Boosting Vision-Language Models with Transduction

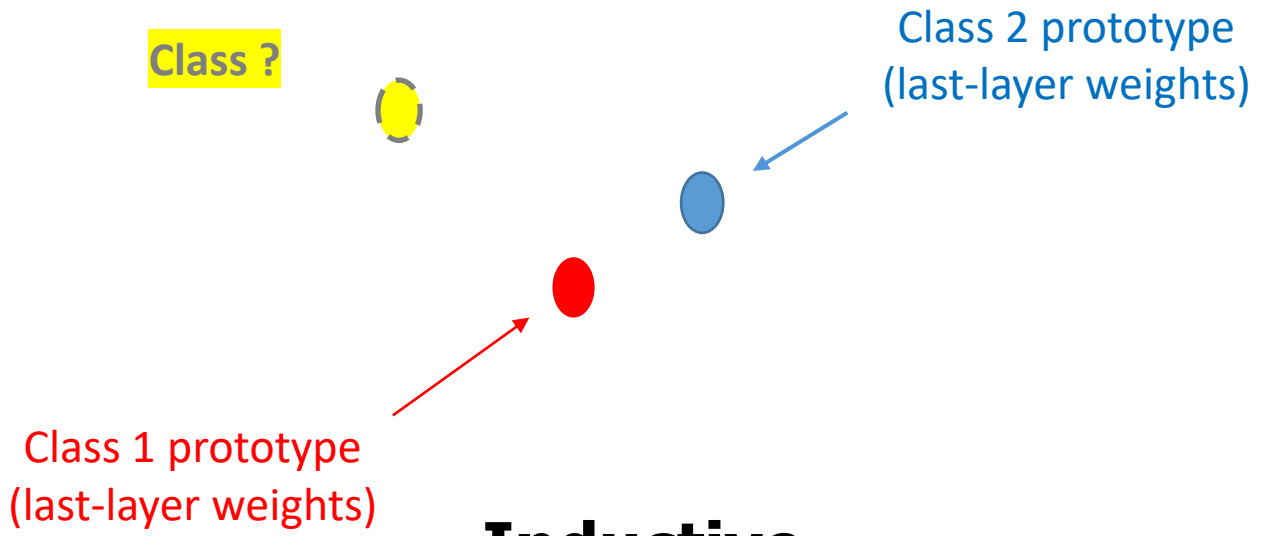
NeurIPS '24 - Spotlight

A joint work with

Maxime Zanella*, Benoît Gérin*, Ismail Ben Ayed

*Equal contributions and corresponding authors

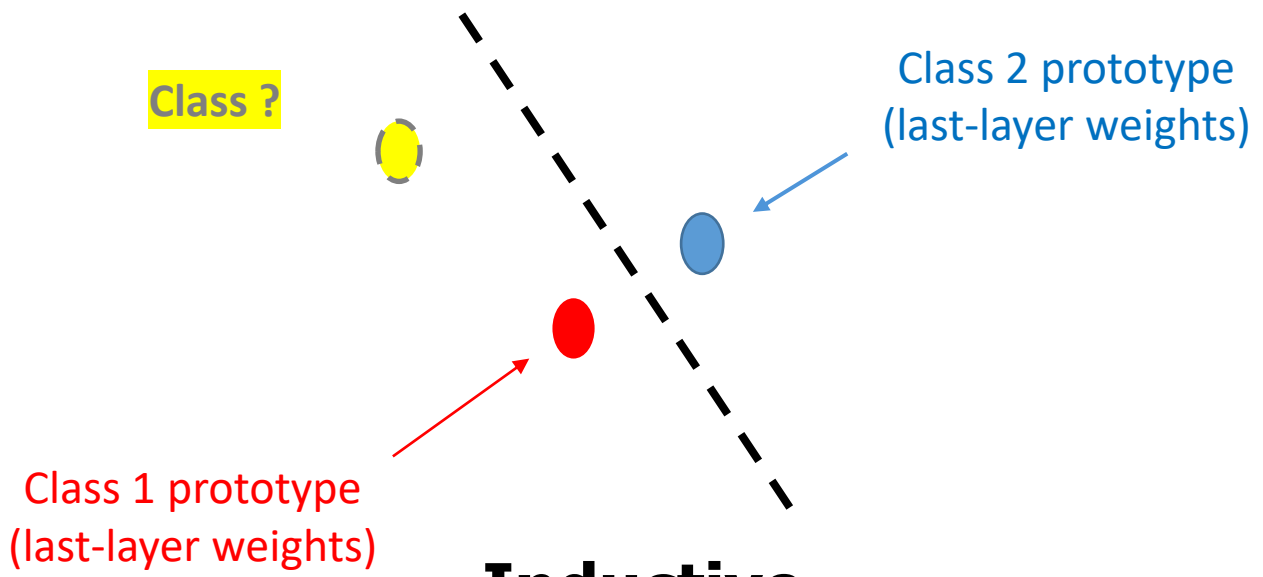
Transductive vs. inductive prediction



Inductive
(one test sample at a time)

