Towards Efficient and Generalizable Multimodal Foundation Model Adaptation

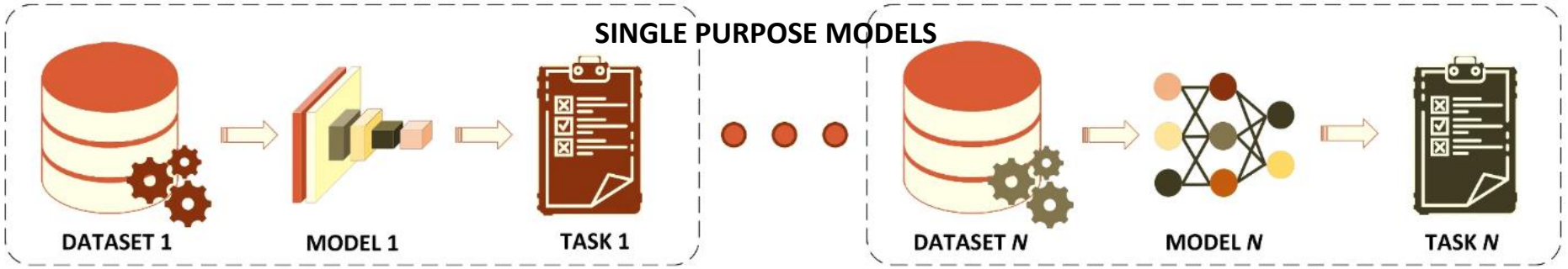
Niloufar Alipour Talemi, PhD

Department of Electrical and Computer Engineering
Clemson University

nalipou@clemson.edu

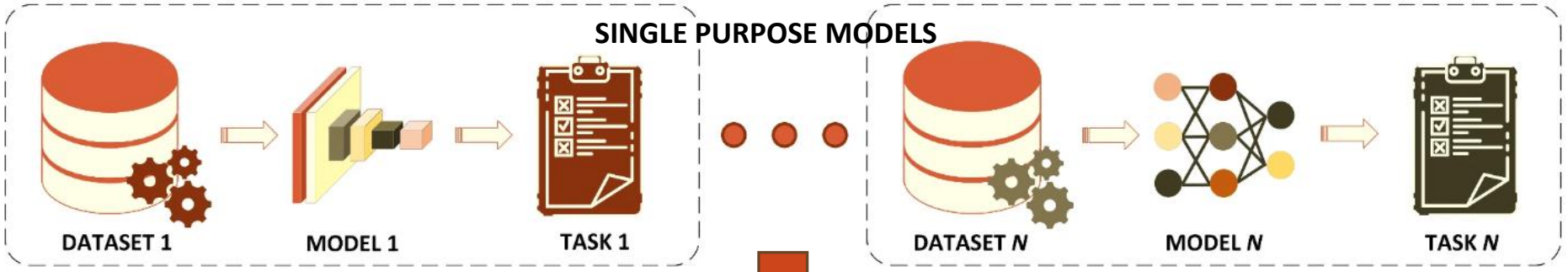
https://nilouap.github.io

Poor generalization ability, Limited training dataset, Unsatisfactory performance



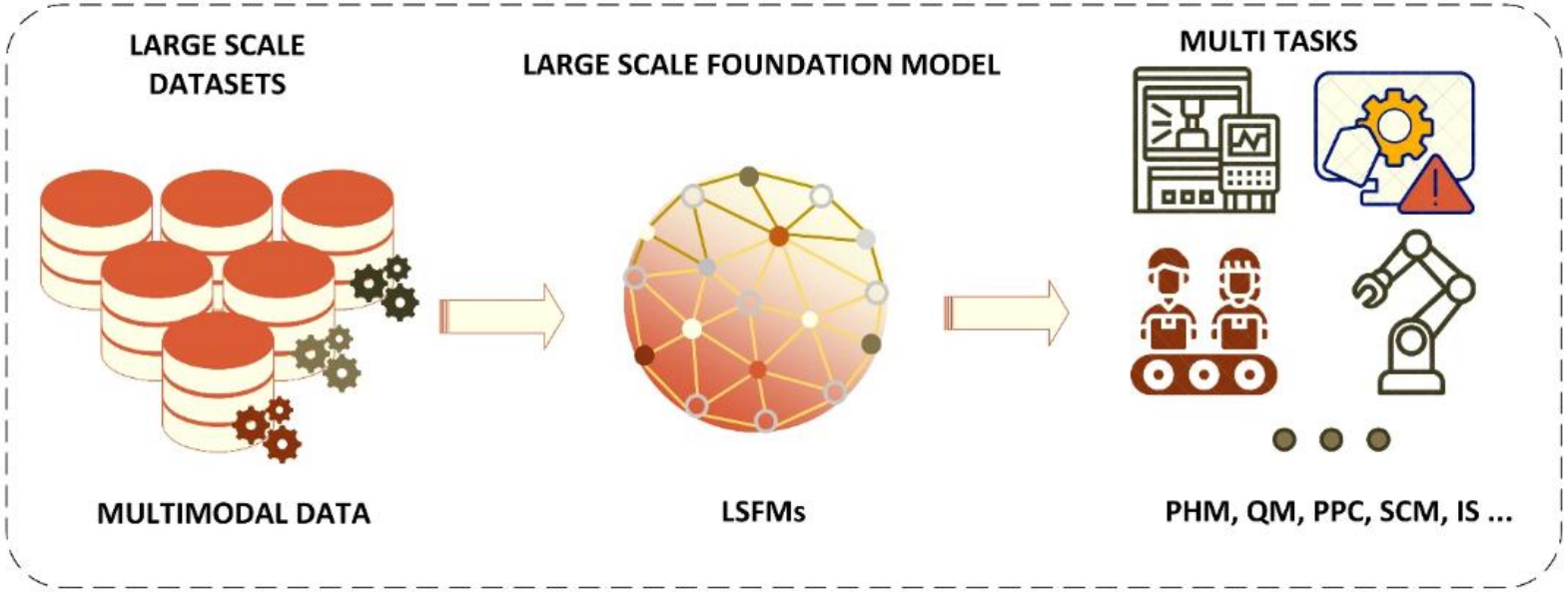
Big Models have arrived!

Poor generalization ability, Limited training dataset, Unsatisfactory performance

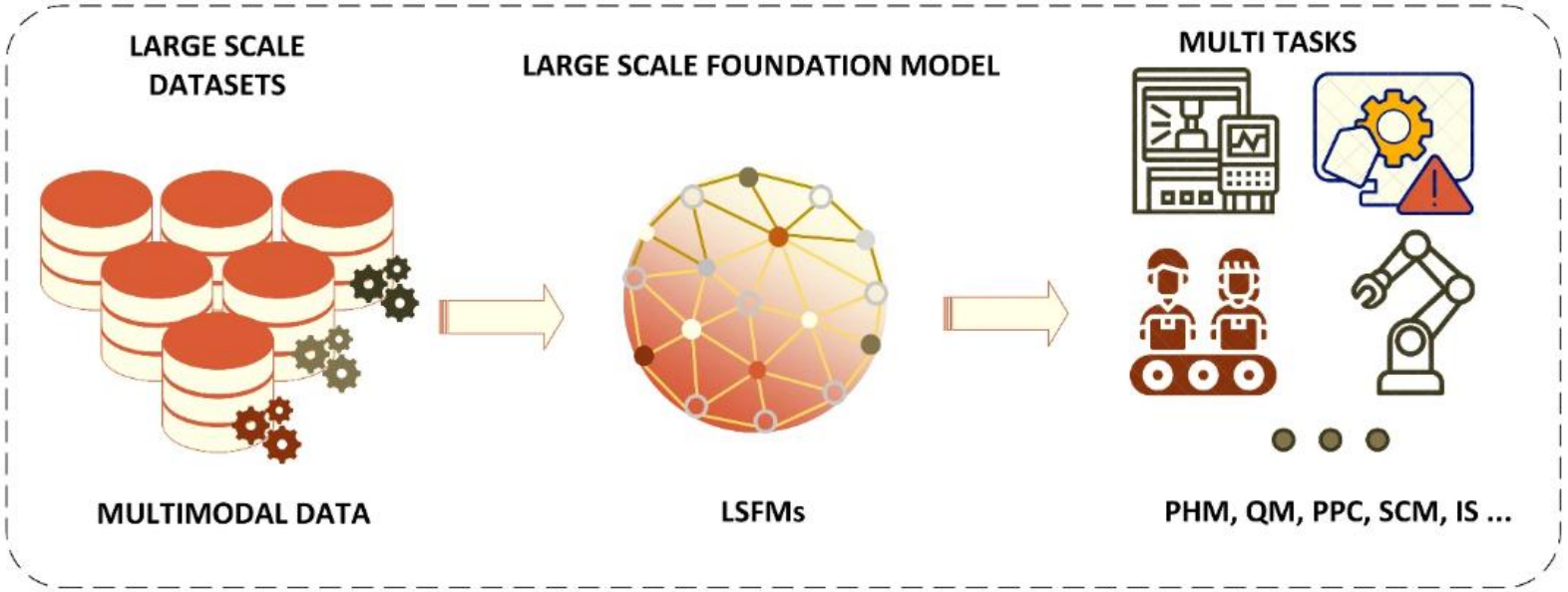


Generalization ability, Efficient data utilization

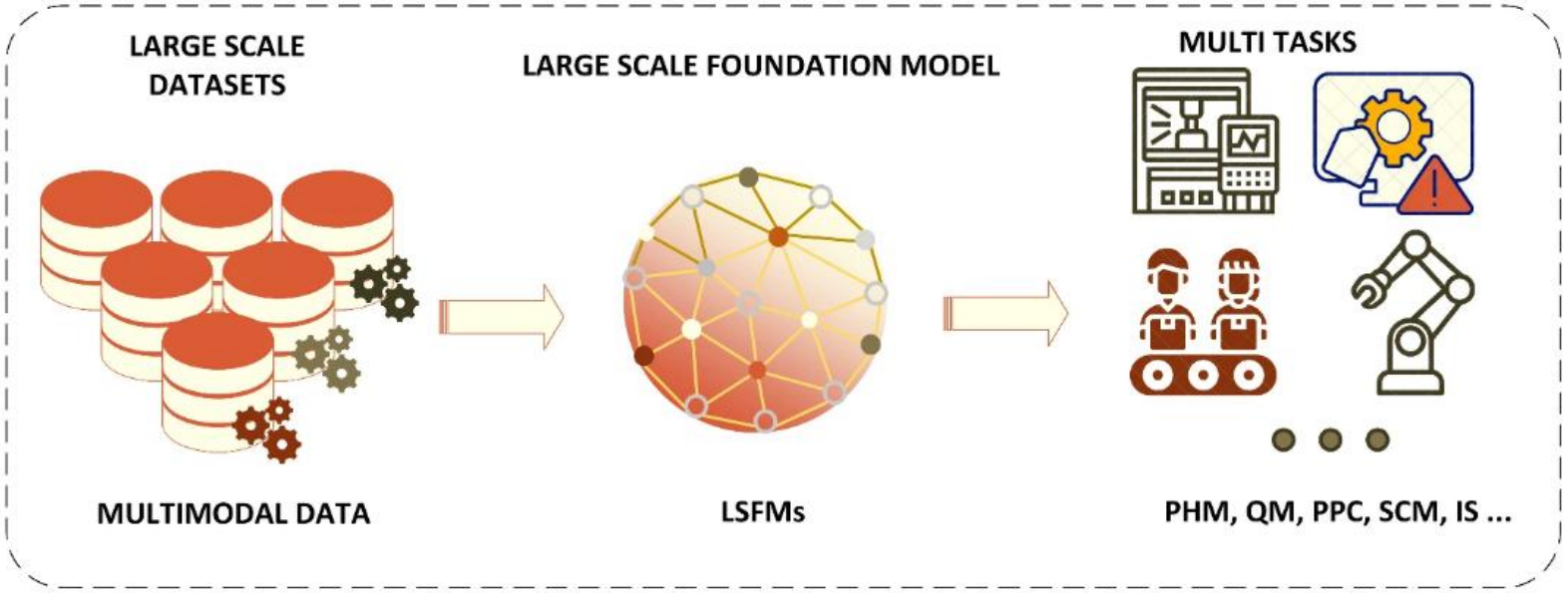
Generative capacity, Superior performance



Big Models have arrived!



