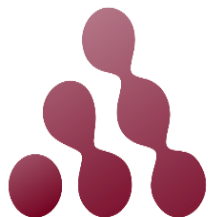


2025년 5월 9일 DMQA 연구실 오픈 세미나

Calibration for vision-language models

고려대학교 산업경영공학부 배진수



DMQA

Presenter

❖ 배진수

- 건국대학교 수학과 학부 졸업
- 고려대학교 산업경영공학과 박사 과정
- 고려대학교 DMQA 연구실 (지도교수:김성범)
- wlstn215@korea.ac.kr



❖ 연구분야

- Calibration for deep neural network
- Semi-supervised learning under class distribution mismatch
- Vision-language models

Content: Calibration for vision–language models

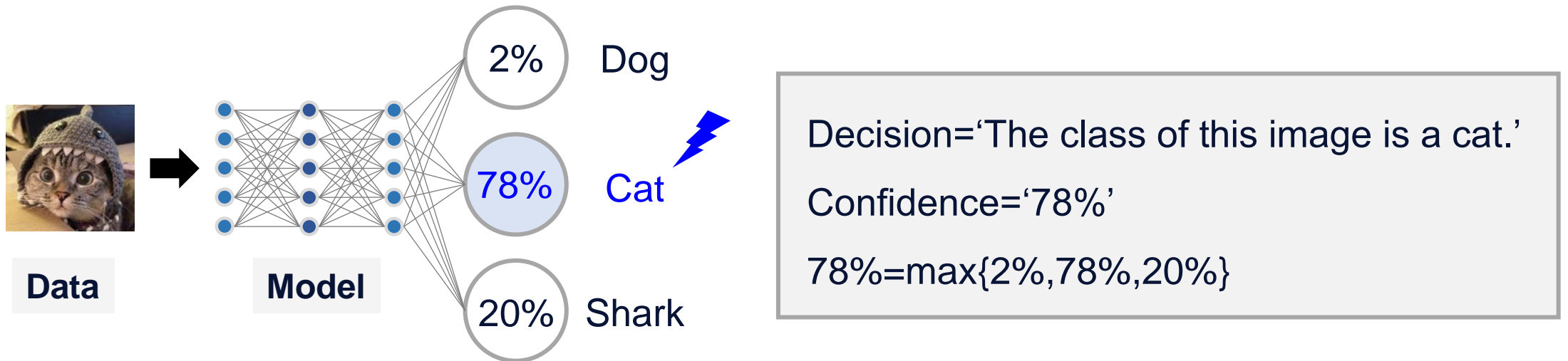
- Introduction
 - Calibration of deep learning models
 - Zero–shot classification using a vision–language model
- Paper review
 - An empirical study into what matters for calibrating vision–language models, 2024, ICML
 - C–tpt: Calibrated test–time prompt tuning for vision–language models via text feature dispersion, 2024, ICLR
- Conclusion

Introduction

Calibration of deep learning models

- **Data → Model → Probability → (Decision, Confidence)**

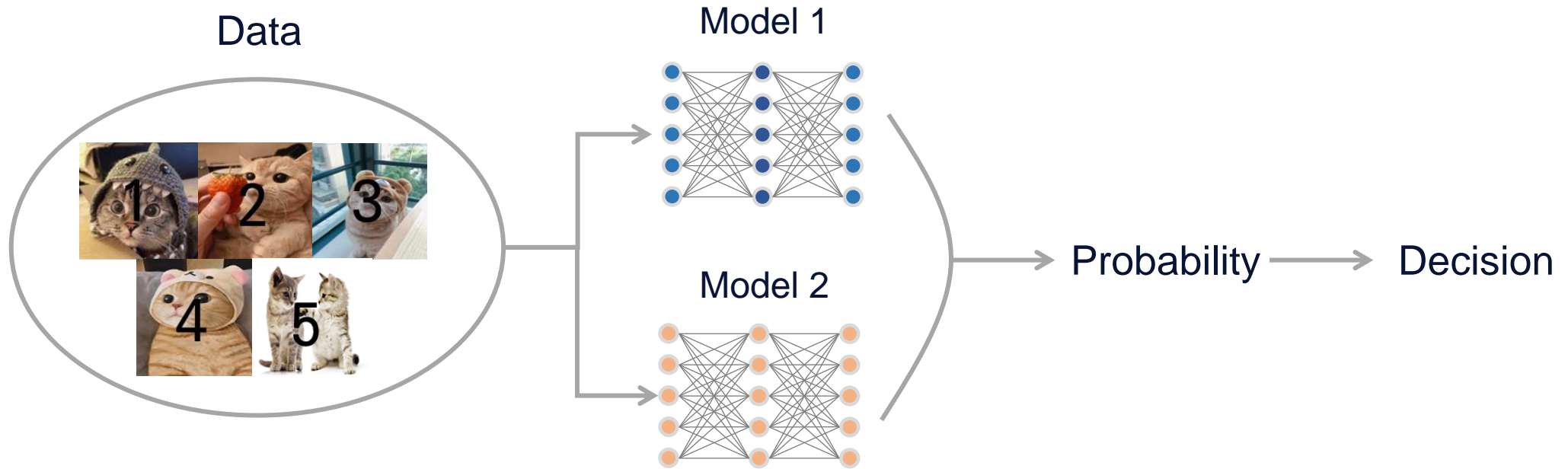
- $\text{Decision} = \text{argmax}\{\text{probability}\}$, $\text{Confidence} = \max\{\text{probability}\}$



Introduction

Calibration of deep learning models

- **Let's suppose the following two neural networks have to classify the below five images**








Introduction

Calibration of deep learning models

- Both models have the same accuracy with different confidence levels of 78.8% and 96.8%






Model 1: Well-calibrated

Accuracy and confidence are similar

	Data	Label	Decision	Confidence
○		Cat	Cat	75%
○		Cat	Cat	80%
✗		Cat	Dog	60%
○		Cat	Cat	80%
○		Cat	Cat	99%
Average confidence				78.8%
Accuracy				80%

Model 2: Poorly-calibrated

Accuracy and confidence are different

	Data	Label	Decision	Confidence
○		Cat	Cat	98%
○		Cat	Cat	97%
✗		Cat	Dog	95%
○		Cat	Cat	95%
○		Cat	Cat	99%
Average confidence				96.8%
Accuracy				80%

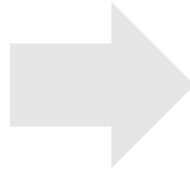
Introduction

Calibration of deep learning models

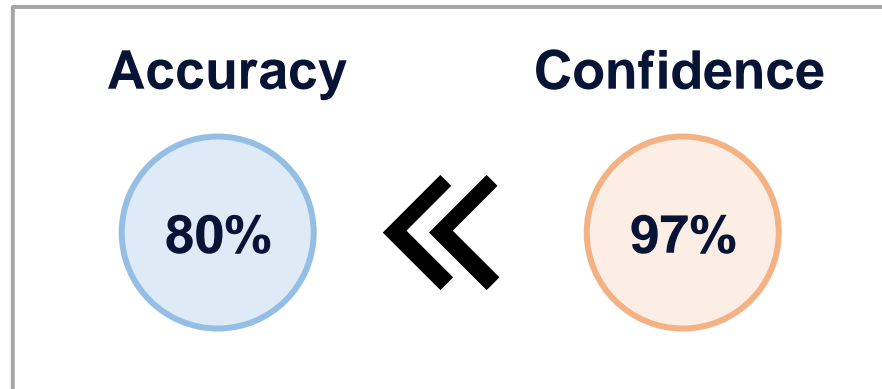
- **Chuan Guo, et al. “On calibration of modern neural networks.” PMLR, (2017)**
 - Most modern neural networks are assessed as overconfident (=poorly calibrated)

“Modern deep neural networks”

such as ResNet, VGG, Visual Transformer ...



“Poorly-calibrated model”



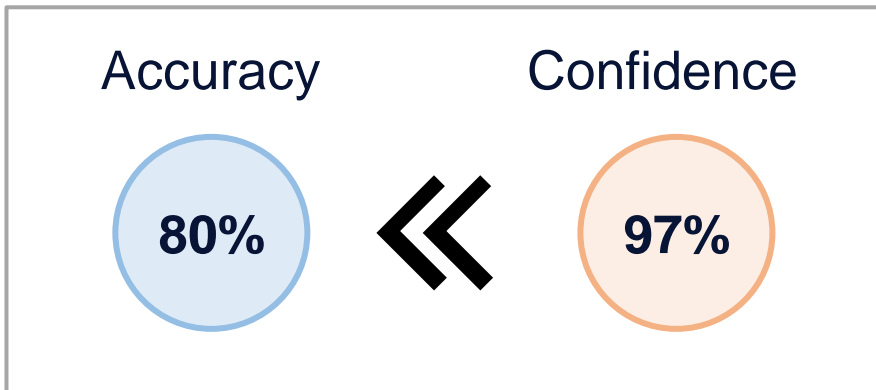
Introduction

Calibration of deep learning models

- **Improving calibration: Aligning the accuracy and confidence to be similar**
 - Improving calibration makes the model to be used safely in real-world applications
 - A well-calibrated model can be aware of their failed prediction

