conf-OT | $\mathbf{66.7}_{+2.9}$ | 0.61 | - | 0.900 | $\mathbf{4.40}_{-1.1}$ | $\mathbf{9.48}_{-0.9}$ |

TIM from [Boudiaf et al., Transductive Information Maximization for Few-Shot Learning, NeurIPS 2020]
TransCLIP from [Zanella et al., Boosting Vision-Language Models with Transduction, NeurIPS 2024]

# Conformal Prediction for Zero-Shot Models

- **Comparison to popular transductive methods**

| Method | Top-1↑ | T | GPU | $\alpha = 0.10$ Cov. | Size↓ | CCV↓ |
|---|---|---|---|---|---|---|
| LAC [42] | 63.8 | **0.42** | - | 0.899 | 5.52 | 10.37 |
| $\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [6] | $64.7_{+0.9}$ | 11.05 | 8.24 | 0.899 | $8.30_{+2.8}$ | $10.41_{+0.0}$ |
| $\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [6] | $65.0_{+1.2}$ | 11.03 | 8.24 | 0.898 | $7.73_{+2.2}$ | $10.89_{+0.5}$ |
| TransCLIP [74] | $65.1_{+1.3}$ | 12.00 | 12.2 | 0.892 | $5.76_{+0.2}$ | $11.02_{+0.7}$ |
| Conf-OT | $\mathbf{66.7}_{+2.9}$ | 0.61 | - | 0.900 | $\mathbf{4.40}_{-1.1}$ | $\mathbf{9.48}_{-0.9}$ |

no improvement

training-free

Better than SoTA even in the discriminative aspect!

TIM from [Boudiaf et al., Transductive Information Maximization for Few-Shot Learning, NeurIPS 2020]
TransCLIP from [Zanella et al., Boosting Vision-Language Models with Transduction, NeurIPS 2024]

# Conformal Prediction for Zero-Shot Models

- **Evaluation of the data-efficiency**

| 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 |

**Robustness to small calibration sets**

| 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 |

**Robustness to small query inputs**

# Conformal Prediction for Zero-Shot Models

- **Evaluation of the data-efficiency**

| Method | Ratio Calib - Test | Top-1↑ | Cov. ($\alpha = 0.10$) | Size↓ ($\alpha = 0.10$) | CCV↓ ($\alpha = 0.10$) |
|---|---|---|---|---|---|
| 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 |

**Robustness to small calibration sets**

| Method | M | Top-1↑ | Cov. ($\alpha = 0.10$) | Size↓ ($\alpha = 0.10$) | CCV↓ ($\alpha = 0.10$) |
|---|---|---|---|---|---|
| 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 |

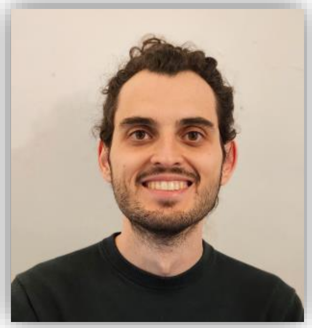
**Robustness to small query inputs**

# Conformal Prediction for Zero-Shot Models


Julio
Silva-Rodríguez

Ismail
Ben Ayed

Jose Dolz


ÉTS Montréal