Big Models have arrived!



- **Foundation models** are large AI models trained on massive datasets.
- They are **general-purpose**, meaning they can adapt to different tasks with little tweaking.
- A **multimodal foundation model** is a general-purpose model trained on multiple data modalities so it can understand and connect information **across** them.

Zhang, Haotian, et al. "Large scale foundation models for intelligent manufacturing applications: a survey," *Journal of Intelligent Manufacturing, 2025*.

What are Vision-Language Tasks?

Image Captioning

A man standing next to a red car in a parking lot.

Image Classification

whippet

Visual Question Answering

Which vegetable in this soup is heart shaped?

carrot

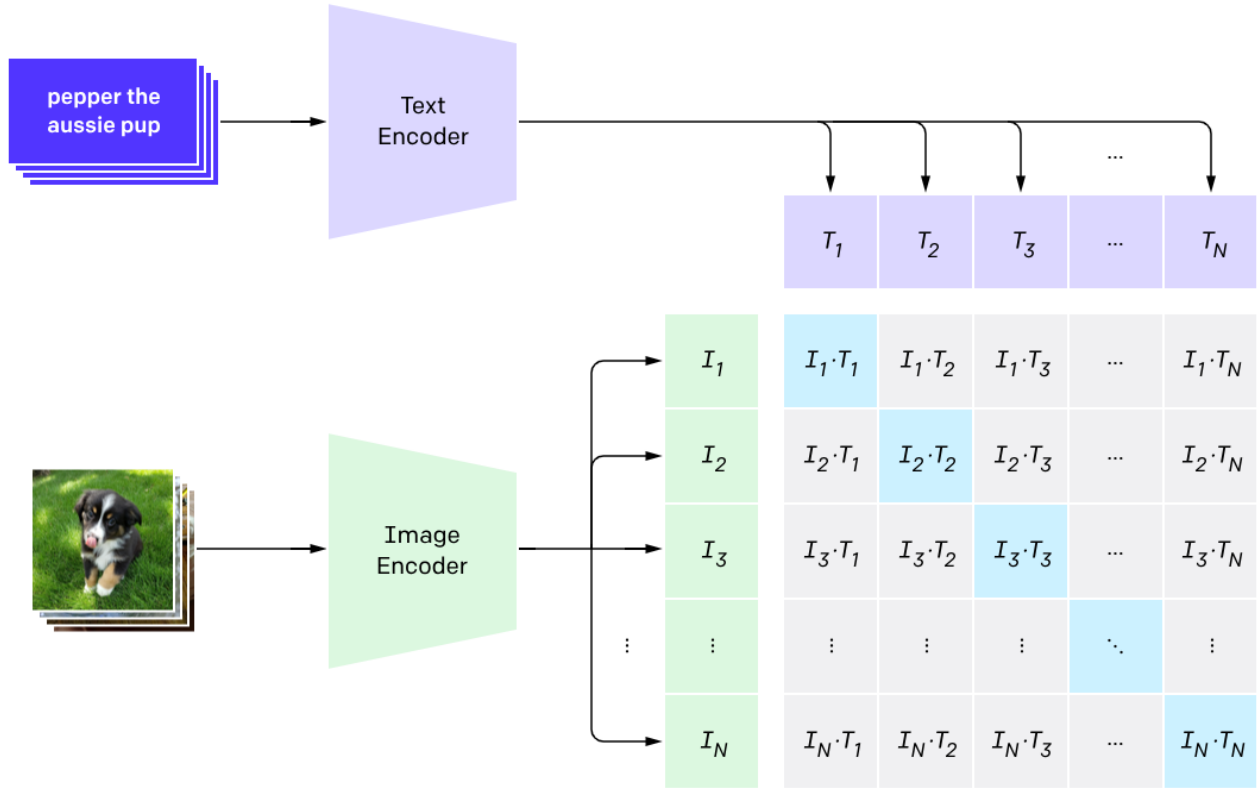
Object Localization

Which regions does the text “giraffe” describe ?

Lu, Jiasen, et al. "Unified-io: A unified model for vision, language, and multi-modal tasks." *arXiv preprint arXiv*, 2022.

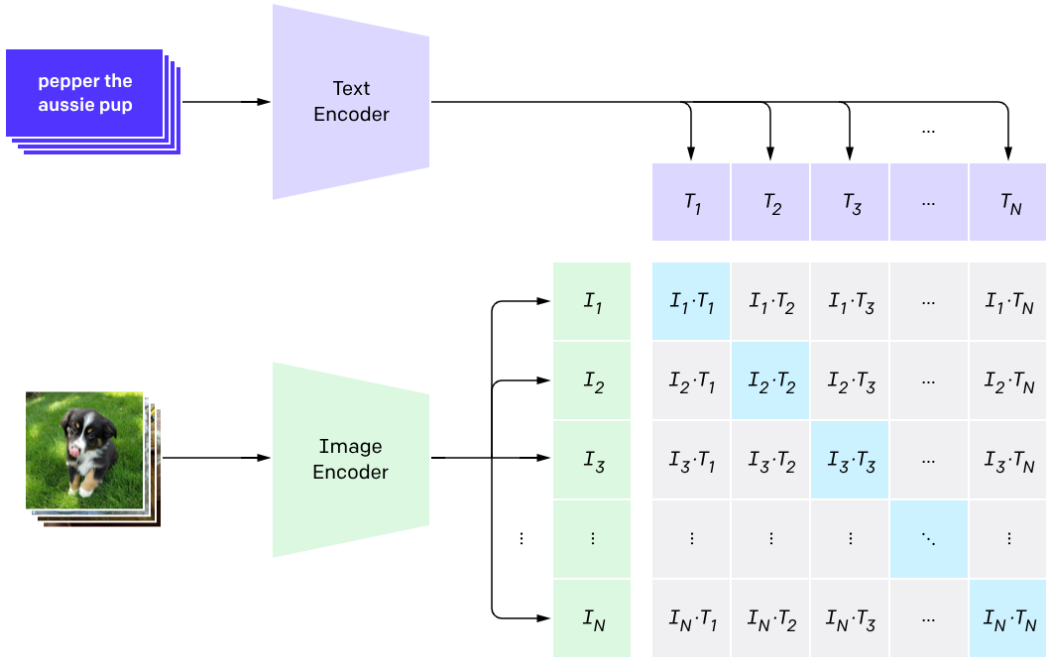
Contrastive Language-Image Pre-training (CLIP)

- CLIP leverages **a contrastive objective** that pairs images with corresponding natural language descriptions.
- CLIP is trained on a dataset of 400 million pairs collected form a variety of publicly available sources on the Internet.



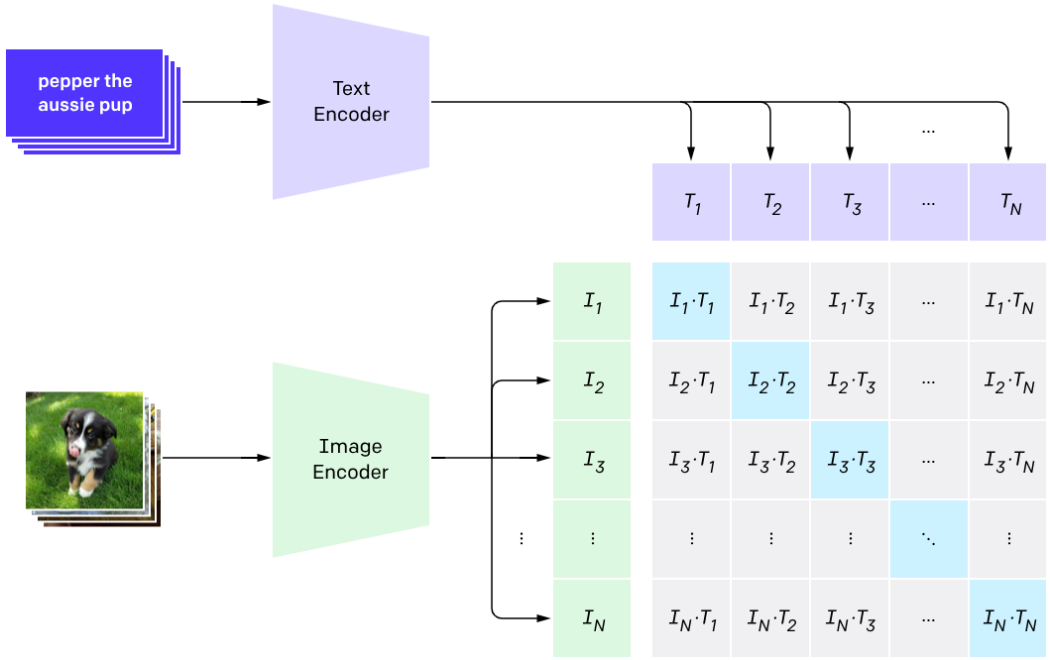
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*, 2021.

Contrastive Language-Image Pre-training (CLIP)

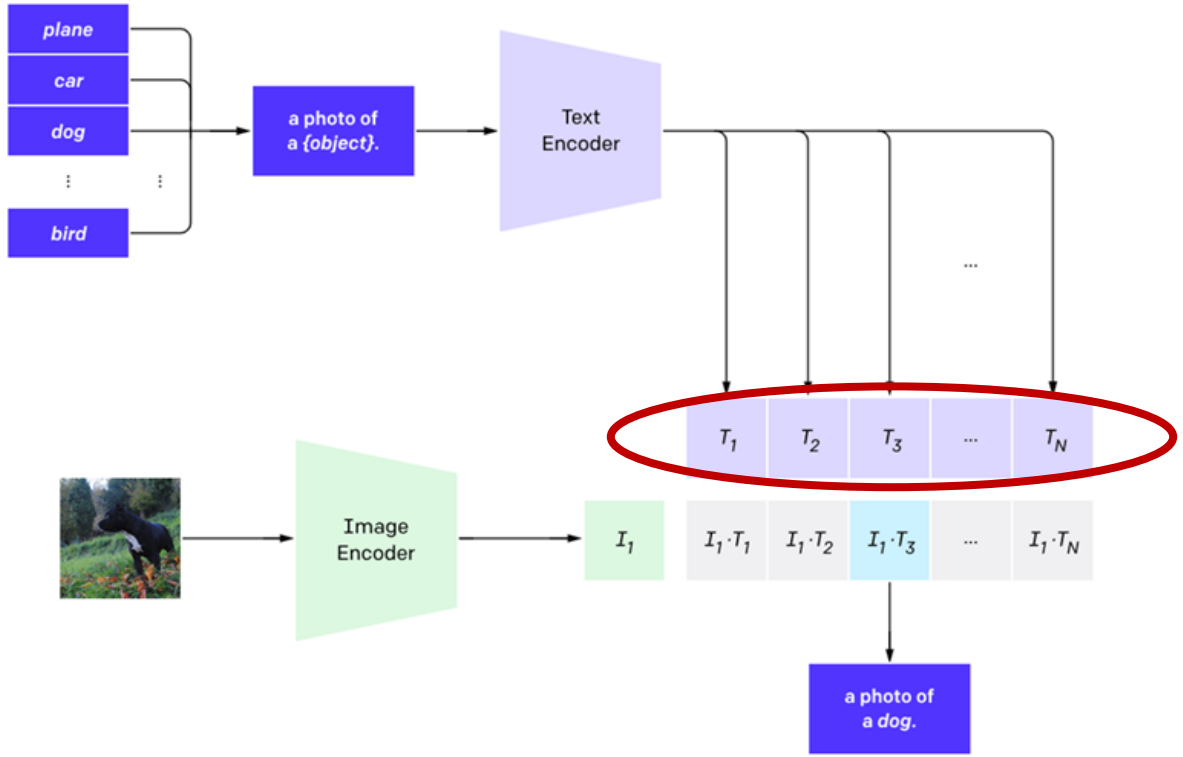


Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*, 2021.

