Uncertainty Quantification & Adaptation of Vision-Language Models (VLMs)

FRQS Workshop
in Digital Pathology and Vision-Language models
27/10/25

Nicolas Thome

CNRS, ILLS, Sorbonne Université, ISIR

Success of AI

ChatGPT

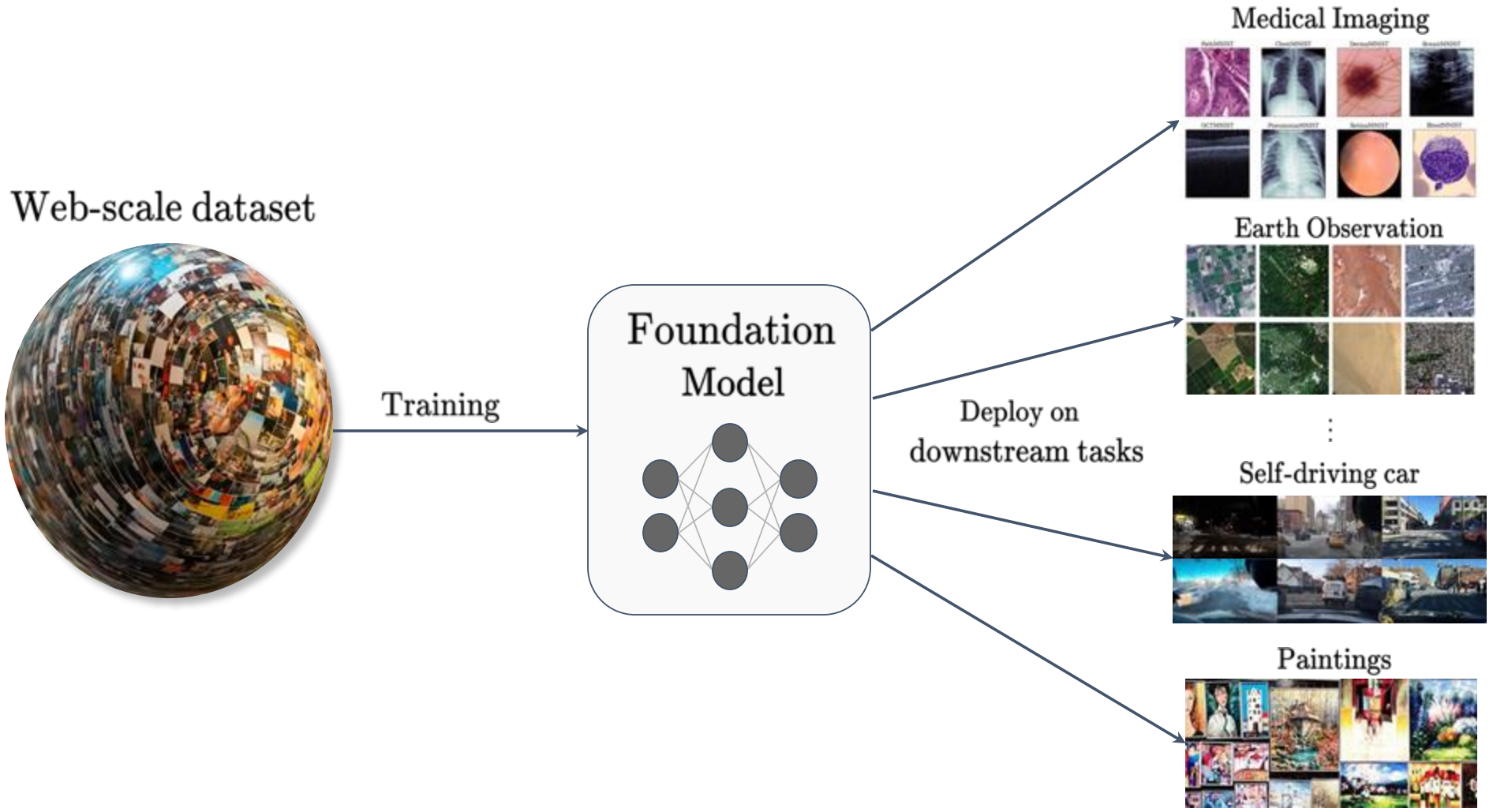
DALL-E

Segment Anything Model (SAM)

Deep Learning: underlying principle powering these breakthroughs

Ramesh, A., et al. “Zero-Shot Text-to-Image Generation”, ICML 2021
ChatGPT OpenAI, 2022, powered by GPT-3.5/4
Kirillov, A., et al. “Segment anything”, ICCV 2023

Foundation models



Vision perception & Vision Language Models (VLMs)

similar?

"Two playful puppies lying on the grass, each holding a colorful woven ball toy, gazing curiously at the camera"

- **Shared representation space** for image and text
- Contrastive loss, e.g., CLIP, SigLIP

- **Zero-shot classification**

Robustness of VLMs

Adversarial attacks

Lion + Noise → Elephant

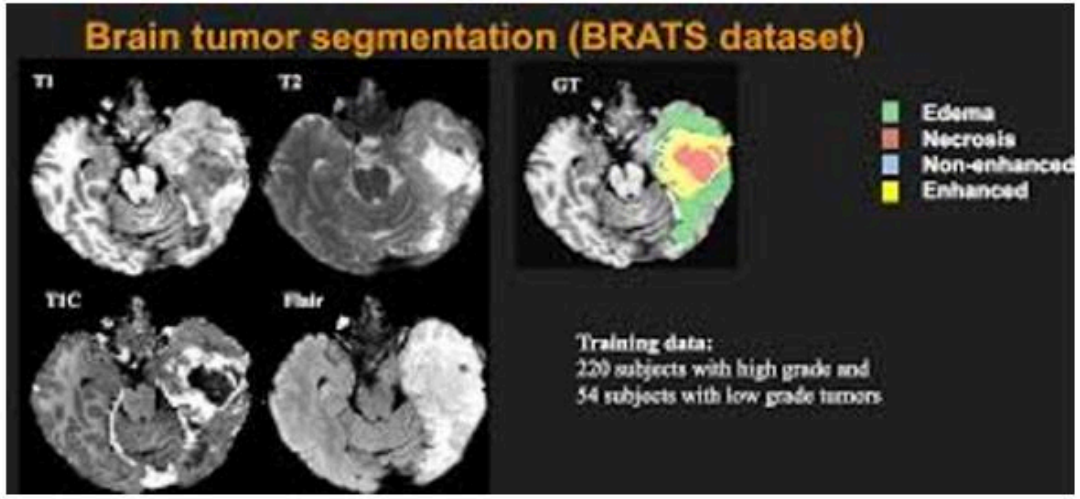
Spurrious correlations

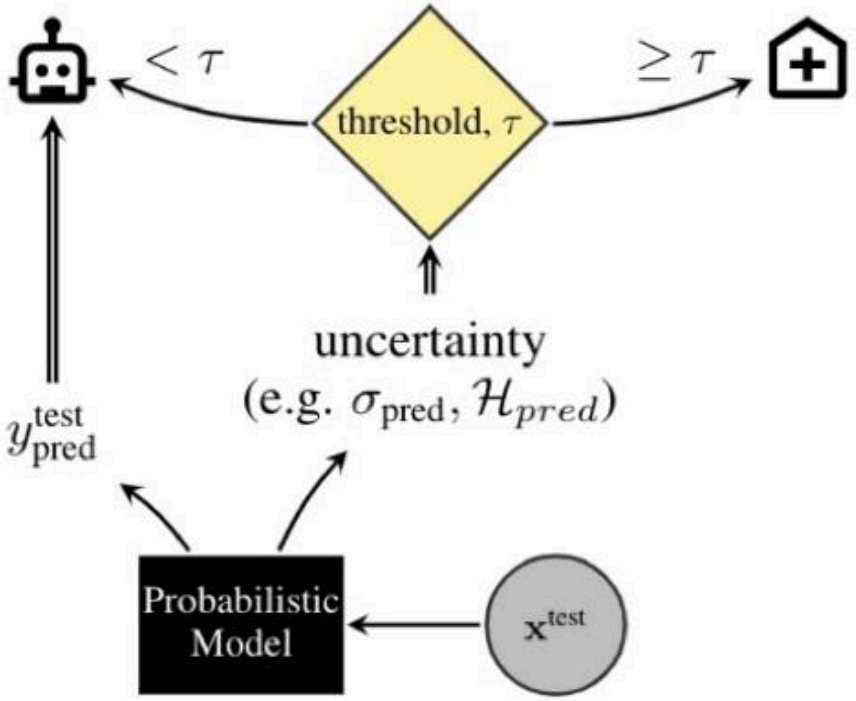
A nurse | A firefighter

Uncertain inputs

Moose on road | Blurry image | Snowy road

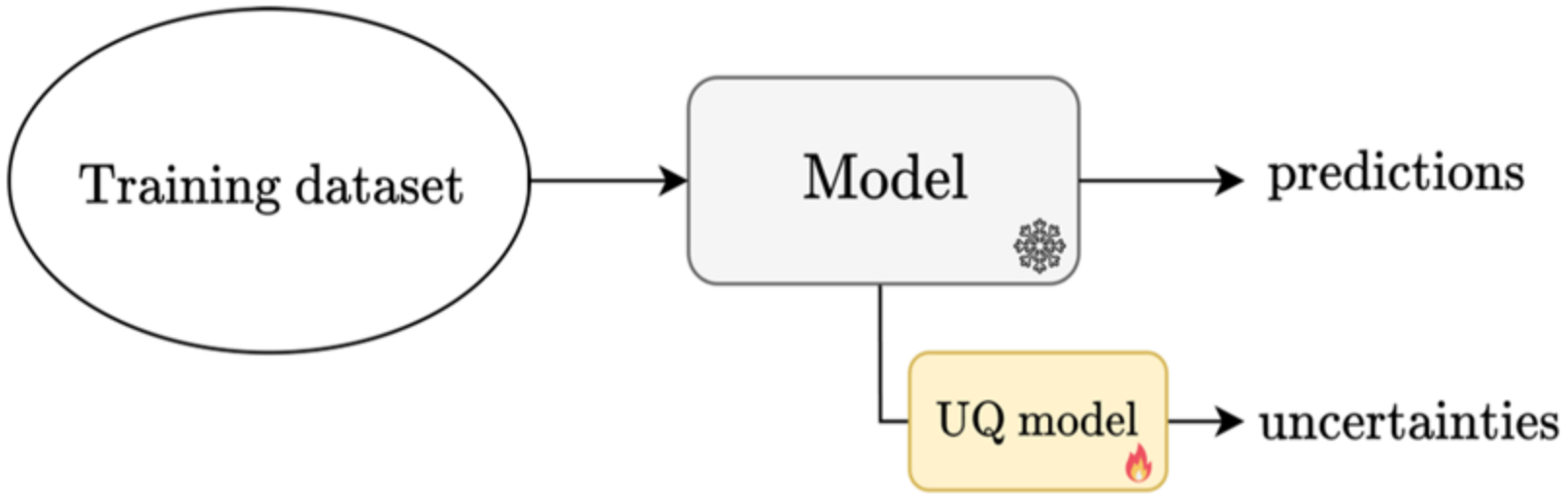
Post-hoc uncertainty quantification





- **Safety-critical applications**, e.g., medical domain
- **Selective classification:** reject uncertain inputs

Post-hoc uncertainty quantification



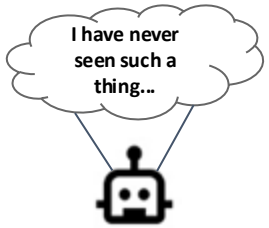
- Foundation model re-training: **computationally expensive**
- **Post-hoc**: predictive model frozen
- **Auxiliary uncertainty module**