Modern deep neural networks

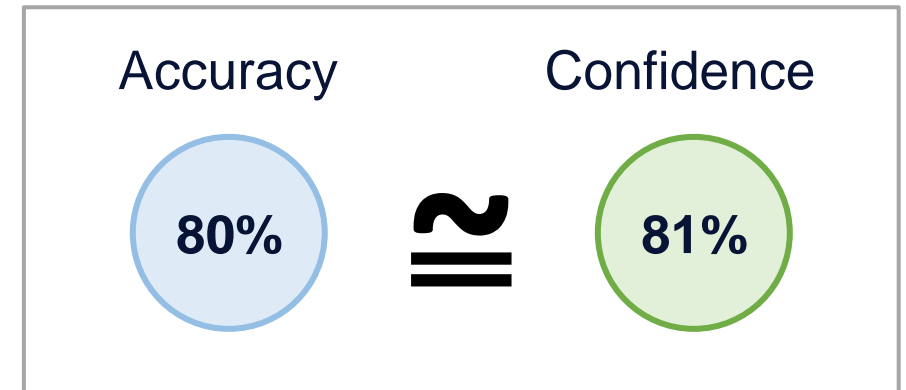
Poorly-calibrated model



Improving
calibration



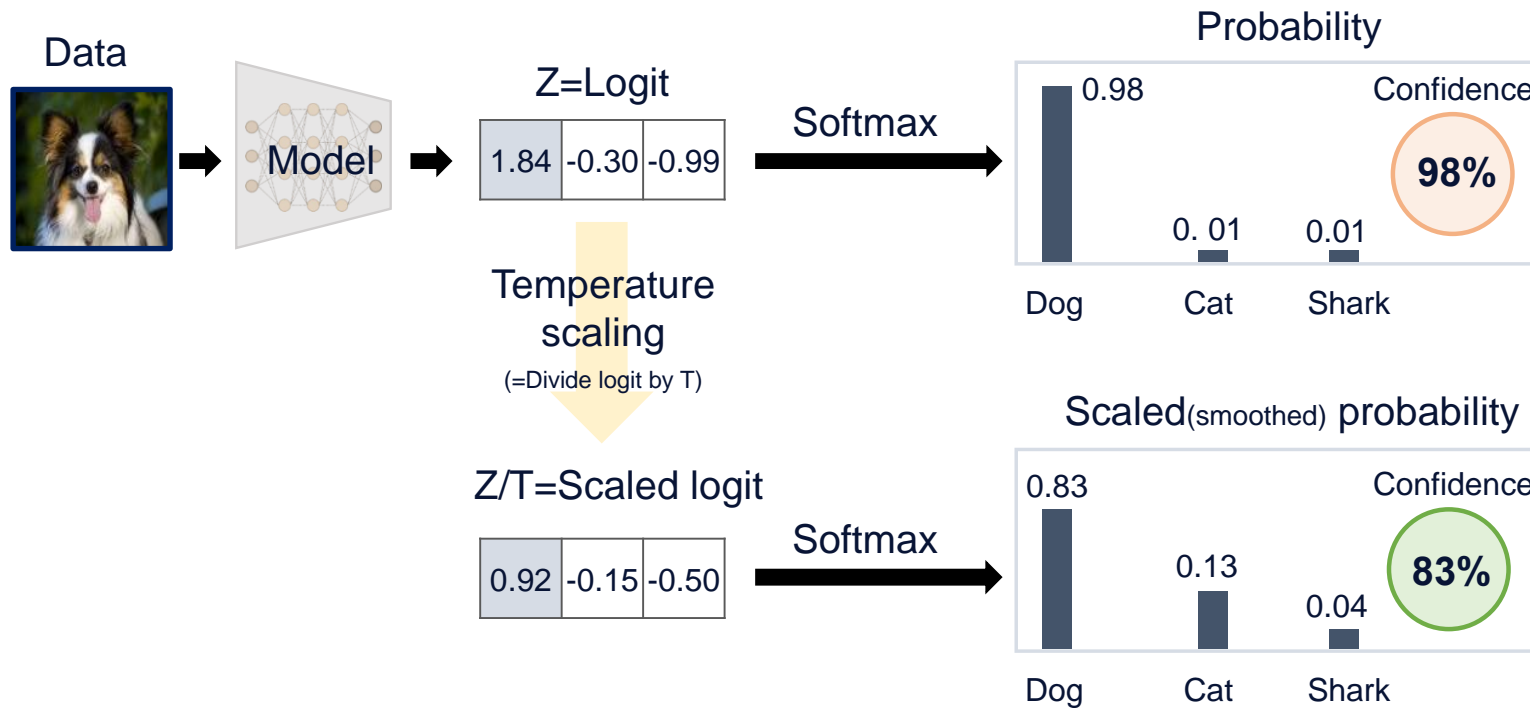
Well-calibrated model



Introduction

Calibration of deep learning models

- **Chuan Guo, et al. “On calibration of modern neural networks.” PMLR, (2017)**
 - Scaling logit vector to mitigate overconfidence in prediction → method name = temperature scaling (TS)
 - TS investigates which scaling factor value (T) can attain the most optimal calibration result



Introduction

Zero-shot classification using a vision-language model (VLMs)

- **VLM can understand visual and linguistic information simultaneously [1]**
 - Therefore, achieving various downstream tasks such as visual question answering, image captioning, etc.

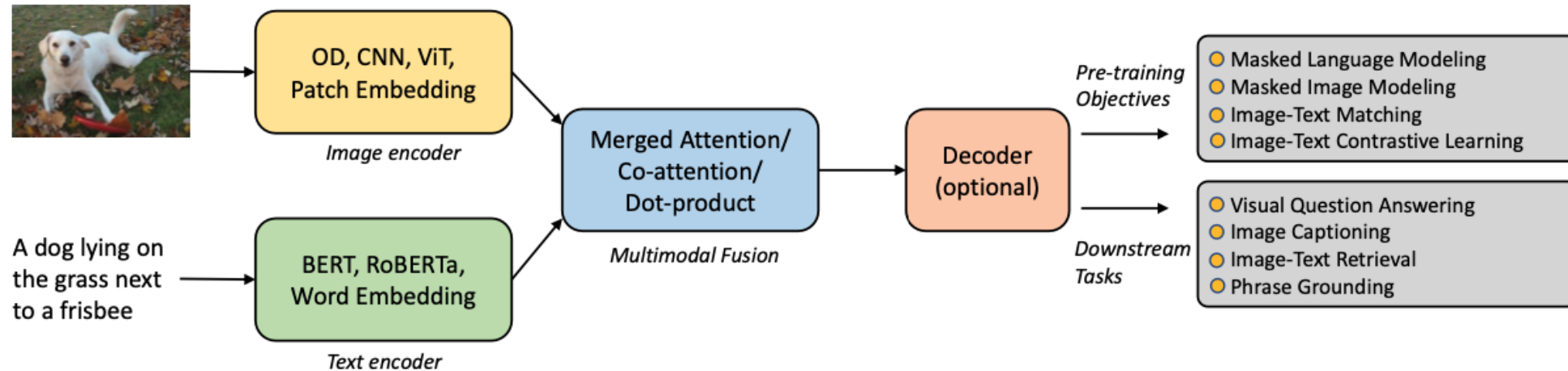
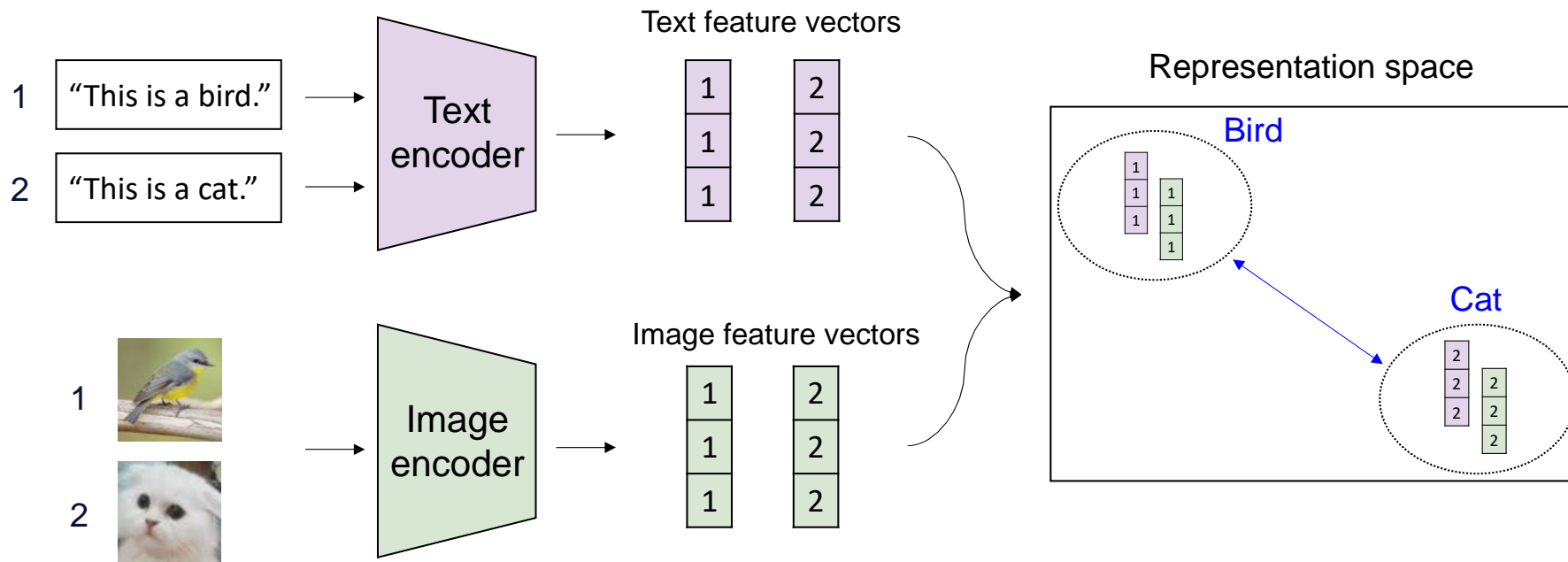


Figure 3.3: Illustration of a general framework for Transformer-based vision-language models.

Introduction

Zero-shot classification using a vision-language model (VLMs)

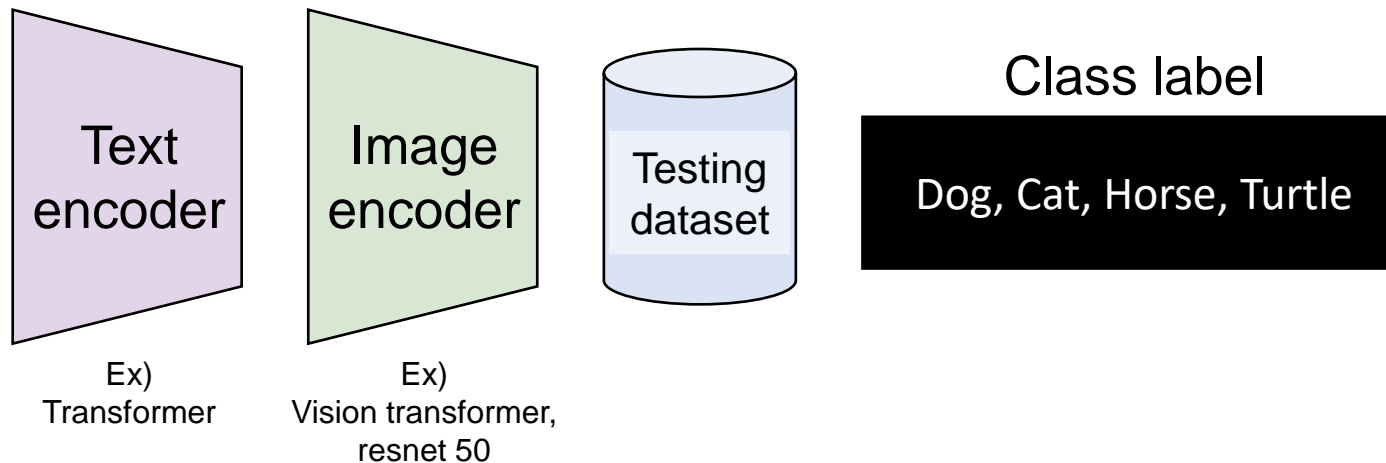
- **VLM can understand visual and linguistic information simultaneously [1]**
 - Ex) Text and image features related to birds are similar to each other, and text and image features related to cats are similar to each other.



