Deep Learning

Lecture 5: Prompting

National University of Singapore
CEG5304/EE6934
Instructor: Joey Tianyi Zhou

Outline

I. Motivation for Prompting
II. Development of Prompting
III. Prompting & Prompt Engineering in NLP
IV. Prompting & Prompt Engineering in CV
V. Prompting & Prompt Engineering on VLM
VI. Summary

Prompting in recent years

- Prompt can be of different forms in various modalities/tasks.

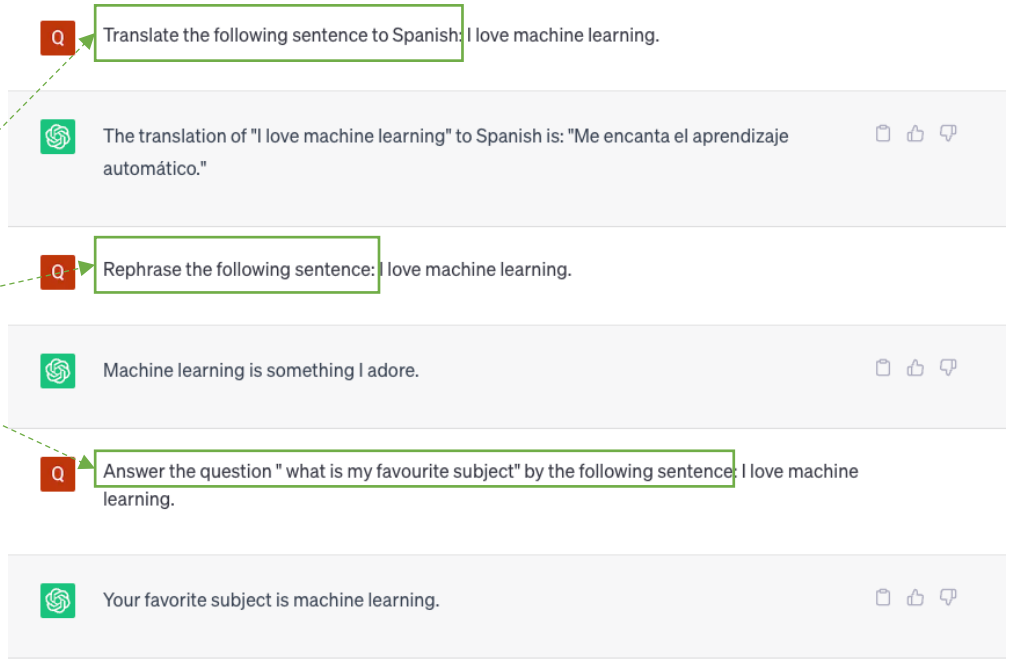
Modality (Task)	Prompt form
Language (text generation)	Text.
Vision (segmentation)	Point, box, mask.
Multimodal(VL) (image generation)	Text.
Multimodal(VL) (image editing)	Text, mask.
Multimodal(VL) (video editing)	Text.
…	…

Prompting in recent years

- **ChatGPT**:

  Prompt the language model to do various tasks.

Given the text prompt



https://chat.openai.com/

Prompting in recent years

- **Segment Anything Model:**

  Prompt the model to segment the desired part of the image.

Given the bounding box prompt.

https://segment-anything.com/

Prompting in recent years

- **DALL·E-2:**

  Prompt the model to generate images.

Examp

Given a text prompt:

Koala bear riding on a motorcycle.

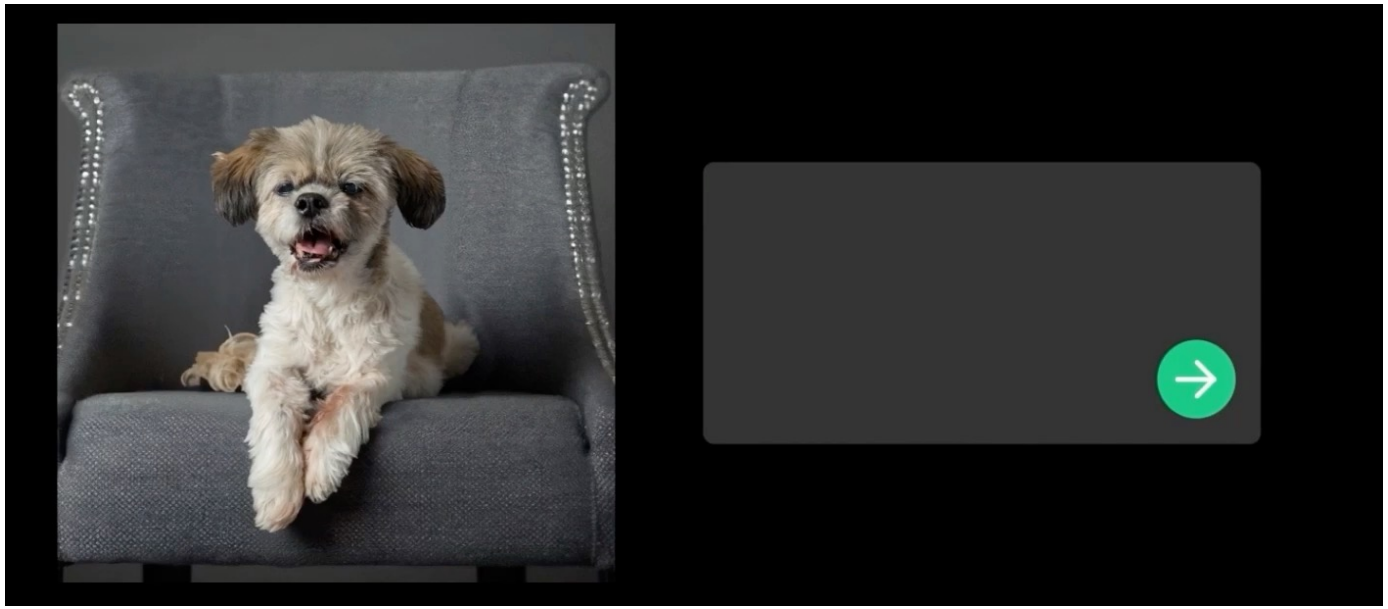
https://openai.com/dall-e-2

Prompting in recent years

- **DALL·E-2:**

  Prompt the model to edit images.

  Given prompts of the region to be edited and the description of desired editing.

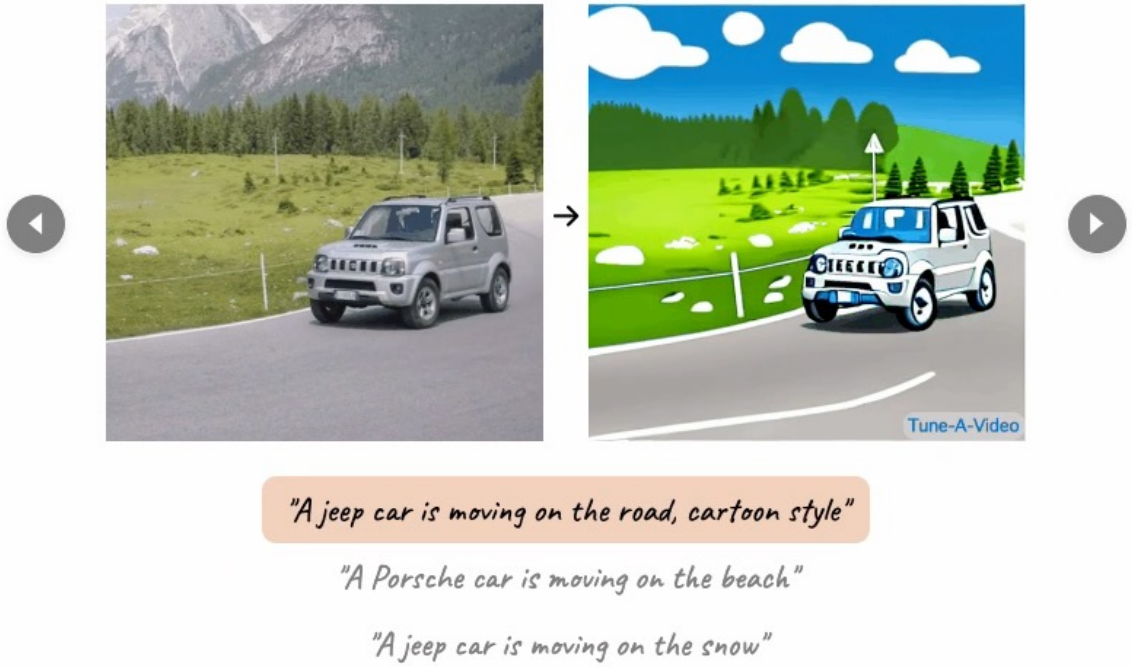


https://openai.com/dall-e-2

Prompting in recent years

- **Tune-A-Video**

  Prompt the model to transfer the video style to a desired one.



https://tuneavideo.github.io/

Given that prompting is getting popular, we may want to ask…

Why prompting?

Why prompting?

[Current Trend – Large model]

- Nowadays the **model scales are booming**.

BERT:
110M – 340M params

Google T5:
220M – 11B params

GPT: 110M params
GPT-2: 1.5B params
GPT-3: 175B params

…

Why prompting?

[Current Trend – Big Data]

- The large foundation models are pre-trained on web-scale data.
- They can be regarded as a **knowledge base** and are **potential to perform many tasks.**

GPT-3 training datasets.

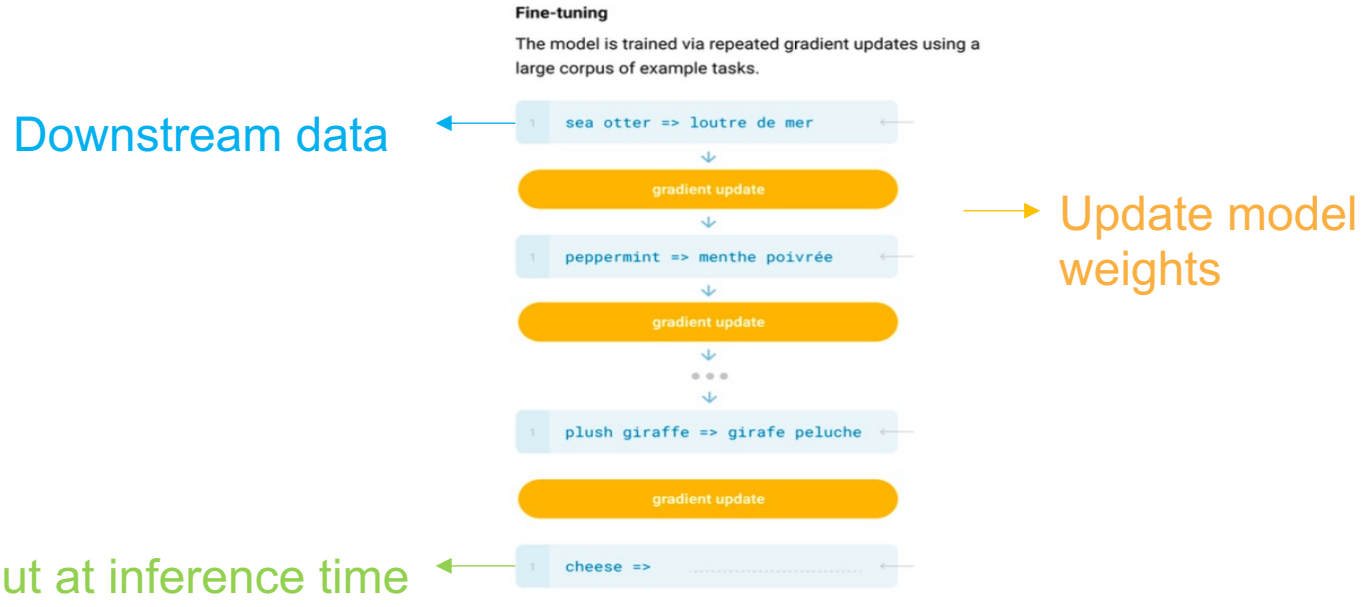
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Brown, Tom B. et al. "Language Models are Few-Shot Learners." ArXiv abs/2005.14165 (2020).

Why prompting?

[Recap] Traditional pre-training and fine-tuning paradigm

- Given a pre-trained model, update the pre-trained model parameters to adapt the model for downstream tasks.

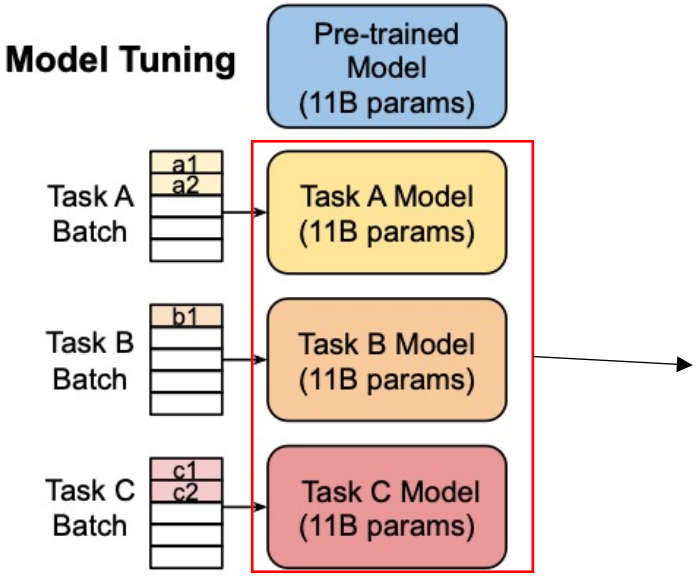


Brown, Tom B. et al. "Language Models are Few-Shot Learners." ArXiv abs/2005.14165 (2020).

Why prompting?

[Challenge 1]

- Fine-tuning can be **costly** and result in **data hunger** issue in previous training paradigms.



High saving and computational cost.

High labeled data demand.

Brown, Tom B. et al. "Language Models are Few-Shot Learners." ArXiv abs/2005.14165 (2020).

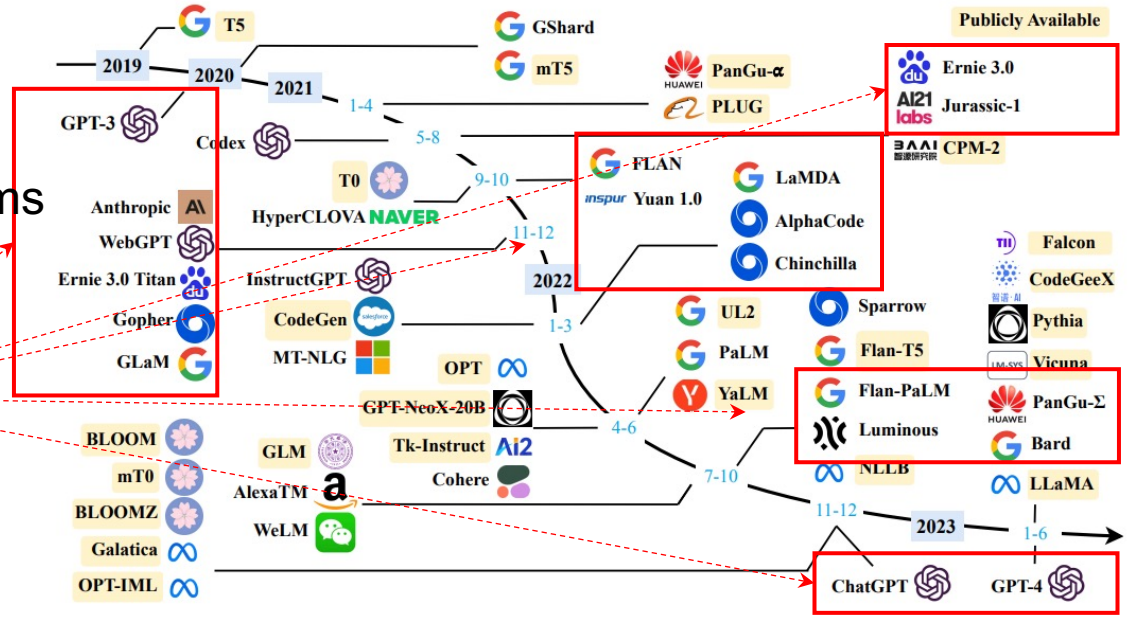
Why prompting?

[Challenge 2]

- Large models tend to be served as **APIs (no access to model architecture and parameters)**, making fine-tuning impractical.

Models with over 10B params

Not publicly released!