Post-hoc uncertainty quantification

novel
class of target

different
imaging view

Epistemic uncertainty

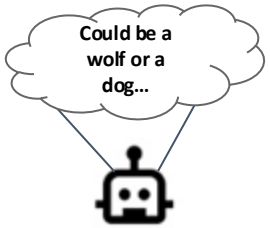


Moose on the road

- Out-of-distribution (OOD) (unknown)
- Reducible

→ Out-of-distribution detection

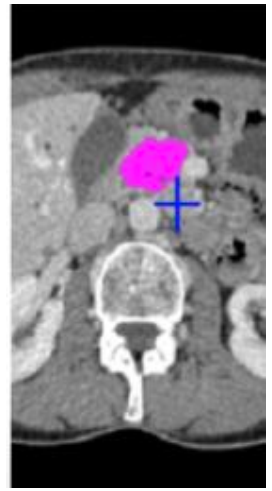
Aleatoric uncertainty



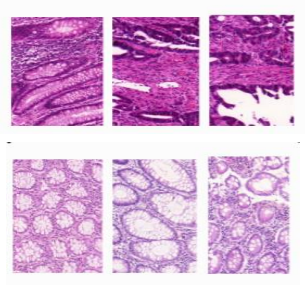
Czechoslovakian wolfdog

- Ambiguous example
- Irreducible

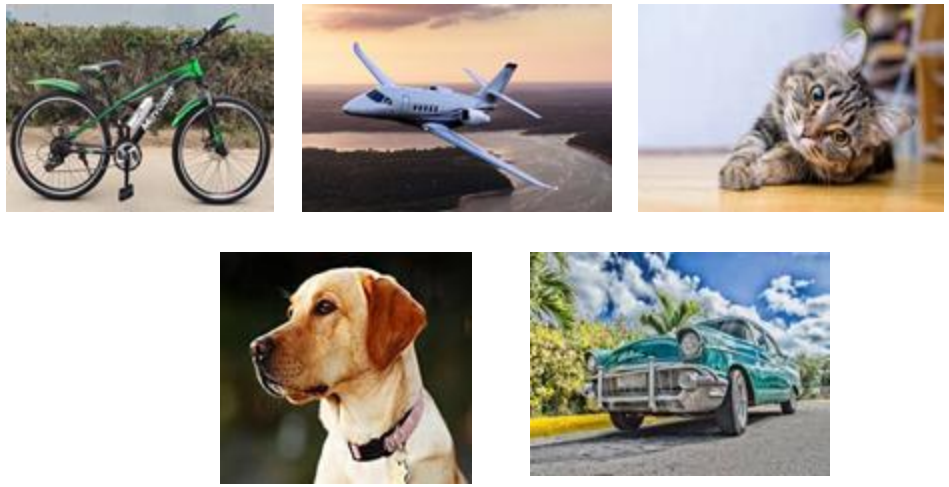
→ Failure prediction

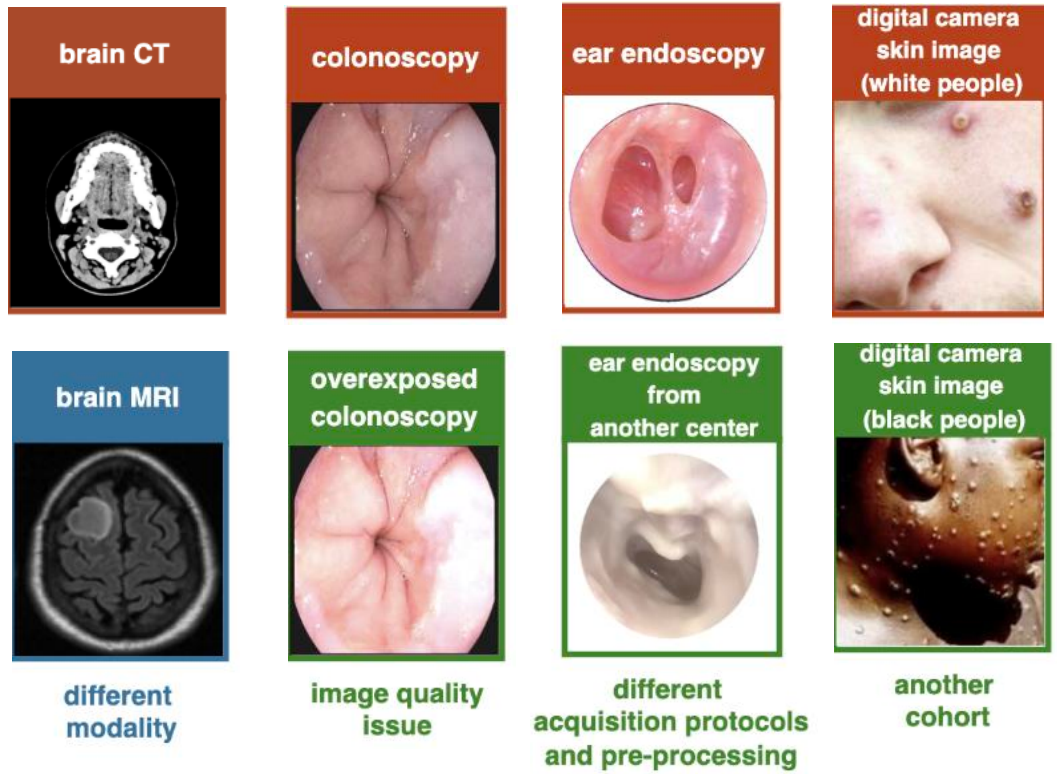


Adaptation of VLMs: domain shifts



General knowledge



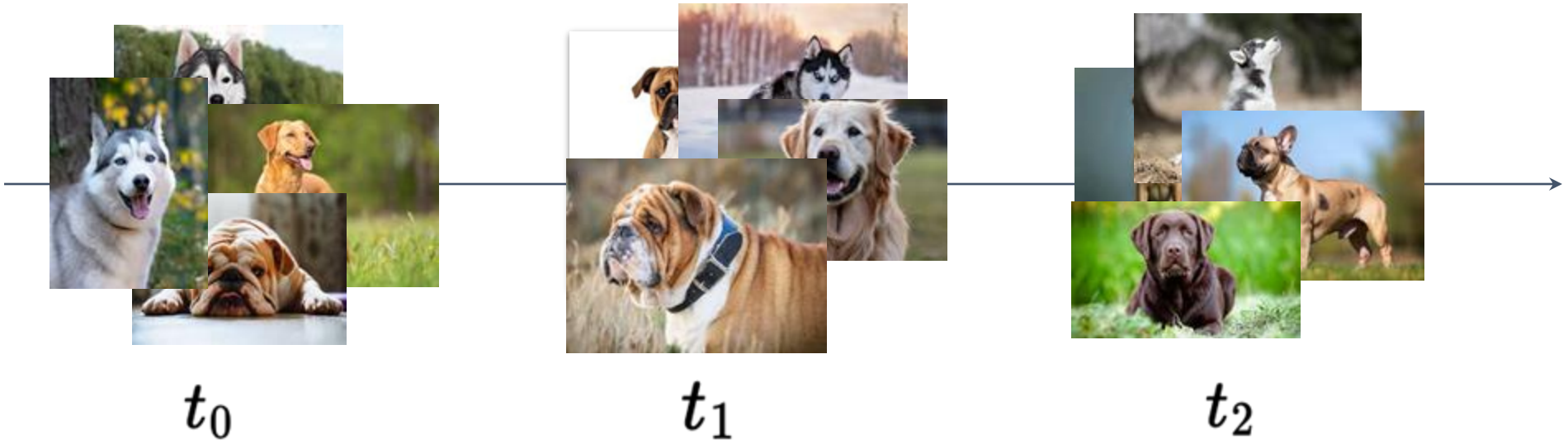


Domain shifts: common in the medical domain: acquisition devices, centers, populations, etc

→ **Adaptation** needed for **highly specialized** applications

Test-Time-Adaptation (TTA)

- No annotated samples
- Exploit incoming test-samples (online)



Additional requirement: **Maintain** or **improve robustness** after **adaptation**

Today's talk

- ViLU [A]: UQ for VLMs, failure prediction
  - Fine text-image interactions for UQ
  - Large-scale training & generalization

- CLIP-TTA [B]: test-time adaptation for CLIP
  - Training loss adapted to CLIP
  - Robustness to pseudo-label errors & class collapse

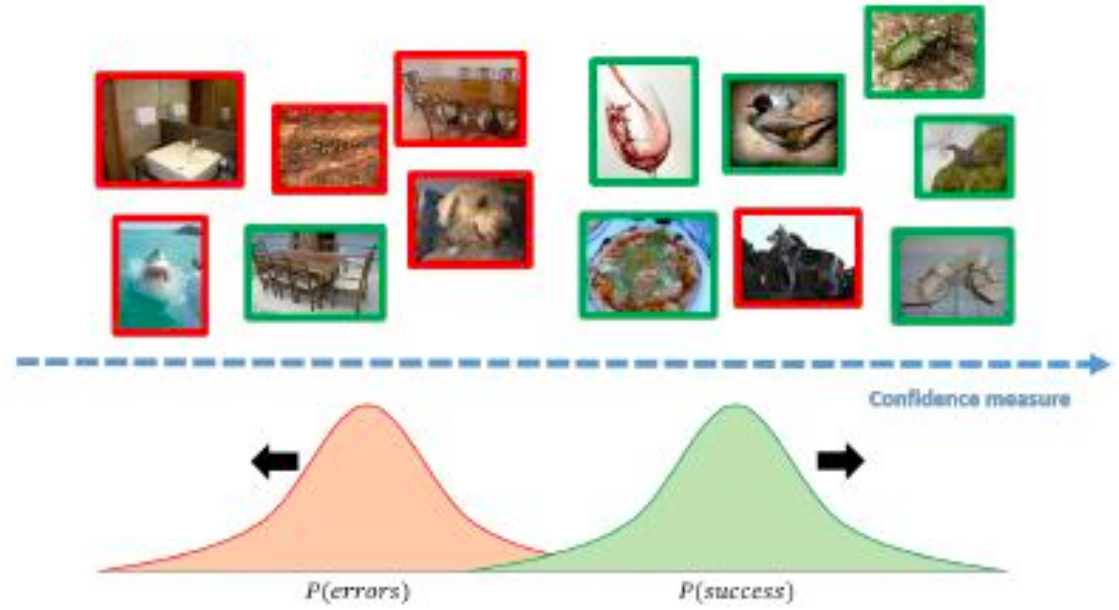
[A] **ViLU: Learning Vision-Language Uncertainties for Failure Prediction.**
M. Lafon, Y. Karmim, J. Silva-Rodriguez, P. Couairon, C. Rambour, R. Fournier, I. Ben Ayed, J. Dolz, N. Thome. ICCV 2025.

[B] **CLIPTTA: Robust Contrastive Vision-Language Test-Time Adaptation.**
M. Lafon, G. Vargas Hakim, C. Rambour, C. Desrosiers, N. Thome. NeurIPS 2025.