Unlabeled instance

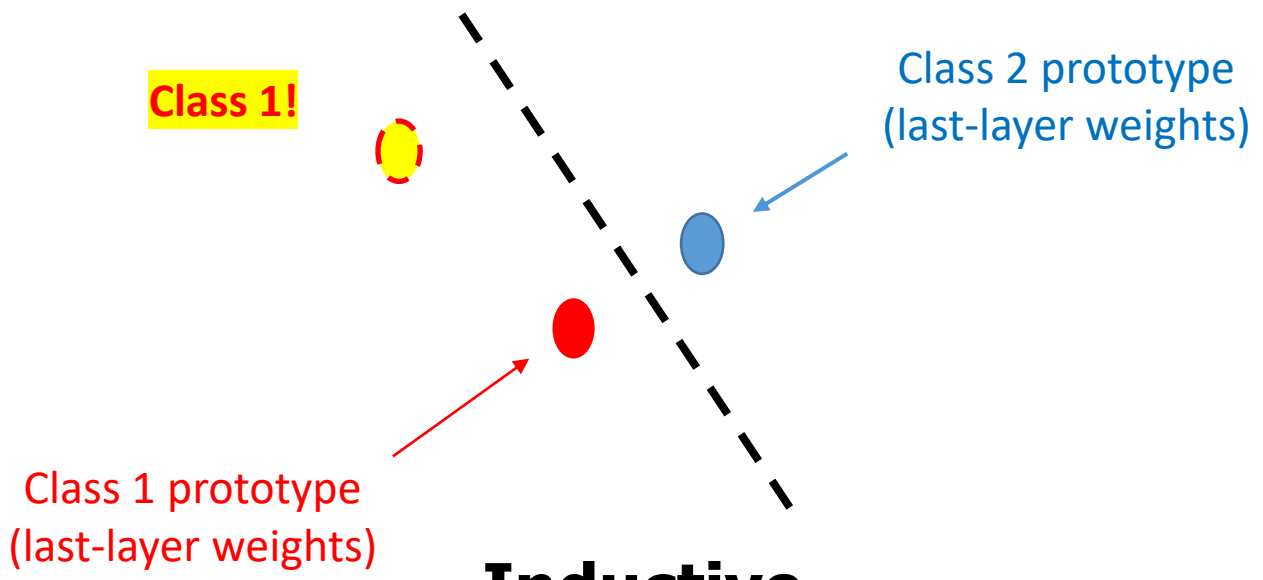
Transductive vs. inductive prediction



Inductive
(one test sample at a time)

Unlabeled instance

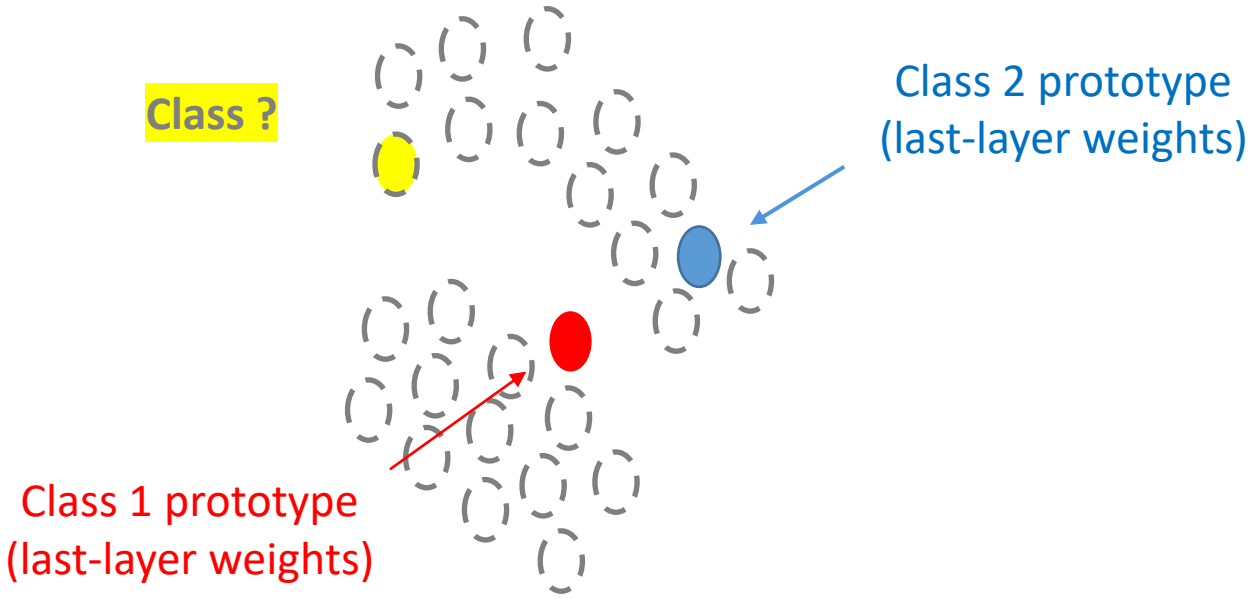
Transductive vs. inductive prediction



Inductive
(one test sample at a time)