Contrastive Language-Image Pre-training (CLIP)

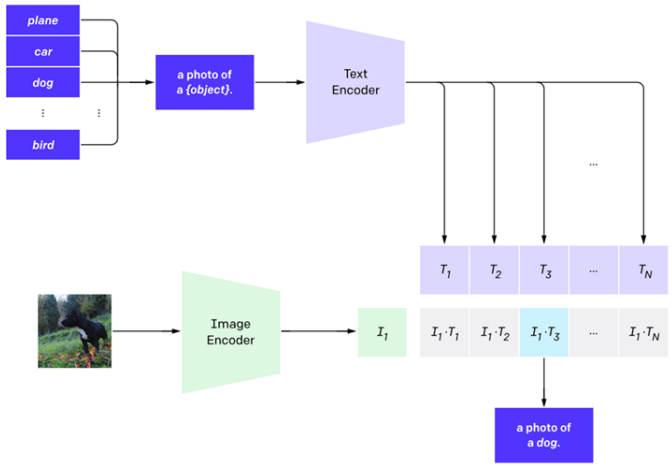


Create dataset classifier from label text



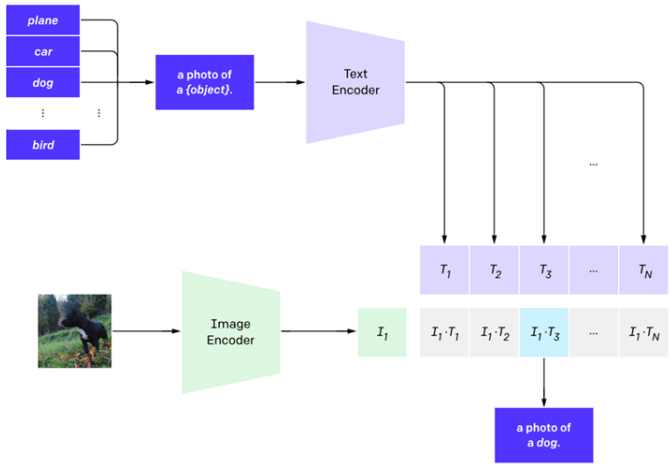
Challenges in CLIP Adaptation

- A major challenge in deploying CLIP is **prompt engineering**.
- Effective prompt engineering often requires **substantial domain expertise**.
- Prompt engineering is highly **time-consuming**.



Challenges in CLIP Adaptation

- A major challenge in deploying CLIP is **prompt engineering**.
- Effective prompt engineering often requires **substantial domain expertise**.
- Prompt engineering is highly **time-consuming**.



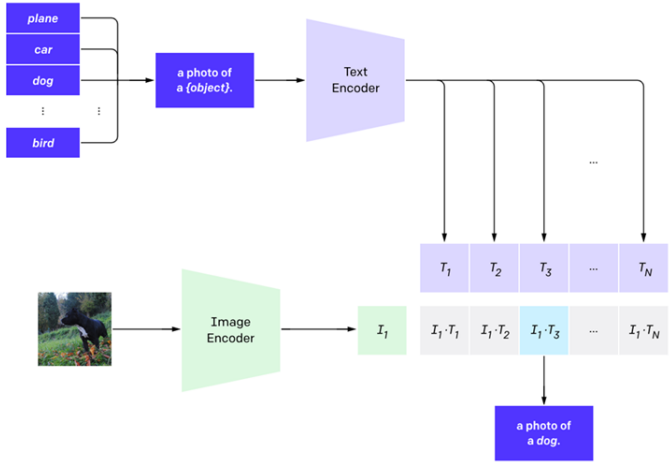
Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International journal of computer vision,* 2022.

Challenges in CLIP Adaptation

- A major challenge in deploying CLIP is **prompt engineering**.
- Effective prompt engineering often requires **substantial domain expertise**.
- Prompt engineering is highly **time-consuming**.



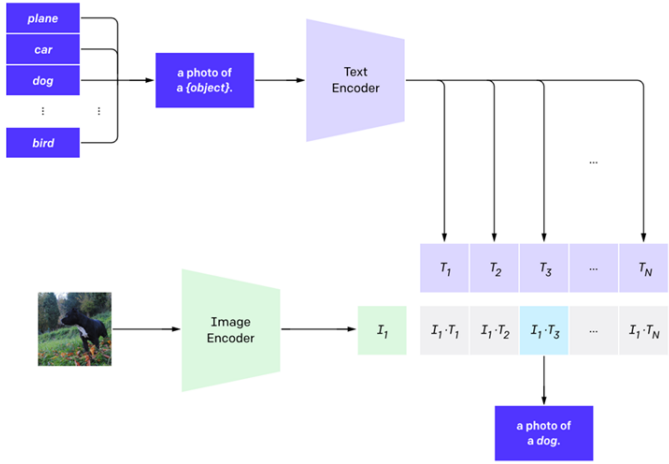
Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

+ 5.48

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International journal of computer vision,* 2022.

Challenges in CLIP Adaptation

- A major challenge in deploying CLIP is **prompt engineering**.
- Effective prompt engineering often requires **substantial domain expertise**.
- Prompt engineering is highly **time-consuming**.



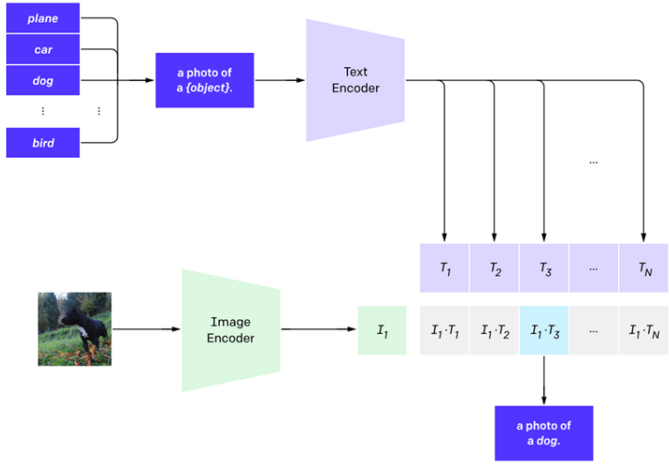
Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

+ 5.48

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International journal of computer vision,* 2022.

Challenges in CLIP Adaptation

- A major challenge in deploying CLIP is **prompt engineering**.
- Effective prompt engineering often requires **substantial domain expertise**.
- Prompt engineering is highly **time-consuming**.



Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1\ [V]_2\ \ldots\ [V]_M$ [CLASS].	**91.83**

+ 5.48

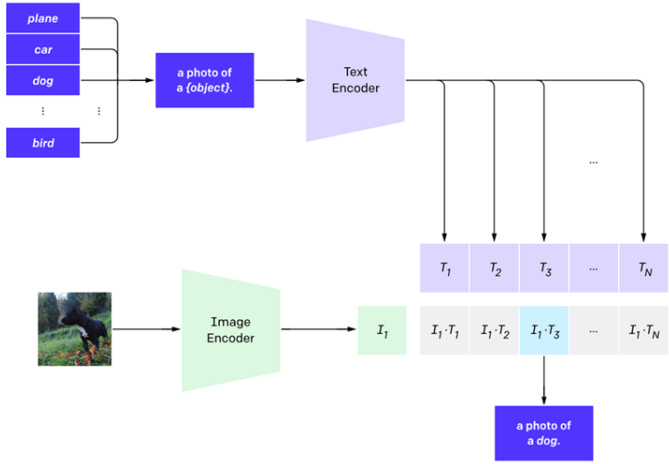
Describable Textures (DTD)

Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1\ [V]_2\ \ldots\ [V]_M$ [CLASS].	**63.58**

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International journal of computer vision,* 2022.

Challenges in CLIP Adaptation

- A major challenge in deploying CLIP is **prompt engineering**.
- Effective prompt engineering often requires **substantial domain expertise**.
- Prompt engineering is highly **time-consuming**.



Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

+ 5.48

Describable Textures (DTD)

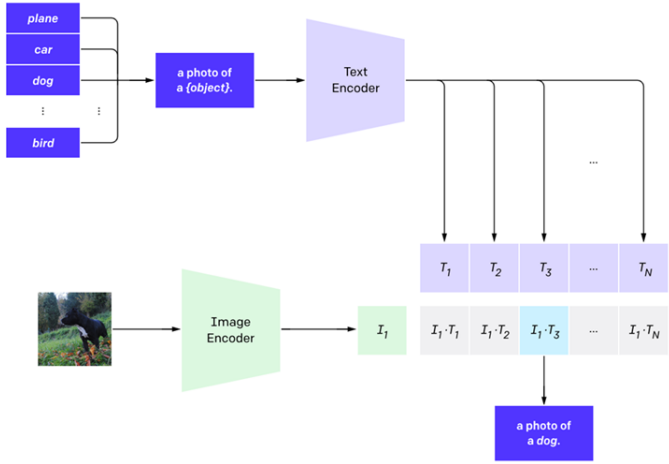
Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

+ 2.07

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International journal of computer vision,* 2022.

Challenges in CLIP Adaptation

- A major challenge in deploying CLIP is **prompt engineering**.
- Effective prompt engineering often requires **substantial domain expertise**.
- Prompt engineering is highly **time-consuming**.



Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1\ [V]_2\ \ldots\ [V]_M$ [CLASS].	**91.83**

Prompt Learning

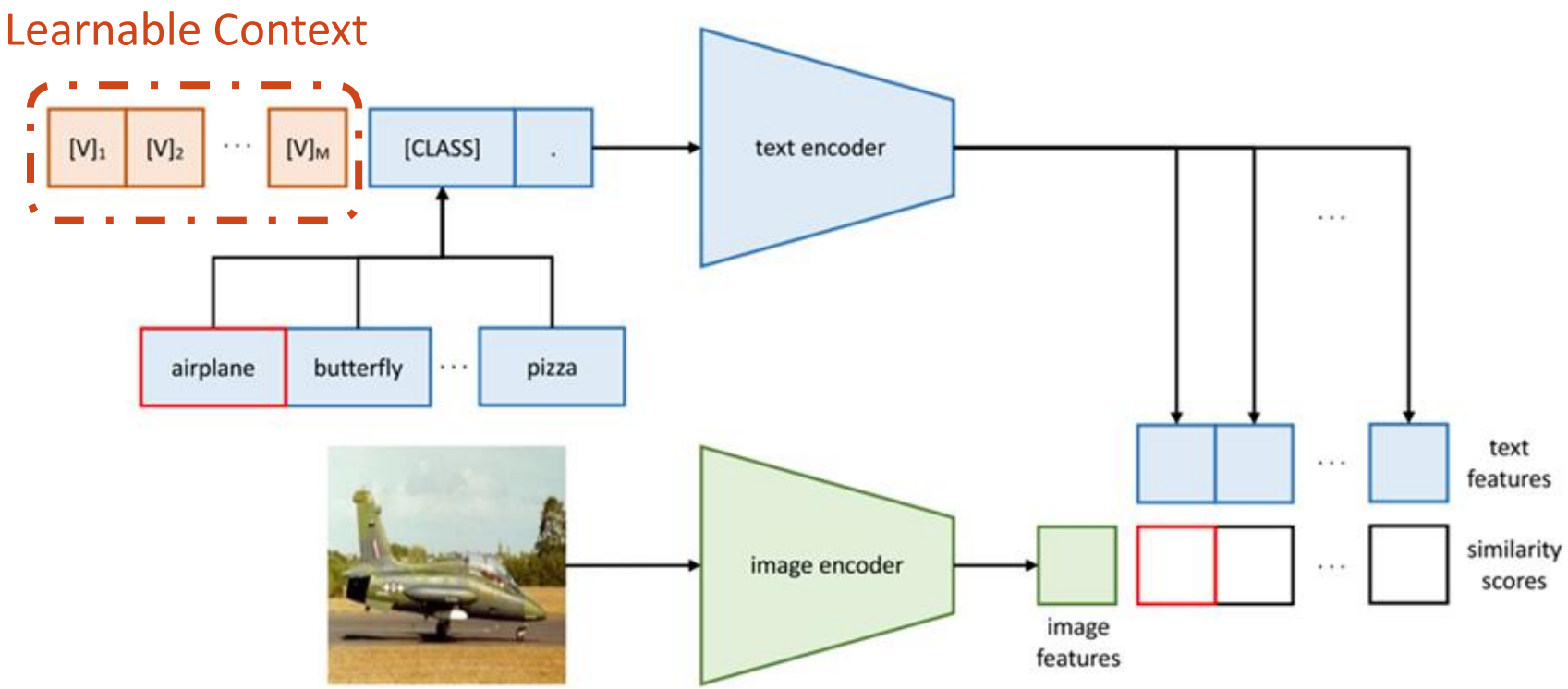
Describable Textures (DTD)

Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1\ [V]_2\ \ldots\ [V]_M$ [CLASS].	**63.58**

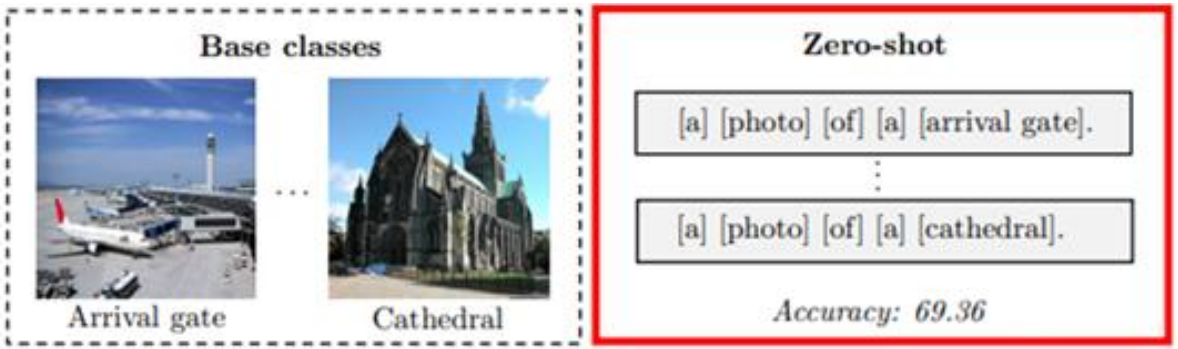
Prompt Learning

Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International journal of computer vision,* 2022.

Prompt Learning for Few-Shot Adaptation

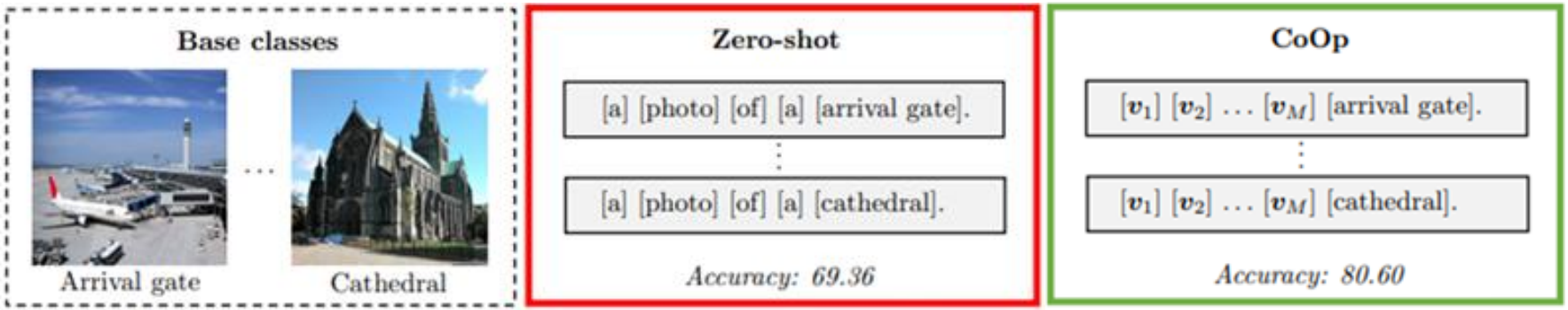


Overfitting in Prompt Learning



Overfitting in Prompt Learning

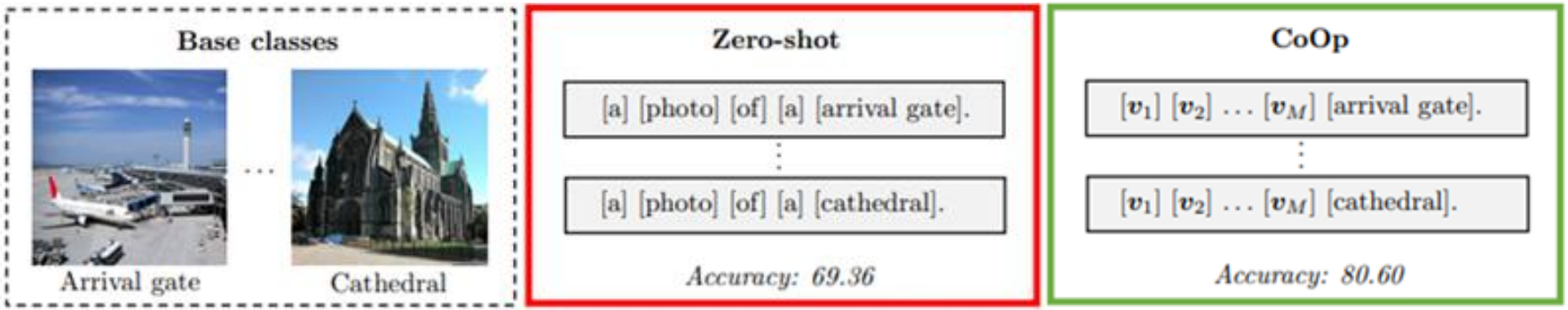
+ 11.24

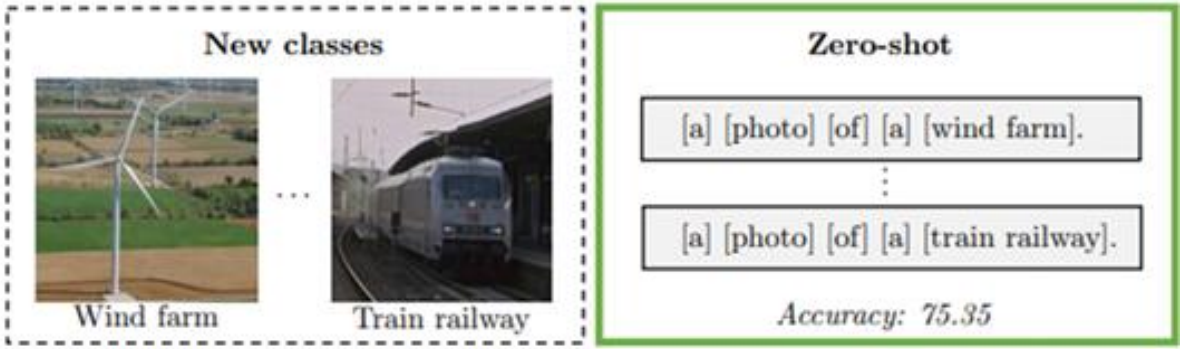


Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Overfitting in Prompt Learning

+ 11.24

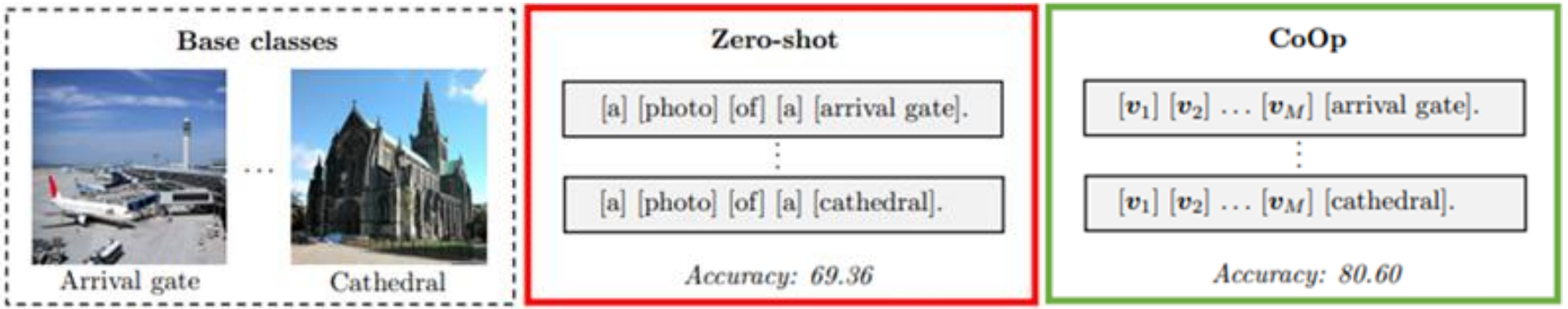




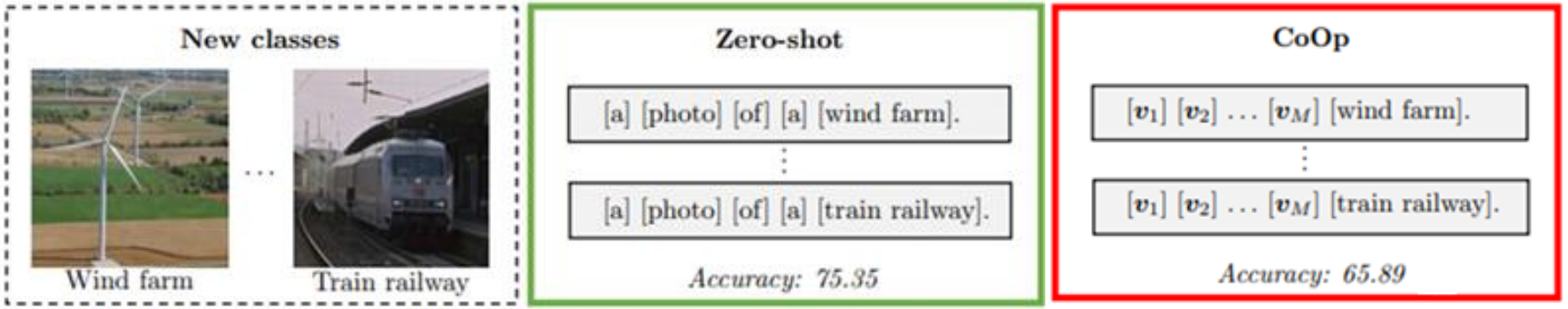
Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Overfitting in Prompt Learning