Outline

1. ViLU
2. CLIPTTA

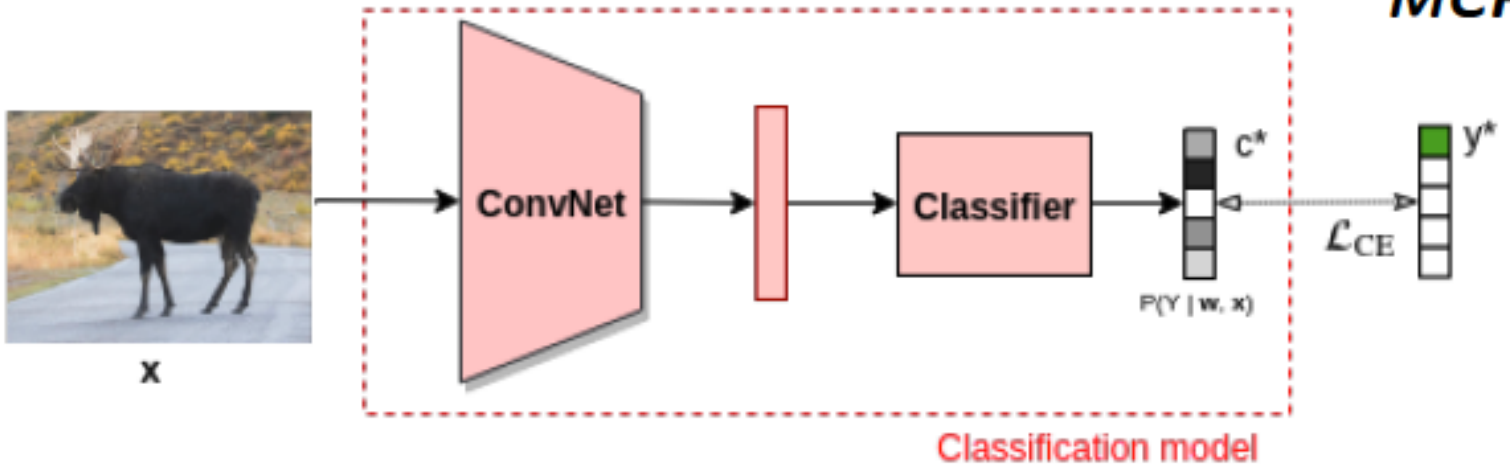
Failure Prediction with deep neural nets



- Confidence criterion C(x): separate correct from incorrect prediction

- **Simple baseline:** Maximum Class Probability (MCP)

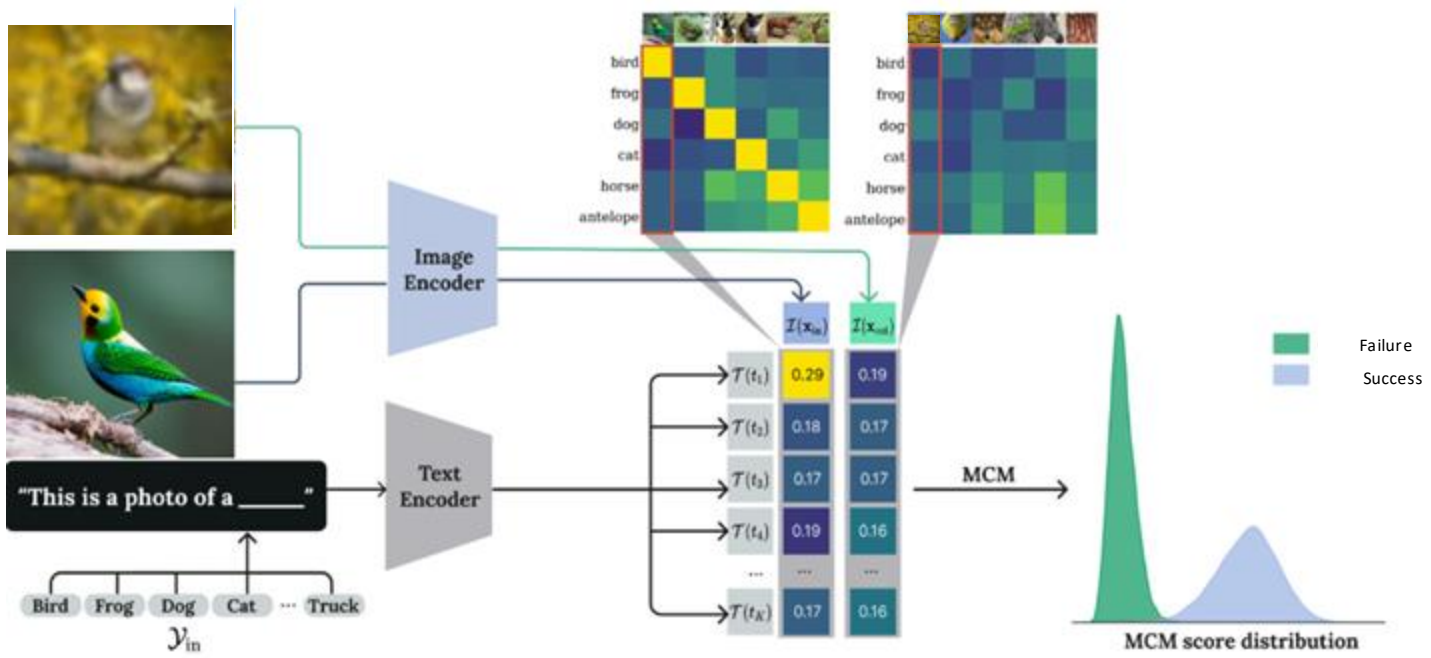
$$MCP(x) = \max_{k \in \mathcal{Y}} p(Y = k | \boldsymbol{w}, \boldsymbol{x})$$



Failure Prediction with VLMs

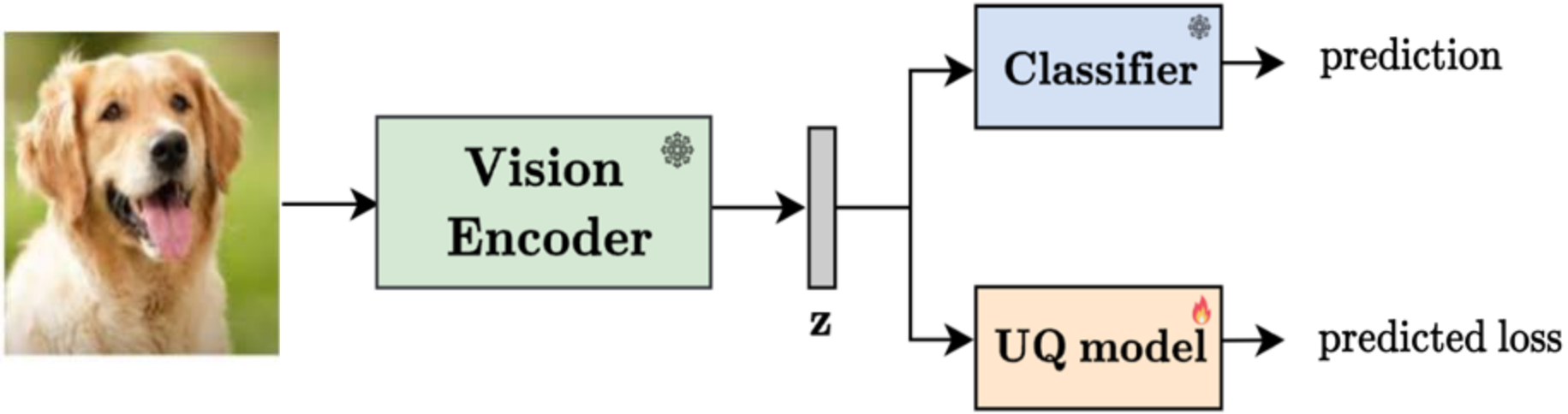
Maximum Concept Matching (MCM) = Probability of predicted class with VLMs

- MCP Extension to VLMs



- <u>Pros:</u>
  - Strong baseline
  - No training required
- <u>Cons:</u>
  - Overconfident by design for errors
  - Limited adaptability

Failure Prediction for vision models: loss prediction



- Learning to predict the training loss
- High predicted loss = likely incorrect prediction
- Learning Visual Uncertainties (LVU): ConfidNet, PVU

  → How to adapt loss prediction for VLMs ?

ConfidNet, Corbière, C., et al. “Addressing failure prediction by learning model confidence”. NeurIPS 2019

D. Yoo and In So Kweon. Learning loss for active learning. CVPR 2019

Kirchhof, Michael, et al. "Pretrained visual uncertainties." arXiv preprint 2024

Taking task complexity into account

What is the uncertainty associated with this image ?

→ It depends on the task

<u>Task 1:</u> cat *vs* dog classification

- Low class confusion
- Low uncertainty

Taking task complexity into account

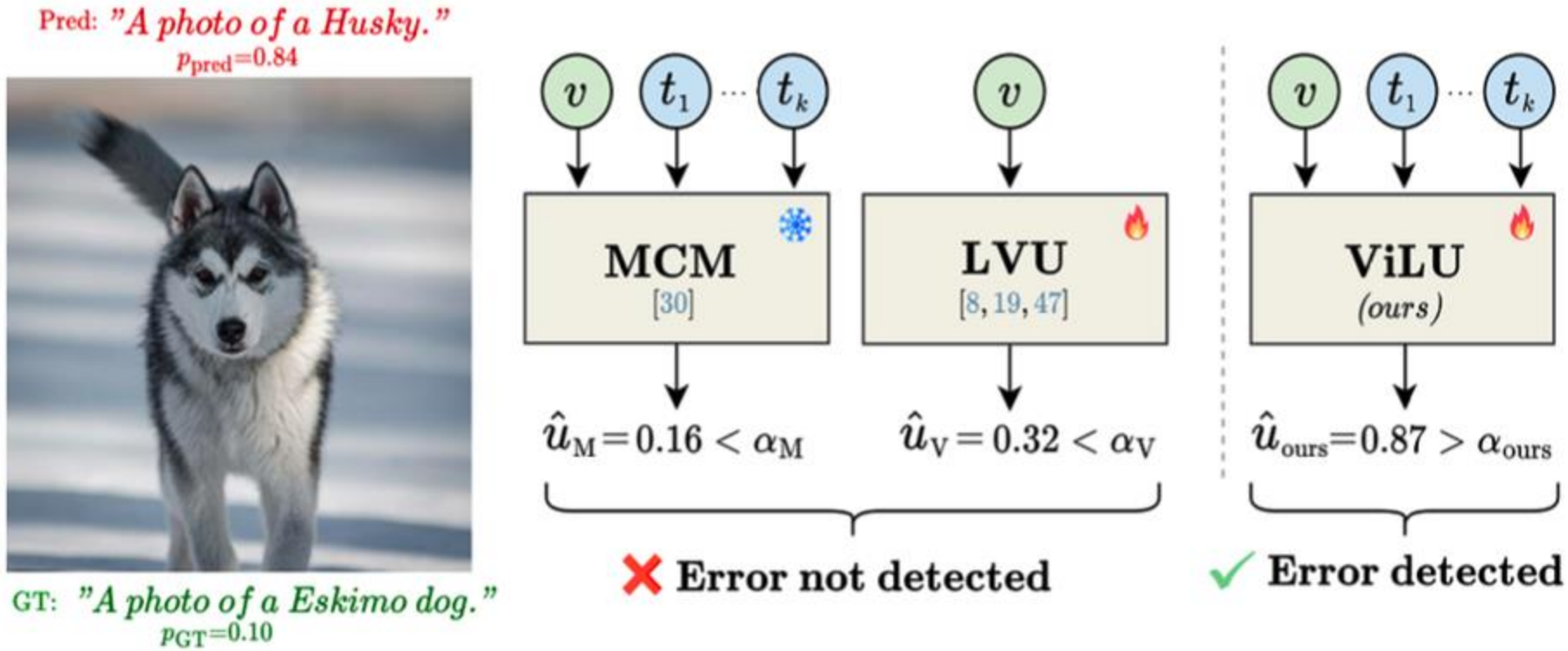
What is the uncertainty associated with this image ?

→ It depends on the task

Task 2: Dog breed classification

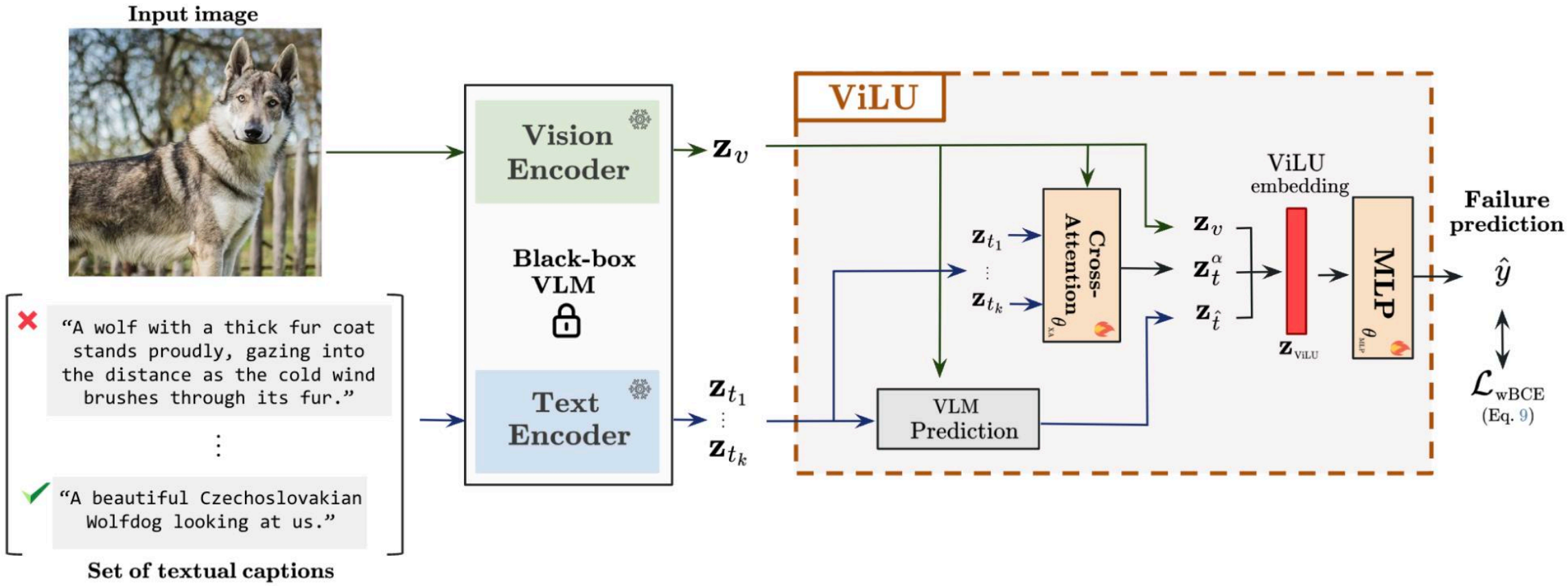
- Higher class confusion
- High uncertainty