Unlabeled instance

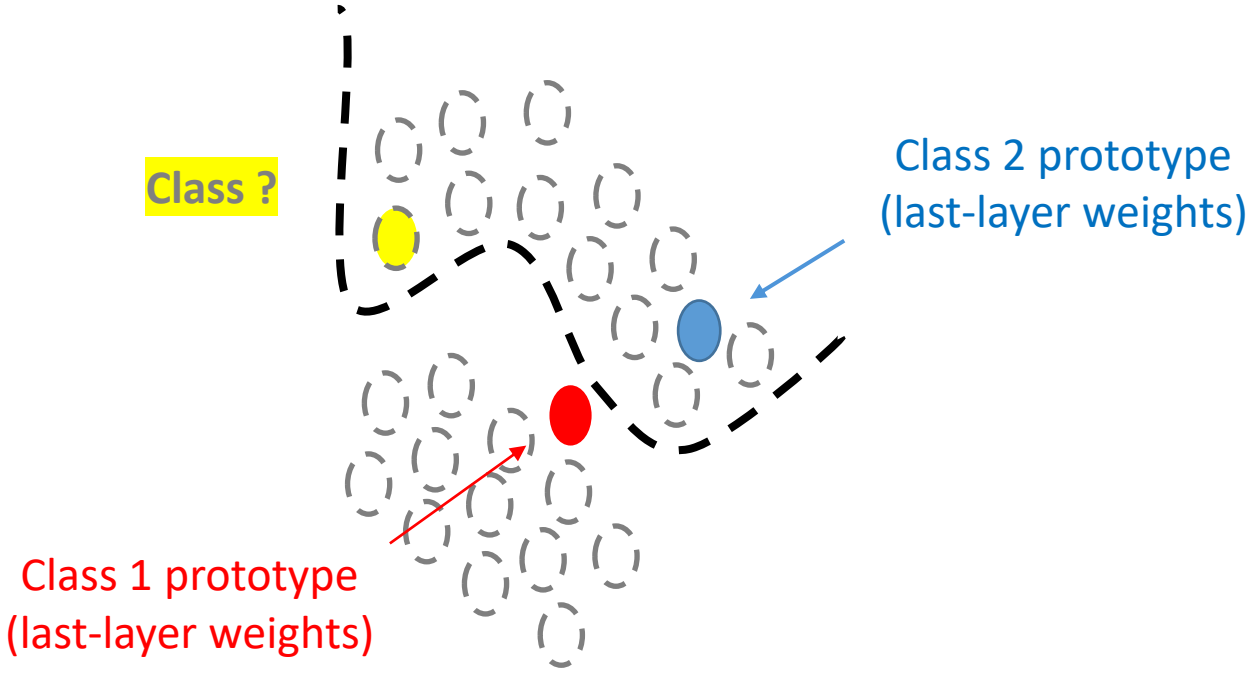
Transductive vs. inductive prediction



Transductive
(joint test-time prediction)

Unlabeled instance

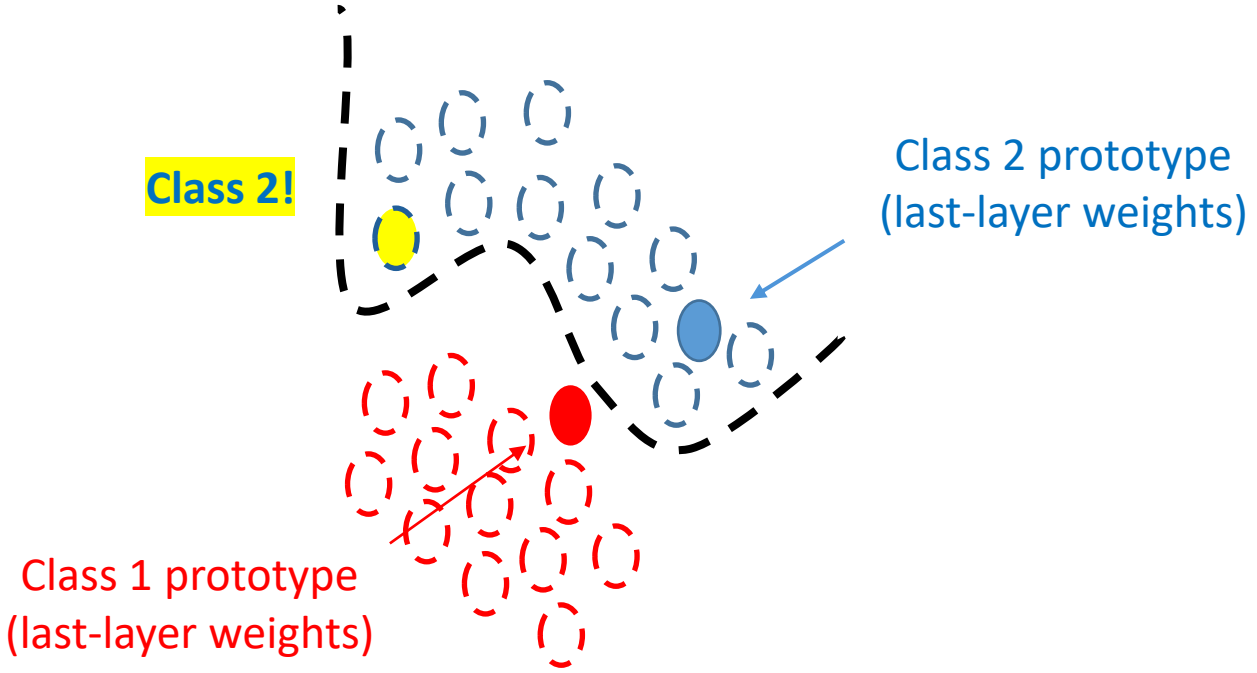
Transductive vs. inductive prediction



Transductive
(joint test-time prediction)

Unlabeled instance

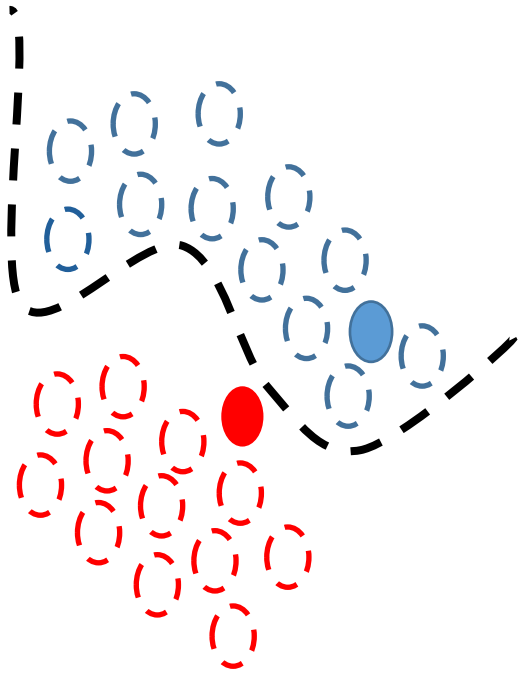
Transductive vs. inductive prediction



Transductive
(joint test-time prediction)