+ 11.24

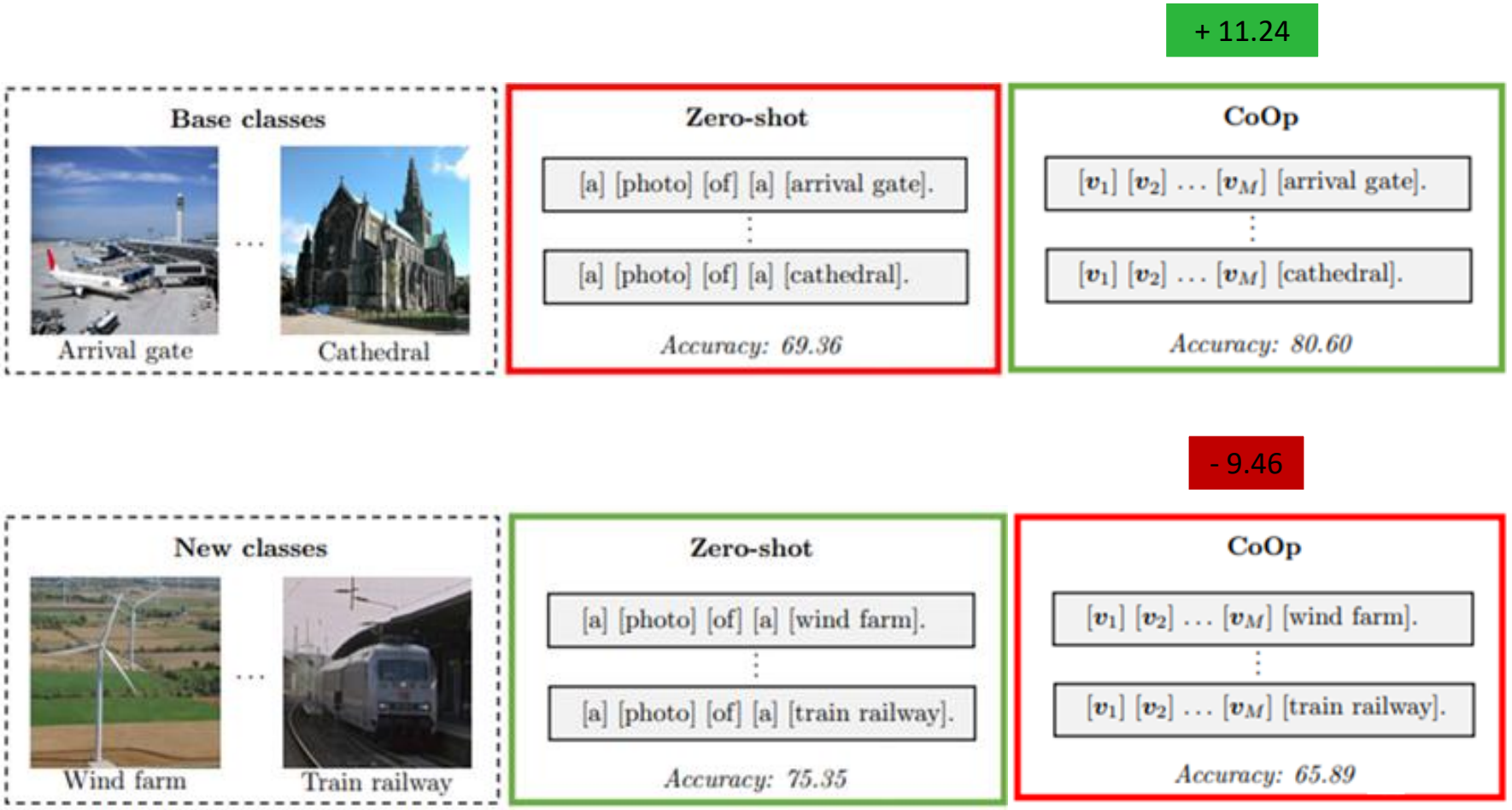


- 9.46



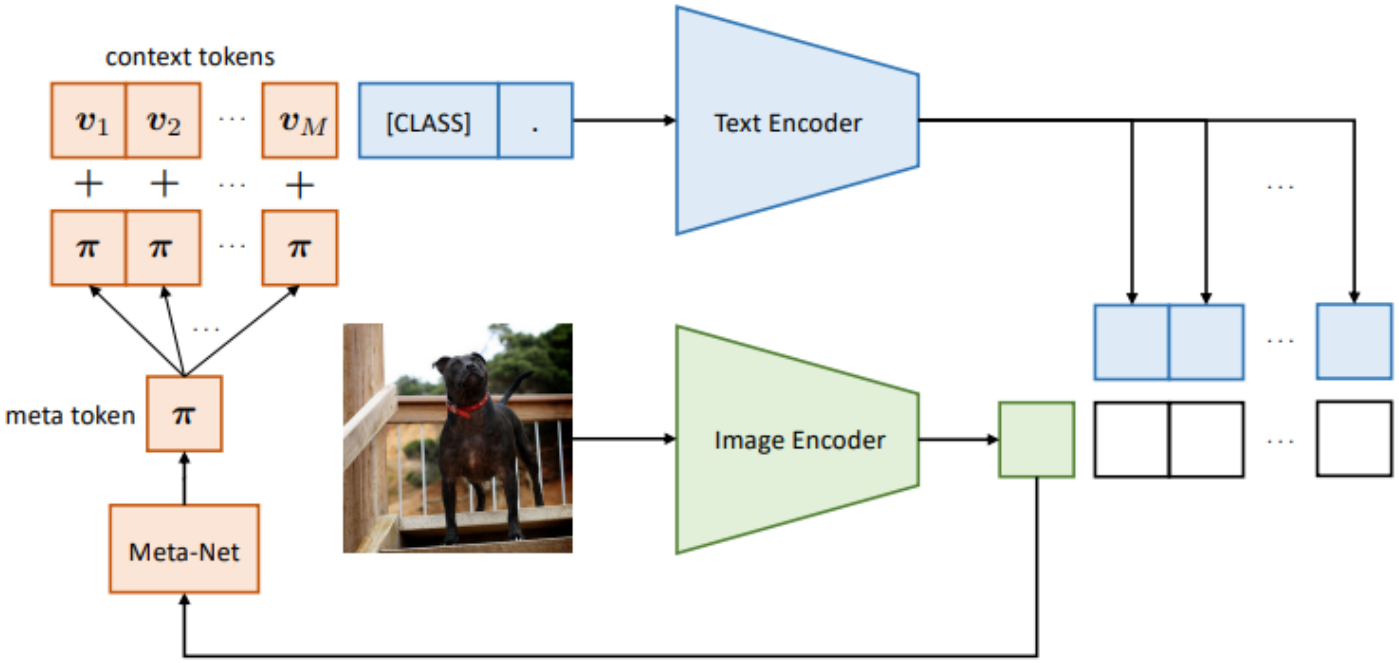
Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Overfitting in Prompt Learning



Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Conditional Prompt Learning

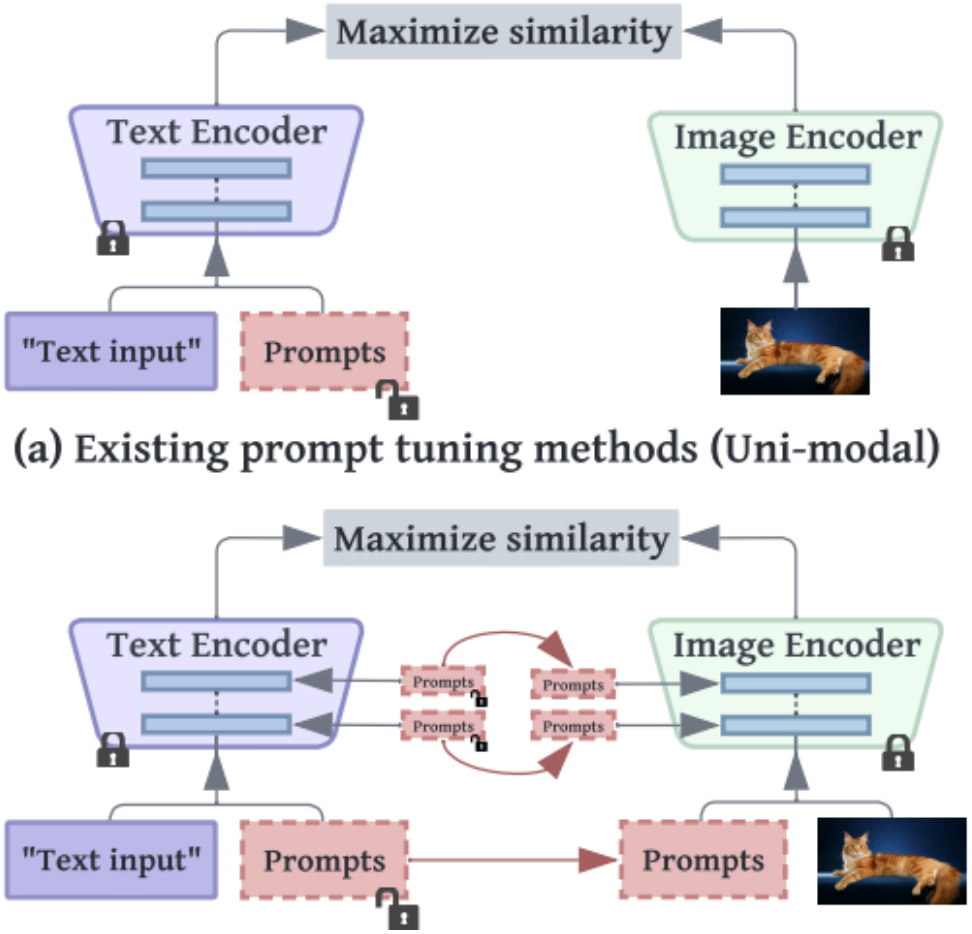


- It extends CoOp by further learning a lightweight neural network to generate for each image an input-conditional token.
- The Meta-Net is built with a two-layer bottleneck structure (Linear-ReLU-Linear).

Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Multi-modal Prompt Learning

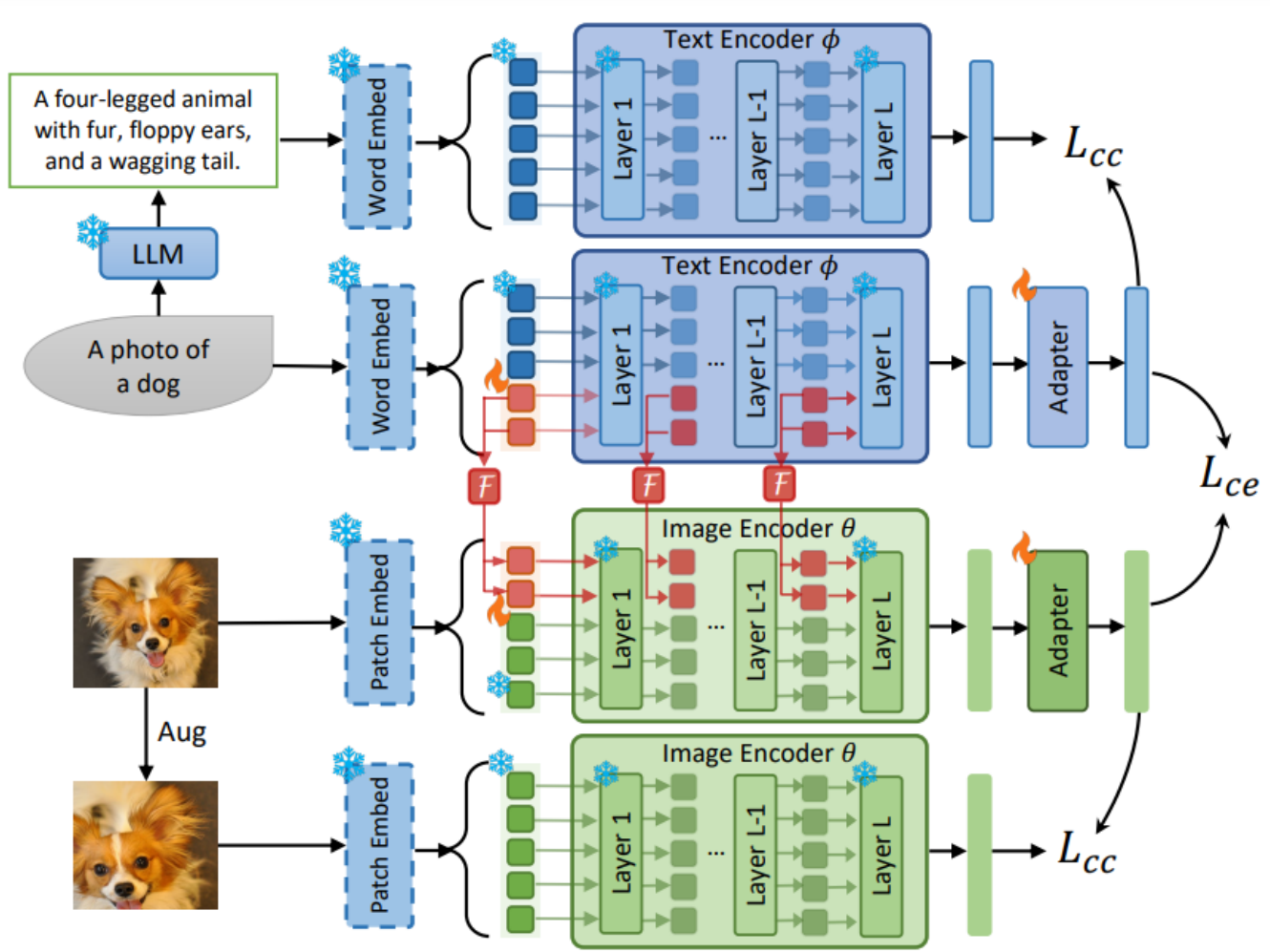
- CoOp and CoCoOp introduced learnable text prompts to adapt CLIP for downstream tasks.
- However, prompt learning only on the language side may provide limited adaptation to visual domain shifts.
- MaPLe extends prompt learning to both the vision and language branches.
- This multi-modal prompting enables better alignment between visual and textual representations.