ViLU: Learning Vision-Language Uncertainty



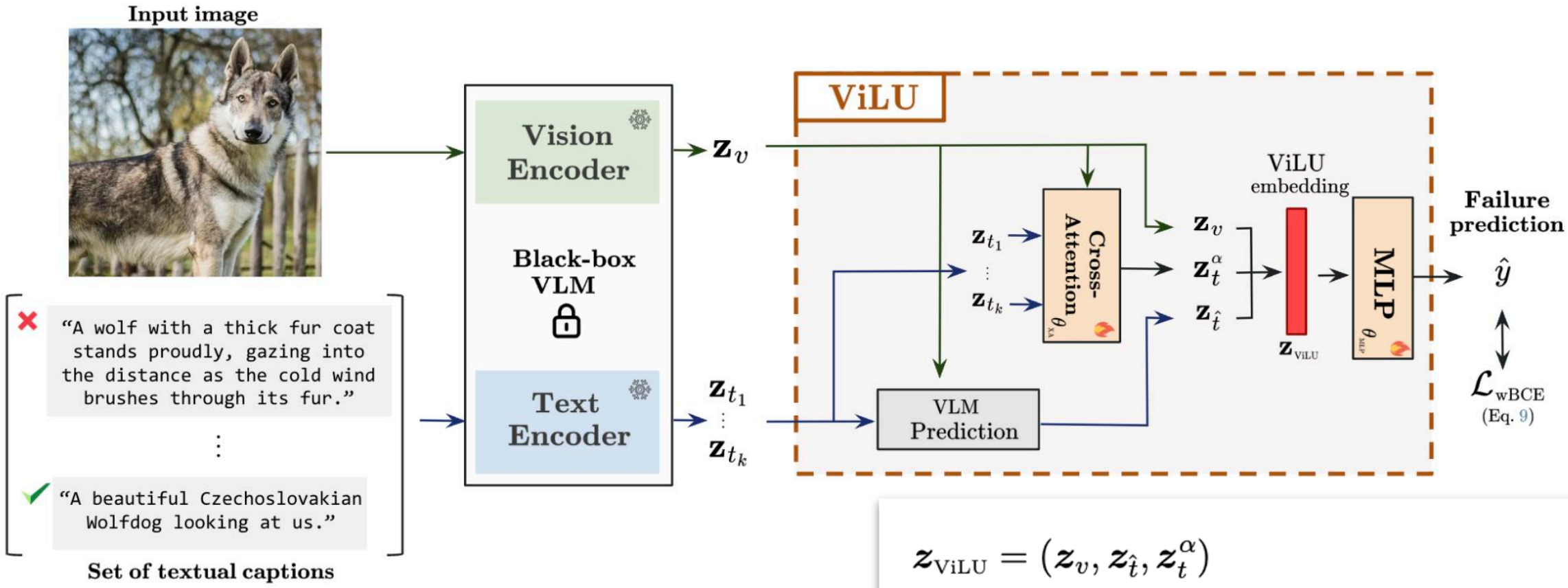
- LVU methods use visual input only
- **ViLU** → Learn uncertainty contextualized with the task information

ViLU: Learning Vision-Language Uncertainty



- Inputs: visual representation z_v + K concept embeddings $Z_t = \{z_{t_j}\}_{1<j<K}$
- ViLU embedding: $z_{\text{ViLU}} = (z_v, z_{\hat{t}}, z_t^\alpha), z_{\hat{t}}$ pred. Class embedding
  - Image-text cross-attention module => z_t^α
  - Query z_v , keys/values Z_t

ViLU: Learning Vision-Language Uncertainty



- **Failure prediction from ViLU embedding:** binary classification
- Consistent generalization of MCM

$$\boldsymbol{z}_{\text{ViLU}} = (\boldsymbol{z}_v, \boldsymbol{z}_{\hat{t}}, \boldsymbol{z}_t^{\alpha})$$

$$g_{\theta_{\text{MLP}}}(\boldsymbol{z}_{\text{ViLU}}) \approx \frac{1}{2}\boldsymbol{z}_{\text{ViLU}}^T \; A \; \boldsymbol{z}_{\text{ViLU}} = z_v^T \; z_{\hat{t}}$$

$$\text{with} \quad A = \begin{pmatrix} \mathbf{0} & \mathrm{I}_d & \mathbf{0} \\ \mathrm{I}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Experiments

- ViLU trained and evaluated on each downstream dataset
- 13 classification datasets (e.g. ImageNet)
- 3 Image-caption datasets (e.g. CC12M)

- Metrics: FPR95 and AUC

- Baselines :
  - MCM, entropy, DOCTOR
  - Data-driven predictors: LVU, Rel-U, BayesVLM

Results

	CIFAR-10		ImageNet-1k		CC12M	
	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓
MCM	89.9	52.1	80.8	71.3	88.8	58.8
Entropy	88.7	59.9	78.3	76.8	80.2	74.0
DOCTOR	89.5	56.5	80.3	72.9	88.6	59.9
BayesVLM	92.6	44.9	81.5	70.3	90.9	53.3
LVU - ConfidNet	96.4	21.8	78.7	77.0	74.0	76.5
LVU + XA	97.9	10.8	88.8	50.1	88.9	48.9
LVU + Pred.	97.7	11.4	86.1	63.5	93.6	37.0
ViLU (ours)	**98.3**	**8.2**	**89.5**	**47.4**	**95.2**	**25.2**

- **LVU competitive** on small datasets (e.g. CIFAR-10)
- **ViLU achieves best performance**

Results

	CIFAR-10		ImageNet-1k		CC12M	
	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓
MCM	89.9	52.1	80.8	71.3	88.8	58.8
Entropy	88.7	59.9	78.3	76.8	80.2	74.0
DOCTOR	89.5	56.5	80.3	72.9	88.6	59.9
BayesVLM	92.6	44.9	81.5	70.3	90.9	53.3
LVU - ConfidNet	96.4	21.8	78.7	77.0	74.0	76.5
LVU + XA	97.9	10.8	88.8	50.1	88.9	48.9
LVU + Pred.	97.7	11.4	86.1	63.5	93.6	37.0
ViLU (ours)	**98.3**	**8.2**	**89.5**	**47.4**	**95.2**	**25.2**

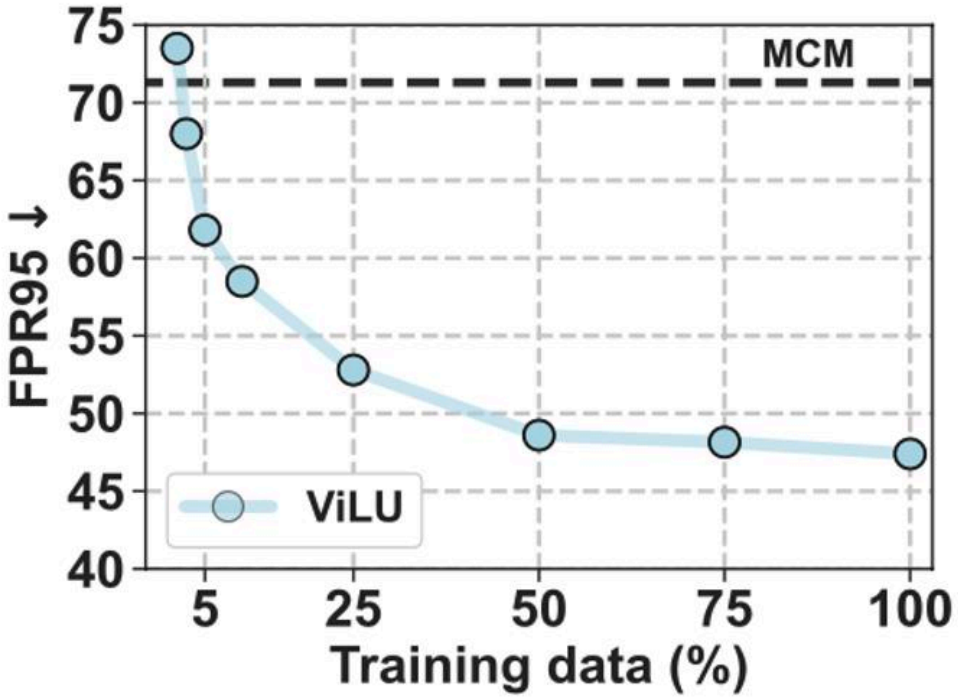
- **LVU struggles** on challenging datasets (e.g. ImageNet)

Results

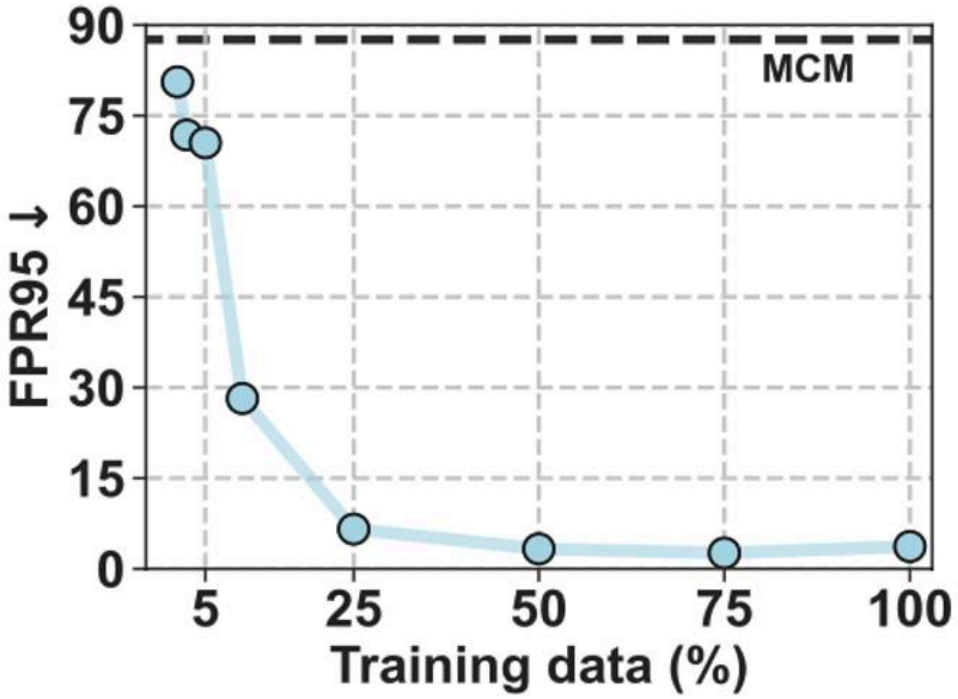
	CIFAR-10		ImageNet-1k		CC12M	
	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓
MCM	89.9	52.1	80.8	71.3	88.8	58.8
Entropy	88.7	59.9	78.3	76.8	80.2	74.0
DOCTOR	89.5	56.5	80.3	72.9	88.6	59.9
BayesVLM	92.6	44.9	81.5	70.3	90.9	53.3
LVU - ConfidNet	96.4	21.8	78.7	77.0	74.0	76.5
LVU + XA	97.9	10.8	88.8	50.1	88.9	48.9
LVU + Pred.	97.7	11.4	86.1	63.5	93.6	37.0
ViLU (ours)	**98.3**	**8.2**	**89.5**	**47.4**	**95.2**	**25.2**