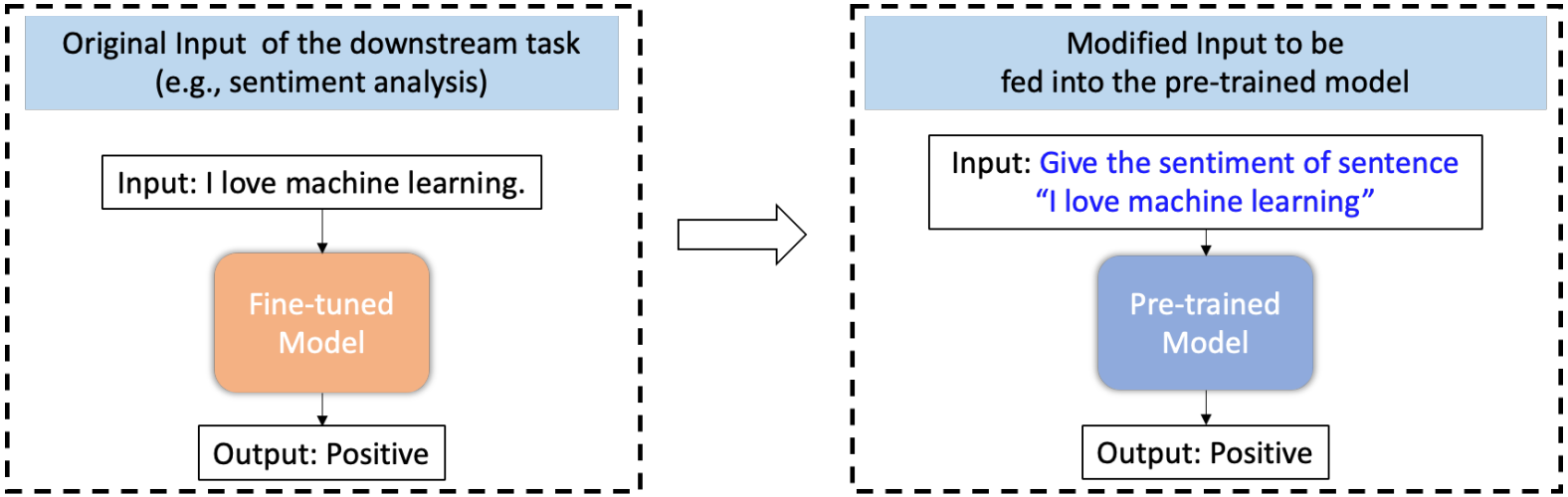


Zhao, Wayne Xin et al. "A Survey of Large Language Models." ArXiv abs/2303.18223 (2023).

Why prompting?

[Solution] **Prompting method**

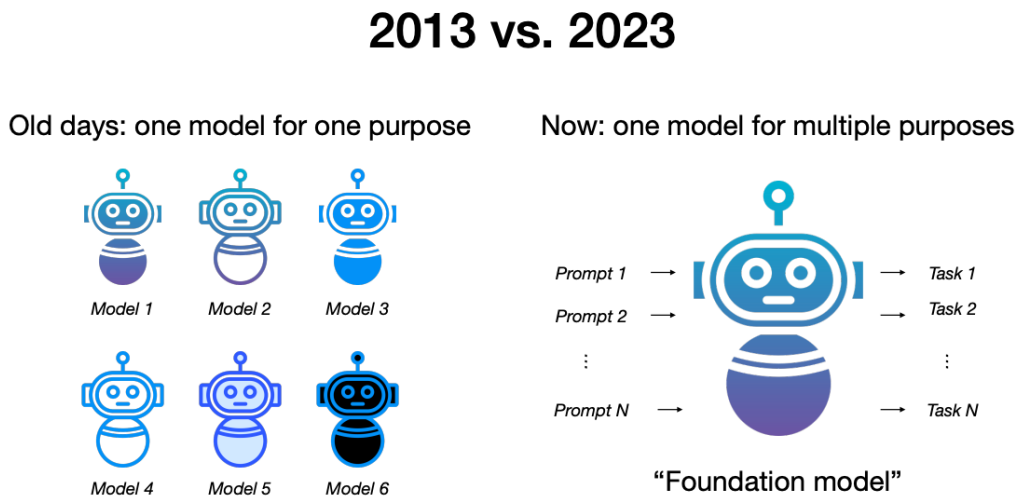
Given a pre-trained model, modify inputs of downstream tasks, e.g., add prompt, and feed them into the pre-trained model to adapt the pre-trained model for downstream tasks without updating its parameters.



Fine-tuning method **Prompting method**

Why prompting?

- Given a large model, we can adapt it to various tasks via giving different prompts.
- Prompting helps us get closer to the unified and general "**foundation model**".

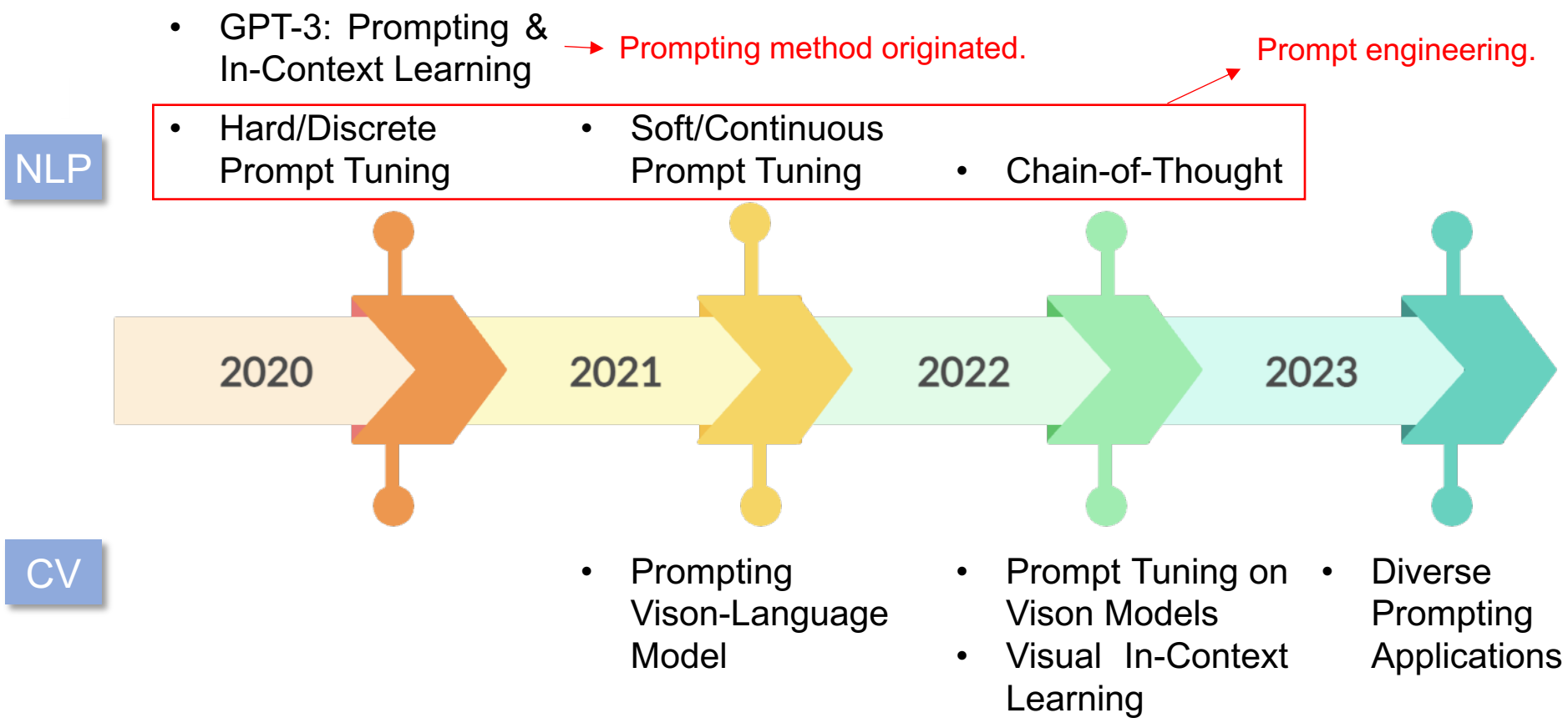


Zhou Kaiyang, "Prompting in Visual Intelligence" on CVPR2023 tutorial.

Outline

I. Motivation for Prompting
II. Development of Prompting
III. Prompting & Prompt Engineering in NLP
IV. Prompting & Prompt Engineering in CV
V. Prompting & Prompt Engineering on VLM
VI. Summary

Development of Prompting

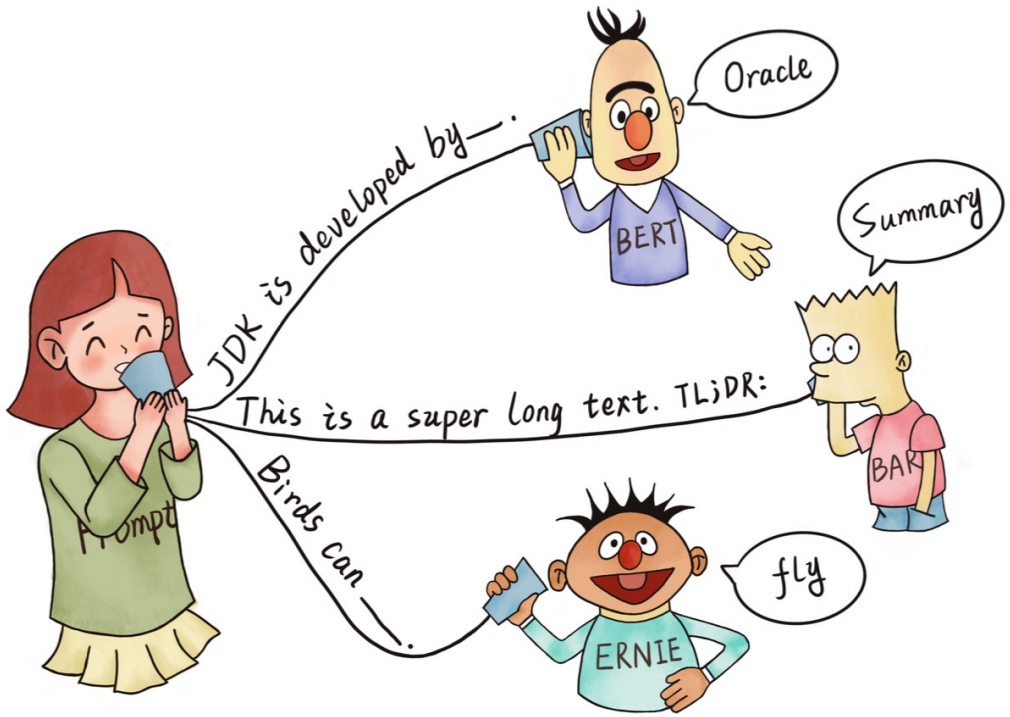


Outline

I. Motivation for Prompting
II. Development of Prompting
III. Prompting & Prompt Engineering in NLP
IV. Prompting & Prompt Engineering in CV
V. Prompting & Prompt Engineering on VLM
VI. Summary

Prompting in NLP

- Prompting paradigm: pre-train, prompt, predict.



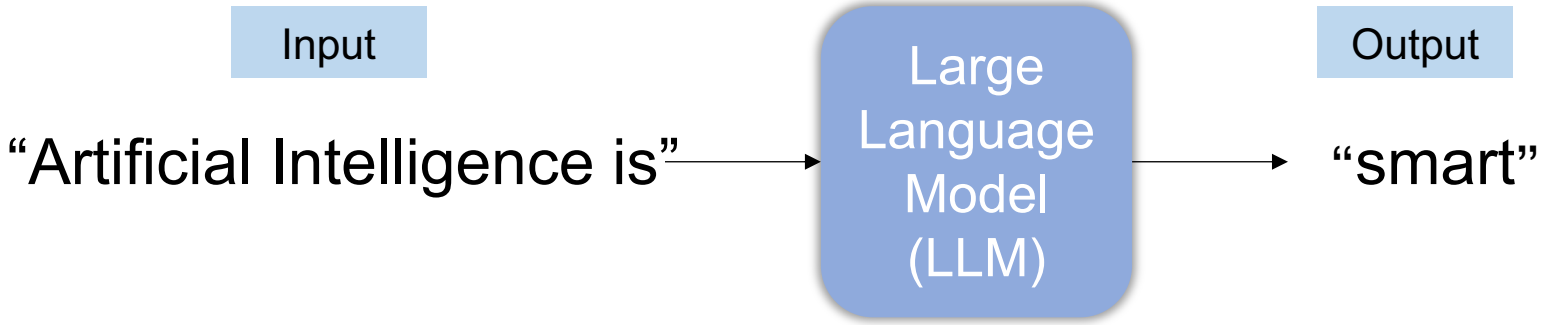
Liu, Pengfei et al. "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." ACM Computing Surveys 55 (2021): 1 - 35.

Prompting in NLP

- **Pre-train**

Pre-train the model to perform some preliminary tasks via self-supervised learning, for example, **next word prediction** and masked language modelling.

Pre-training task example
1: Next word prediction

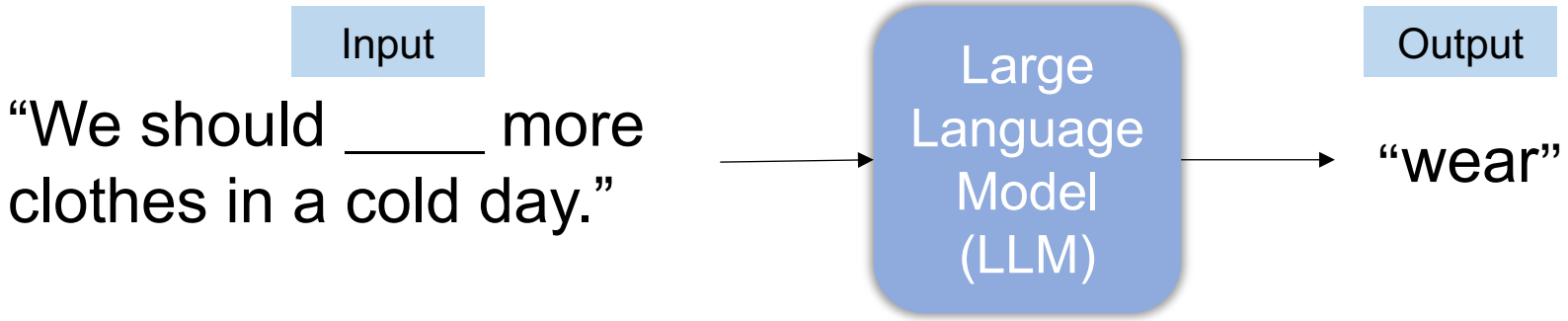


Prompting in NLP

- **Pre-train**

Pre-train the model to perform some preliminary tasks via self-supervised learning, for example, next word prediction and **masked language modelling**.

Pre-training task example 2:
Masked language modelling



Prompting in NLP

- **Prompt**

Modify the original input (query) using a prompt template to reformulate the [downstream task] format into the [pre-training task] format.

Downstream task example:
Sentiment analysis

When the pre-training task is next word prediction.

Original Input

"I love machine learning" → A specific model predicts sentiment directly.

Modified Input (Prompt)

"The sentiment of
'I love machine learning' is" → A generic LLM predicts sentiment by predicting what comes next

Prompting in NLP

- **Predict**

Feed the modified input with prompt into the pre-trained model, predicting **without updates on the model**.

Downstream task example:
Sentiment analysis

Modified Input (Prompt)

"The sentiment of 'I love machine learning' is"

→ LLM →

Pre-trained on next word prediction.

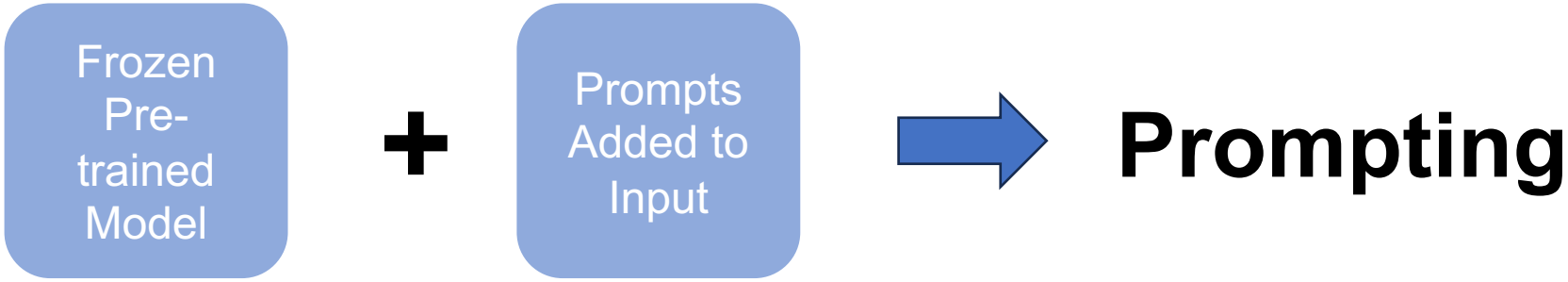
Output

"positive"

Prompting in NLP

- Key components of prompting:

- Frozen pre-trained model.
- Modification on input instead of model, adding prompt to the original input (query).



What is prompting?

Prompting:

- Employs a pre-trained model for desired tasks ***without* any gradient updates or fine-tuning on the model**, converting downstream tasks into the format of pre-training task[1,2]

1. *Brown et al. Language Models are Few-Shot Learners, 2020.*
2. *Bahng et al. Exploring Visual Prompts for Adapting Large-Scale Models, 2022.*

What is prompting?