Introduction

Zero-shot classification using a vision-language model (VLMs)

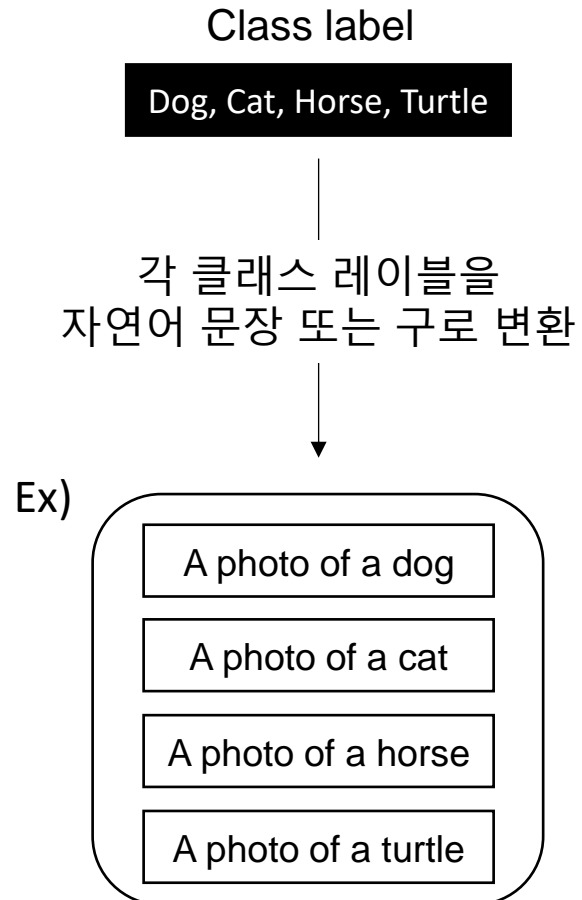
- For zero-shot classification tasks, we need VLM, a testing dataset, and class labels (text)



Introduction

Zero-shot classification using a vision-language model (VLMs)

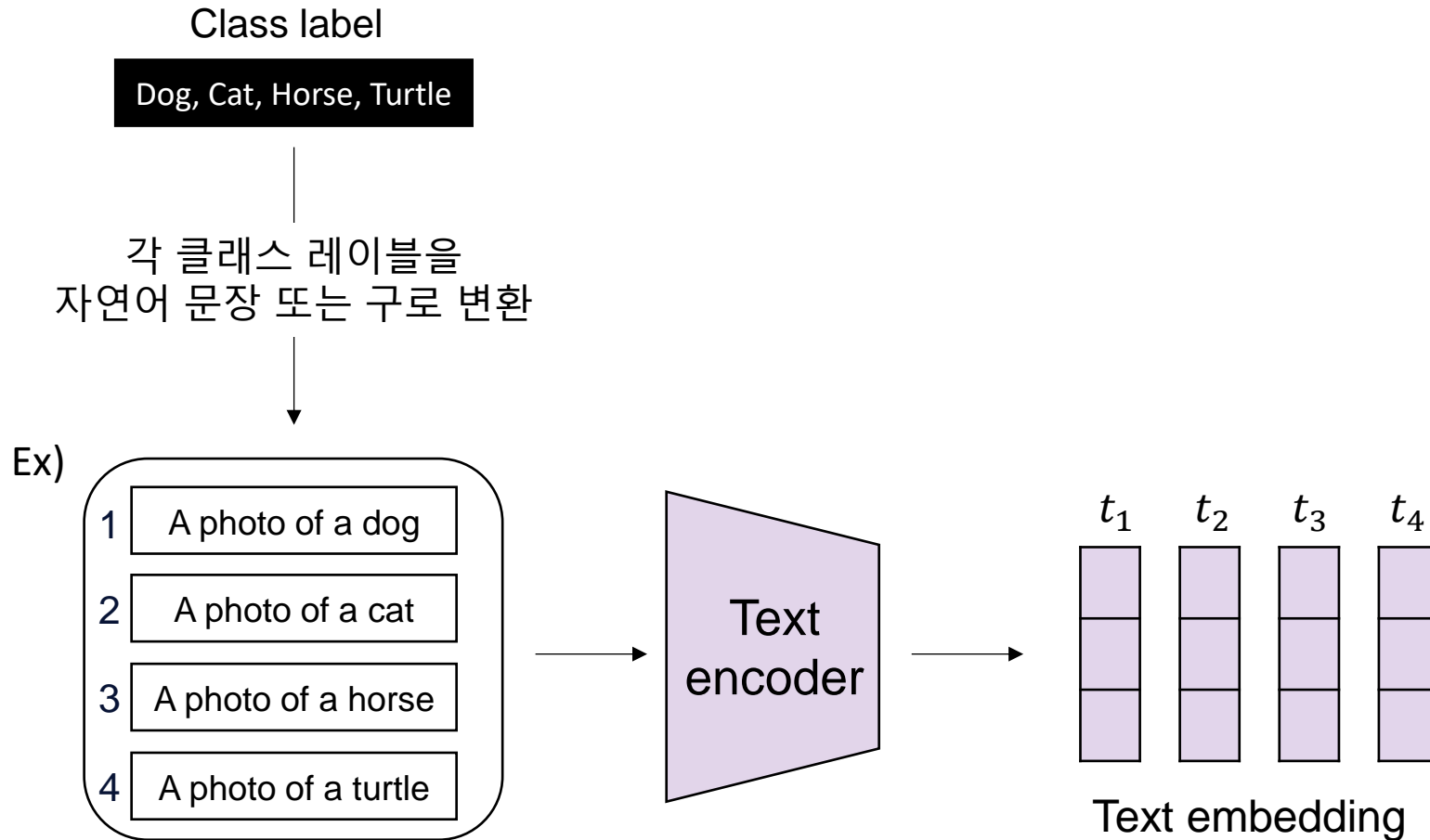
1. Prompt engineering: a photo of a + <class>



Introduction

Zero-shot classification using a vision-language model (VLMs)

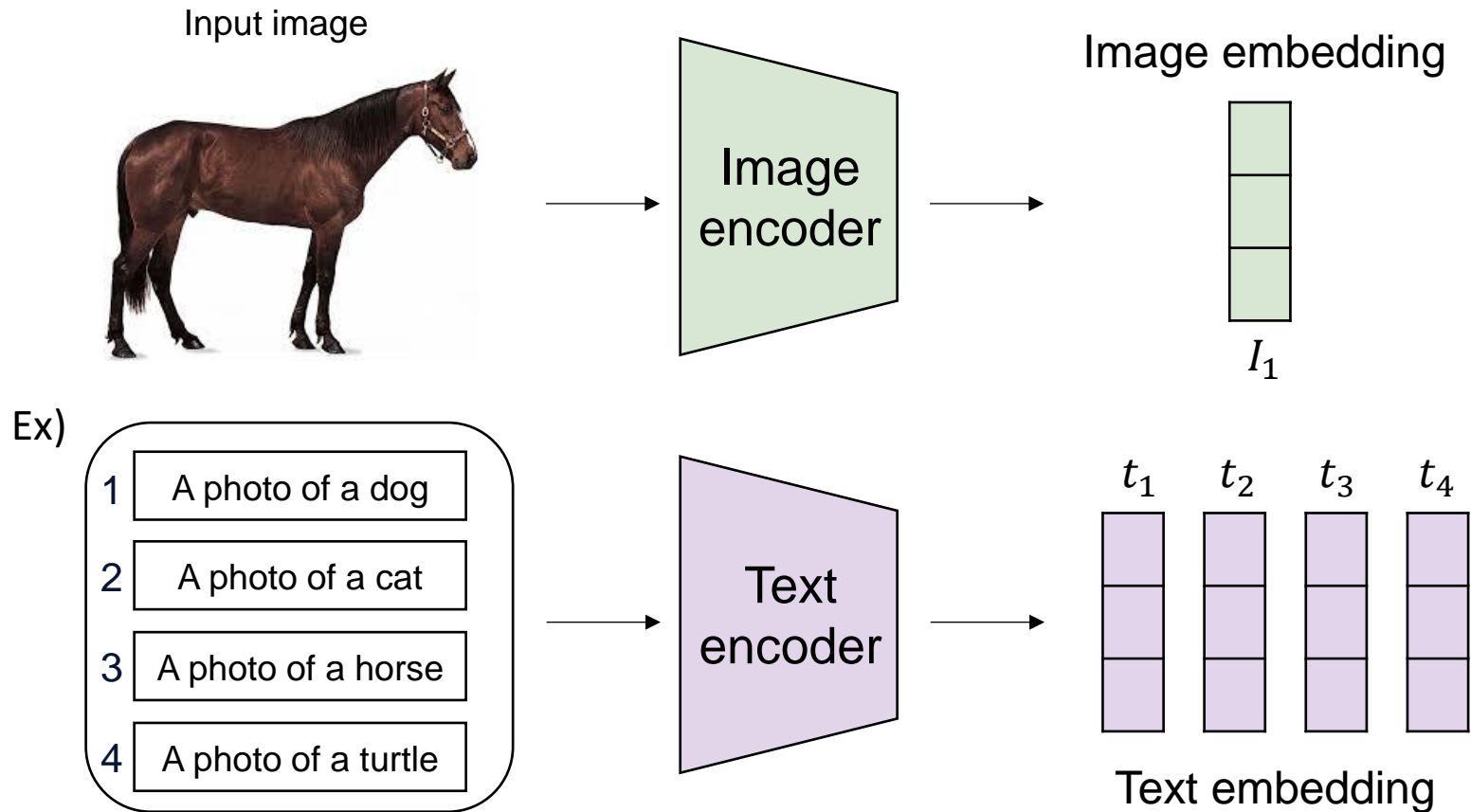
2. Extract text feature vectors from the generated prompts



Introduction

Zero-shot classification using a vision-language model (VLMs)

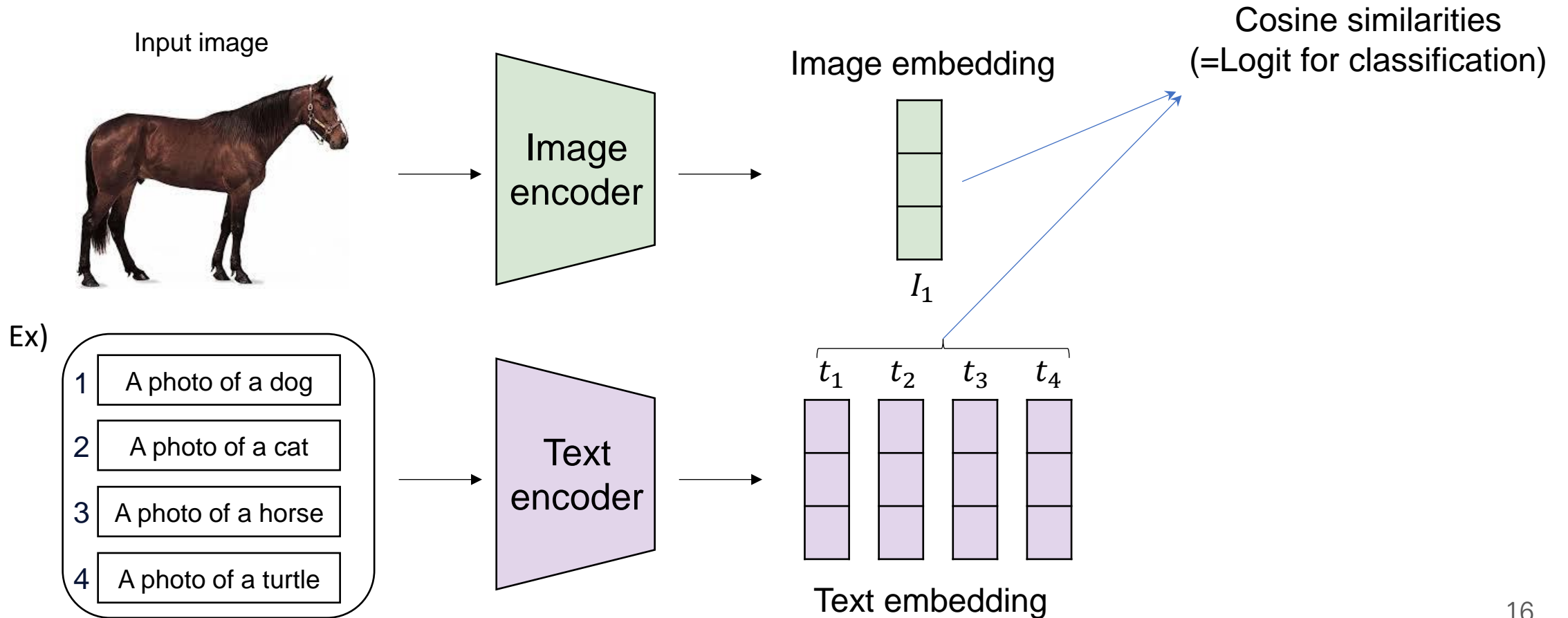
3. Extract an image feature vector from the input image