- **Importance of ViLU embeddings: XA + predicted class**

Model Analysis



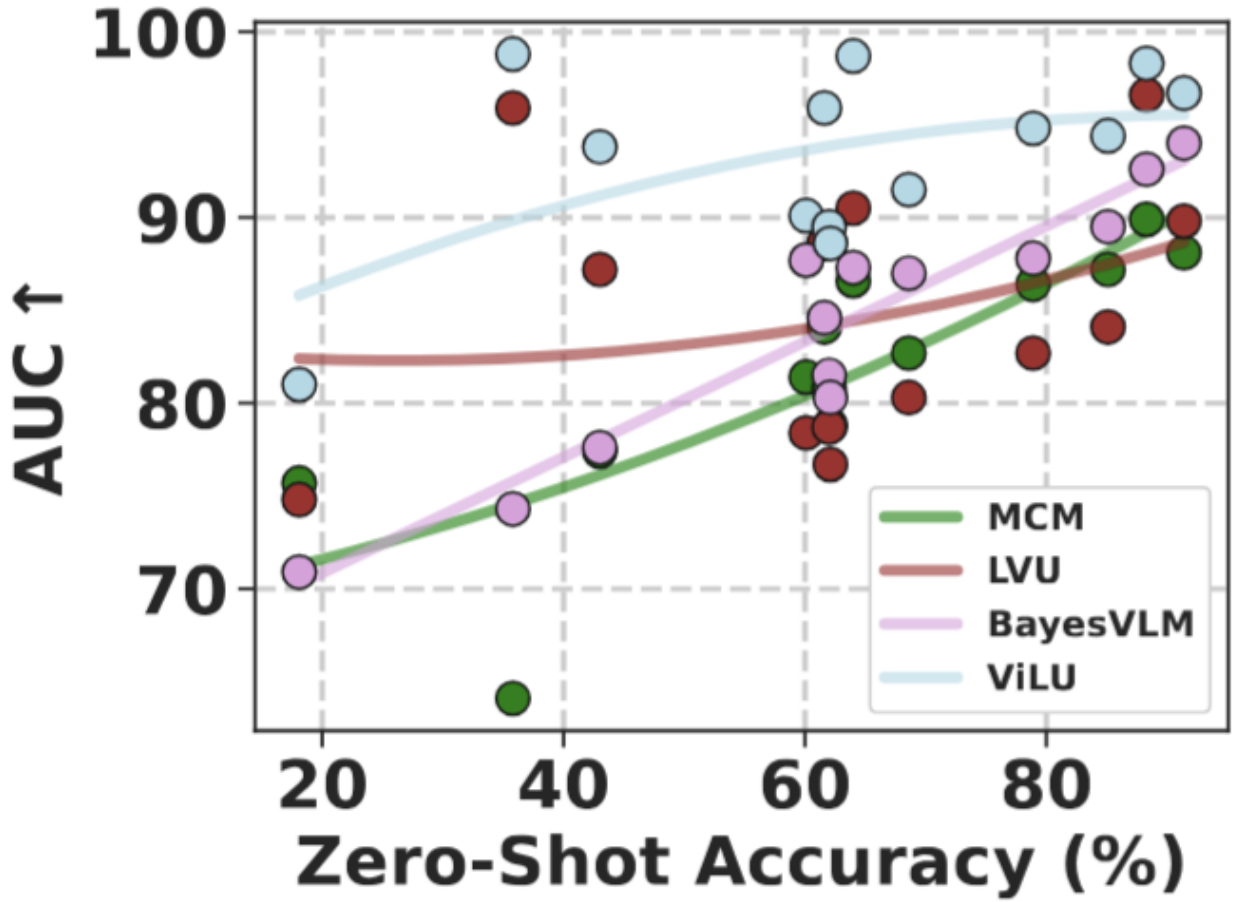
(a) ImageNet-1k



(b) EuroSAT

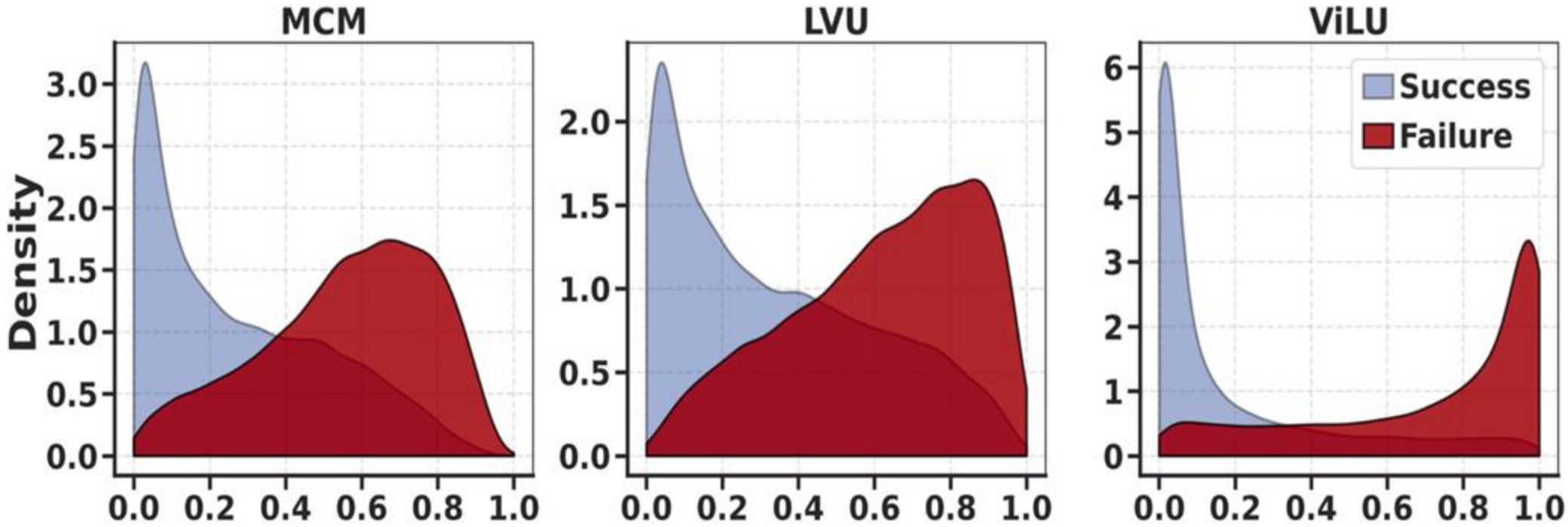
Data efficiency: ViLU effective even with a small fraction of train set

Model Analysis



Effective with low ZS accuracy
≠ MCM, BayesVLM

Model Analysis



- ViLU better separation

Model Analysis

GT: *"container ship"*
Pred: *"ocean liner"*

GT: *"mailbox"*
Pred: *"birdhouse"*

- Misclassification detection with ViLU ≠ MCM or LVU

Towards zero-shot uncertainty quantification

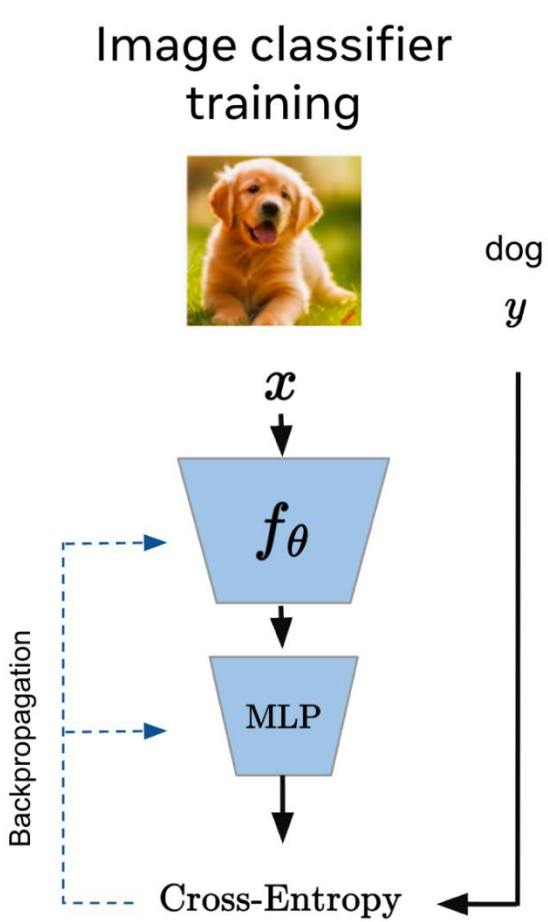
	FPR95↓		
Dataset	**MCM**	**LVU**	**ViLU**
CIFAR-10	**52.1**	77.2	54.2
CIFAR-100	67.3	83.8	**59.9**
Caltech101	68.7	82.5	**48.8**
Flowers102	68.0	96.8	**67.4**
OxfordPets	59.9	93.1	**58.1**
Food101	**63.3**	87.2	67.4
FGVCAircraft	82.9	94.5	**82.3**
EuroSAT	87.6	88.2	**85.7**
DTD	**77.9**	93.1	78.2
SUN397	75.9	90.1	**72.7**
StanfordCars	**73.4**	92.6	84.1
UCF101	68.9	90.4	**63.8**
Average	70.5	89.1	**68.6**

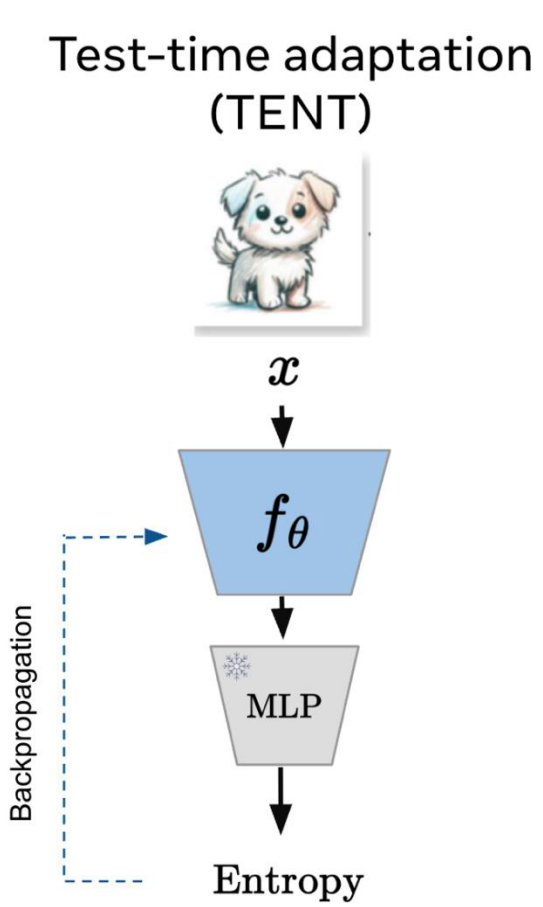
- ViLU pre-trained on CC12M transfers well on downstream datasets

Outline

1. ViLU
2. CLIPTTA

Test-Time-Adapation (TTA) methods of vision models





- TENT-like methods: pseudo-labels, reinforce predictions

Wang, Dequan, et al. "Tent: Fully test-time adaptation by entropy minimization." ICLR 2021.