Prompting:

- Employs a pre-trained model for desired tasks ***without*** **any gradient updates or fine-tuning on the model**, converting downstream tasks into the format of pre-training task. [1,2]
- Adapts a frozen pre-trained model by **modifying the *input* space**[2], injecting prompt as task specification.

1. *Brown et al. Language Models are Few-Shot Learners, 2020.*
2. *Bahng et al. Exploring Visual Prompts for Adapting Large-Scale Models, 2022.*

What is prompting?

Prompting:

- Employs a pre-trained model for desired tasks ***without*** **any gradient updates or fine-tuning on the model**, converting downstream tasks into the format of pre-training task. [1,2]
- Adapts a frozen pre-trained model by **modifying the *input* space**[2], injecting prompt as task specification.
- **Elicits knowledge** from pre-trained models.

1. Brown et al. Language Models are Few-Shot Learners, 2020.
2. Bahng et al. Exploring Visual Prompts for Adapting Large-Scale Models, 2022.

Prompting in NLP

What about performance of prompting?

What if compared to compared to fine-tuning?

Prompting in NLP

[Example] GPT-3

- GPT-3 is a large language model for text generation pre-trained on next word prediction.
- GPT-3 is adapted to various tasks via prompting.

Zero-shot | One-shot | Few-shot

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

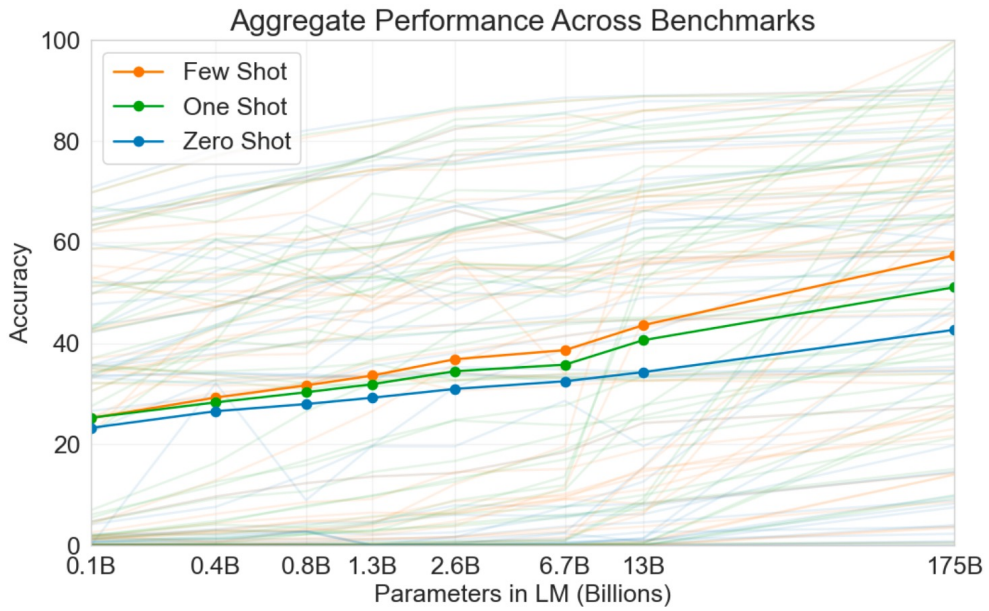
```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

Fed into GPT-3

Brown, Tom B. et al. "Language Models are Few-Shot Learners." ArXiv abs/2005.14165 (2020).

Prompting in NLP

- Prompting performance depends on model scale. The larger pre-trained model is, the more effective prompting method is.



Brown, Tom B. et al. "Language Models are Few-Shot Learners." ArXiv abs/2005.14165 (2020).

Prompting in NLP

- Prompting 175B GPT-3 can outperform fine-tuning 340M Bert-Large on SuperGLUE benchmark, where prompting GPT-3 uses 32 examples to form prompt and fine-tuning Bert-Large requires 125K examples as training data.

SuperGLUE* Performance

	SuperGLUE Average
Fine-tuned SOTA	**89.0**
Fine-tuned BERT-Large	69.0
GPT-3 Few-Shot	71.8

Brown, Tom B. et al. "Language Models are Few-Shot Learners." ArXiv abs/2005.14165 (2020).

**SuperGLUE – overall evaluation: an NLP benchmark consisting of 8 different language tasks. Refer to https://super.gluebenchmark.com/*

Prompting in NLP

- There is a gap between the performance of the state-of-the-art traditional fine-tuning method and prompting method.
- One possible reason can be the prompt used is manually designed and fixed.

SuperGLUE* Performance

	SuperGLUE Average
Fine-tuned SOTA	**89.0**
Fine-tuned BERT-Large	69.0
GPT-3 Few-Shot	71.8

Large gap!

Brown, Tom B. et al. "Language Models are Few-Shot Learners." ArXiv abs/2005.14165 (2020).

*SuperGLUE – overall evaluation: an NLP benchmark consisting of 8 different language tasks. https://super.gluebenchmark.com/

Prompting in NLP

[Limitation of prompting method]

- Prompting performance is unstable and sensitive to the prompt: when different prompts are used for the same downstream task, the performance varies.

Prompt	P@1
[X] is located in [Y]. *(original)*	31.29
[X] is located in which country or state? [Y].	19.78
[X] is located in which country? [Y].	31.40
[X] is located in which country? In [Y].	51.08

Best prompt

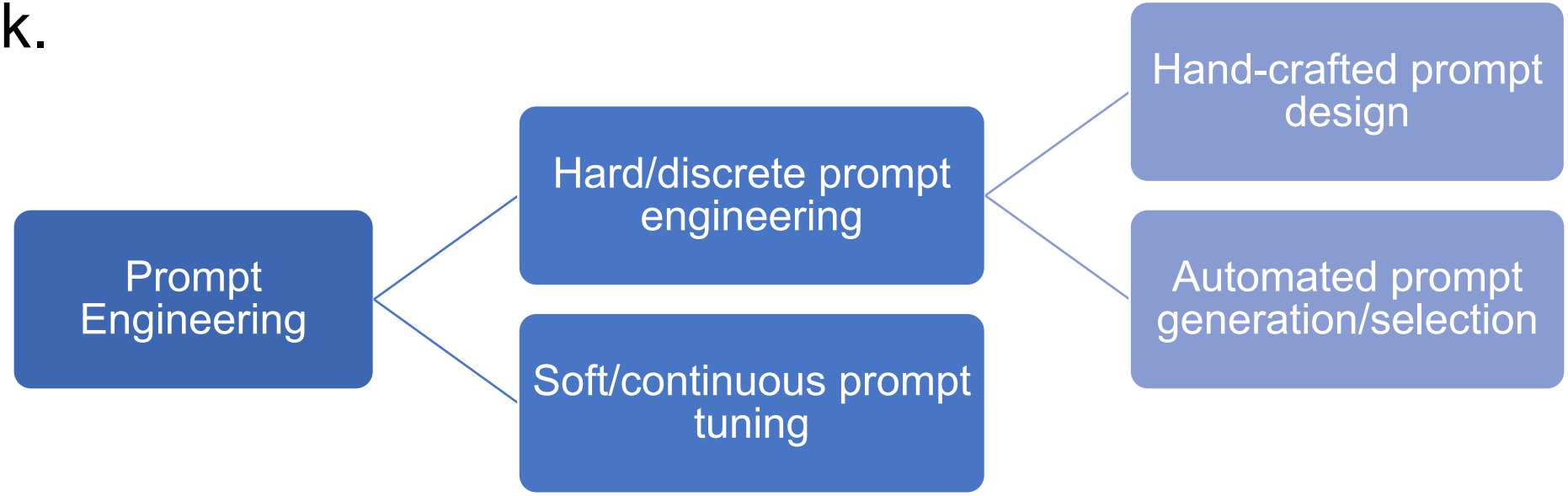
Table 1. Case study on LAMA-TREx P17 with bert-base-cased. A single-word change in prompts could yield a drastic difference.

Liu, Xiao et al. "GPT Understands, Too." ArXiv abs/2103.10385 (2021).