Khattak, Muhammad Uzair, et al. "Maple: Multi-modal prompt learning." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

Consistency-Guided Prompt Learning

- MaPLe improves adaptation through vision-language prompting.
- CoPrompt further focuses on preserving CLIP's generalization during adaptation.
- It regularizes the model using consistency with frozen CLIP.



So far:

So far:

- ✓ Efficient generalizable adaptation for score-based VLMs.

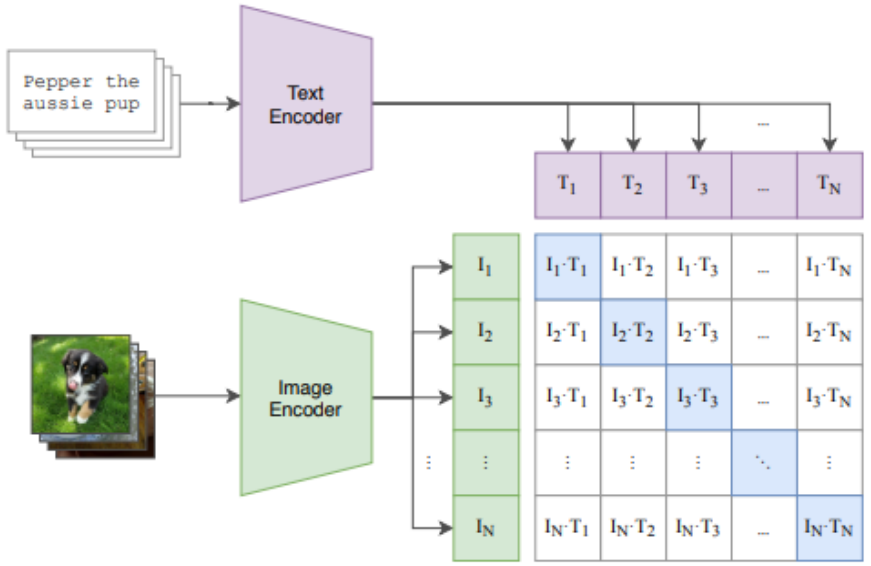
So far:

- ✓ Efficient generalizable adaptation for score-based VLMs.

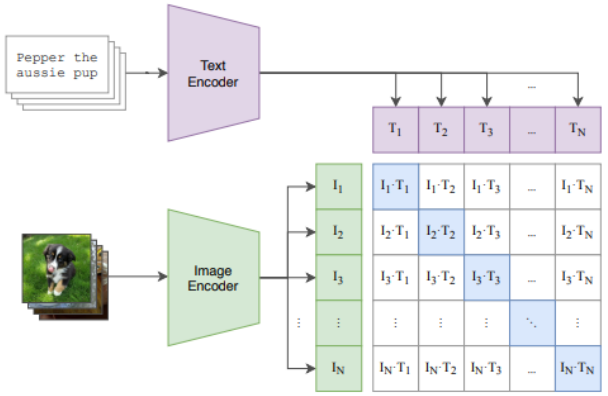
What about **generative** VLMs?

Let’s step back and take a closer look at some representative vision-language models!

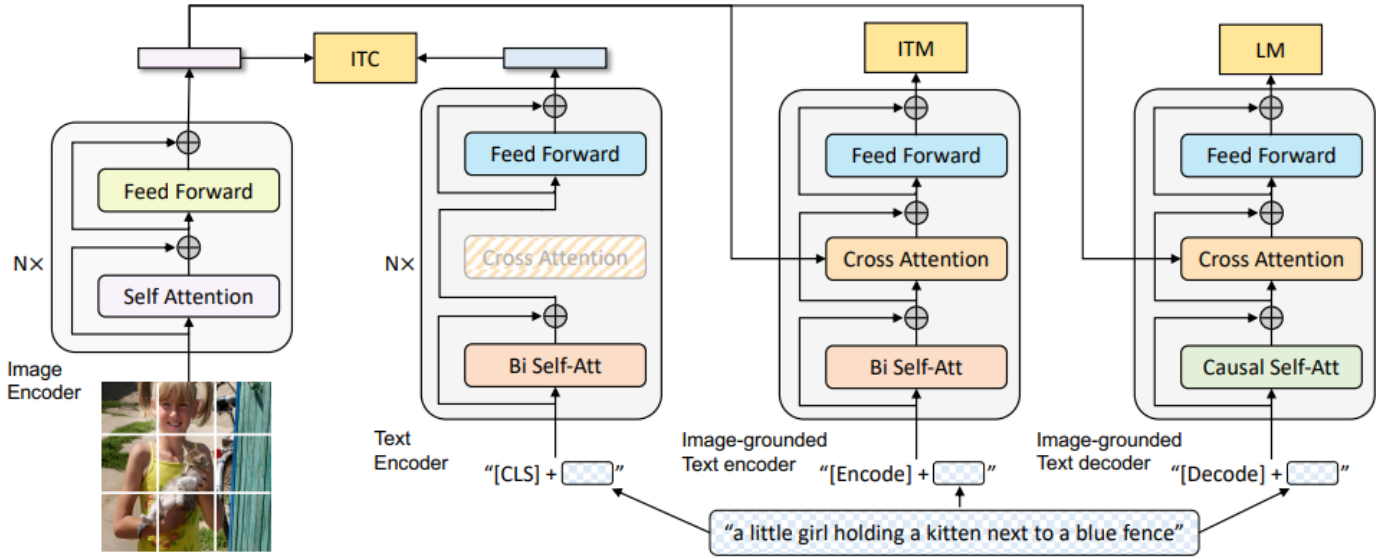
CLIP: Two-stream, Contrastive Learning



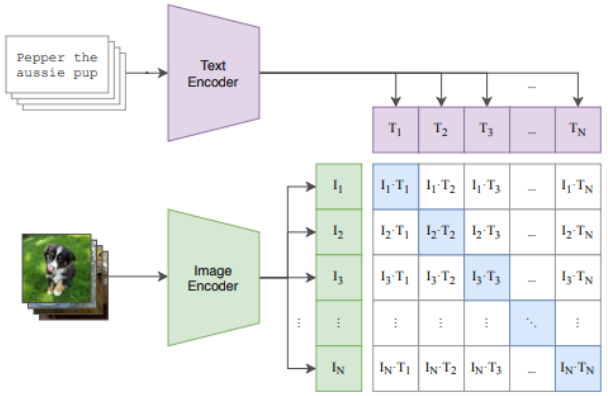
CLIP: Two-stream, Contrastive Learning



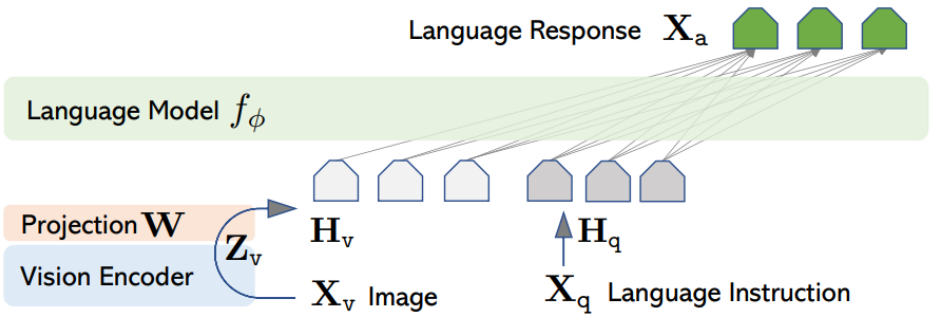
BLIP: Encoder–Decoder



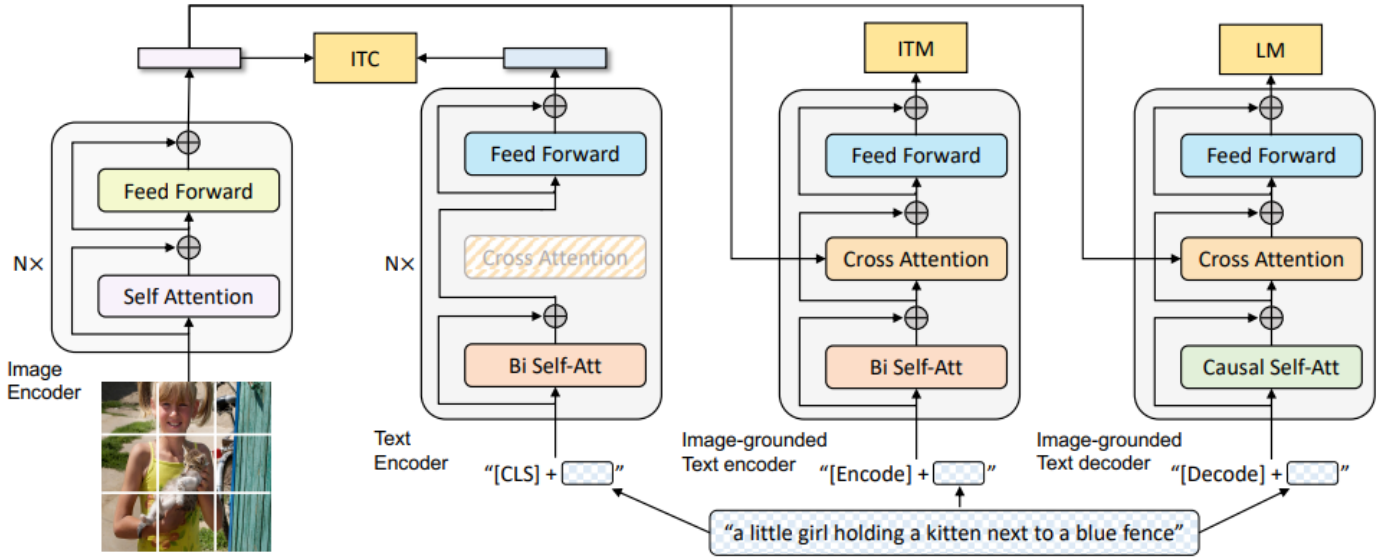
CLIP: Two-stream, Contrastive Learning



LLaVA: Modular VLMs



BLIP: Encoder–Decoder



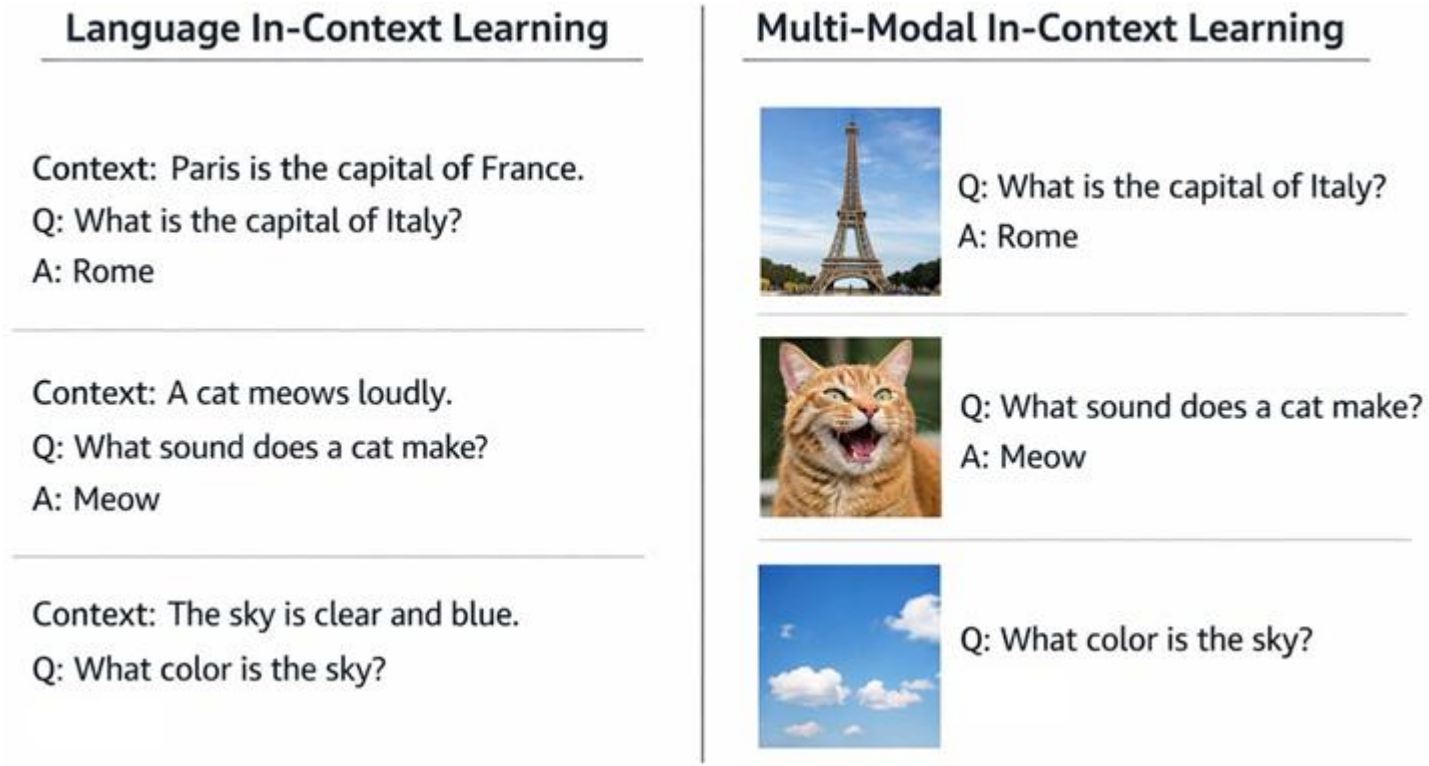
What is In-Context Learning (ICL)?