Introduction

Zero-shot classification using a vision-language model (VLMs)

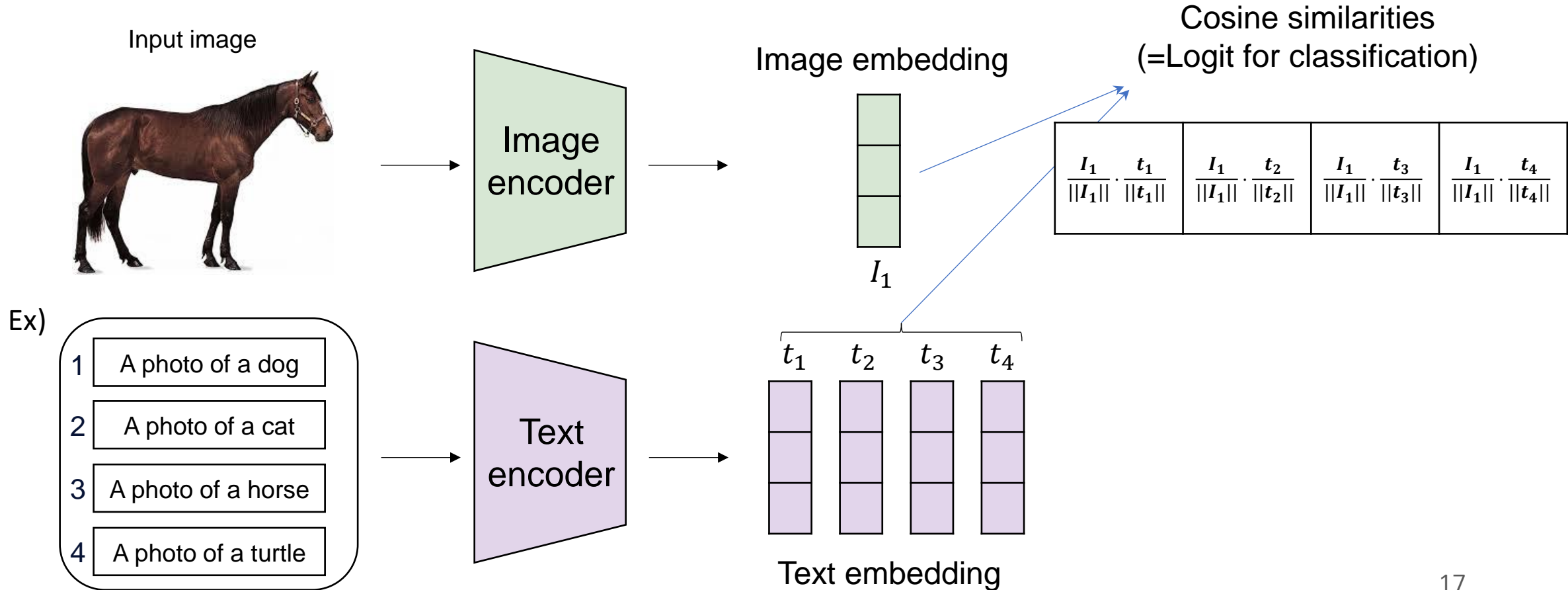
4. Calculate the cosine similarities between the image feature and the text features



Introduction

Zero-shot classification using a vision-language model (VLMs)

4. Calculate the cosine similarities between the image feature and the text features

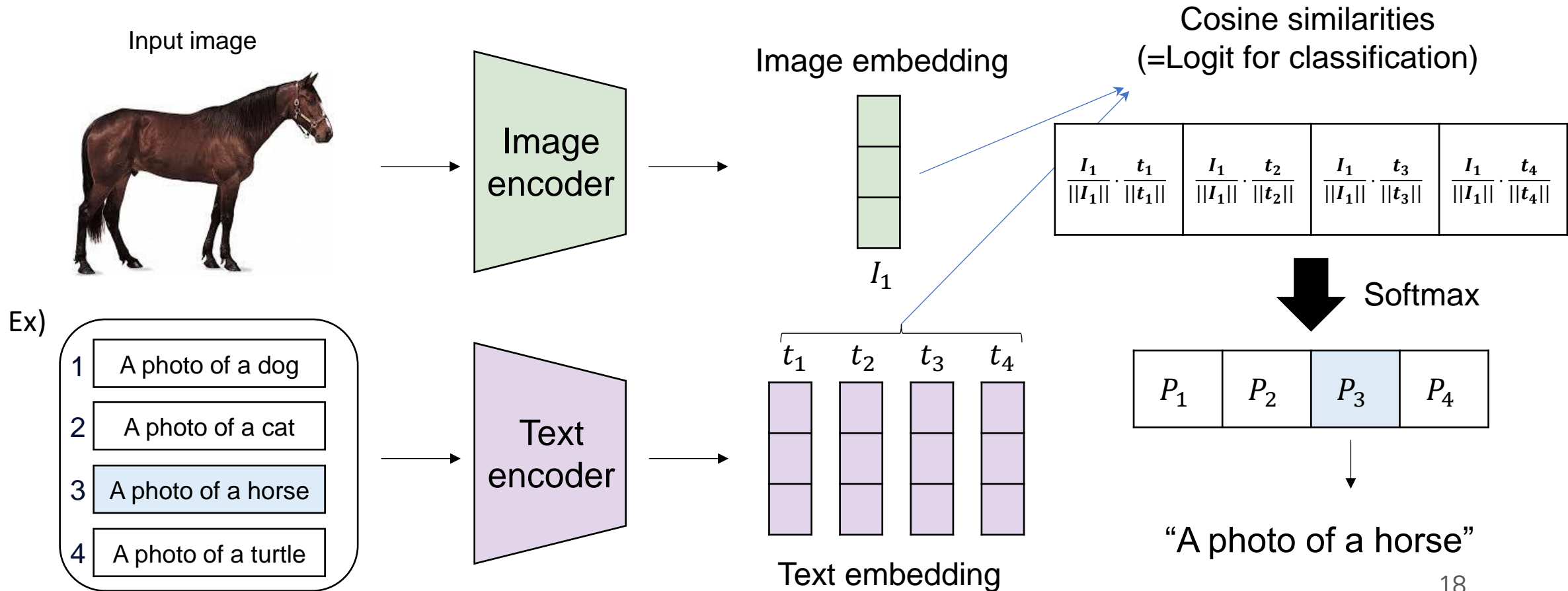


Introduction

Zero-shot classification using a vision-language model (VLMs)

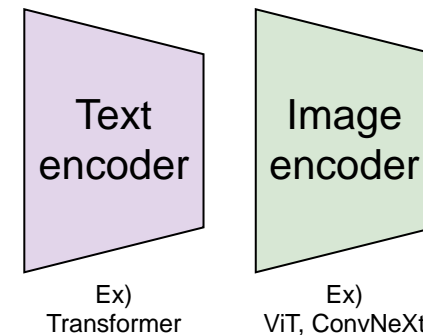
5. Transform the cosine similarities into probabilities over the all classes using softmax

- Max prob = confidence / Class with the max prob = decision



Paper review (1)

An empirical study into what matters for calibrating VLMs, 2024, ICML

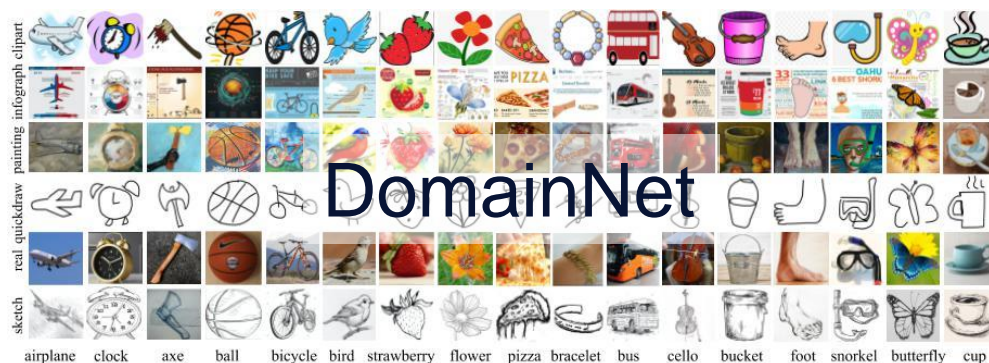


- **Models:** 35 VLMs (including CLIP, BLIP) with diverse training data and architectures
 - They have various image-text pre-training frameworks, such as CLIP and BLIP
 - They also have different visual encoder architectures (e.g., ViT and ConvNeXt) and training dataset distributions and quantities
- **Baseline:** Non-VLM models
 - ImageNet-trained CNNs (ResNet) and vision transformers (ViT)

Paper review (1)

An empirical study into what matters for calibrating VLMs, 2024, ICML

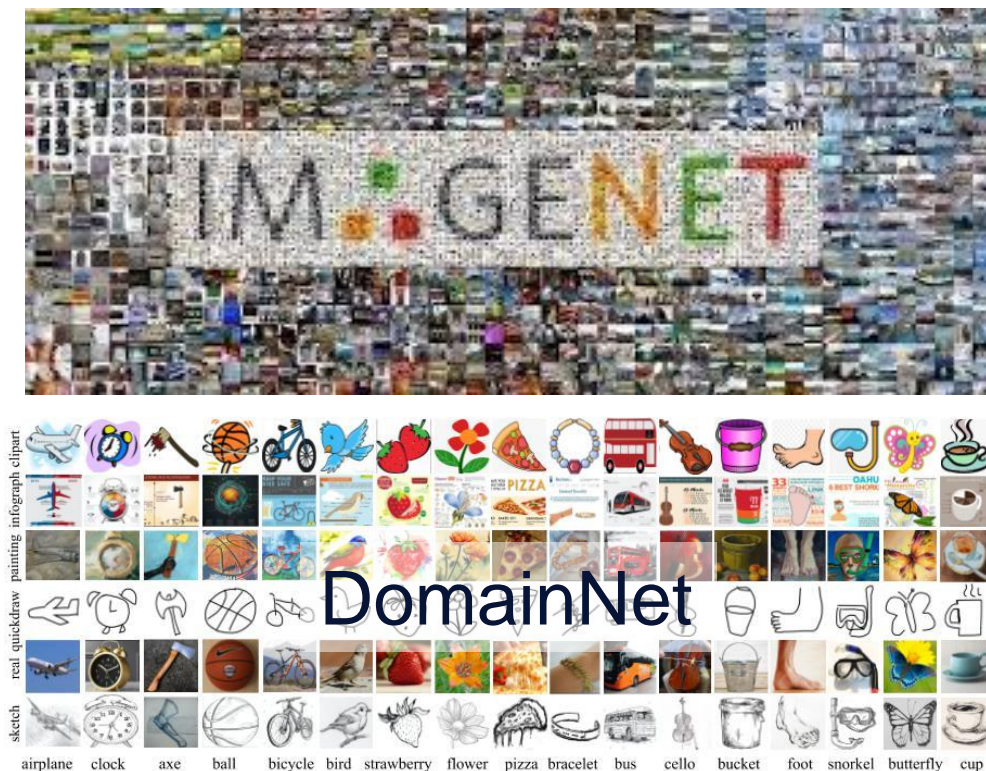
- **Datasets:** ImageNet, CIFAR-10, DomainNet, including out-of-distribution scenarios
 - Three standard image classification benchmarks