TENT-like TTA methods for vision models

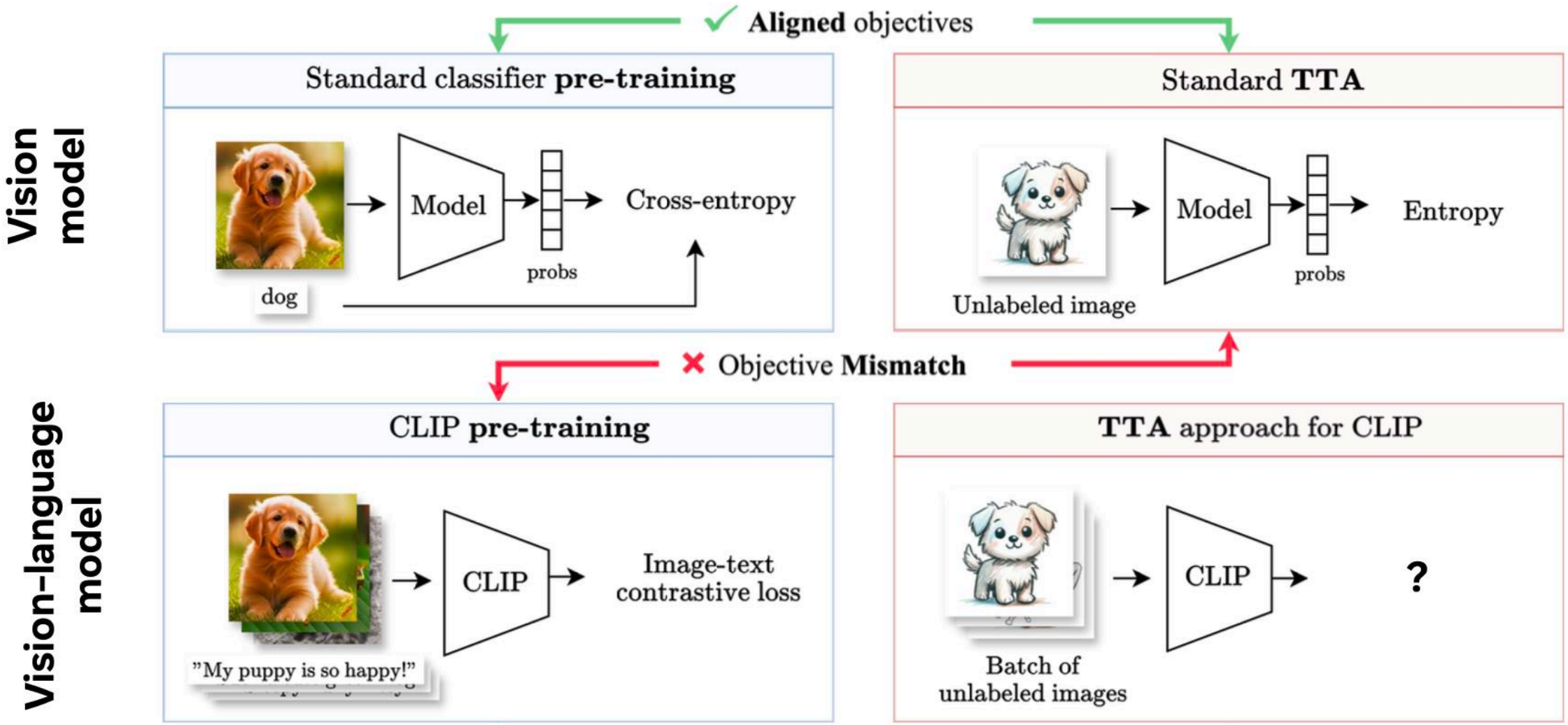
Vision model

- Entropy minimization: soft version of cross-entropy

$$\mathcal{L}_{\text{XE}} = -\log p(y|\boldsymbol{x}) \quad \Leftrightarrow \quad \mathcal{L}_{\text{Entropy}} = -\sum_c p(y_c|\boldsymbol{x}) \log p(y_c|\boldsymbol{x})$$

$$\text{when} \quad p(y_c|\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad y_c = y \\ 0 & \text{otherwise} \end{cases}$$

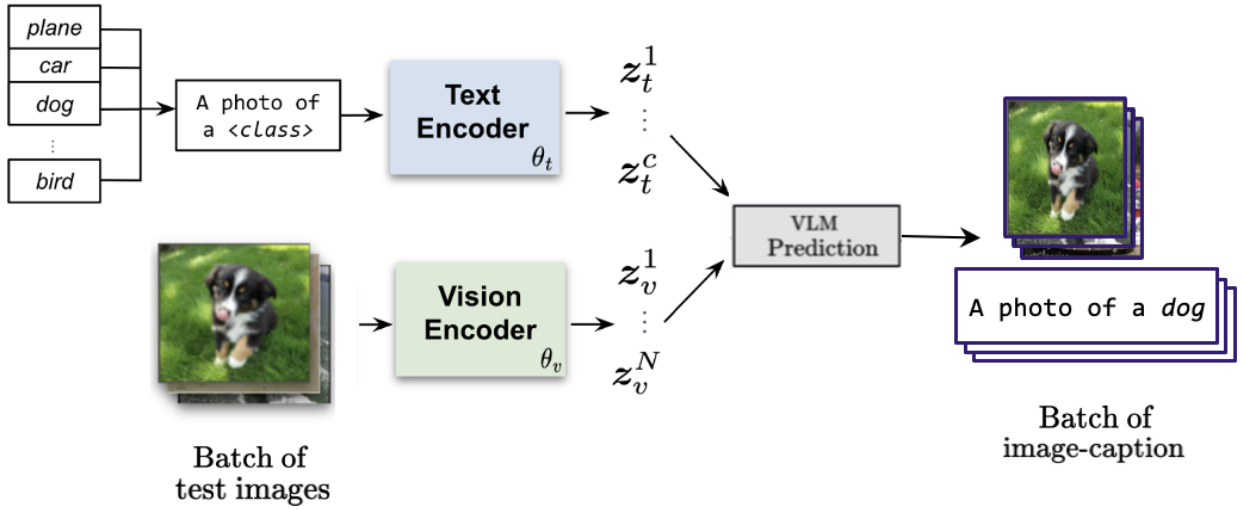
TTA on VLMS, e.g., CLIP: objective mismatch



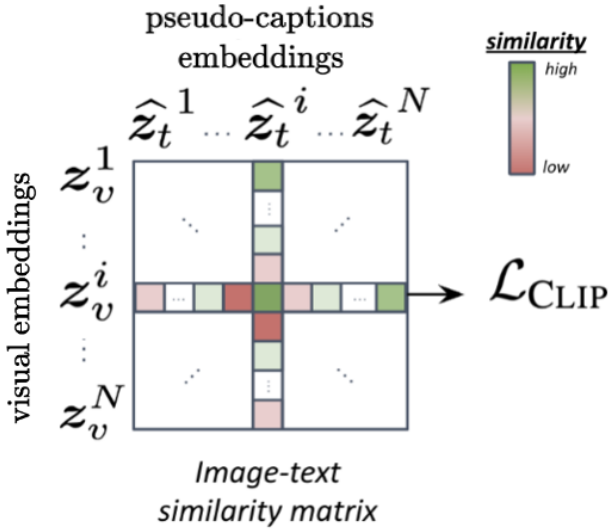
=> CLIP-TTA

CLIP-TTA: how to use CLIP objective at test-time?

1) Pseudo-captioning

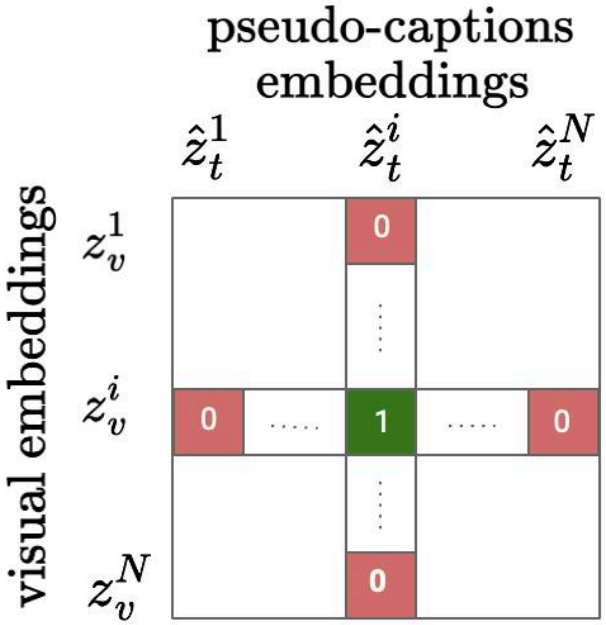


2) Contrastive loss



- Associate a pseudo-caption to each test image
- Compute CLIP loss using batch of paired image-caption

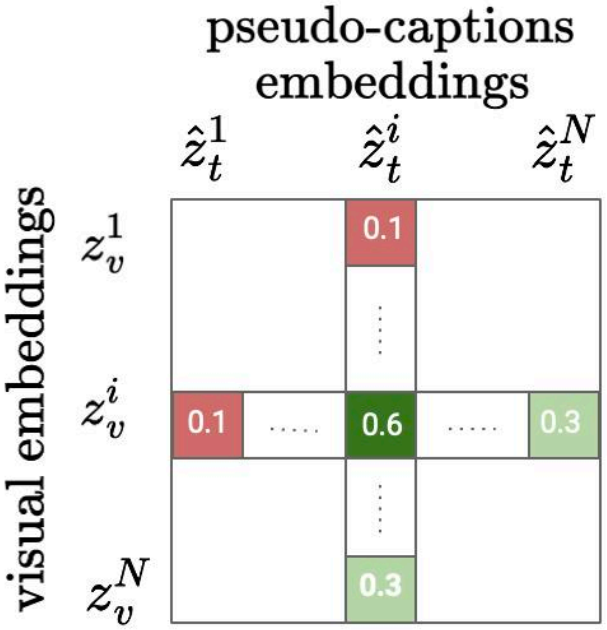
CLIP-TTA: soft-contrastive loss



$$\mathcal{L}_{\text{CLIP}}(\theta) := -\sum_{i=1}^{N} \Big[\underbrace{\log p(\hat{\boldsymbol{t}}_i|\boldsymbol{x}_i)}_{\text{image}\to\text{text}} + \underbrace{\log p(\boldsymbol{x}_i|\hat{\boldsymbol{t}}_i)}_{\text{text}\to\text{image}} \Big]$$

- Use CLIP loss for the predicted text / caption
- **Does not account** for the **uncertainty** among the pseudo-captions

CLIP-TTA: soft-contrastive loss



$$\mathcal{L}_{\text{CLIP}}(\theta) := -\sum_{i=1}^{N}\Big[\underbrace{\log p(\hat{\boldsymbol{t}}_i|\boldsymbol{x}_i)}_{\text{image}\to\text{text}} + \underbrace{\log p(\boldsymbol{x}_i|\hat{\boldsymbol{t}}_i)}_{\text{text}\to\text{image}}\Big]$$

$$\mathcal{L}_{\text{s-cont}}(\theta) := \sum_{i=1}^{N}\Big[\underbrace{-\sum_{j=1}^{N} p(\hat{\boldsymbol{t}}_j|\boldsymbol{x}_i)\log p(\hat{\boldsymbol{t}}_j|\boldsymbol{x}_i)}_{\text{image}\to\text{text}} \underbrace{-\sum_{j=1}^{N} p(\boldsymbol{x}_j|\hat{\boldsymbol{t}}_i)\log p(\boldsymbol{x}_j|\hat{\boldsymbol{t}}_i)}_{\text{text}\to\text{image}}\Big]$$

- Still **aligned with** CLIP loss, e**xploit uncertainty** among the pseudo-captions
- **=> Entropy version** of CLIP loss

- Loss computed on batch + on a memory batch $\mathcal{M}$ + regularization loss $\mathcal{L}_{\text{reg}} = -\sum_{c=1}^{C} \bar{q}_c \log \bar{q}_c$

$$\mathcal{L}_{\text{CLIPTTA}}(\theta) = \frac{1}{2}\Big[\mathcal{L}_{\text{s-cont}}(\theta) + \mathcal{L}^{\mathcal{M}}_{\text{s-cont}}(\theta)\Big] + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}(\theta)$$

CLIP-TTA soft-contrastive loss: properties

Gradient analysis

Entropy loss

$$\nabla_{z_i}\mathcal{L}_{\text{TENT}} = -\sum_{k=1}^{C}\left[\sum_{c=1}^{C}\log\frac{q_{ik}}{q_{ic}}\,q_{ic}\right]q_{ik}\,\boldsymbol{z}_t^k$$