Unlabeled instance

Popular in few-shot learning



Transductive
(joint test-time prediction)

[Finn et al., ICML'17]
(MAML: Transductive BatchNorm)

[Dhillon et al., ICLR'20]
(Entropy Minimization)

[Ziko et al., ICML'20]
(Laplacian Regularization)

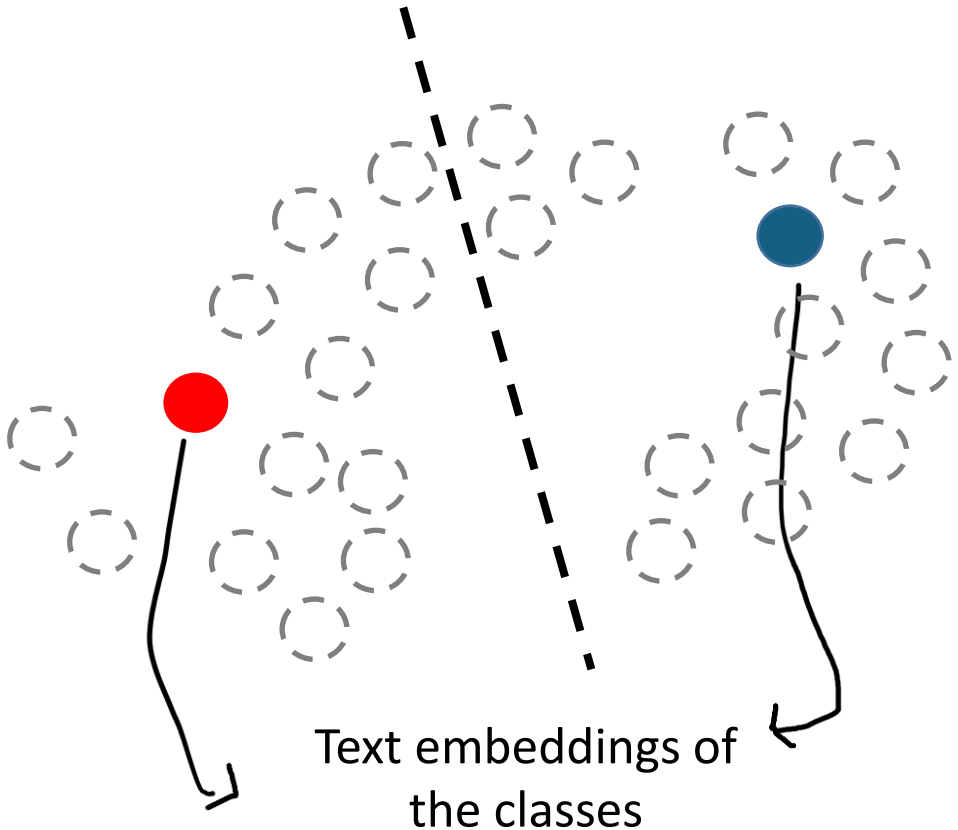
[Boudiaf et al., NeurIPS'20]
(Information Maximization)

TransCLIP: Transductive inference for VLMs

$$\mathcal{L}_{\text{Zero-Shot}}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}} \mathbf{z}_i^\top \log(\mathbf{p}_i) - \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} w_{ij}\mathbf{z}_i\mathbf{z}_j}_{\text{GMM clustering}\qquad\qquad\text{Laplacian reg.}} + \underbrace{\sum_{i\in\mathcal{Q}} \mathrm{KL}_\lambda(\mathbf{z}_i||\hat{\mathbf{y}}_i)}_{\text{Text knowledge}}$$