Paper review (1)

An empirical study into what matters for calibrating VLMs, 2024, ICML

- **Datasets:** ImageNet, CIFAR-10, DomainNet, including out-of-distribution scenarios
 - Three standard image classification benchmarks

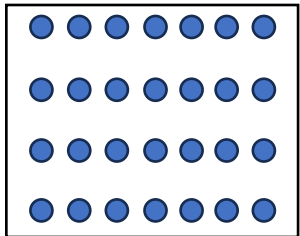


Paper review (1)

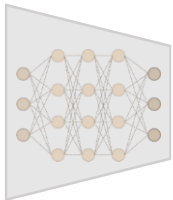
An empirical study into what matters for calibrating VLMs, 2024, ICML

- **Evaluation:** expected calibration error (ECE)
 - Difference between accuracy and confidence

Dataset



Model



● :Data point

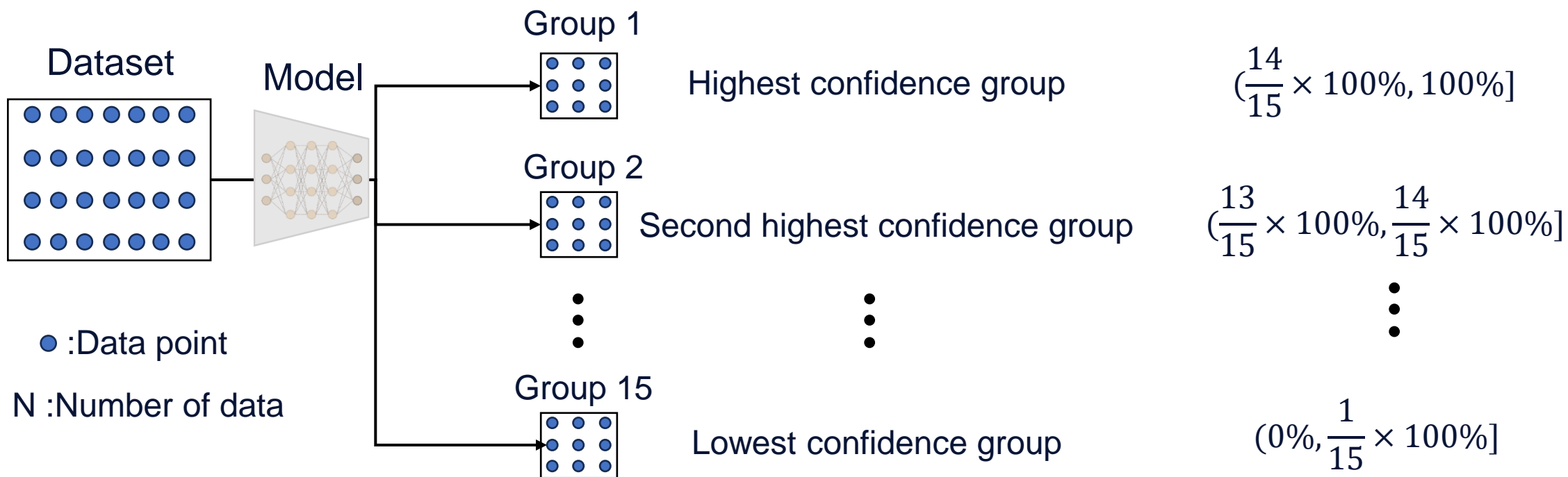
N :Number of data

Paper review (1)

An empirical study into what matters for calibrating VLMs, 2024, ICML

- **Evaluation:** expected calibration error (ECE)

- Difference between accuracy and confidence

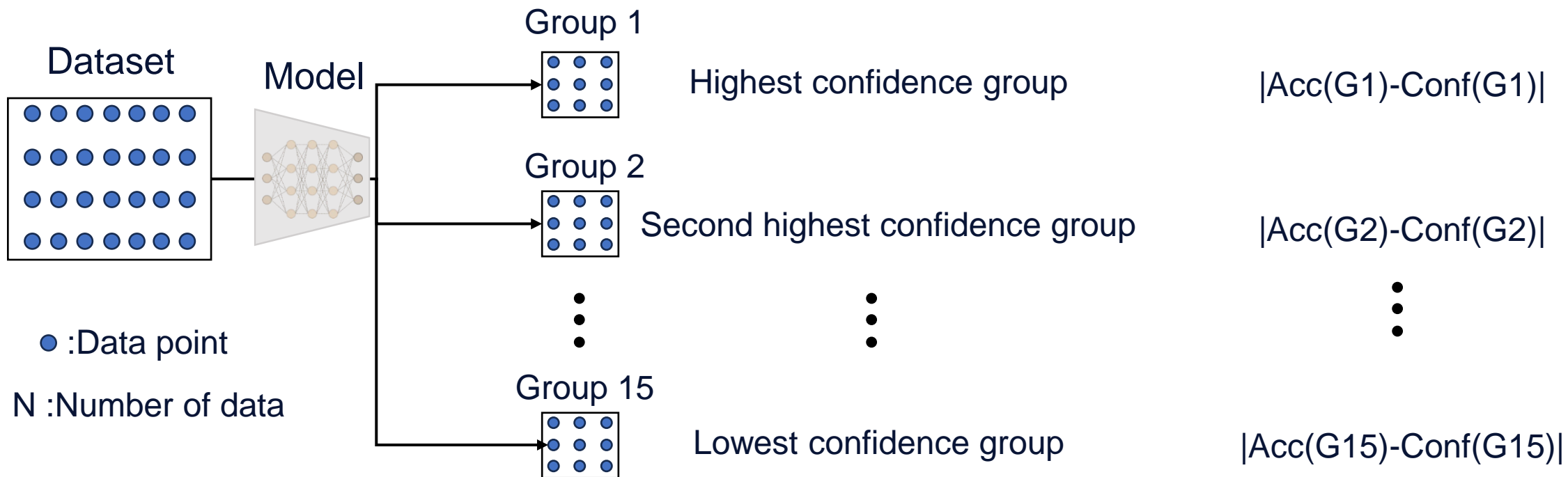


Paper review (1)

An empirical study into what matters for calibrating VLMs, 2024, ICML

- **Evaluation:** expected calibration error (ECE)

- Difference between accuracy and confidence

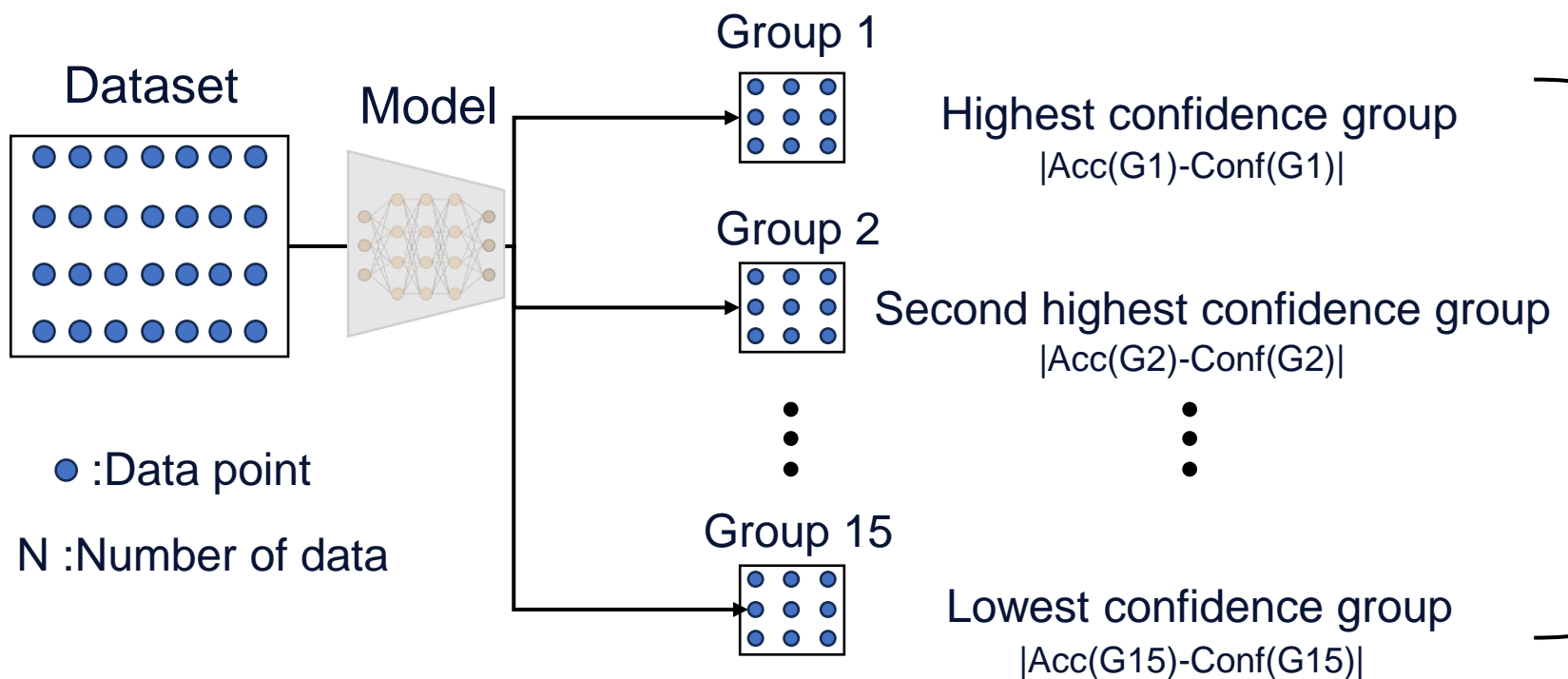


Paper review (1)

An empirical study into what matters for calibrating VLMs, 2024, ICML

- **Evaluation:** expected calibration error (ECE)

- Difference between accuracy and confidence



Expected calibration error (ECE)

$$\sum_{m=1}^{15} \frac{|G_m|}{N} |\text{Acc}(G_m) - \text{Conf}(G_m)|$$

Paper review (1)

An empirical study into what matters for calibrating VLMs, 2024, ICML

1. Before calibration, VLMs have similar or worse ECE compared to ImageNet-trained models.
2. After temperature scaling, VLM calibration significantly improves (average ECE reduced to ~ 0.05).
3. Stable calibration performance under distribution shifts.