- Always points towards predicted class
- Always Reinforces errors

Soft contrastive loss

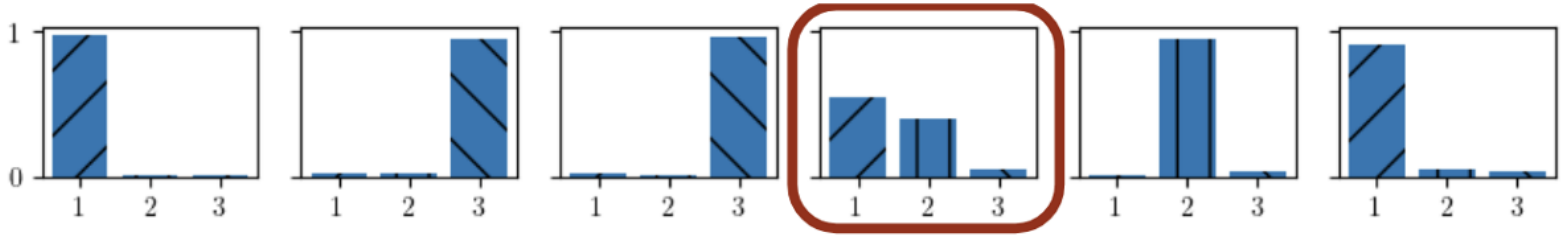
$$\nabla_{z_i}\mathcal{L}_{\text{CLIPTTA}} = \sum_{j=1}^{N}\beta_{i,j}\left[-\widehat{\boldsymbol{z}_t}^{j} + \sum_{k=1}^{C} w_{k,i}\;\boldsymbol{z}_t^k\right]$$

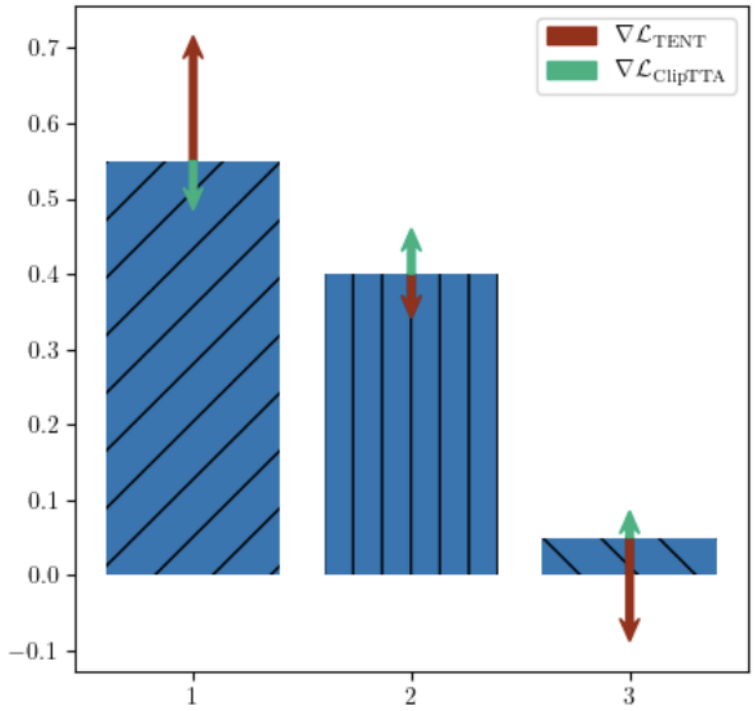
- Can points towards a class ≠ prediction
- Batch-aware gradient: leverage correct predictions to correct errors
- Mitigate pseudo-label drift

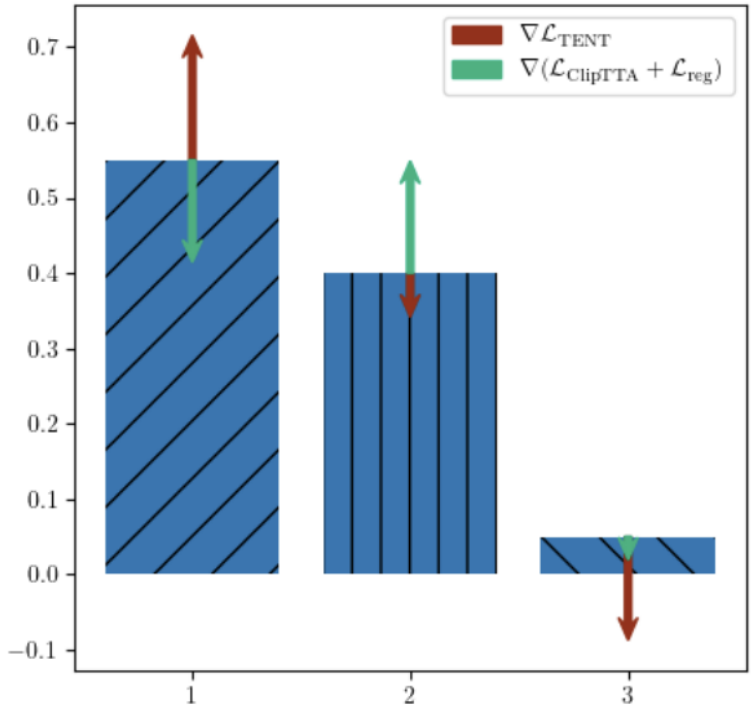
CLIP-TTA soft-contrastive loss: properties

- **Toy example:**

3-class classification problem, batch of size 6, sample 4 wrongly classified







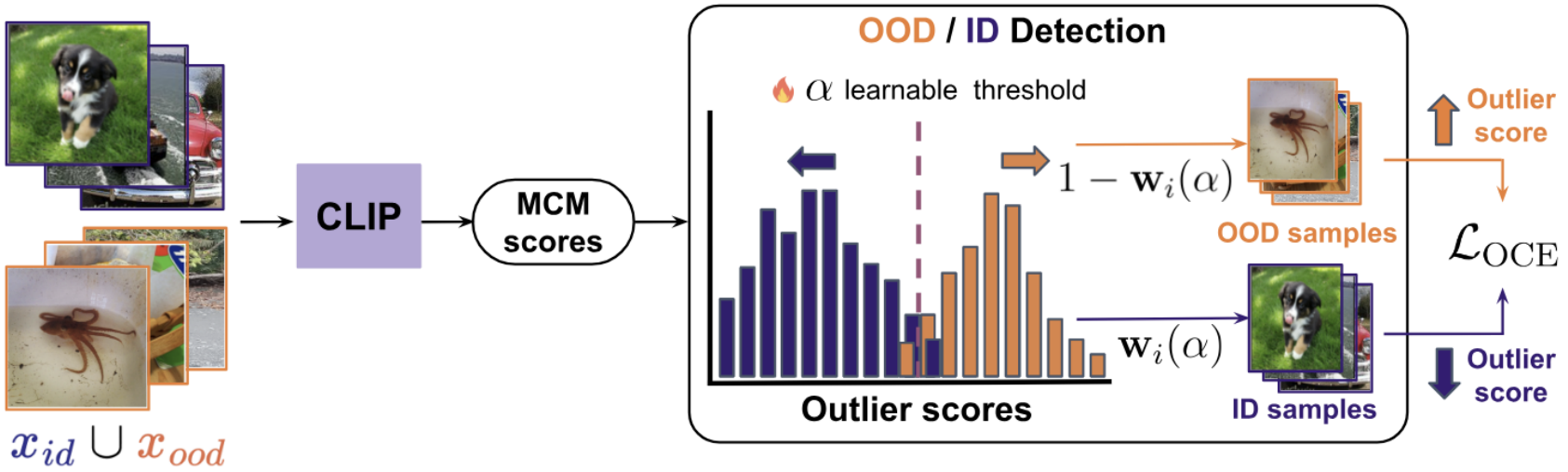
- CLIPTTA gradient points towards correct class

- Can be combined with a regularization loss

$$\mathcal{L}_{\text{reg}} = -\sum_{c=1}^{C} \bar{q}_c \log \bar{q}_c$$

Extension to open-set TTA

- Deal with In-Distribution (ID) & Out-Of-Distribution (OOD) samples in a batch



- Use MCM confidence, learn a separation threshold α => $w_i = \text{sigmoid}(s_i - \alpha)$
- Outlier Contrastive Exposure (OCE) loss to increase ID/OOD separability

$$\mathcal{L}_{\text{OCE}} = -\left[\underbrace{\frac{\sum_{i=1}^{N} w_i s_i}{\sum_{i=1}^{N} w_i}}_{\mu_{\text{id}}} - \underbrace{\frac{\sum_{i=1}^{N}(1-w_i)s_i}{\sum_{i=1}^{N}(1-w_i)}}_{\mu_{\text{ood}}}\right]^2.$$

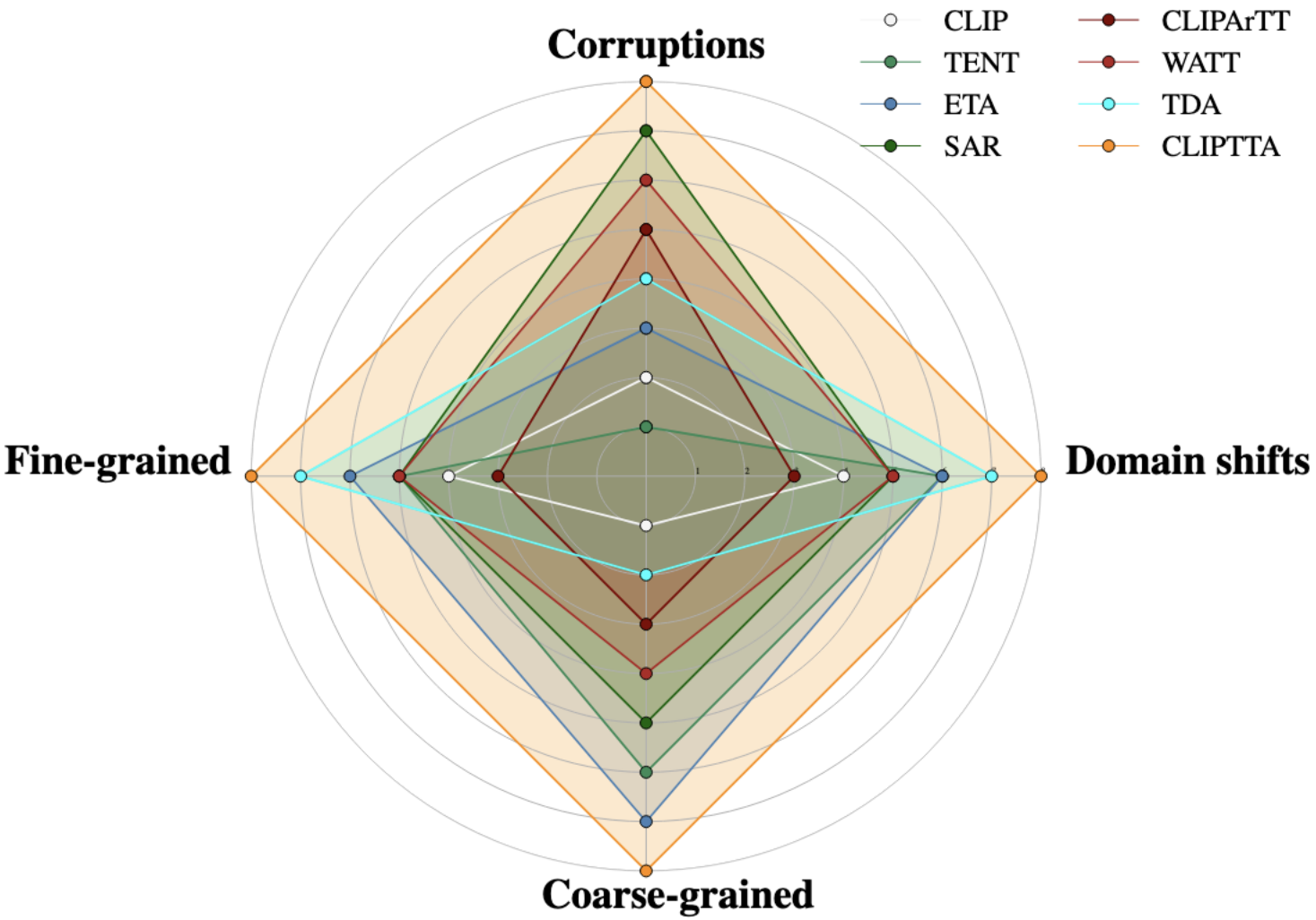
Experiments

- **Experiments on 75 datasets**
  - Corruptions: 15 corruptions applied to CIFAR-10, CIFAR-100, ImageNet
  - Domain shifts: VisDA-C, PACS, OfficeHome, Imagenet-Domains
  - Coarse grained classification: CIFAR-10, CIFAR-100
  - Fine-grained classification: Imagenet, + 10 datasets from the CLIP zero-shot suite