Prompt Engineering in NLP

[Solution]

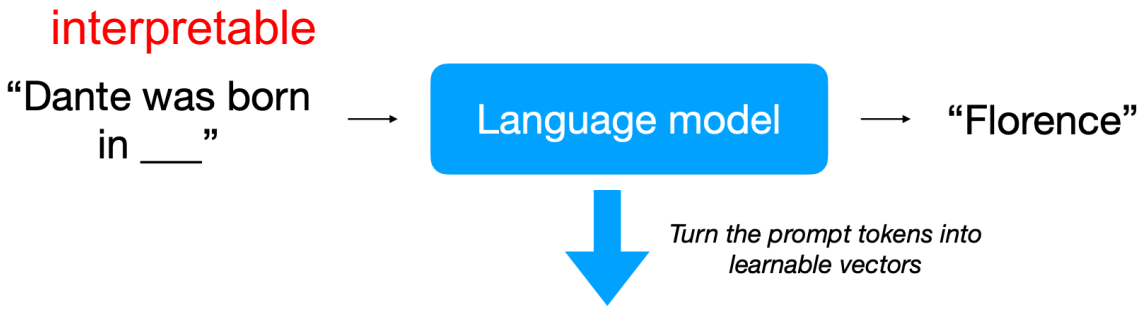
- **Prompt engineering**: manually or automatically find the most appropriate prompt to allow the model to solve the downstream task.



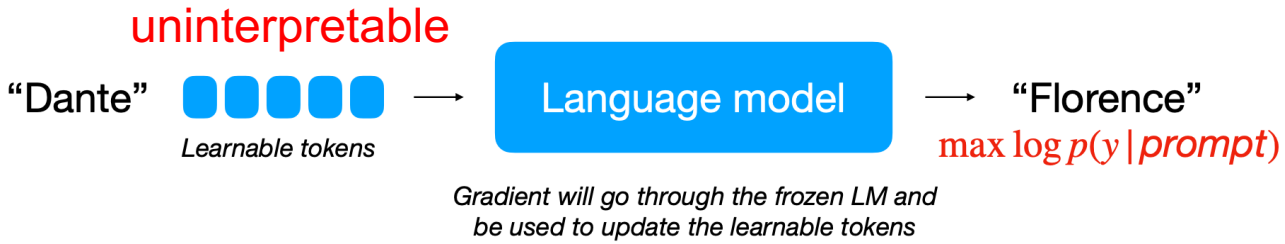
Liu, Pengfei et al. "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." ACM Computing Surveys 55 (2021): 1 - 35.

Prompt Engineering in NLP

- **Hard/discrete prompt**: natural language text string.

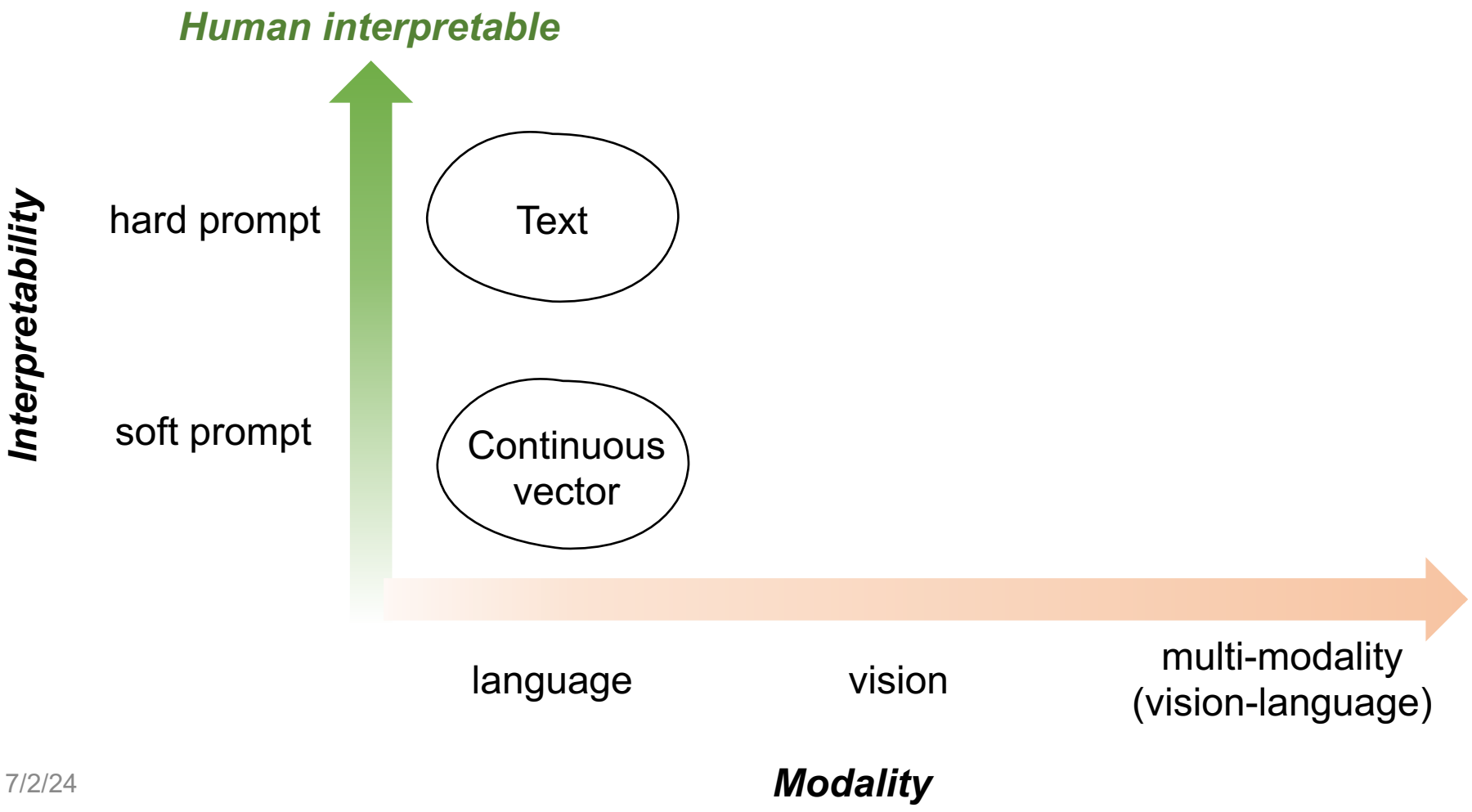


- **Soft/continuous prompt**: continuous vector in embedding space.



Zhou Kaiyang, "Prompting in Visual Intelligence" on CVPR2023 tutorial.

Prompt Engineering in NLP

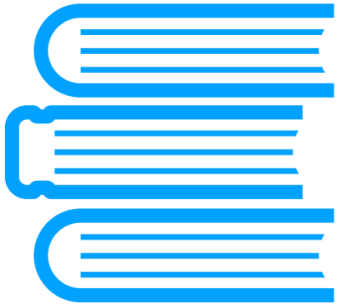


Hard Prompt Tuning in NLP

- Mining-based prompt generation

Relation: subclass of
Manual: x is a subclass of y
Mined: x is a type of y
Gain: +22.7

y is the target to be predicted



A large database (e.g., Wikipedia) containing both subjects (x) and objects (y)

Mining

Mined candidates:
x is a type of y (92.7%)
x belongs to y (80.2%)
…
x is a subclass of y (70.0%)

Zhou Kaiyang, "Prompting in Visual Intelligence" on CVPR2023 tutorial.

Hard Prompt Tuning in NLP

- Paraphrasing-based prompt generation

x <u>is a subclass of</u> y

A seed prompt (manual or mined)

Paraphrasing →

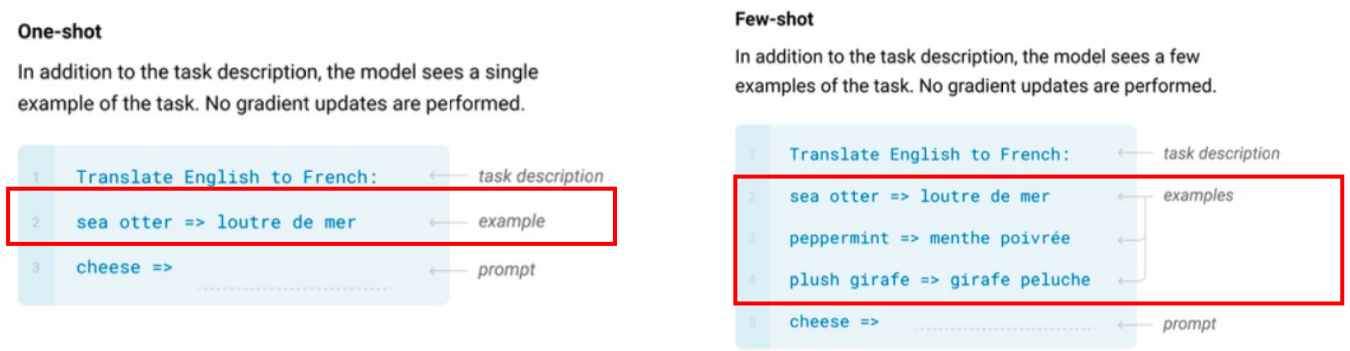
- *Back translation (Jiang et al., 2020)*
- *Neural prompt rewriter (Haviv et al., 2021)*
- *etc.*

x <u>is a type of</u> y (92.7%)
x <u>belongs to</u> y (80.2%)
…

Zhou Kaiyang, "Prompting in Visual Intelligence" on CVPR2023 tutorial.

Special Case: In-Context Learning

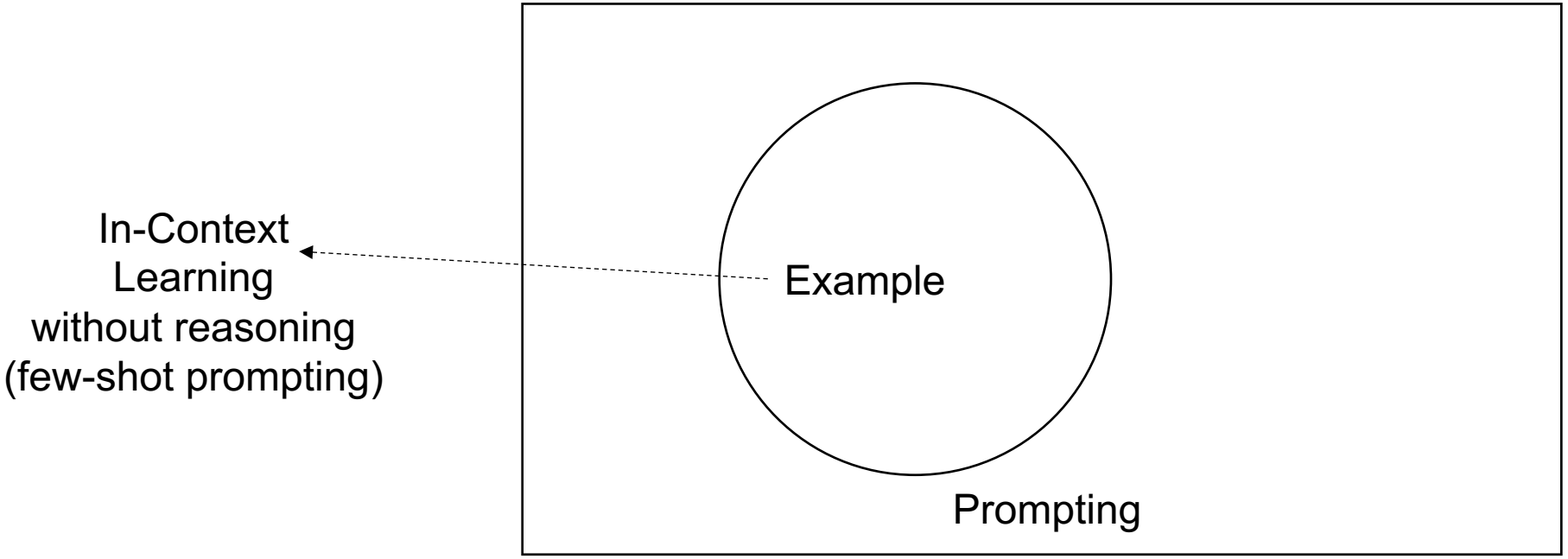
- In-context learning (ICL) is **a special case of hard prompt engineering,** involving no additional training on the pre-trained model.
- In ICL, the prompt is designed to be provided with **demonstration examples** of downstream task as context.



Dong, Qingxiu et al. "A Survey on In-context Learning." ArXiv abs/2301.00234 (2022).
Brown, Tom B. et al. "Language Models are Few-Shot Learners." ArXiv abs/2005.14165 (2020).

Special Case: In-Context Learning

- In-context learning is one of prompting methods, with example demonstration provided in the prompt as context.



Special Case: Chain-of-Thought Prompting

- Chain-of-thought (CoT) prompting is **a special case of hard prompt engineering**, involving no additional training on the pre-trained model.
- CoT prompting **enables complex reasoning capabilities** through intermediate reasoning steps.

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

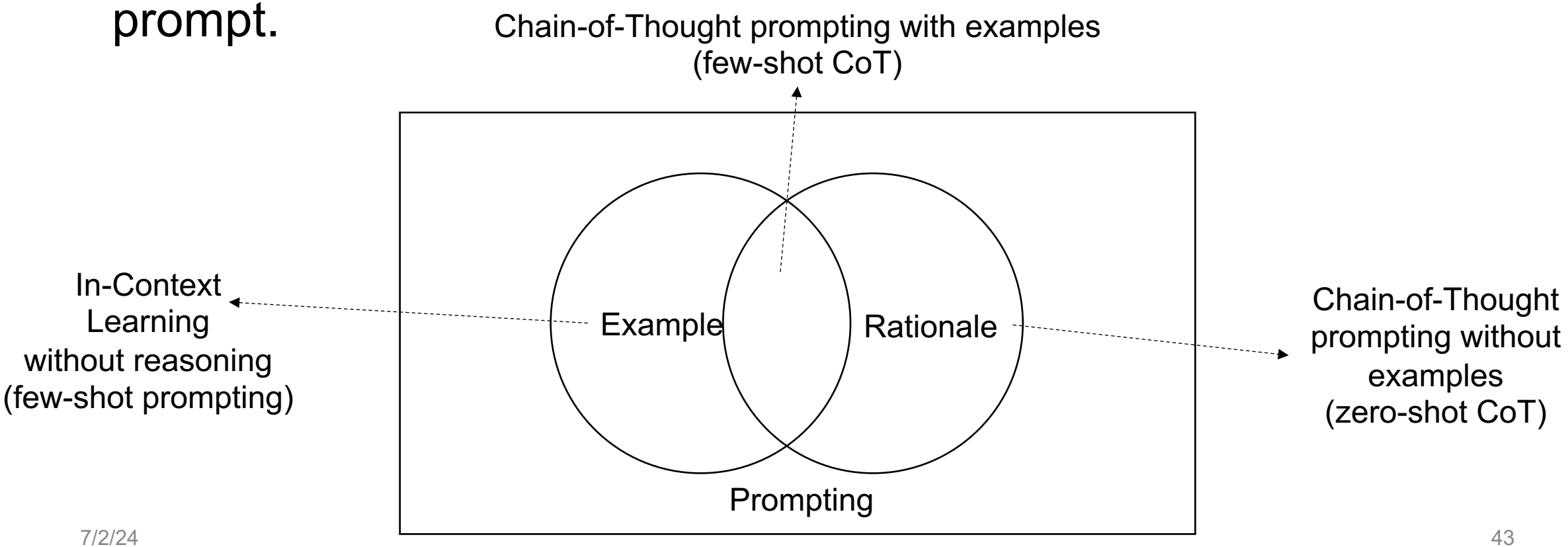
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

Wei, Jason et al. "Chain of Thought Prompting Elicits Reasoning in Large Language Models." ArXiv abs/2201.11903 (2022).
Kojima, Takeshi et al. "Large Language Models are Zero-Shot Reasoners." ArXiv abs/2205.11916 (2022).

Special Case: Chain-of-Thought Prompting

- Chain-of-thought prompting is one of prompting methods, with reasoning rationale or instruction to generate rationale in the prompt.



Hard prompt may lead to **sub-optimal** results since **constraining the prompts to be text strings**.

Soft Prompt Tuning in NLP

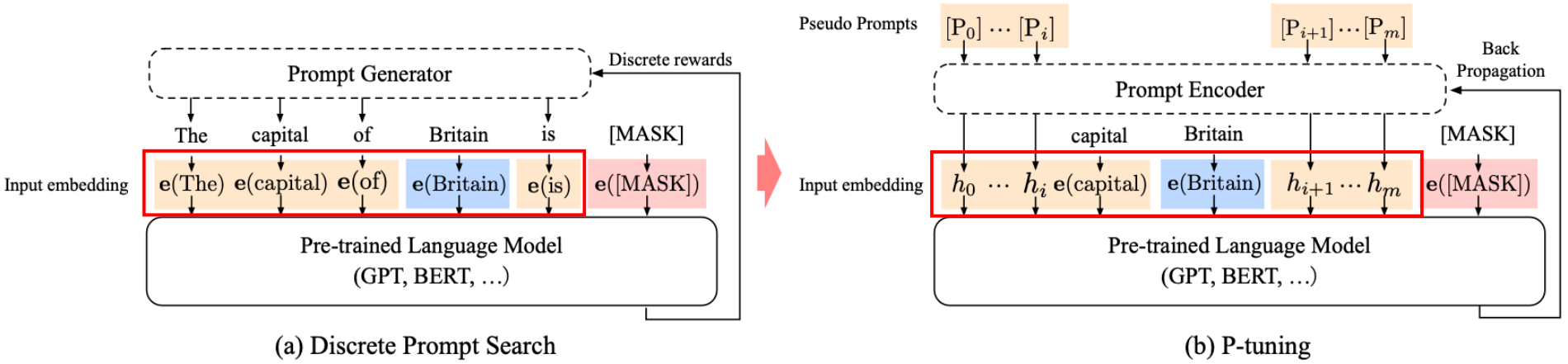
There are **concurrent works** studying soft prompt tuning: p-tuning, prefix-tuning, prompt-tuning.

- They add **trainable soft prompts** to the input embeddings and **learn the prompts** on downstream dataset.
- Their results show that **prompt tuning helps reduce the gap** between fine-tuning and prompting.

Soft Prompt Tuning in NLP