What is In-Context Learning (ICL)?

- It is a way for an LLM or MLLM to learn **from examples placed inside the input prompt**, instead of updating its weights.
- It uses **the pattern** in the examples to infer the task and answer a new input in the same way.

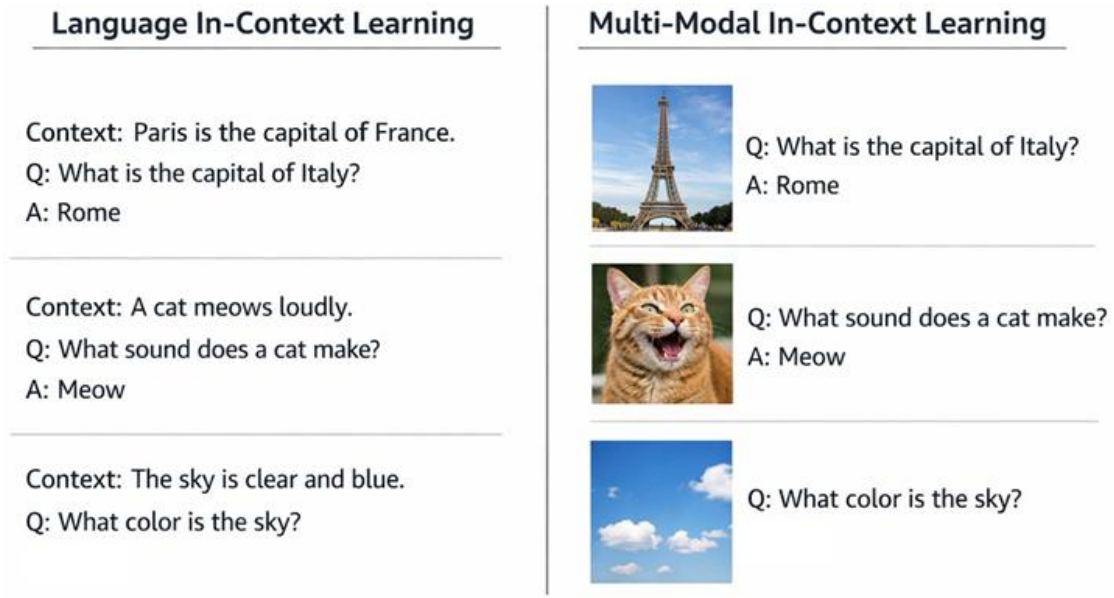
What is In-Context Learning (ICL)?

- It is a way for an LLM or MLLM to learn **from examples placed inside the input prompt**, instead of updating its weights.
- It uses **the pattern** in the examples to infer the task and answer a new input in the same way.



What is In-Context Learning (ICL)?

- It is a way for an LLM or MLLM to learn **from examples placed inside the input prompt**, instead of updating its weights.
- It uses **the pattern** in the examples to infer the task and answer a new input in the same way.

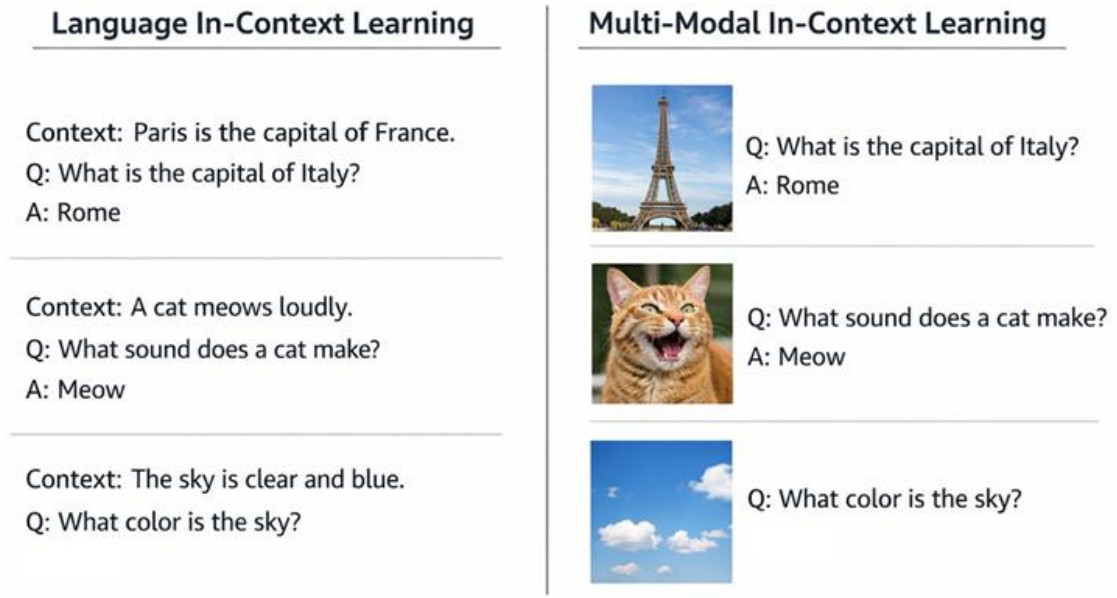


What is In-Context Learning (ICL)?

- It is a way for an LLM or MLLM to learn **from examples placed inside the input prompt**, instead of updating its weights.
- It uses **the pattern** in the examples to infer the task and answer a new input in the same way.

Challenges:

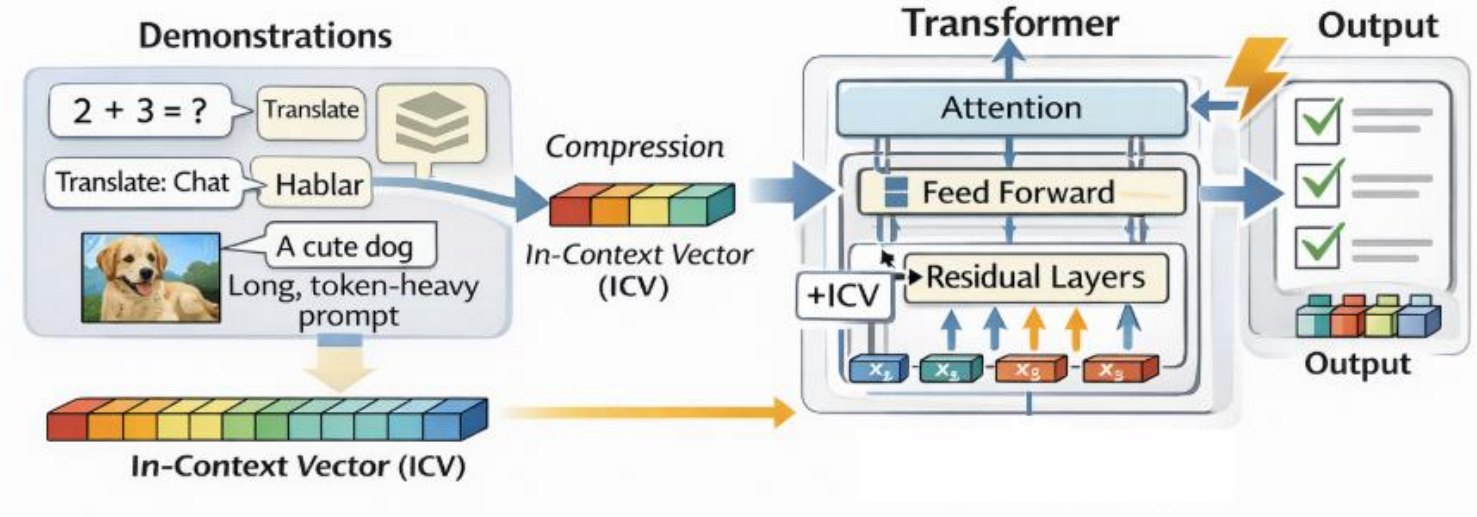
- Sensitive to **formatting** and example **order**.
- Multimodal ICL is **expensive at inference time**.
- **Cross-modal alignment** increases instability.
- Models may **mimic outputs** instead of learning mappings.



Takeaway: ICL enables adaptation without weight updates, but it is sensitive and computationally expensive.

Vector-Based ICL

- The core idea is to **compress** multiple demonstrations into a single In-Context Vector (ICV).
- The task representation is injected into the model as a **shift in hidden activations**.

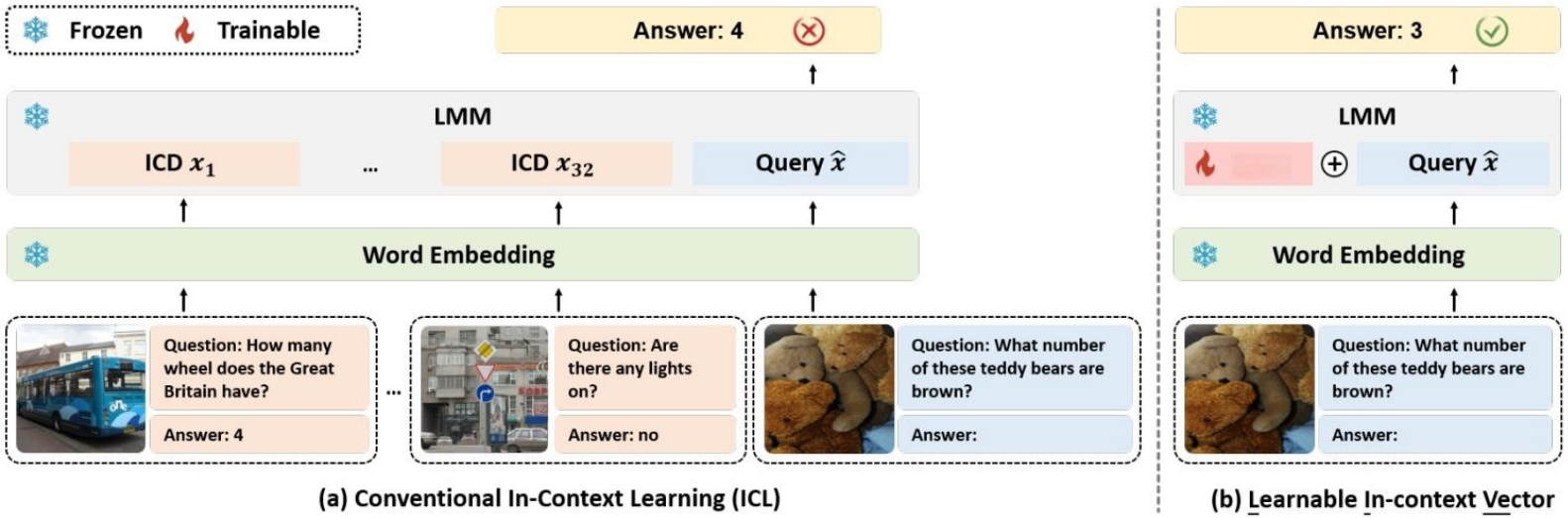


- ICV-based methods improve **inference efficiency** and **reduce sensitivity** to demonstration selection and ordering.