- **Baselines**
  - TENT-like: ETA, SAR, RoTTA, + CLIP TENT-like : CLIPArTT, WATT
  - More generic methods:
    - Test-Prompt-Tuning (TPT)
    - Parameter-free: Training-free Dynamic Adapter (TDA)

- Results in non-episodic setting (no reset)

Overall results



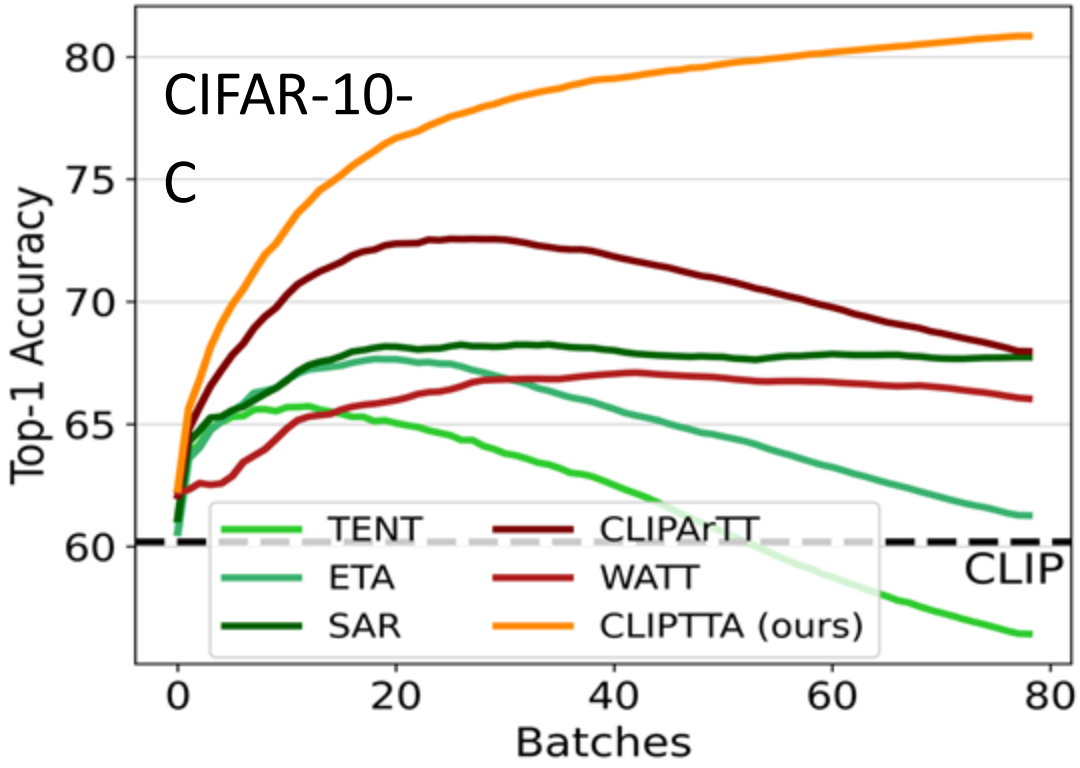
- Outperforms SOTA results on each class of dasets
- ≠ Specific methods effective only on some datasets, e.g., TDA on ImageNet

CLIPTTA Results: detailed results

	Corruptions				Domain shifts				
	CIFAR-10-C	CIFAR-100-C	Imagenet-C	Average	VisDA-C	PACS	OfficeHome	ImageNet-D	Average
CLIP	60.2	35.2	25.5	40.3	87.1	96.1	82.5	59.4	81.3
TPT (NeurIPS '22)	58.0	33.6	24.6	38.7	85.0	94.0	81.7	62.4	80.8
TDA (CVPR '24)	63.4	37.4	26.8	42.5	86.6	96.1	83.0	**65.0**	<u>82.8</u>
TENT (ICLR '21)	56.4	31.4	17.6	35.1	<u>89.3</u>	96.6	83.4	60.2	82.3
ETA (ICML '22)	61.3	38.9	26.8	42.3	88.3	<u>96.7</u>	<u>84.1</u>	59.9	82.3
SAR (ICLR '23)	67.8	43.2	<u>33.6</u>	48.2	87.8	96.2	83.8	60.6	82.1
RoTTA (CVPR '23)	58.0	33.6	24.6	38.7	83.7	95.8	82.5	61.6	80.9
CLIPArTT (WACV '25)	<u>68.1</u>	<u>48.0</u>	33.3	<u>49.8</u>	84.1	96.3	82.0	60.7	80.8
WATT (NeurIPS '24)	66.0	38.5	26.0	43.5	87.7	96.2	83.4	61.8	82.1
CLIPTTA (ours)	**80.7**	**52.5**	**41.1**	**58.1**	**89.6**	**97.5**	**84.2**	<u>63.4</u>	**83.7**

- Huge performance boost on corrupted datasets, where CLIP's ZS performance low

CLIP-TTA: robustness to pseudo-label drift

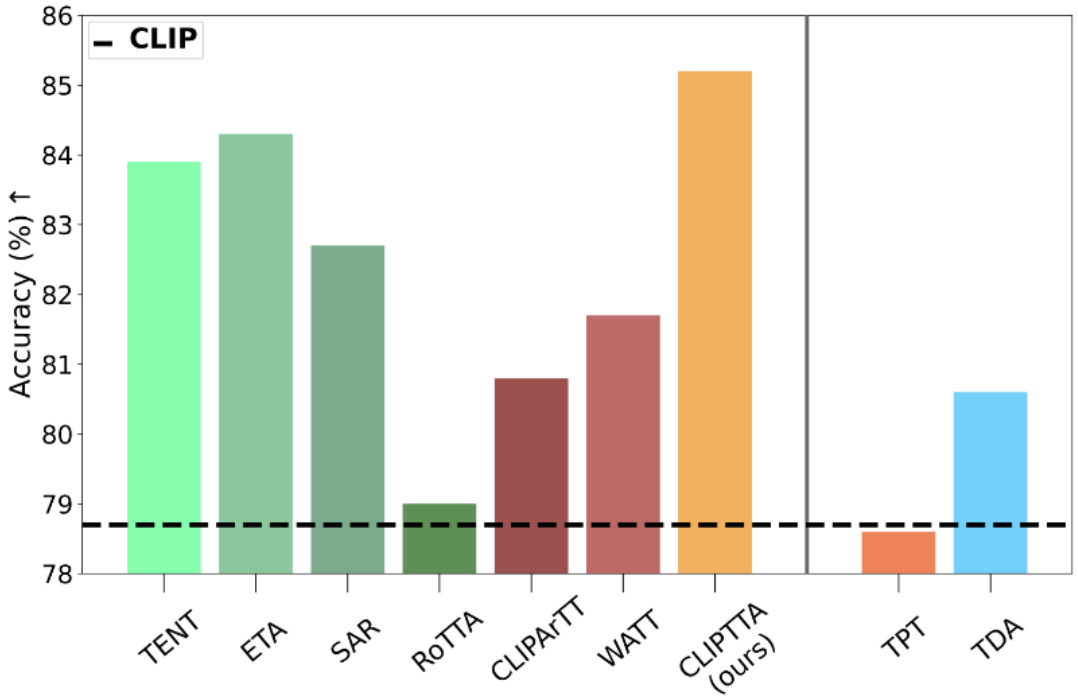


	C-100	C-100-C	IN	IN-C	Avg.
CLIP	68.1	35.2	66.7	25.5	48.9
TENT	72.9	31.4	66.5	17.6	47.1
$\mathcal{L}_{\text{s-cont}}$	74.2	50.8	68.8	40.3	58.5
$\mathcal{L}_{\text{s-cont}} + \mathcal{L}_{\text{reg}}$	74.9	52.4	69.1	38.6	58.8
$\mathcal{L}_{\text{s-cont}} + \mathcal{L}_{\text{reg}} + \mathcal{M}$	**75.3**	**52.6**	**69.6**	**41.1**	**59.6**

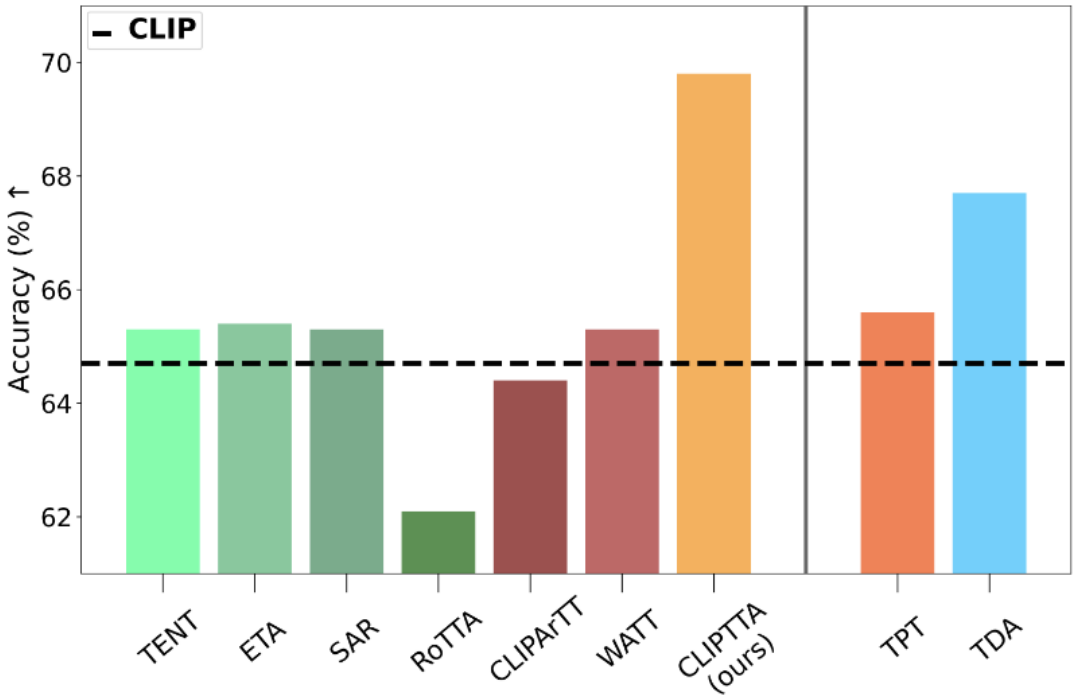
- CLIPTTA improves during adaptation
- TENT-based methods collapse

- Big improvements due to $\mathcal{L}_{\text{s-cont}}$
- Reg and memory can further boost results

CLIP-TTA: Coarse/fine-grained classification



(a) Coarse-grained results



(b) Fine-grained results

CLIP-TTA: open-set TTA

	ACC↑	AUC↑	FPR95↓
CLIP [1]	66.7	90.1	43.8
TENT [11]	12.4	49.9	89.4
ETA [12]	67.1	89.6	46.1
SAR [13]	58.8	62.0	75.7
CLIPArTT [6]	31.2	61.1	87.5
WATT [7]	67.1	87.4	53.4
TDA [4]	66.8	82.1	59.8
CLIPTTA (ours)	**67.6**	93.5	25.7
OSTTA [17] †	66.9	84.9	59.2
SoTTA[26] †	66.7	89.3	47.1
STAMP [27] †	29.7	63.0	80.2
UniEnt [28] †	65.2	95.4	17.1
CLIPTTA + OCE (ours) †	**67.6**	**97.7**	**9.7**

- ImageNet ID, Places OOD
- OOD detection: baselines << CLIP ≠ CLIPTTA (ours)
- CLIPTTA + OCE boost OOD detection with maintained accuracy

Conclusion & perspectives

- ViLU: failure prediction for VLMs
- CLIPTTA: open-set TTA for CLIP, OOD detection

Perspectives

- ViLU: generalization & combining failure, OOD + calibration
- Extension to semantic segmentation & theoretical guarantees (conformal)
- CLIPTTA: detecting ID failures before adaptation, UQ to overcome low ZS accuracy & pseudo-label drift

Thank you for your attention!

- Questions?