- P-Tuning v1

Prompt is actual text string → Prompt is continuous embedding vector.



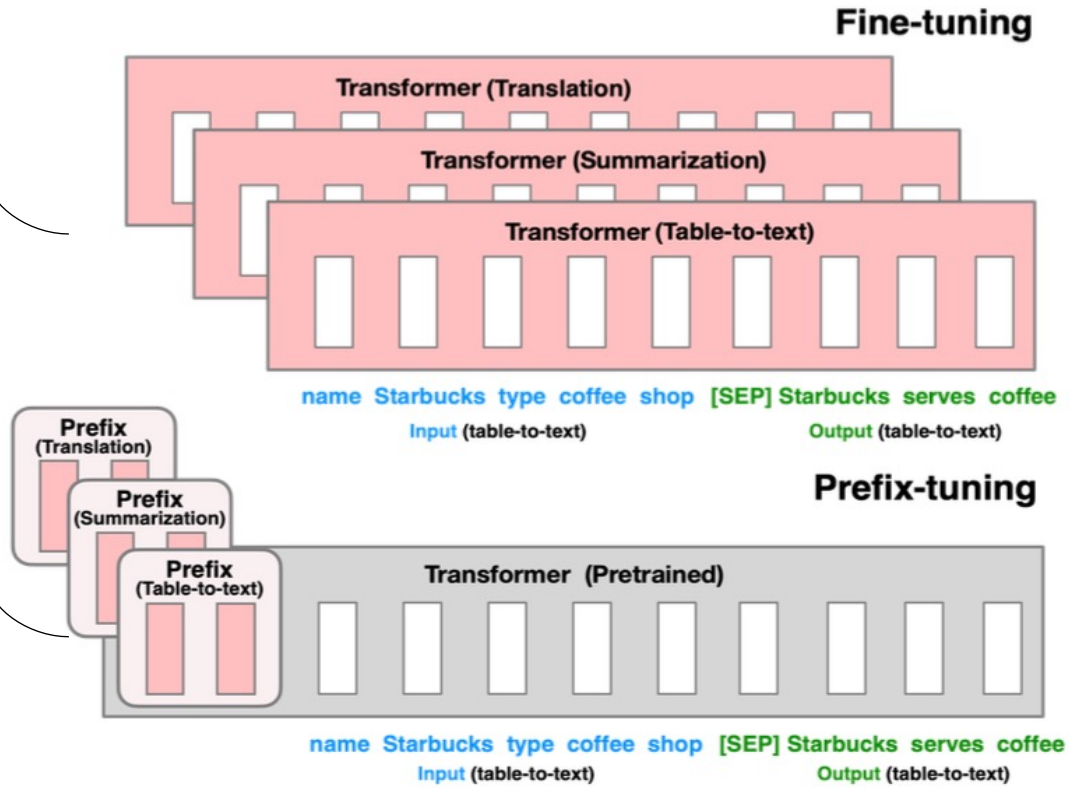
Liu, Xiao et al. "GPT Understands, Too." ArXiv abs/2103.10385 (2021).

Soft Prompt Tuning in NLP

- Prefix-Tuning

Update the whole pre-trained model.

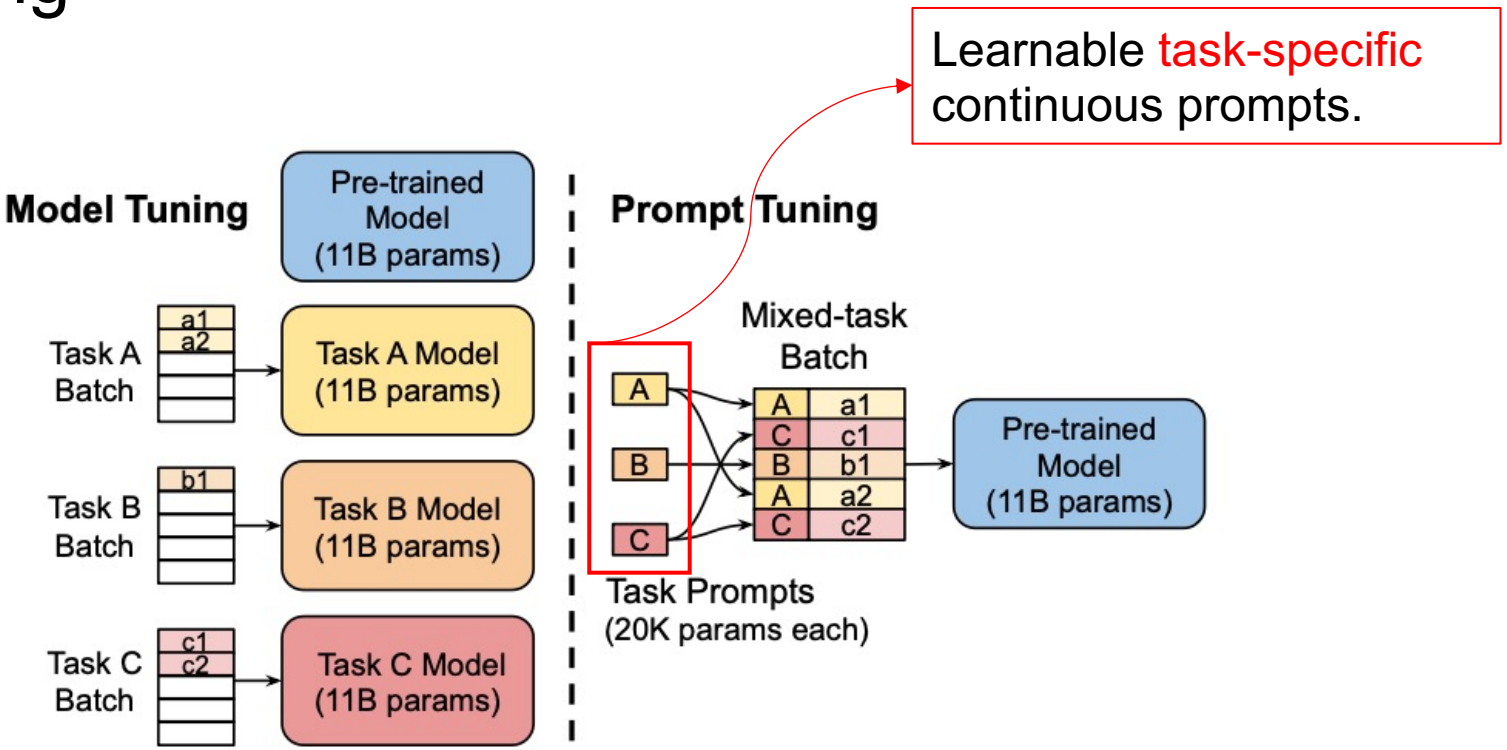
Only update task-specific prefix tokens added to the input.



Li, Xiang Lisa and Percy Liang. "Prefix-Tuning: Optimizing Continuous Prompts for Generation." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) abs/2101.00190 (2021).

Soft Prompt Tuning in NLP

- Prompt-Tuning



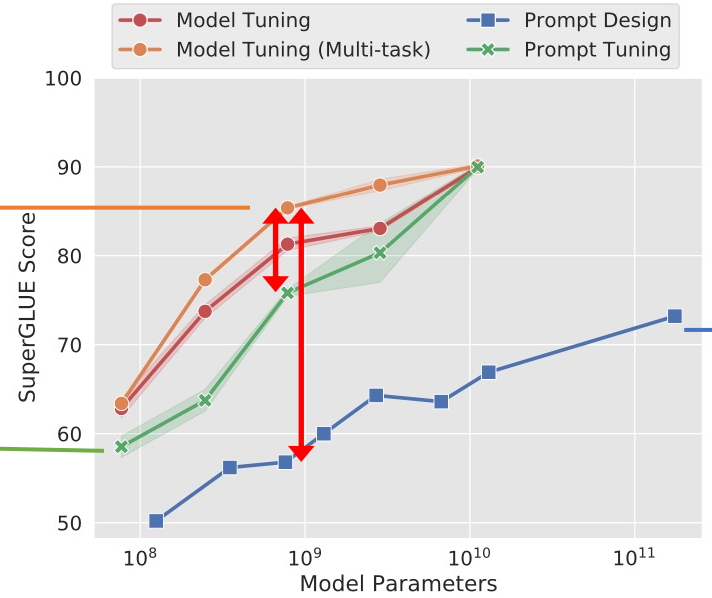
Lester, Brian et al. "The Power of Scale for Parameter-Efficient Prompt Tuning." Conference on Empirical Methods in Natural Language Processing (2021).

Soft Prompt Tuning in NLP

- Prompt tuning outperforms few-shot prompting (without additional training) with designed prompt and reduces the gap between fine-tuning and few-shot prompting.

Fine-tuning based on T5 with different sizes.

Prompt tuning based on T5 with different sizes.

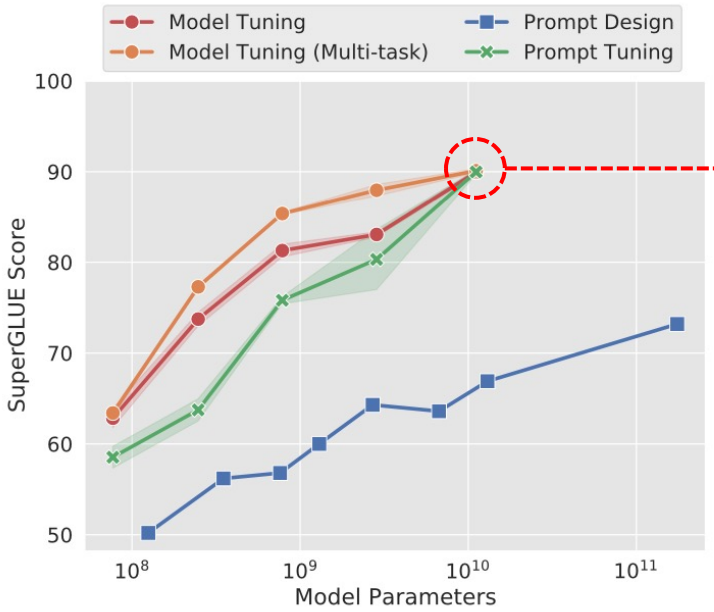


Manually designed prompt based on GPT-3.

Lester, Brian et al. "The Power of Scale for Parameter-Efficient Prompt Tuning." Conference on Empirical Methods in Natural Language Processing (2021).

Soft Prompt Tuning in NLP

- Prompt tuning can achieve comparable performance compared to full model fine-tuning, with much less tunable parameters.



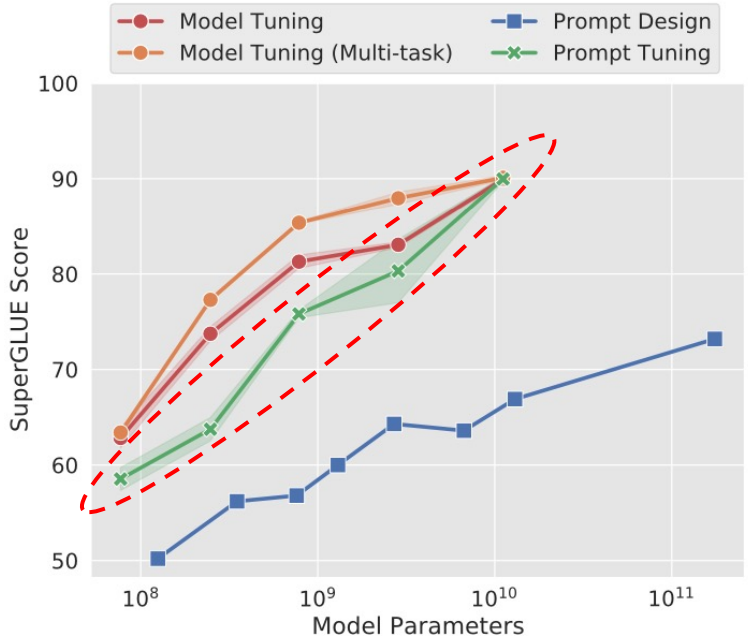
Full model fine-tuning
Fine-tune all model parameters.
# of tunable params: $\sim 10^{10}$

Prompt tuning
Fine-tune prompt only.
# of tunable params: < 0.01% of model parameters.

Lester, Brian et al. "The Power of Scale for Parameter-Efficient Prompt Tuning." Conference on Empirical Methods in Natural Language Processing (2021).

Soft Prompt Tuning in NLP

- Prompt tuning benefits from model size scaling up, and the increase does not plateau (green line in the figure).



Lester, Brian et al. "The Power of Scale for Parameter-Efficient Prompt Tuning." Conference on Empirical Methods in Natural Language Processing (2021).

Soft Prompt Tuning in NLP

As mentioned,

the performance of both prompting with fixed prompts and prompt tuning depends on model scale.

The larger the model is,

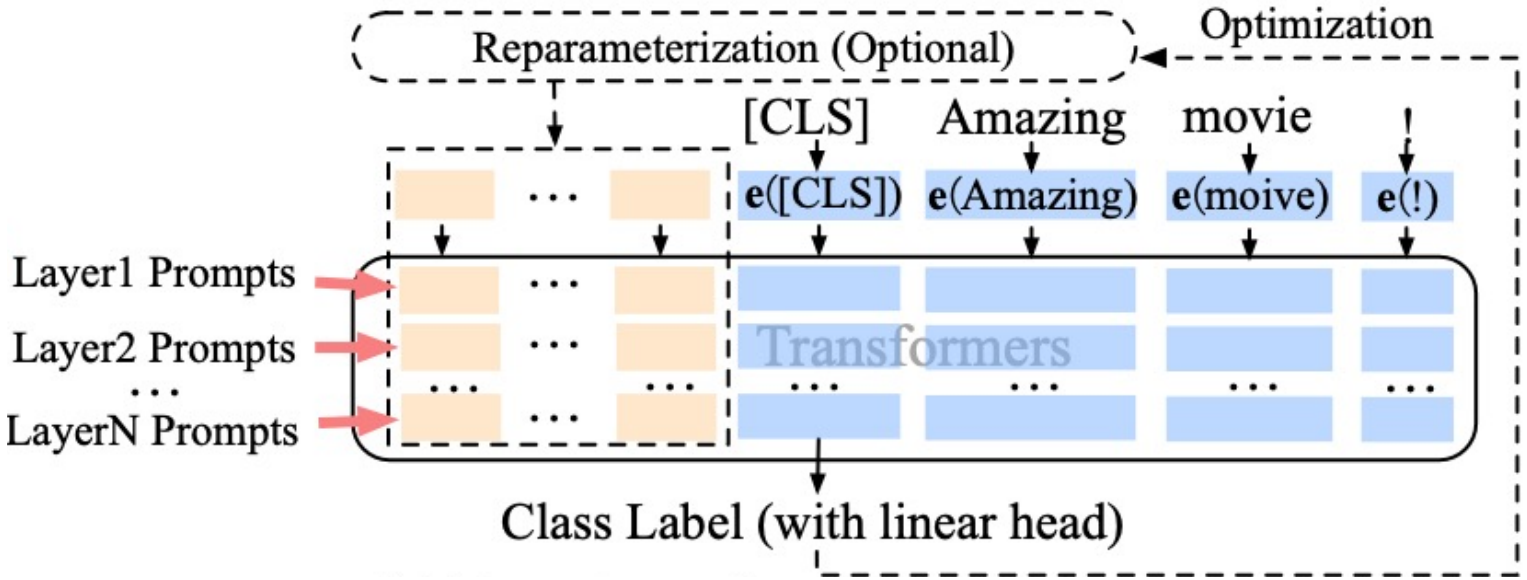
the more improvements the prompting brings.

Can prompting be applied to small models?

Soft Prompt Tuning in NLP

[Solution]

- P-tuning v2: insert learnable continuous prompts in multiple layers.

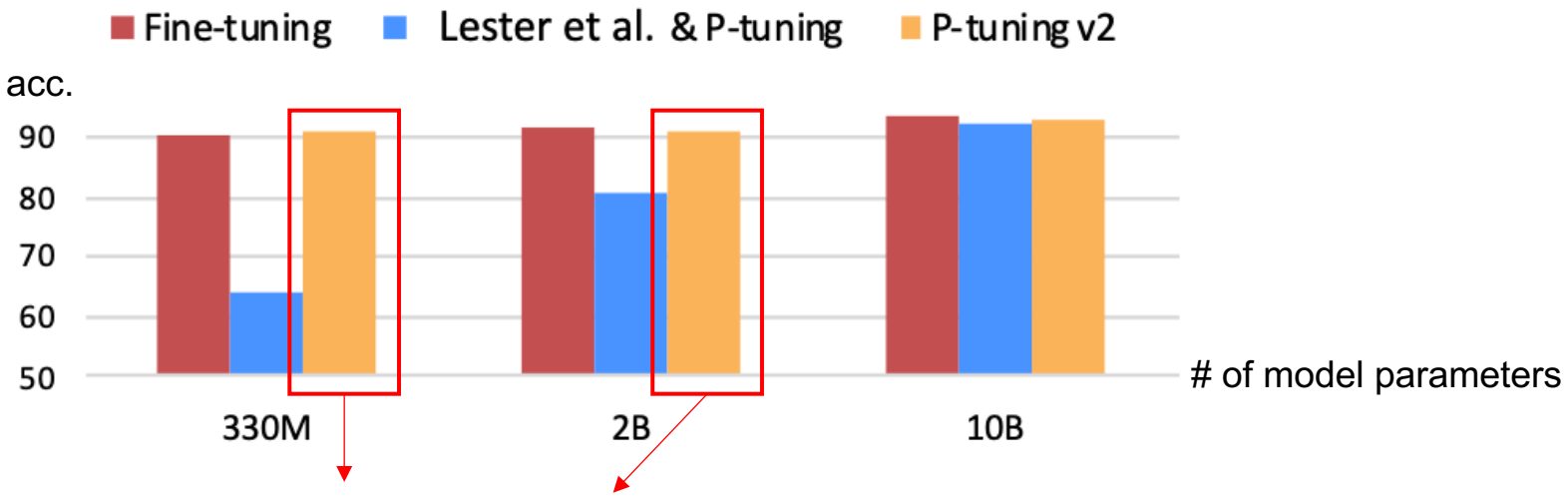


Liu, Xiao et al. "P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks." Annual Meeting of the Association for Computational Linguistics (2022).

Soft Prompt Tuning in NLP

[Performance]

- Prompts in multiple latent spaces reduce the gap between models with various scales.



Prompt tuning also works for small or medium models.

Liu, Xiao et al. "P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks." Annual Meeting of the Association for Computational Linguistics (2022).

Bio Break– 15 Mins

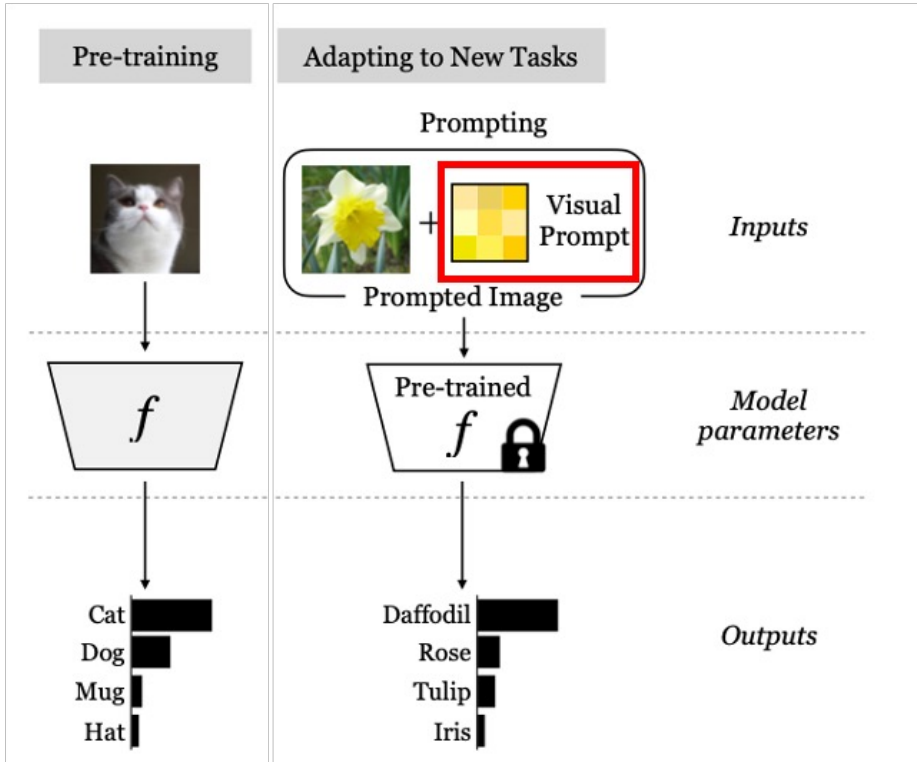
Outline

I. Motivation for Prompting
II. Development of Prompting
III. Prompting & Prompt Engineering in NLP
IV. Prompting & Prompt Engineering in CV
V. Prompting & Prompt Engineering on VLM
VI. Summary

Prompting in CV

[Example 1] Image classification

- **Pre-training task:** classify animal images.
- **Downstream task:** classify flower images.
- The downstream input is added with a **prompt (pixel patch)**, and then fed into the pre-trained model to conduct downstream task.