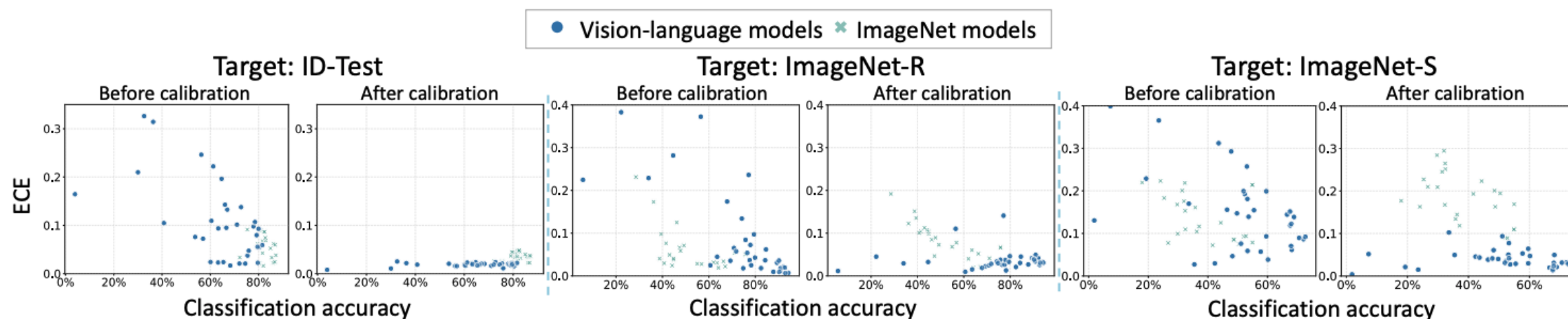
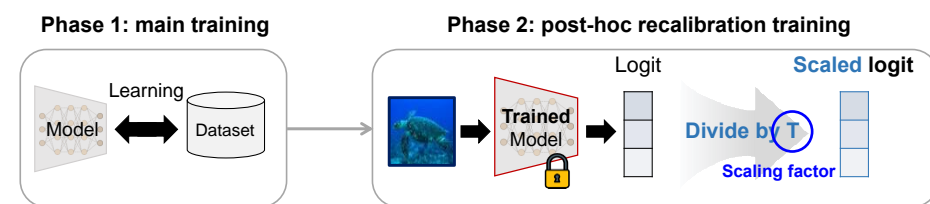


Figure 1. Comparing the calibration performance of ImageNet-trained models and VLMs.

Paper review (1)

An empirical study into what matters for calibrating VLMs, 2024, ICML



1. VLMs can effectively be calibrated even if the calibration set has different labels from the target test set.
2. High correlation (R^2 , Spearman's $\rho > 0.90$) between calibrated probabilities and actual accuracy, even with label mismatches.

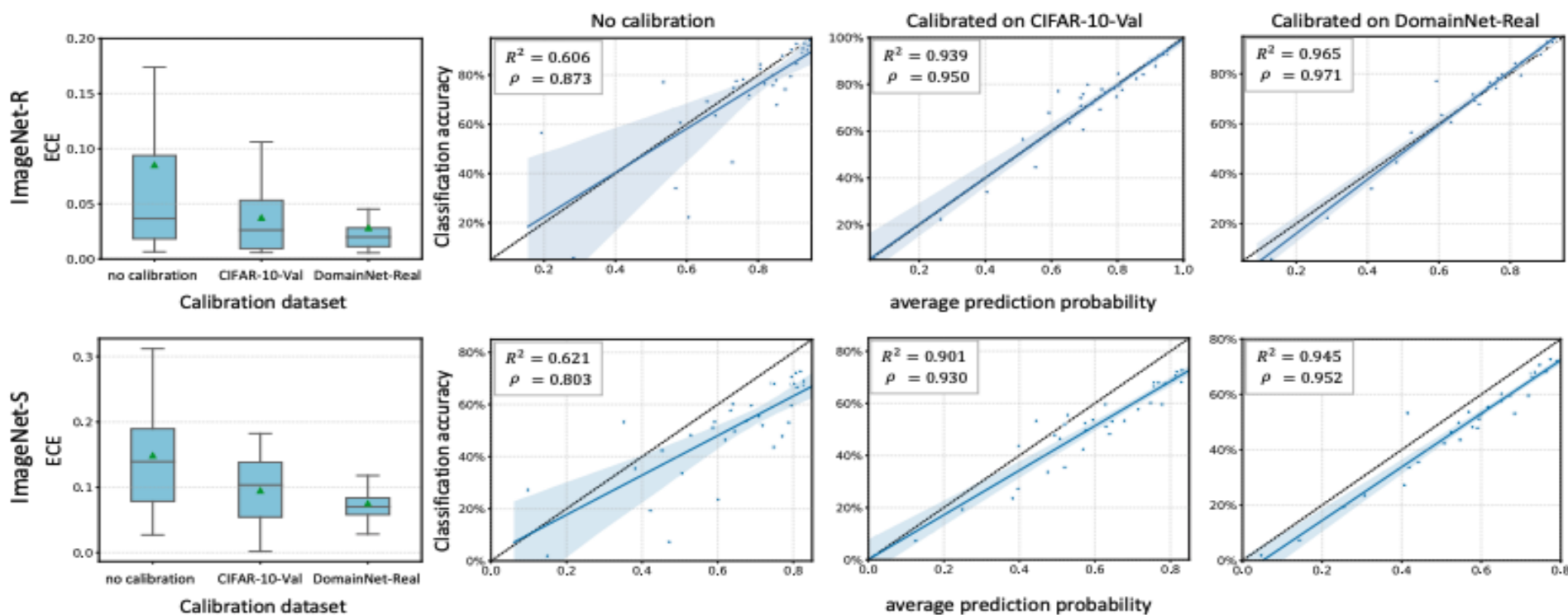


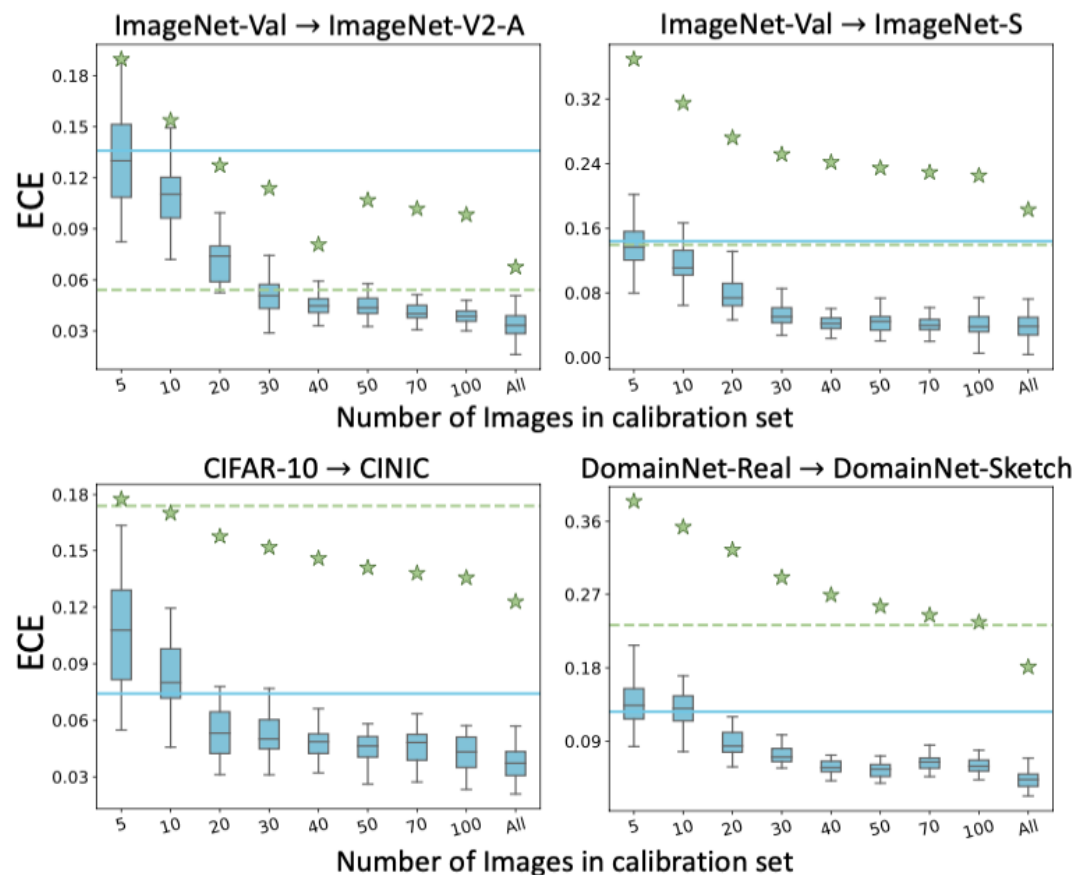
Figure 2. Adaptability of VLMs to different calibration label sets.

Paper review (1)

An empirical study into what matters for calibrating VLMs, 2024, ICML

1. VLM calibration requires very few examples (~40-50 samples) to reach optimal calibration performance
2. Effective even in high-class-count scenarios (e.g., DomainNet(=345), ImageNet(=1000))

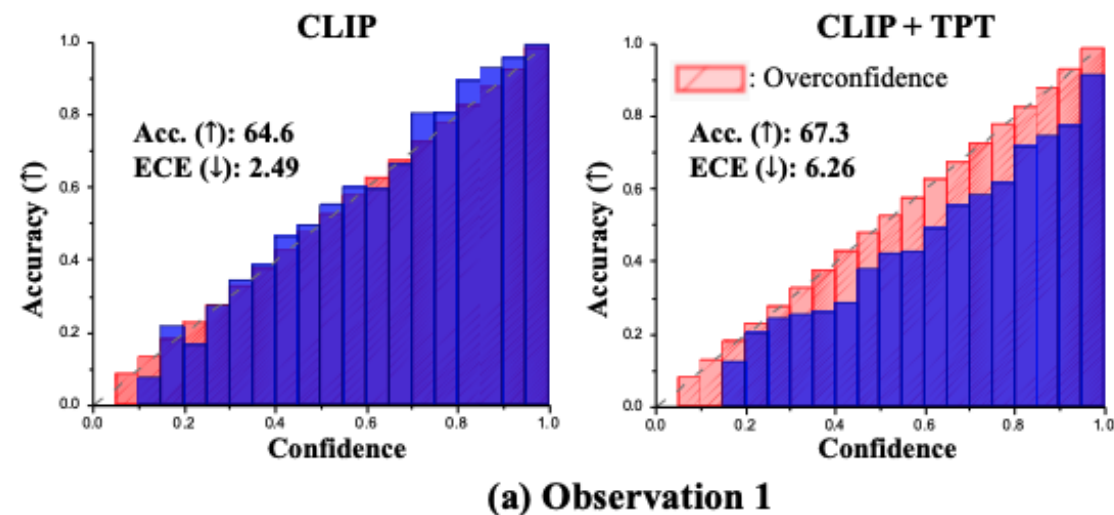
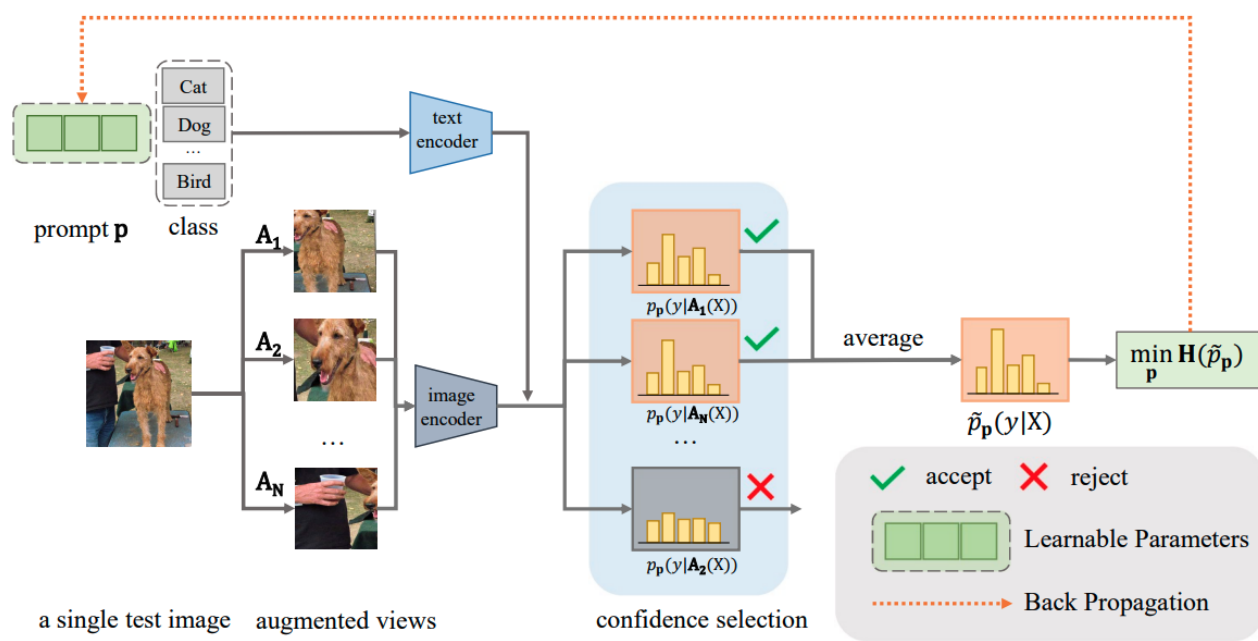
Blue: VLM models
Green: Non-VLM models
Line: Before calibration