TransCLIP: Transductive inference for VLMs

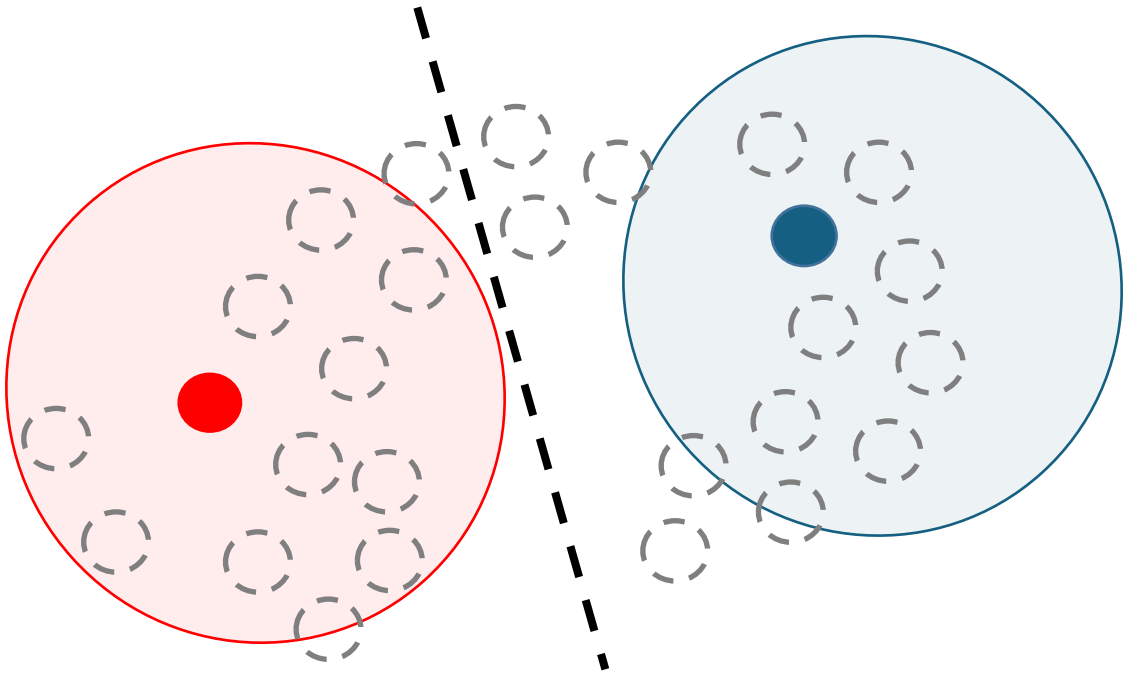
$$\mathcal{L}_{\text{Zero-Shot}}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}}\mathbf{z}_i^\top \log(\mathbf{p}_i) - \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} w_{ij}\mathbf{z}_i\mathbf{z}_j}_{\text{GMM clustering}\qquad\qquad\text{Laplacian reg.}} + \underbrace{\sum_{i\in\mathcal{Q}}\mathrm{KL}_\lambda(\mathbf{z}_i||\hat{\mathbf{y}}_i)}_{\text{Text knowledge}}$$



Unlabeled instance

TransCLIP: Transductive inference for VLMs

$$\mathcal{L}_{\text{Zero-Shot}}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}}\mathbf{z}_i^\top \log(\mathbf{p}_i) - \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} w_{ij}\mathbf{z}_i\mathbf{z}_j}_{\text{GMM clustering}\quad\quad\text{Laplacian reg.}} + \underbrace{\sum_{i\in\mathcal{Q}}\text{KL}_\lambda(\mathbf{z}_i||\hat{\mathbf{y}}_i)}_{\text{Text knowledge}}$$

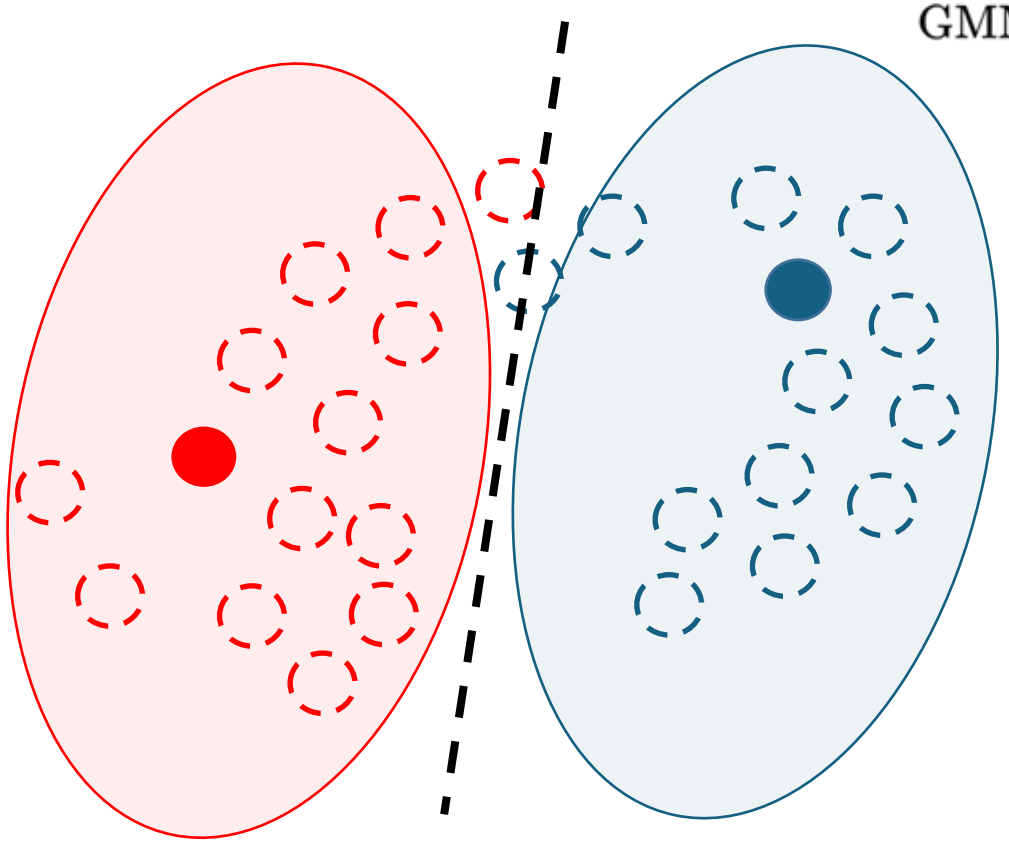


Initialization

Unlabeled instance

TransCLIP: Transductive inference for VLMs

$$\mathcal{L}_{\text{Zero-Shot}}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}} \mathbf{z}_i^\top \log(\mathbf{p}_i) - \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} w_{ij}\mathbf{z}_i\mathbf{z}_j}_{\text{GMM clustering}\quad\quad\text{Laplacian reg.}} + \underbrace{\sum_{i\in\mathcal{Q}} \mathrm{KL}_\lambda(\mathbf{z}_i||\hat{\mathbf{y}}_i)}_{\text{Text knowledge}}$$