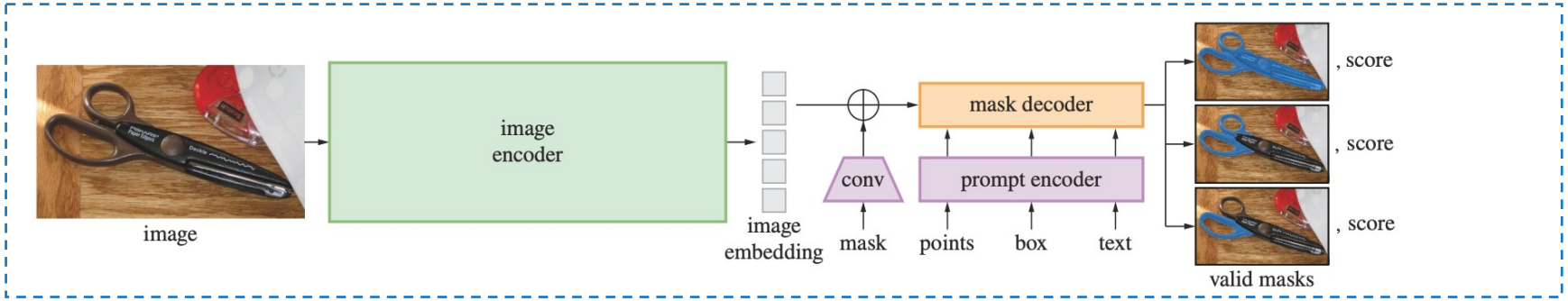
Bahng, Hyojin et al. "Exploring Visual Prompts for Adapting Large-Scale Models." ArXiv abs/2203.17274 (2022).

Prompting in CV

[Example 2] Segment Anything

- **Pre-training task**: promptable segmentation, aiming to predict valid segmentation masks given segmentation prompt specifying what to segment in an image.
- **Downstream task**: any segmentation task which outputs masks, e.g., foreground segmentation, instance segmentation, etc.

Model pre-trained on promptable segmentation task

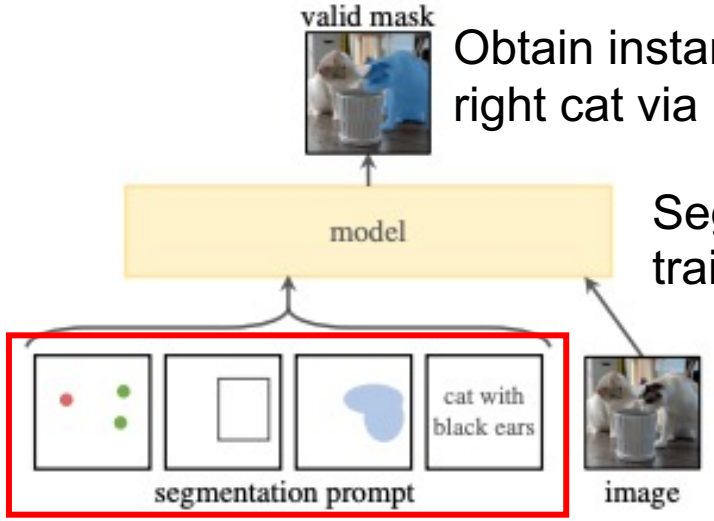


Kirillov, Alexander et al. "Segment Anything." ArXiv abs/2304.02643 (2023).

Prompting in CV

[Example 2] Segment Anything

- **Prompting**: **visual elements (points, bounding boxes, masks) and text can serve as prompt**, specifying the downstream task, i.e., what to segment.



Obtain instance segmentation mask for the right cat via prompting the pre-trained model.

Segment Anything Model (SAM) pre-trained on promptable segmentation.

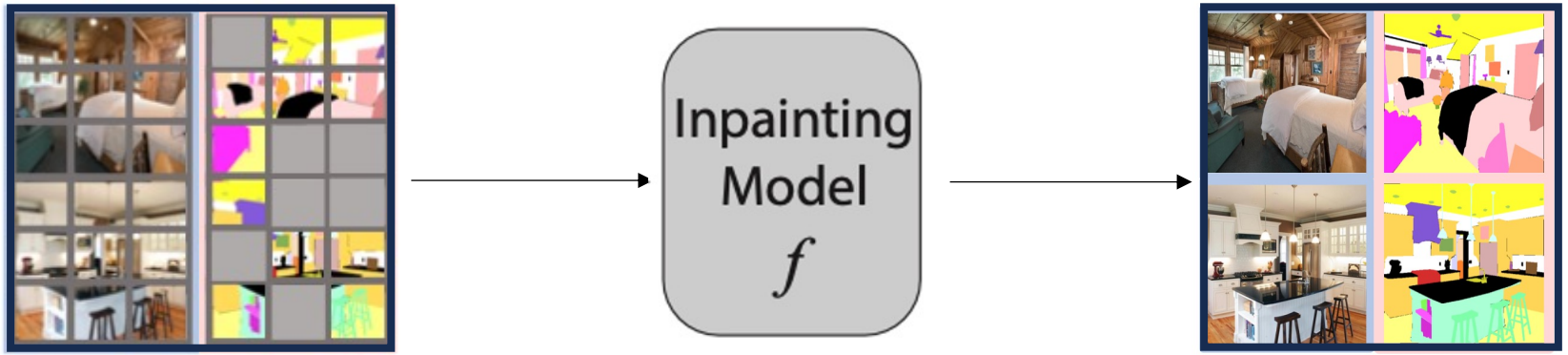
Prompts specify to segment the right cat (instance segmentation).

Kirillov, Alexander et al. "Segment Anything." ArXiv abs/2304.02643 (2023).

Prompting in CV

[Example 3] Visual in-context learning

- **Pre-training task**: image inpainting.



Input: an image with randomly masked patches

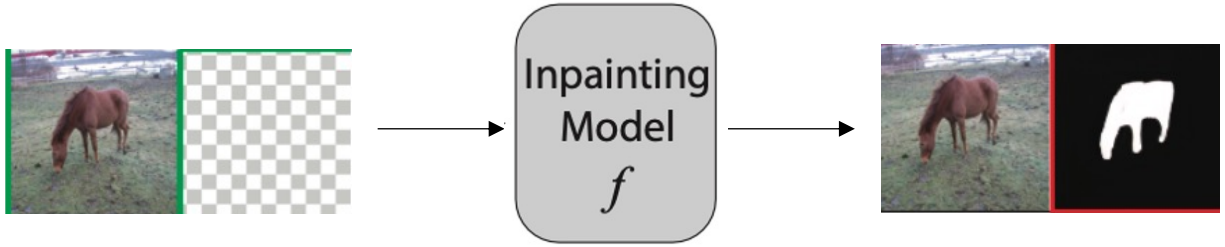
Output: an intact image with prediction of masked patches

Bar, Amir et al. "Visual Prompting via Image Inpainting." NeurIPS (2022).
Wang, Xinlong et al. "Images Speak in Images: A Generalist Painter for In-Context Visual Learning." 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023).

Prompting in CV

[Example 3] Visual in-context learning

- **Reformulate the downstream task**: 1) represent the input and output of downstream task as images; 2) reformulate the downstream task (e.g., instance segmentation) as image inpainting, that is, combine the input and output together, and regard the output as masked part which is to be predicted by image inpainting.



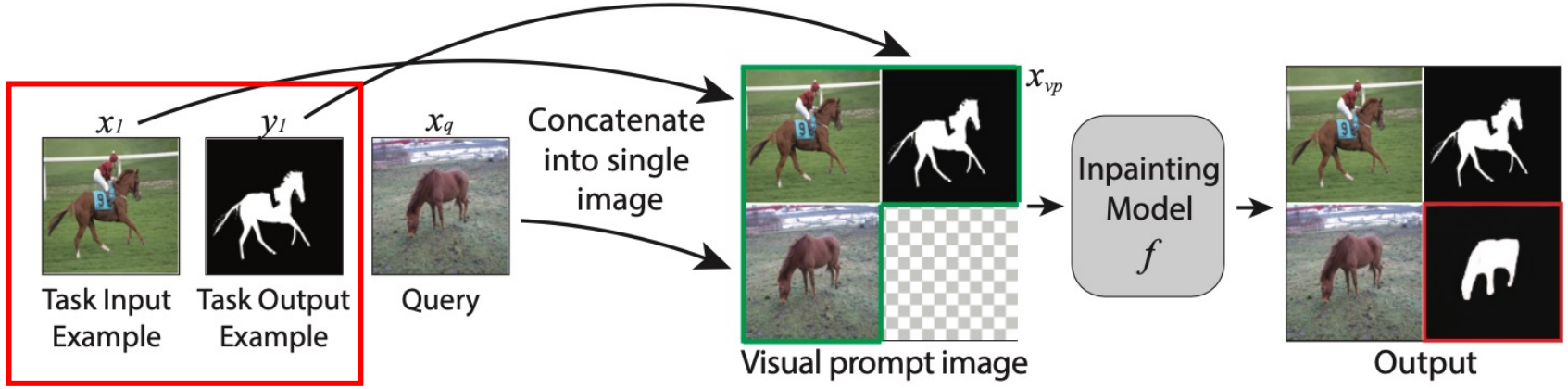
Bar, Amir et al. "Visual Prompting via Image Inpainting." NeurIPS (2022).
Wang, Xinlong et al. "Images Speak in Images: A Generalist Painter for In-Context Visual Learning." 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023).

Prompting in CV

[Example 3] Visual in-context learning

- **Prompting**: combine downstream input (query) with a prompt (a pair of downstream input and output images as task example), feed the combined image into the pre-trained model, prompting it to perform downstream task by inpainting.

Visual prompt



Bar, Amir et al. "Visual Prompting via Image Inpainting." NeurIPS (2022).
Wang, Xinlong et al. "Images Speak in Images: A Generalist Painter for In-Context Visual Learning." 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023).

Prompting in CV

[Example 3] Visual in-context learning

- **Applicable downstream task**: tasks of which output can be represented as images, e.g., semantic segmentation, image denoising, and etc.

Downstream task	Downstream input	Downstream output
Semantic segmentation		
Image denoising		

Bar, Amir et al. "Visual Prompting via Image Inpainting." NeurIPS (2022).
Wang, Xinlong et al. "Images Speak in Images: A Generalist Painter for In-Context Visual Learning." 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023).

Prompt Engineering in CV

[Limitation] An example in visual in-context learning