Paper review (2)

C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion, 2024, ICLR

- Test-time prompt tuning (TPT) improves CLIP accuracy without labels.
- But: TPT often worsens calibration, leading to overconfident predictions.

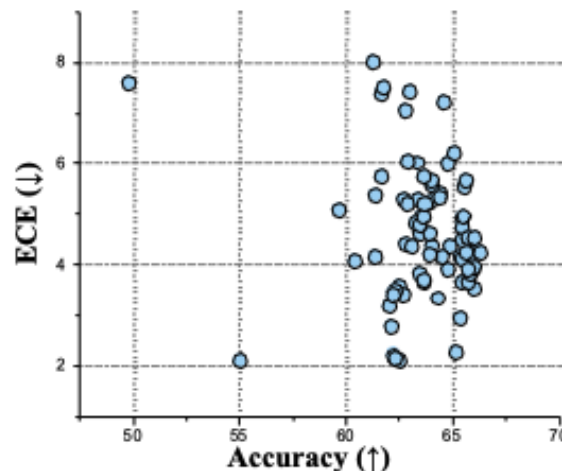


Paper review (2)

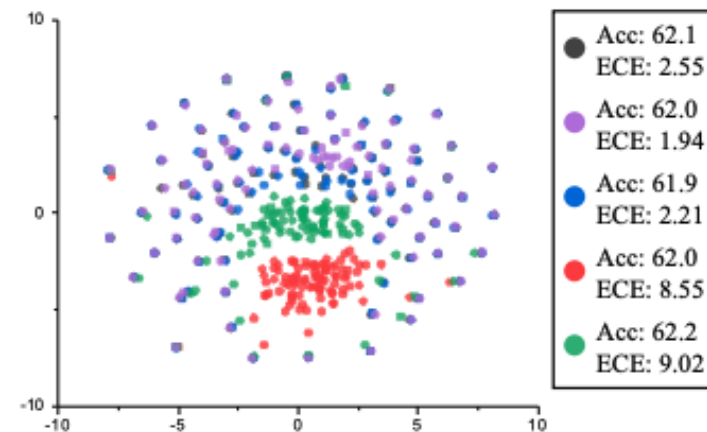
C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion, 2024, ICLR

- Discover that **prompt choice strongly influences calibration**.
- **Well-calibrated prompts show high dispersion** in class-embedded text features.

1. A photo of a <class>
2. A painting of a <class>
3. A photo of a clean <class>
4. A drawing of a <class>
- ⋮
80. A sketch of a <class>



(b) Observation 2



(c) Observation 3

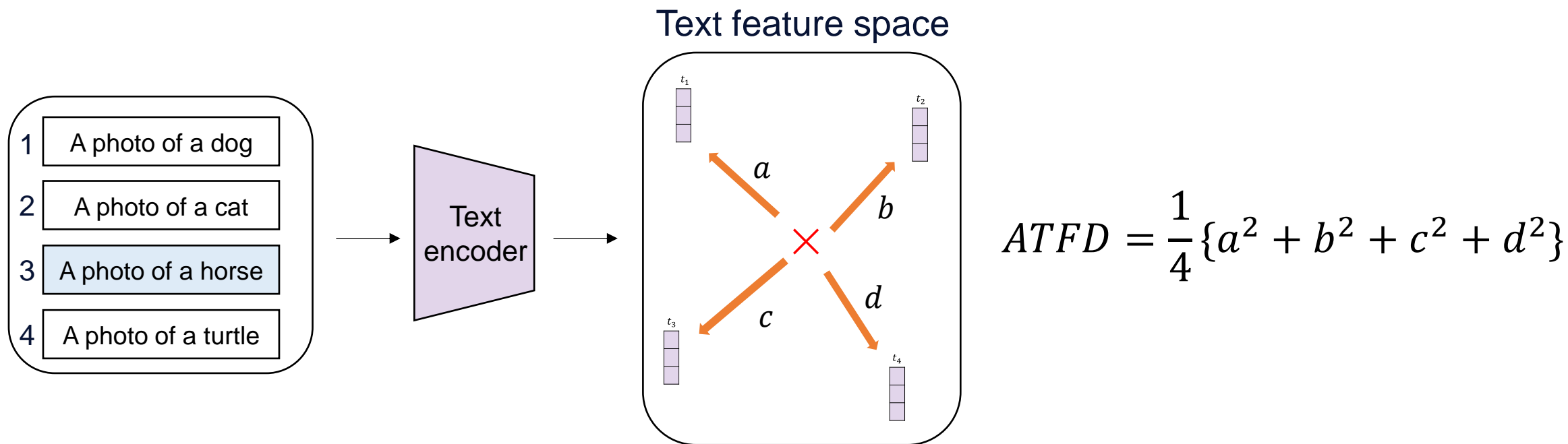
Paper review (2)

Dispersion \propto Calibration

C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion, 2024, ICLR

- How to measure dispersion in class-embedded text features: **Average text feature dispersion (ATFD)**

$$t_{centroid} = \frac{1}{K} \sum_{k=1}^K t_k \quad // \quad ATFD = \frac{1}{K} \sum_{k=1}^K ||t_{centroid} - t_k||^2$$



Paper review (2)

C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion, 2024, ICLR

- Empirically observed strong negative correlation between ATFD and ECE ($\rho \approx -0.7$ to -0.76)

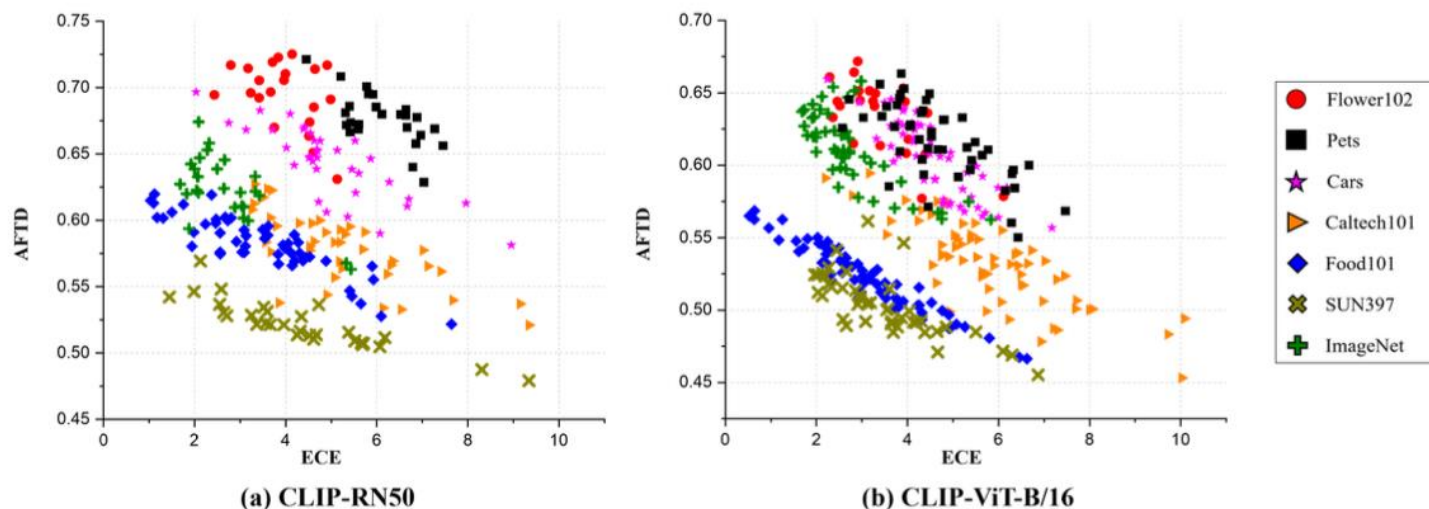
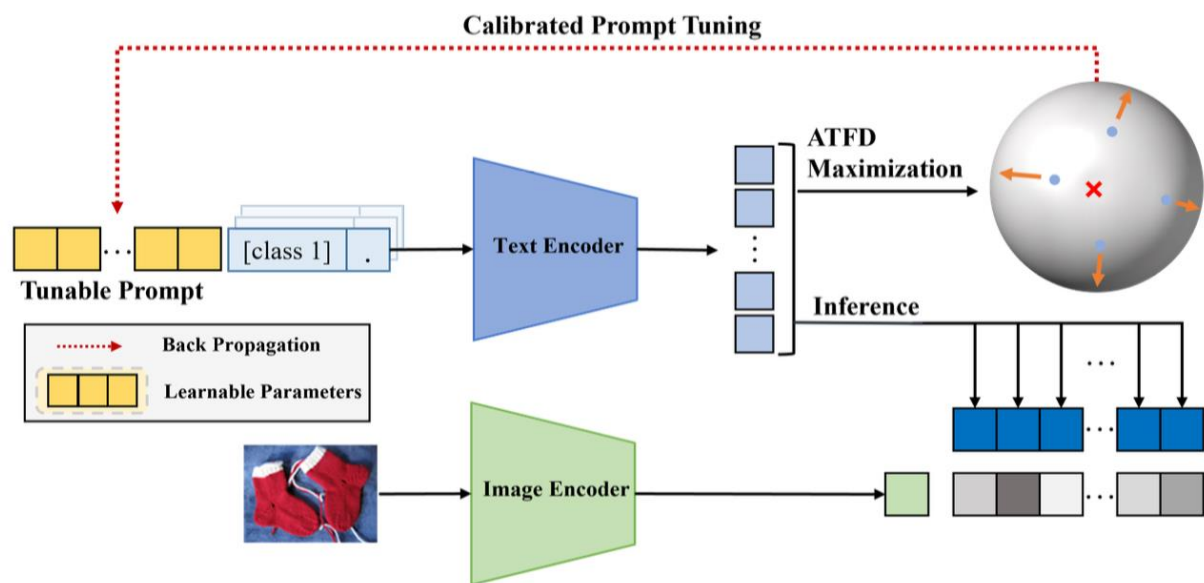


Figure 2: **Plot illustrating the correlation between ECE and ATFD** for hard prompts that achieve accuracies within 3% of the highest accuracy observed for each dataset. A notable negative association is observed for CLIP-RN50 and CLIP-ViT-B/16 across different datasets, with Pearson correlation coefficients (Freedman et al., 2007) averaging -0.70 and -0.76, respectively.