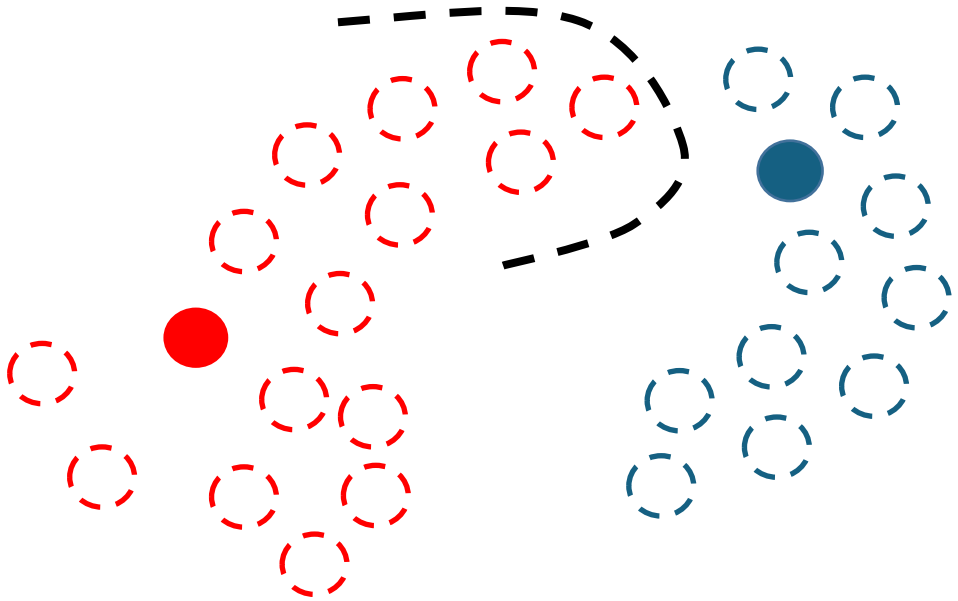
Parameter/assignment updates

Stay close to the Zero-shot predictions
No label needed!

Unlabeled instance

TransCLIP: Transductive inference for VLMs

$$\mathcal{L}_{\text{Zero-Shot}}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}}\mathbf{z}_i^\top \log(\mathbf{p}_i) - \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} w_{ij}\mathbf{z}_i\mathbf{z}_j}_{\text{GMM clustering}\quad\quad\text{Laplacian reg.}} + \underbrace{\sum_{i\in\mathcal{Q}}\mathrm{KL}_\lambda(\mathbf{z}_i||\hat{\mathbf{y}}_i)}_{\text{Text knowledge}}$$



Nearby samples (in the input space)
Are encouraged to have similar predictions

Unlabeled instance

Implementation

Just a **few lines of code**, check our Github repository:
https://github.com/MaxZanella/transduction-for-vlms

Runs in **few seconds** on ImageNet!

	Performance	Runtime
UPL*	69.8	>150 min
TransCLIP-ZS	**70.3**	**14.4 sec**

Algorithm 1 TransCLIP

Require: A set of image embeddings $(\mathbf{f}_i)_{1\leq i\leq N}$, a set of textual class embeddings $(\mathbf{t}_k)_{1\leq k\leq K}$, τ the temperature of the CLIP model.

$w_{i,j} \leftarrow \mathbf{f}_i^\top \mathbf{f}_j \quad \forall\, i,j$ ▷ Affinity measure, truncated with top-3 values

$\hat{\mathbf{y}}_i \leftarrow \varphi(\tau \mathbf{f}_i^\top \mathbf{t})) \quad \forall\, i$ ▷ Initial predictions, φ the softmax function

$\boldsymbol{\mu}_k \leftarrow \text{mean}\{\mathbf{f}_i \text{ s.t } y = k, i \in \mathcal{S}\}$[8] $\quad \forall\, k$ ▷ Class centroids initialization

$\text{diag}(\boldsymbol{\Sigma}) \leftarrow \mathbf{1}\frac{1}{d}$ ▷ Covariance matrix initialization, d is the emb. dim.

$\mathbf{z}_i \leftarrow \hat{\mathbf{y}}_i \quad \forall\, i$ ▷ Initial assignments

while ***(1)***, ***(2)*** and ***(3)*** not converged **do** ▷ Block-wise updates loop

  while ***(1)*** not converged **do** ▷ z-update loop

$$z_{i,k} \leftarrow \frac{\hat{y}_{i,k}^{\lambda} \exp(\log(p_{i,k}) + \sum_{j\in\mathcal{D}} w_{ij} z_{j,k})}{\sum_{k'} \hat{y}_{i,k'}^{\lambda} \exp(\log(p_{i,k'}) + \sum_{j\in\mathcal{D}} w_{ij} z_{j,k'})} \quad \forall\, i\, \forall\, k$$

▷ ***(1)*** z-step

  end while

$$\boldsymbol{\mu}_k \leftarrow \frac{\frac{\gamma}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} z_{i,k}\mathbf{f}_i + \frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}} z_{i,k}\mathbf{f}_i}{\frac{\gamma}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} z_{i,k} + \frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}} z_{i,k}} \quad \forall\, k$$