Given a query image (left-bottom), prediction result (right-bottom) largely varies from different prompts (top row) and is affected by:

1) The annotation quality (task output example) in the prompt;
2) The similarity between the task input example in the prompt and the query.

Zhang, Yuanhan et al. "What Makes Good Examples for Visual In-Context Learning?" ArXiv abs/2301.13670 (2023).

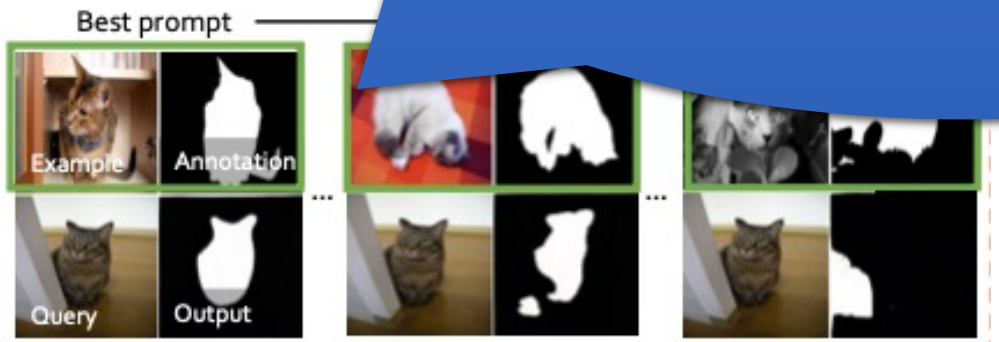
Prompt Engineering in CV

[Limitation] An example in visual in

Given a query image (left-bo
largely vary from different

1) The annotation quality
2) The similarity between
and the query.

While adapting a pre-trained model to a new downstream task via prompting, the **downstream performance depends on the prompt**, necessitating to find the best prompt for the task.

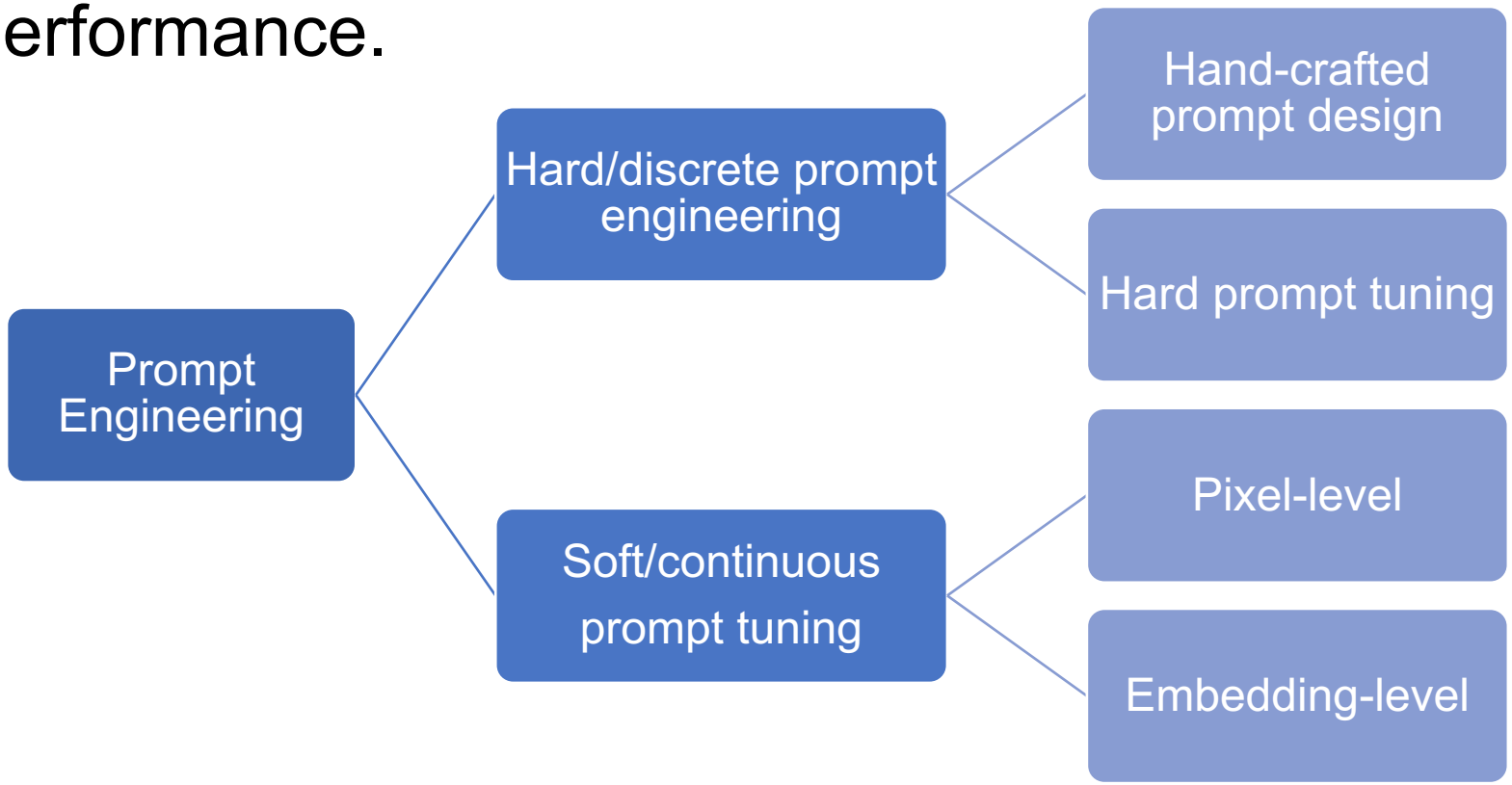


Zhang, Yuanhan et al. "What Makes Good Examples for Visual In-Context Learning?" ArXiv abs/2301.13670 (2023).

Prompt Engineering in CV

[Solution]

- Prompt engineering helps to build appropriate prompts to obtain better performance.

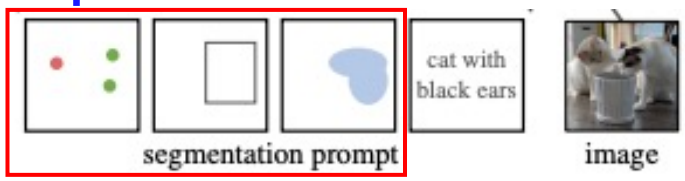


Prompt Engineering in CV

Similar to prompt in NLP:

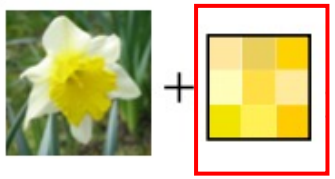
- **Hard/discrete visual prompt**:

  human interpretable visual elements.

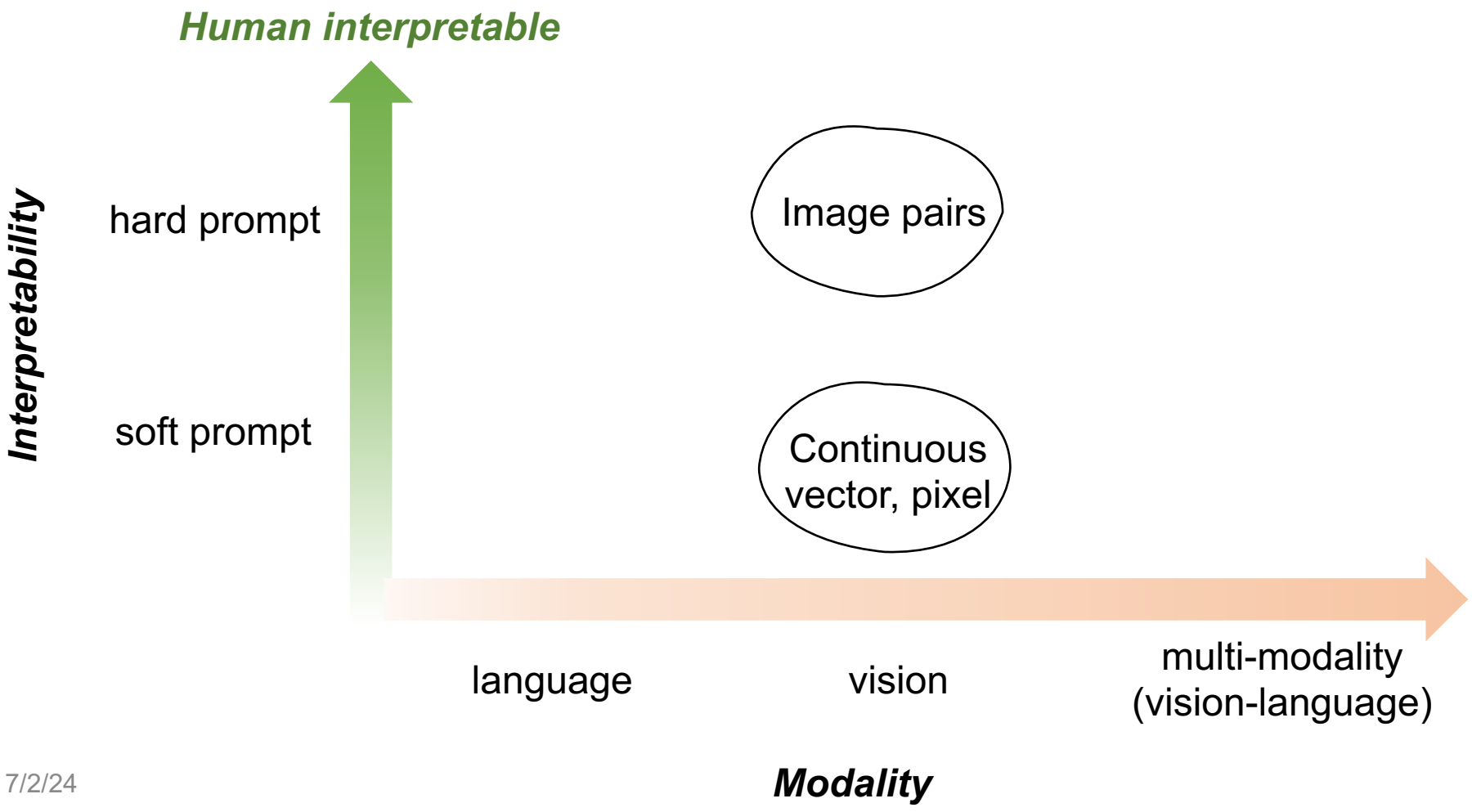


- **Soft/continuous visual prompt**:

  continuous values in lack of human interpretability.



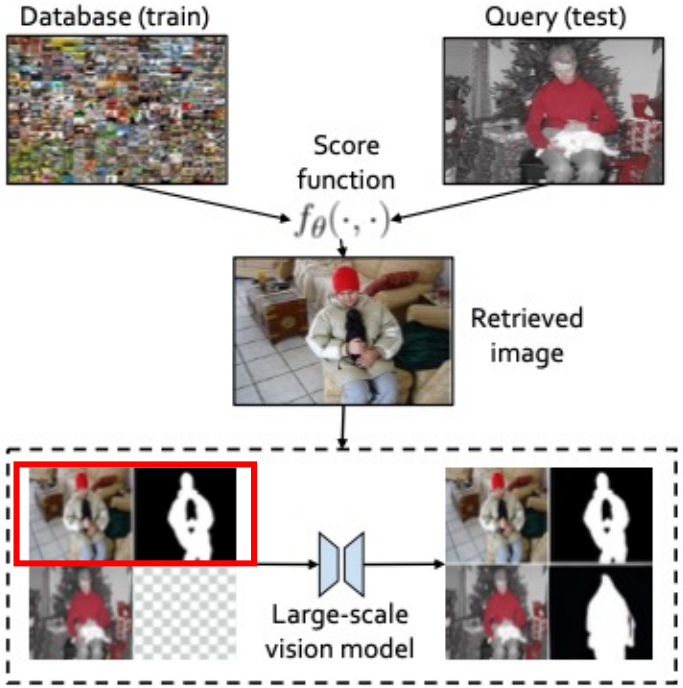
Prompt Engineering in CV



Hard Prompt Engineering in CV

- Retrieving-based method: Search the training dataset and retrieve the best image pair as prompt.

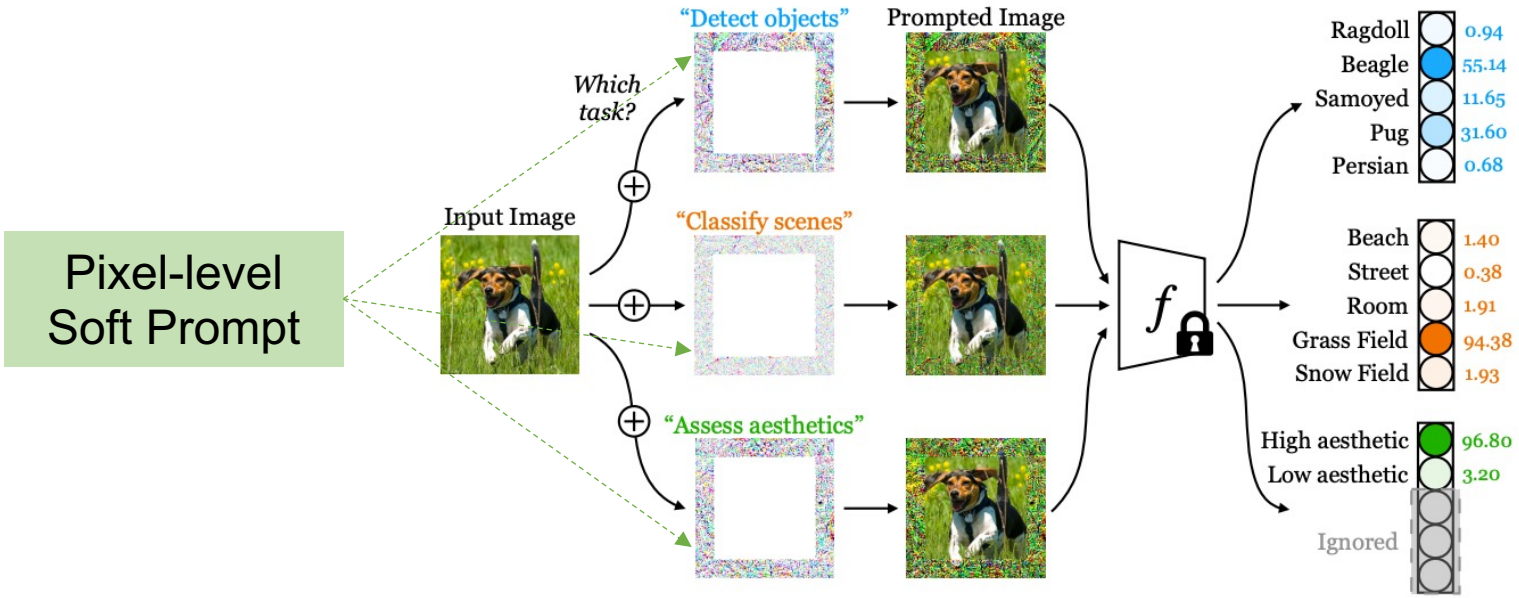
Hard Prompt



Zhang, Yuanhan et al. "What Makes Good Examples for Visual In-Context Learning?" ArXiv abs/2301.13670 (2023).

Soft Prompt Tuning in CV

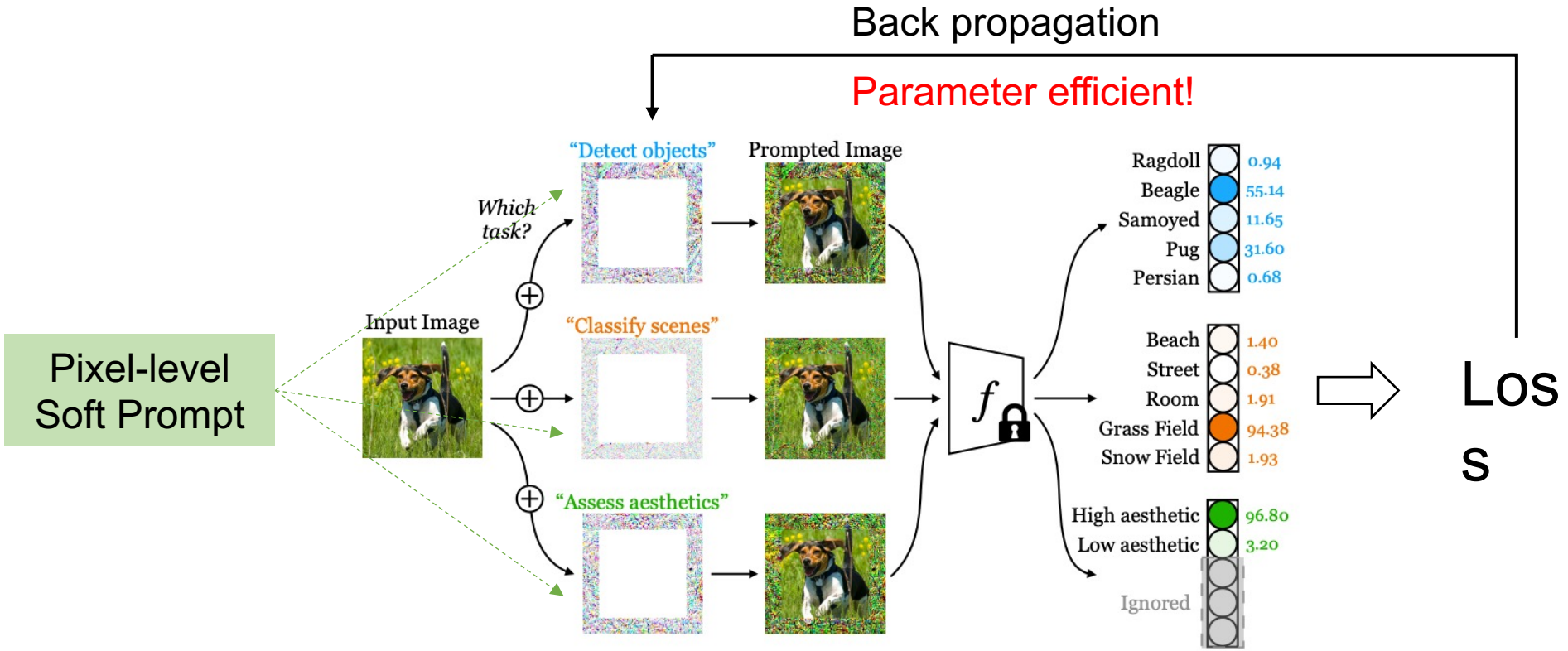
- Pixel-level learnable visual prompt: add perturbation as prompt to the input image.
- Prompt plays a role of task specification.



Bahng, Hyojin et al. "Exploring Visual Prompts for Adapting Large-Scale Models." ArXiv abs/2203.17274 (2022).

Soft Prompt Tuning in CV

- Task-specific pixel-level prompt tuning.



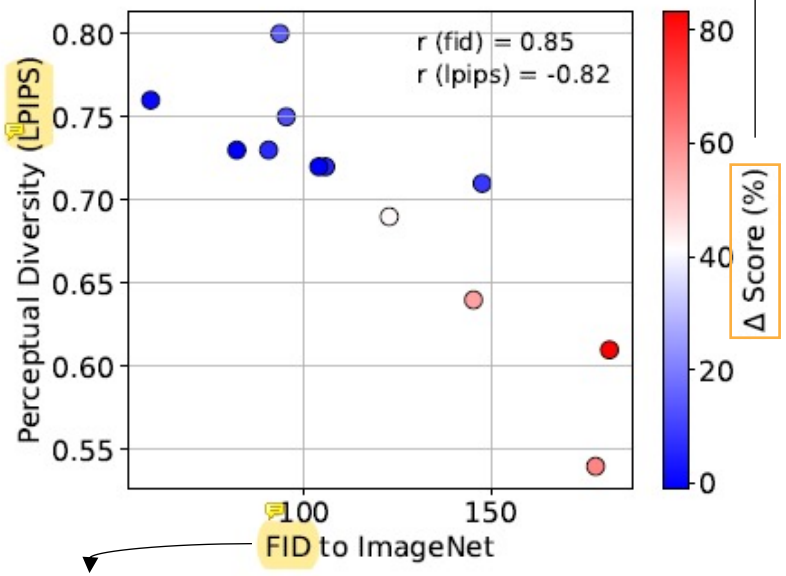
Bahng, Hyojin et al. "Exploring Visual Prompts for Adapting Large-Scale Models." ArXiv abs/2203.17274 (2022).

Soft Prompt Tuning in CV

- Pixel-level visual prompt tuning performance.

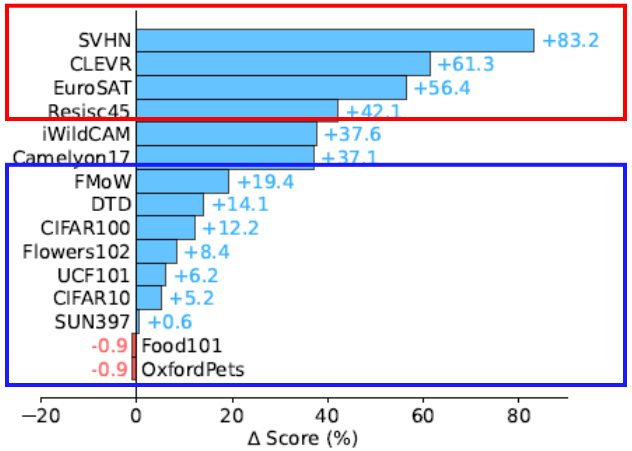
Diversity of downstream dataset.

performance gain



Discrepancy between the downstream and pre-training datasets.

Effective for reducing the distribution gap

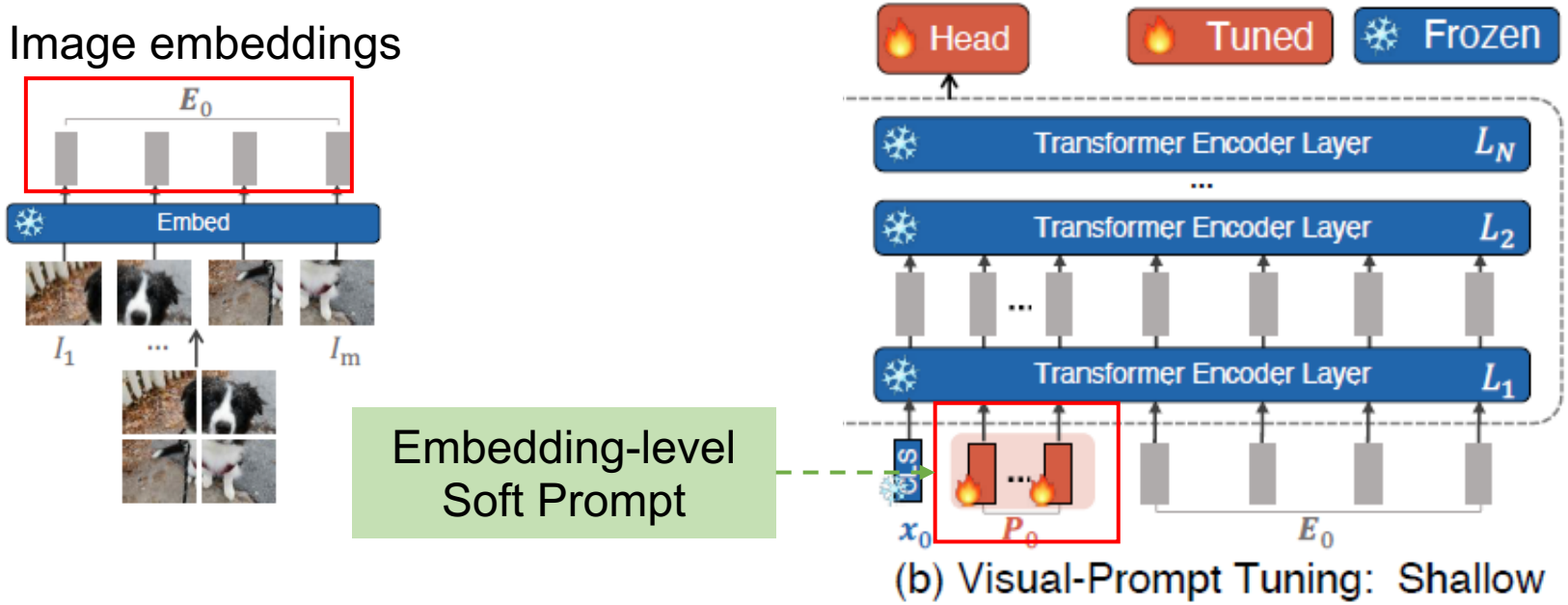


Struggles with difficult tasks (diverse datasets)
Some works study this problem…

Bahng, Hyojin et al. "Exploring Visual Prompts for Adapting Large-Scale Models." ArXiv abs/2203.17274 (2022).

Soft Prompt Tuning in CV

- Embedding-level learnable visual prompt: add prefix tokens to the image embeddings as prompt.



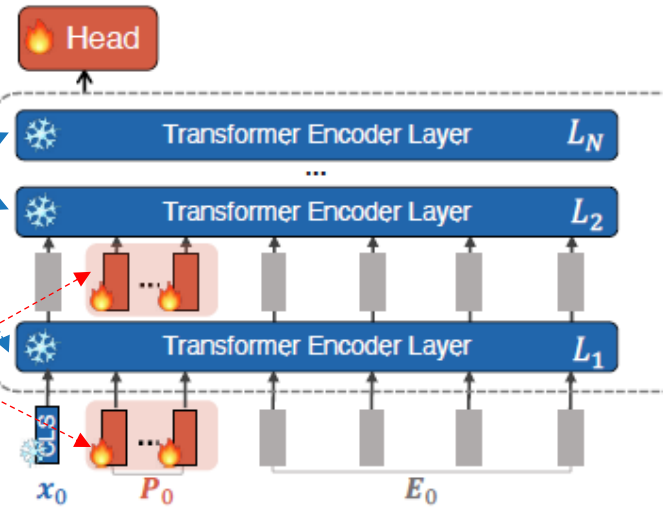
Jia, Menglin et al. "Visual Prompt Tuning." ECCV (2022).

Soft Prompt Tuning in CV

- Embedding-level learnable visual prompt: add prefix tokens to the image embeddings as prompt.
- Visual prompts can be added in multiple layers.

Visual model with many layers.

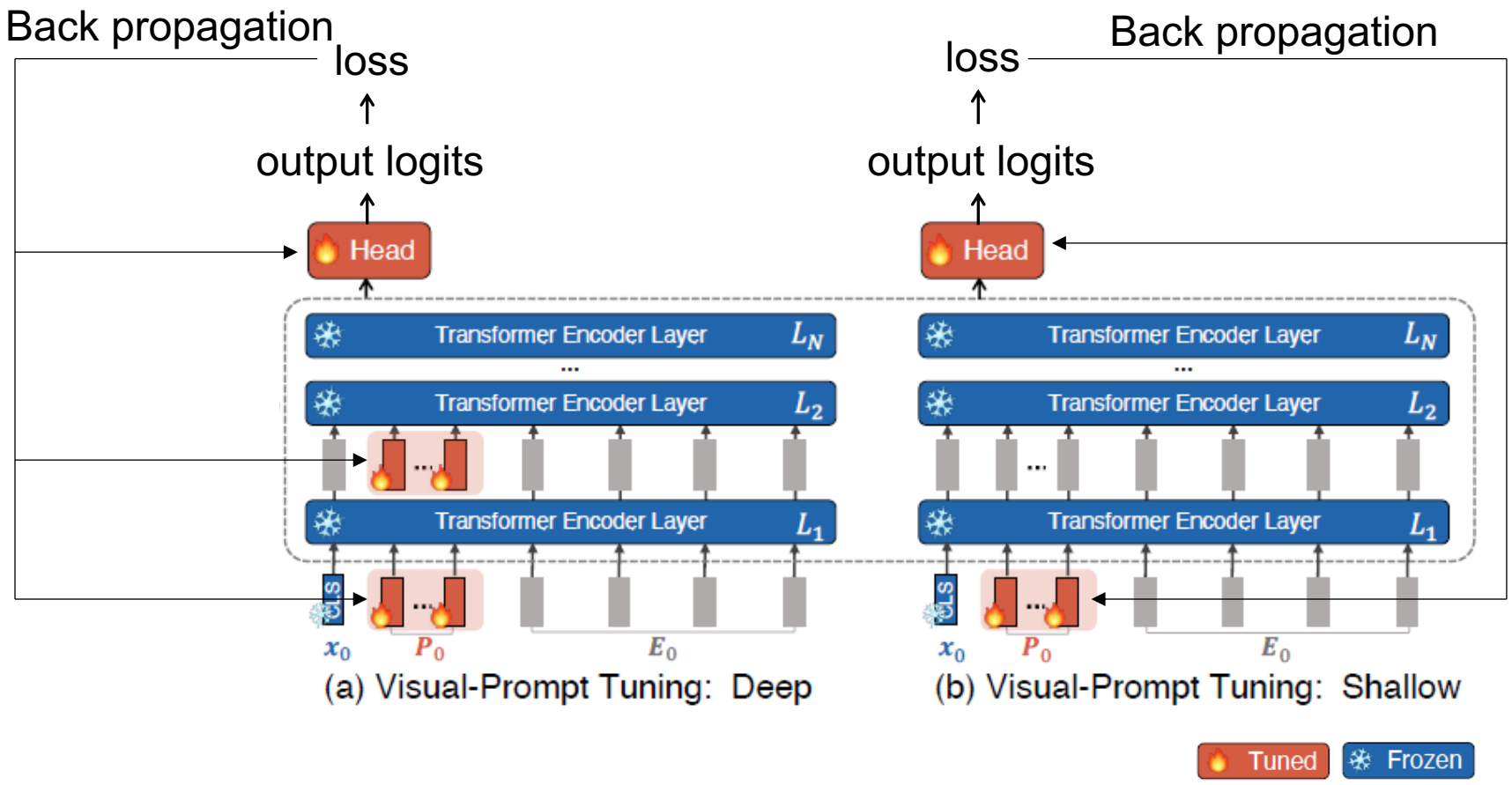
Learnable prompts are added to the inputs of multiple layers.



(a) Visual-Prompt Tuning: Deep

Jia, Menglin et al. "Visual Prompt Tuning." ECCV (2022).

Soft Prompt Tuning in CV



Jia, Menglin et al. "Visual Prompt Tuning." ECCV (2022).

Soft Prompt Tuning in CV

- Embedding-level prompt tuning performance.