Challenges

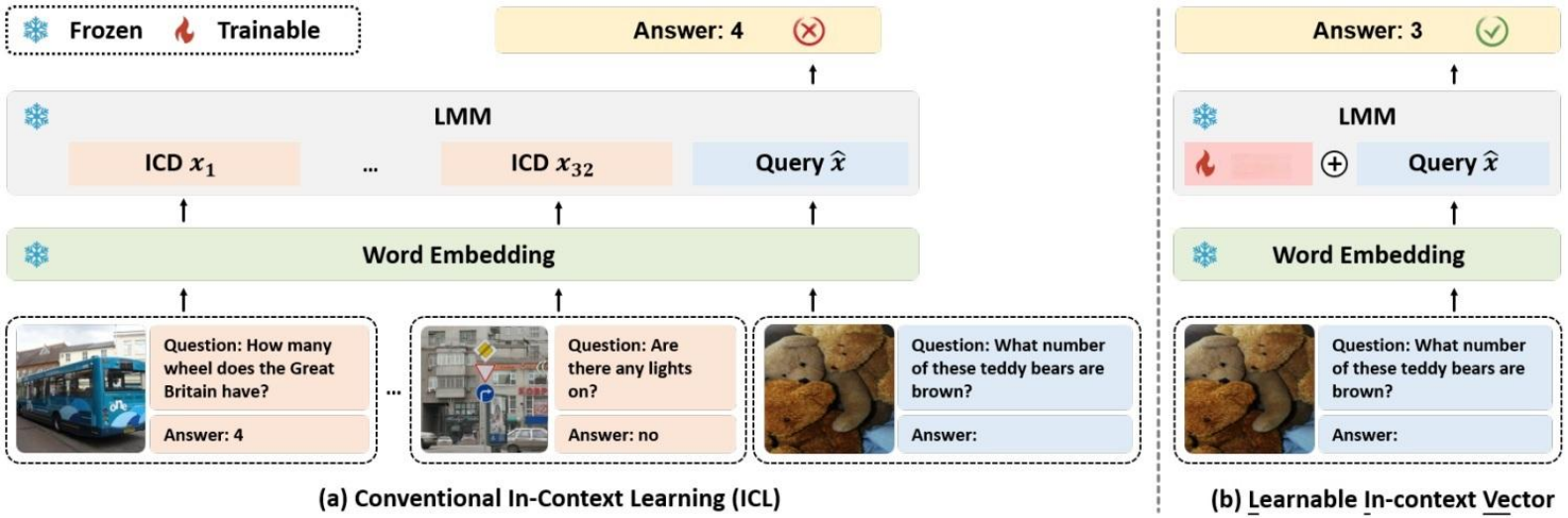
- Heuristic ICV extraction is often insufficient for complex multimodal tasks such as visual question answering.
- Compared with language-only settings, multimodal tasks require cross-modal reasoning and image-text alignment.

Learnable In-Context Vector for Visual Question Answering



- LIVE **learns** a richer shift vector from a large supporting set using training.
- It distills information from many randomly sampled demonstration sets into layer-wise ICVs and performs better than heuristic methods.

Learnable In-Context Vector for Visual Question Answering

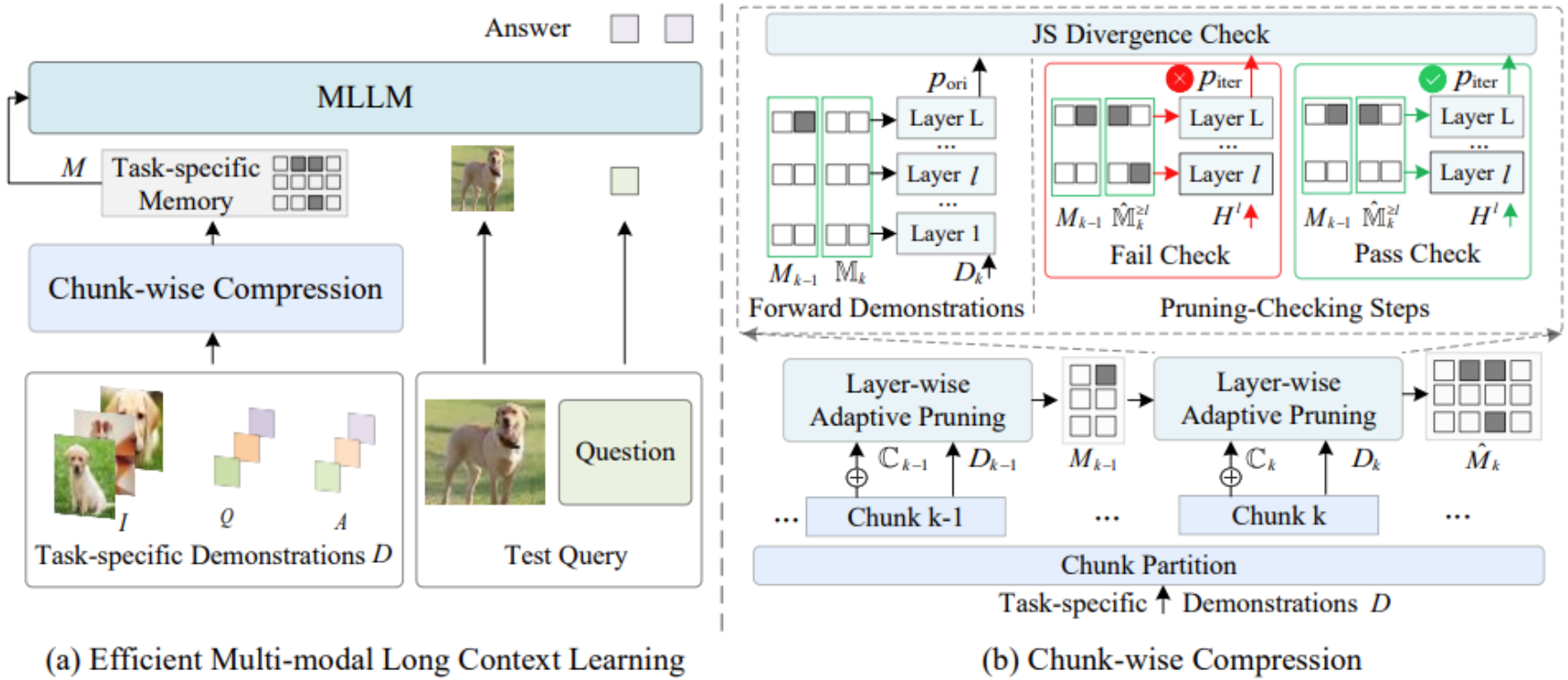


Challenges

- LIVE captures demonstration effects only at the **layer level**, while multimodal ICL exhibits substantially more fine-grained behavior.
- LIVE modifies hidden representations **after attention**, rather than controlling where the model attends.
- LIVE also uses the **same adaptation** for all inputs, which can hurt performance when different queries need different amounts of adjustment.

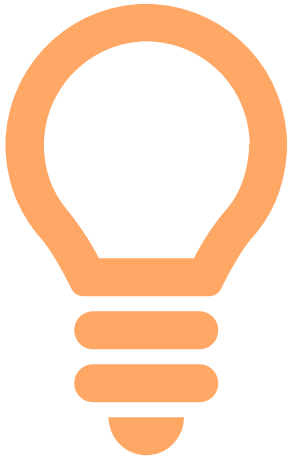
Peng, Yingzhe, et al. "Live: Learnable in-context vector for visual question answering." *Advances in Neural Information Processing Systems*, 2024.

Efficient Multi-modal Long Context Learning for Training-free Adaptation



- It adapts MLLMs using multimodal demonstrations, without updating model parameters.

- Compresses long image-text demonstrations into compact KV-cache memory through chunk-wise and layer-wise pruning.
- Retains tokens most relevant to the answers and controls information loss using JS divergence.



Future Directions

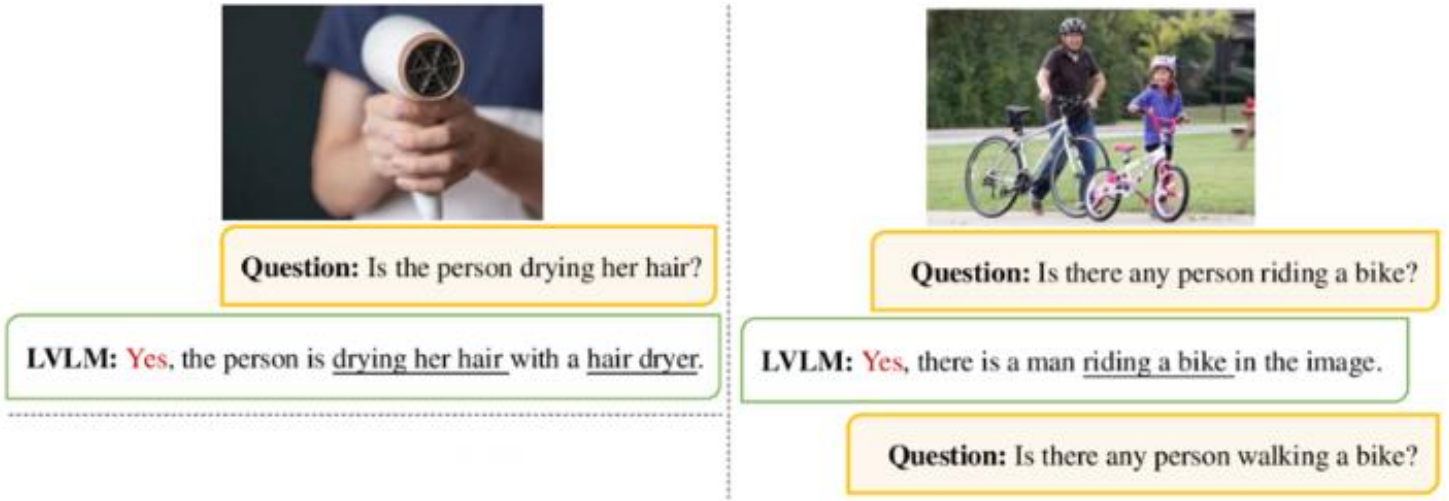
CHALLENGES AND OPEN PROBLEMS IN VLMS

Challenges and Open Problems in VLMs

Challenges specifically for CLIP-style VLMs:

- Overfitting in Adaptation: Prompt tuning can overfit to seen classes; generalization still weak.
- Lack of Local Grounding: CLIP aligns whole images with texts but lacks region-level grounding.
- Weak Attribute Binding: CLIP often ignores word order and fails to correctly link attributes with objects. For example, confusing a *yellow bus* with a *blue bus*.

More General Challenges:



- **Hallucination**: For example, in visual hallucination, model mentions objects/attributes not in the image or cites the wrong region.
- **Overreliance on language cues**: answers reflect text bias more than visual evidence.
- **Uncertainty & calibration:** Overconfident wrong answers; no principled notion of "I don't know."
- **Safety, Privacy, and Copyright:** Models can unintentionally expose or misuse sensitive content (like faces, IDs, or private documents).

More General Challenges:

- **Fixed visual resolution:** VLMs trained on fixed-size inputs struggle with large images or small objects, limiting fine-grained reasoning in tasks like VQA.

- **Beyond RGB:** VLMs struggles with depth, thermal, and medical images due to distribution gaps and limited labeled data.

- **Adversarial Vulnerability:** Easily fooled by modified images or hidden text (typographic attacks).
- **Context length & ICL cost**: Long prompts (images + text) are slow and expensive.

Thank you!

nalipou@clemson.edu

https://nilouap.github.io

https://github.com/NilouAP