▷ ***(2)*** $\boldsymbol{\mu}$-step

$$\text{diag}(\boldsymbol{\Sigma}) \leftarrow \frac{\frac{\gamma}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\sum_k z_{i,k}(\mathbf{f}_i - \boldsymbol{\mu}_k)^2 + \frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}}\sum_k z_{i,k}(\mathbf{f}_i - \boldsymbol{\mu}_k)^2}{\gamma + 1}$$

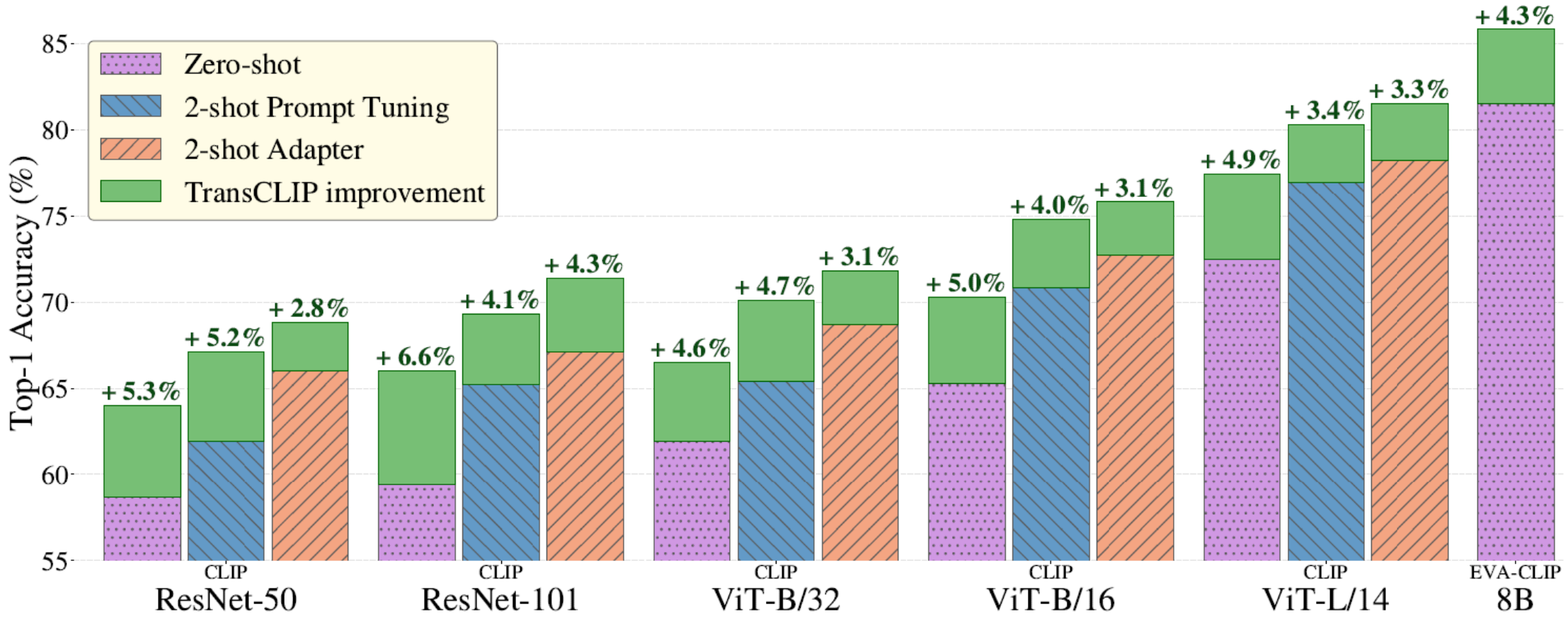
▷ ***(3)*** $\boldsymbol{\Sigma}$-step

end while

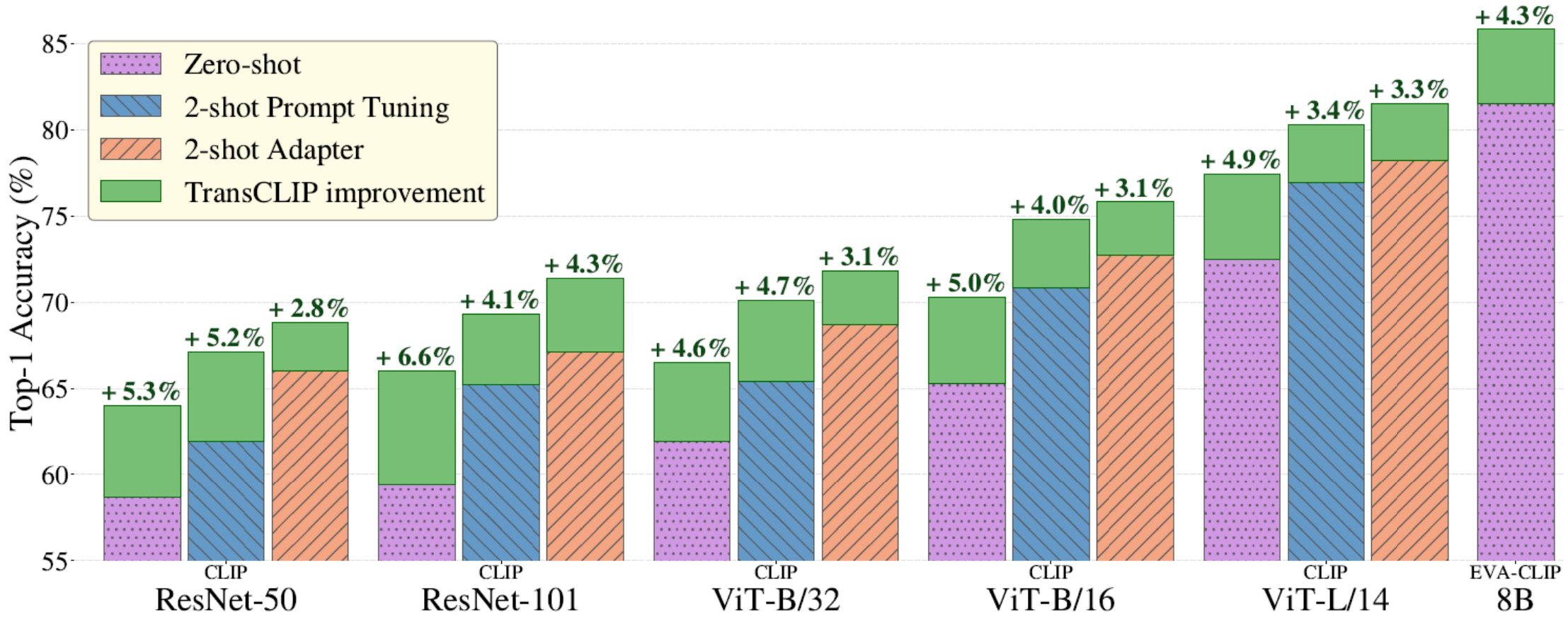
return $\text{argmax}_k(\mathbf{z})$ ▷ Prediction with assignment variables