Paper review (2)

C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion, 2024, ICLR

- Proposed Method = Maximizing text feature dispersion during TPT



$$\min_p [L_{TPT} + \lambda \cdot (-ATFD)]$$

- Jointly optimize for accuracy (TPT) and calibration (via ATFD)
- No labeled data needed
- $\lambda = 50$ (tunable)

Figure 3: **Illustration of the Calibrated Test-time Prompt Tuning (C-TPT)** for zero-shot image classification using CLIP. C-TPT improves calibration by optimizing the prompt so that it maximizes the Average Text Feature Dispersion (ATFD) during test-time prompt tuning.

Paper review (2)

C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion, 2024, ICLR

Table 1: **Fine-Grained Classification.** We present the results of CLIP-RN50 and CLIP-ViT-B/16. For each setting, we report the **Acc.** (\uparrow) and **ECE** (\downarrow) of the initialization, after applying TPT, and after jointly employing TPT and our proposed C-TPT—the values highlighted in **bold** signify the best ECE achieved after test-time prompt tuning. Std. is reported in Appendix [A.7](#)

Method		ImageNet	Caltech	Pets	Cars	Flower	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP-RN50 _{HardPrompt}	Acc.	58.1	85.8	83.8	55.7	61.0	74.0	15.6	58.6	40.0	23.7	58.4	55.9
	ECE	2.09	4.33	5.91	4.70	3.19	3.11	6.45	3.54	9.91	15.4	3.05	5.61
+TPT _{HardPrompt}	Acc.	60.7	87.0	84.5	58.0	62.5	74.9	17.0	61.1	41.5	28.3	59.5	57.7
	ECE	11.4	5.04	3.65	3.76	13.4	5.25	16.1	9.24	25.7	22.5	12.4	11.7
+TPT _{HardPrompt} +C-TPT	Acc.	60.2	86.9	84.1	56.5	65.2	74.7	17.0	61.0	42.2	27.8	59.7	57.8
	ECE	3.01	2.07	2.77	1.94	4.14	1.86	10.7	2.93	19.8	15.1	3.83	6.20
CLIP-RN50 _{Ensemble}	Acc.	59.7	87.1	82.9	55.6	60.5	75.6	16.4	60.2	41.0	29.3	59.8	57.1
	ECE	5.15	6.43	6.46	7.34	5.02	5.04	3.92	6.19	4.54	7.70	3.55	5.58
+TPT _{Ensemble}	Acc.	61.1	87.4	83.2	59.2	61.4	76.2	17.9	62.0	42.8	28.4	60.2	58.2
	ECE	11.2	4.29	4.79	3.08	14.1	5.27	14.6	7.68	22.2	18.9	11.1	10.7
+TPT _{Ensemble} +C-TPT	Acc.	61.2	87.4	84.0	57.3	65.3	76.0	17.5	62.1	43.1	29.4	60.7	58.5
	ECE	4.13	2.15	2.71	1.68	3.60	1.47	10.9	2.96	15.7	8.70	3.27	5.20
CLIP-ViT-B/16 _{HardPrompt}	Acc.	66.7	92.9	88.0	65.3	67.3	83.6	23.9	62.5	44.3	41.3	65.0	63.7
	ECE	2.12	5.50	4.37	4.25	3.00	2.39	5.11	2.53	8.50	7.40	3.59	4.43
+TPT _{HardPrompt}	Acc.	69.0	93.8	87.1	66.3	69.0	84.7	23.4	65.5	46.7	42.4	67.3	65.0
	ECE	10.6	4.51	5.77	5.16	13.5	3.98	16.8	11.3	21.2	21.5	13.0	11.6
+TPT _{HardPrompt} +C-TPT	Acc.	68.5	93.6	88.2	65.8	69.8	83.7	24.0	64.8	46.0	43.2	65.7	64.8
	ECE	3.15	4.24	1.90	1.59	5.04	3.43	4.36	5.04	11.9	13.2	2.54	5.13
CLIP-ViT-B/16 _{Ensemble}	Acc.	68.2	93.4	86.3	65.4	65.7	85.2	23.5	64.0	45.6	43.0	66.1	64.2
	ECE	3.70	6.16	4.88	7.09	6.01	3.78	4.56	4.01	13.8	6.01	4.05	5.82
+TPT _{Ensemble}	Acc.	69.6	94.1	86.1	67.1	67.6	85.1	24.4	66.5	47.2	44.0	68.5	65.5
	ECE	9.82	4.48	5.72	4.00	13.9	4.27	14.6	9.01	18.6	14.1	10.5	9.91
+TPT _{Ensemble} +C-TPT	Acc.	69.3	94.1	87.4	66.7	69.9	84.5	23.9	66.0	46.8	48.7	66.7	65.8
	ECE	4.48	3.14	1.54	1.84	5.77	2.38	6.40	3.09	13.7	5.49	3.04	4.62

- Experimental results
 - 11 datasets (e.g., OxfordPets, Flowers 102,...)
 - TPT increases ECE
 - C-TPT reduces ECE by 47-56% on average
 - Accuracy is mostly preserved (within ~1%)

Paper review (2)

C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion, 2024, ICLR

Table 2: **Natural Distribution Shifts**. We report the **Acc.** (\uparrow) and **ECE** (\downarrow) of the initialization, after applying TPT, and after jointly employing TPT and our proposed C-TPT—the values highlighted in **bold** signify the best ECE achieved after test-time prompt tuning. Std. is reported in Appendix [A.7](#)

Method		ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-Sketch	Average
CLIP-RN50 _{HardPrompt}	Acc.	21.7	51.4	56.0	33.3	40.6
	ECE	21.3	3.33	2.07	3.15	7.46
+TPT _{HardPrompt}	Acc.	25.2	54.6	58.9	35.1	43.5
	ECE	31.0	13.1	9.18	13.7	16.7
+TPT _{HardPrompt} +C-TPT	Acc.	23.4	54.7	58.0	35.1	42.8
	ECE	25.4	8.58	4.57	9.70	12.1
CLIP-RN50 _{Ensemble}	Acc.	22.7	52.5	57.9	34.7	42.0
	ECE	17.0	2.68	5.64	10.9	9.06
+TPT _{Ensemble}	Acc.	26.9	55.0	60.4	35.6	44.5
	ECE	29.1	12.7	7.50	14.0	15.8
+TPT _{Ensemble} +C-TPT	Acc.	25.6	54.8	59.7	35.7	44.0
	ECE	27.0	9.84	5.17	12.2	13.6
CLIP-ViT-B/16 _{HardPrompt}	Acc.	47.8	60.8	74.0	46.1	57.2
	ECE	8.61	3.01	3.58	4.95	5.04
+TPT _{HardPrompt}	Acc.	52.6	63.0	76.7	47.5	59.9
	ECE	16.4	11.1	4.36	16.1	12.0
+TPT _{HardPrompt} +C-TPT	Acc.	51.6	62.7	76.0	47.9	59.6
	ECE	8.16	6.23	1.54	7.35	5.82
CLIP-ViT-B/16 _{Ensemble}	Acc.	50.9	62.0	74.5	46.0	58.4
	ECE	8.85	3.01	2.85	9.70	6.10
+TPT _{Ensemble}	Acc.	54.2	63.9	78.2	48.5	61.2
	ECE	13.5	11.2	3.64	15.3	10.9
+TPT _{Ensemble} +C-TPT	Acc.	52.9	63.4	78.0	48.5	60.7
	ECE	10.9	8.38	1.40	12.6	8.32

- Experimental results
 - Datasets: ImageNet-A, V2, R, Sketch
 - C-TPT again reduces ECE across data shifts
 - Up to 52% ECE reduction
 - Accuracy maintained vs. TPT

Paper review (2)

C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion, 2024, ICLR

- Compared to post-hoc TPT+TS (based on ImageNet):
 - C-TPT consistently achieves better ECE
 - No labeled data needed (unlike temperature scaling)

7.1 COMPARISON WITH PREVIOUS CALIBRATION METHOD

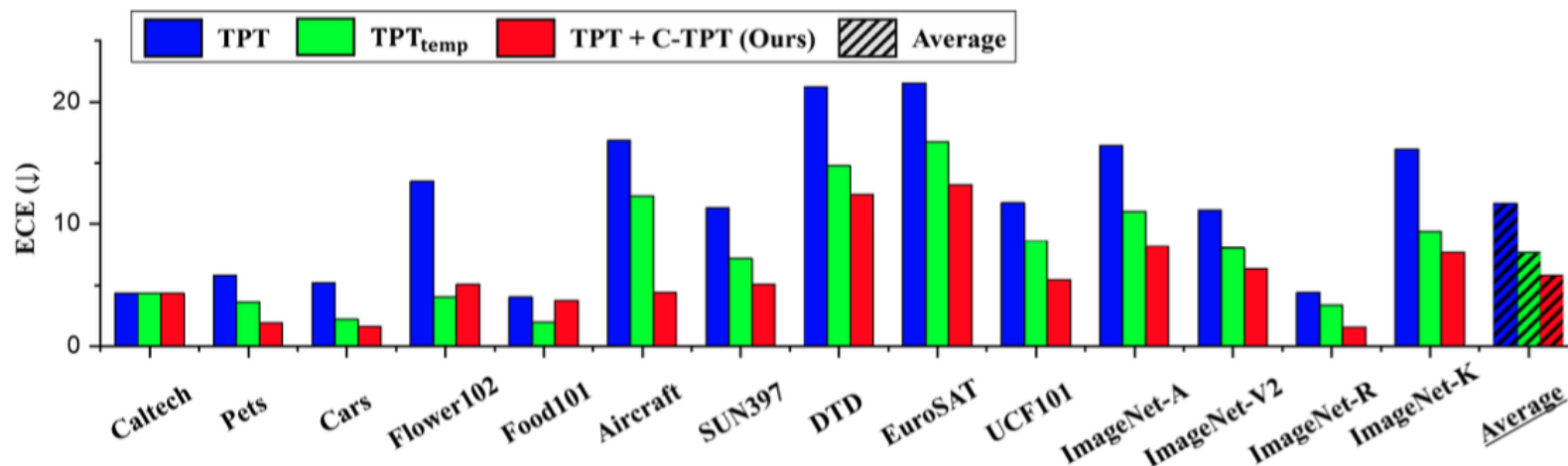


Figure 4: **Comparison of calibration error** between TPT, temperature-scaled TPT (TPT_{temp}), and the joint use of our proposed C-TPT (TPT+C-TPT). Results are based on CLIP-ViT-B/16.

Conclusion

- VLM's calibration matters: it enables safer and more trustworthy predictions
- Empirical study (ICML 2024)
 - Temperature scaling effectively improves calibration.
 - Works across distribution and label shifts with a few samples.
- C-TPT (ICLR 2024)
 - Improves test-time calibration without labeled data.
 - Uses ATFD (text feature dispersion) as a calibration guide.
 - Reduces ECE by up to 50%, accuracy remains stable.