*Unsupervised Prompt Learning (UPL) adapted for transduction

Results in short



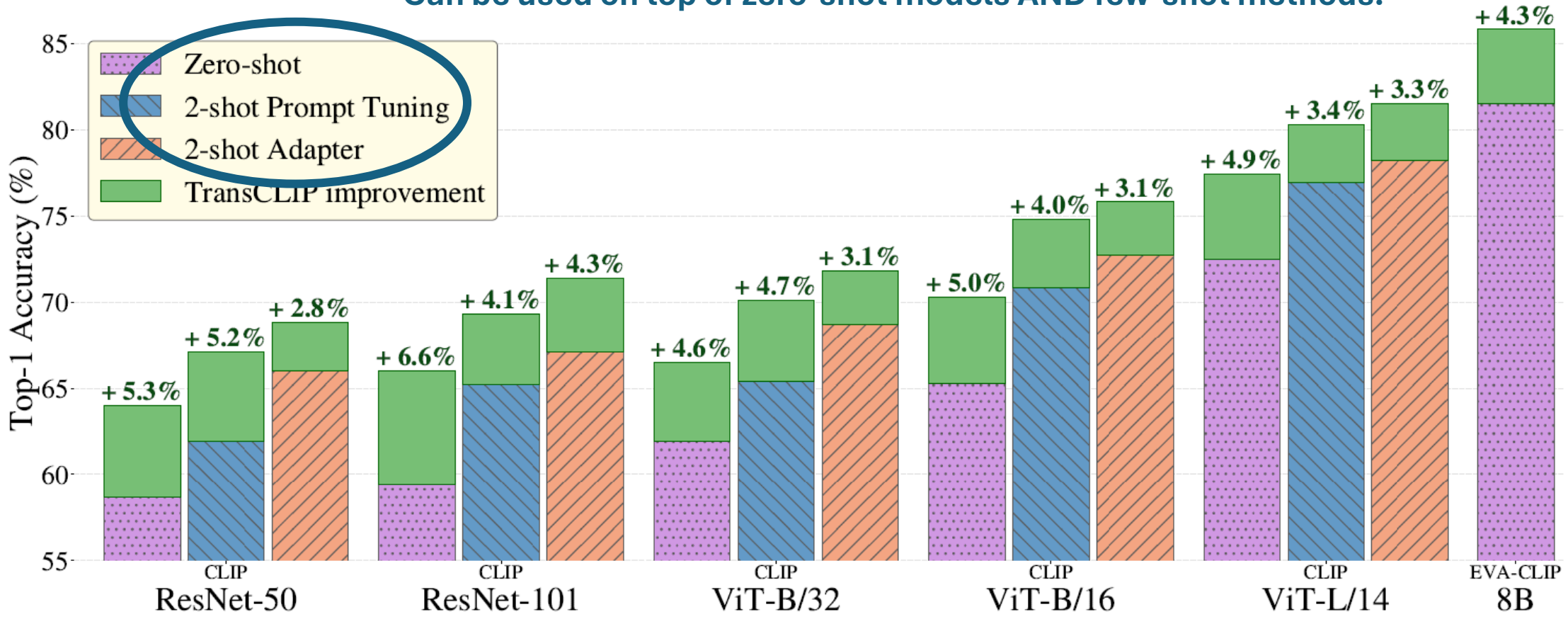
Results in short



Works across various architectures and sizes!

Results in short

Can be used on top of zero-shot models AND few-shot methods!



Also in the main paper

- **Convergence guarantee** of the solving procedure
- Detailed results on various settings (**TransCLIP-ZS**)
  - On top of zero-shot model
  - On top of prompt learning and adapter few-shot methods
  - Cross-dataset transferability
  - Domain generalization on ImageNet and variants
- Extension for few-shot learning (**TransCLIP-FS**)
- **Scaling** to larger VLMs (up to 8 billion parameters)

Thanks for listening!

NEURAL INFORMATION PROCESSING SYSTEMS

Boosting Vision-Language Models with Transduction

NeurIPS '24 - Spotlight

A joint work with

Maxime Zanella*, Benoît Gérin*, Ismail Ben Ayed

UCLouvain UMONS Université de Mons ÉTS

*Equal contributions and corresponding authors