- Visual Prompt Tuning (VPT) outperforms other fine-tuning methods.

	ViT-B/16 (85.8M)	Total params	Scope Input	Scope Backbone	Extra params	FGVC	VTAB-1k Natural	VTAB-1k Specialized	VTAB-1k Structured
	Total # of tasks					5	7	4	8
(a)	FULL	24.02×		✓		88.54	75.88	83.36	47.64
(b)	LINEAR	1.02×				79.32 (0)	68.93 (1)	77.16 (1)	26.84 (0)
	PARTIAL-1	3.00×				82.63 (0)	69.44 (2)	78.53 (0)	34.17 (0)
	MLP-3	1.35×			✓	79.80 (0)	67.80 (2)	72.83 (0)	30.62 (0)
(c)	SIDETUNE	3.69×		✓	✓	78.35 (0)	58.21 (0)	68.12 (0)	23.41 (0)
	BIAS	1.05×		✓		88.41 (3)	73.30 (3)	78.25 (0)	44.09 (2)
	ADAPTER	1.23×		✓	✓	85.66 (2)	70.39 (4)	77.11 (0)	33.43 (0)
(ours)	VPT-SHALLOW	1.04×	✓		✓	84.62 (1)	76.81 (4)	79.66 (0)	46.98 (4)
	VPT-DEEP	1.18×				**89.11 (4)**	**78.48 (6)**	**82.43 (2)**	**54.98 (8)**

Jia, Menglin et al. "Visual Prompt Tuning." ECCV (2022).

Outline

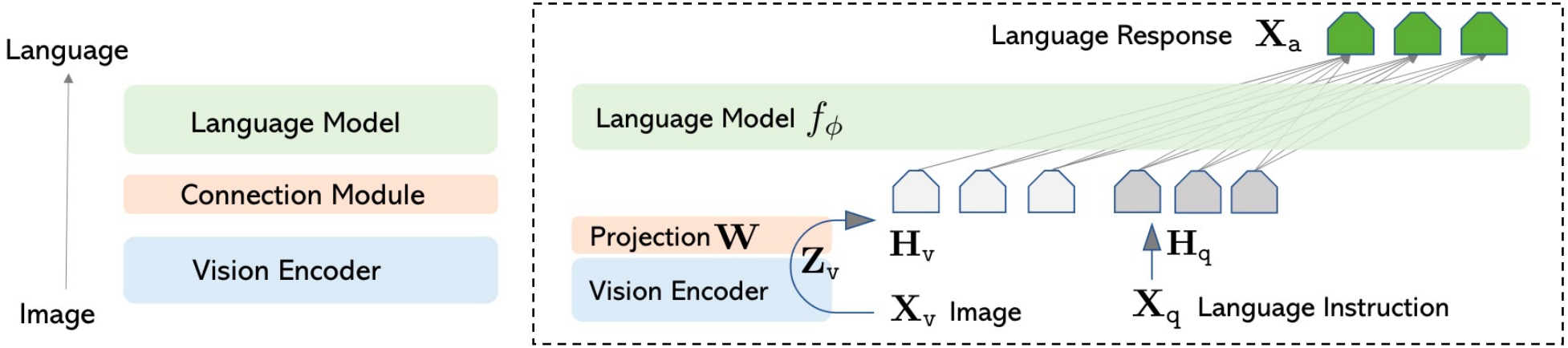
I. Motivation for Prompting
II. Development of Prompting
III. Prompting & Prompt Engineering in NLP
IV. Prompting & Prompt Engineering in CV
V. Prompting & Prompt Engineering on VLM
VI. Summary

Vision-Language Model (VLM)

[Pilot Work for image-to-text generation]

- Large multimodal model (LMM)*: combining large language model (LLM) with visual models allows the original LLM to perceive and understand visual content.

A simple example: LLaVA.

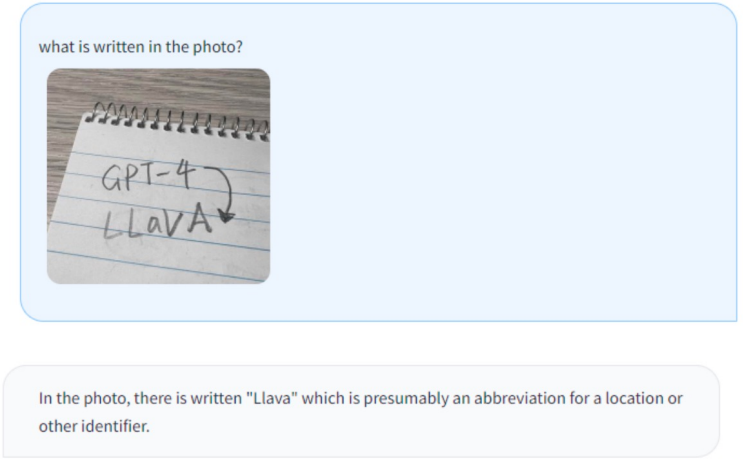


Liu, Haotian et al. "Visual Instruction Tuning." NeurIPS (2023).
**For more LMMs architecture designs, refer to https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models.*

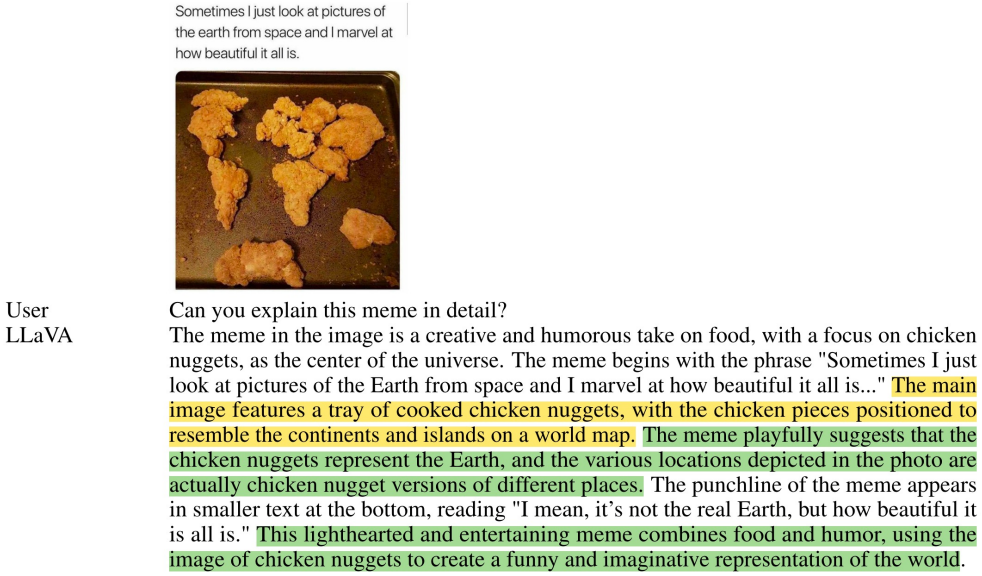
Prompting on LMM

[Zero-shot prompting]

- Prompting LMM can be analogous to prompting LLM, while the visual content can be included.



Visual recognition



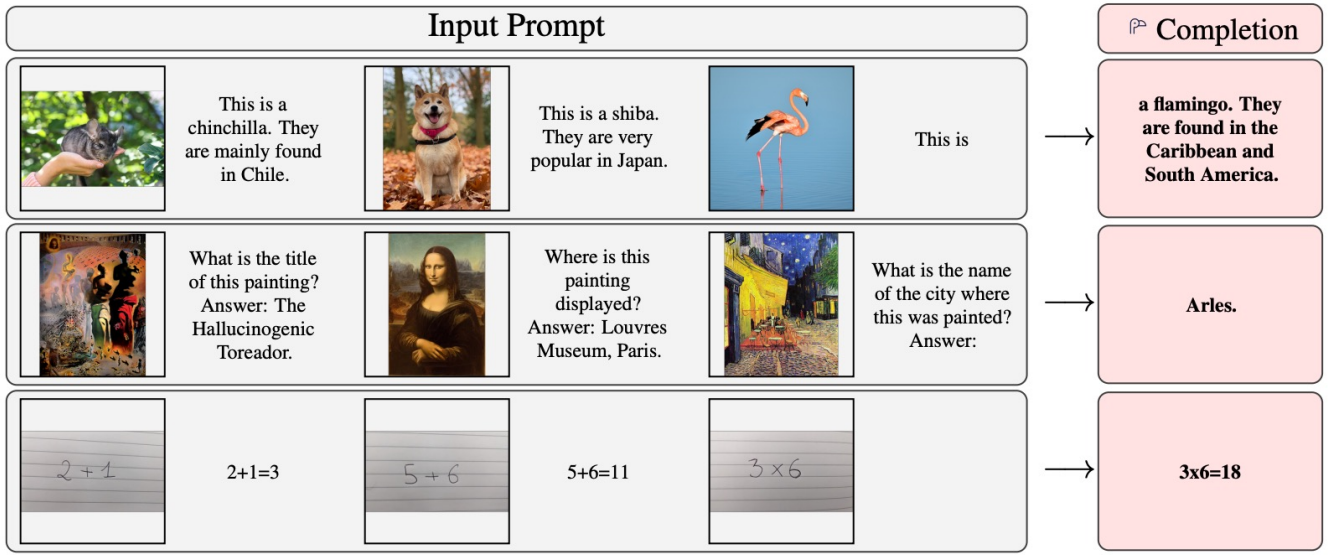
Visual understanding and reasoning

Liu, Haotian et al. "Visual Instruction Tuning." NeurIPS (2023).

Prompting on LMM

[Few-shot prompting]

- In-context learning of LMM: the prompt is provided with input-output pair examples of downstream task as demonstration.



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." NeurIPS(2022).

Vision-Language Model (VLM)

[Pilot Work for image-text matching]

Contrastive **L**anguage-**I**mage **P**re-training (CLIP):

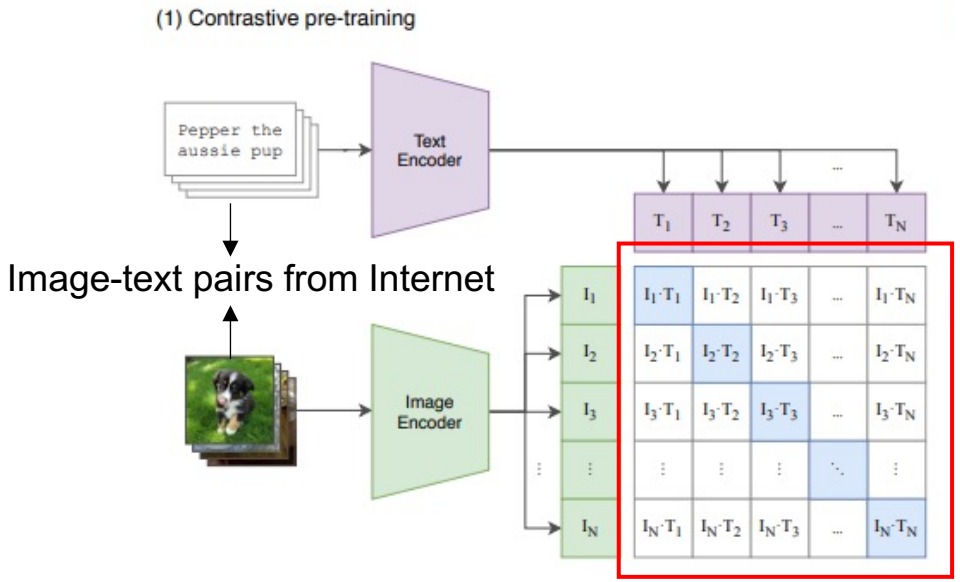
- For text-image matching.
- Incorporate both text and image as inputs.
- Utilize prompt to adapt the pre-trained model to downstream task.

Radford, Alec et al. "Learning Transferable Visual Models From Natural Language Supervision." International Conference on Machine Learning (2021).

Vision-Language Model (VLM)

[CLIP pre-training]

- CLIP is pre-trained on 400M text-image pairs from the Internet (unlabeled data without manual annotation) in a contrastive learning manner.



Contrastive pre-training:

- Calculate similarity matrix between text and image embeddings.
- Train the text and image encoders to maximize the similarity between the paired text and image embeddings (diagonal elements).

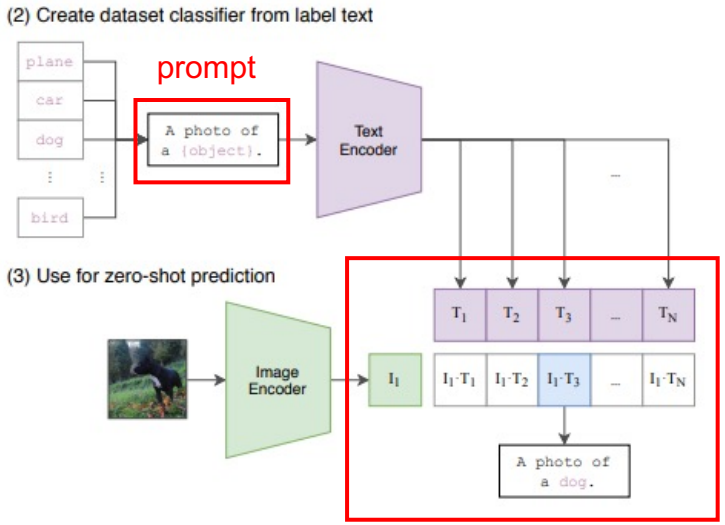
Radford, Alec et al. "Learning Transferable Visual Models From Natural Language Supervision." International Conference on Machine Learning (2021).

Prompting on VLM

[CLIP inference for downstream task]

- CLIP is able to make inference for downstream tasks(i.e., new dataset to be classified) in **zero-shot** manner, without additional training on the pre-trained model.

The downstream labels are combined with a prompt as text inputs.



The label contained in the text with the highest similarity is the final prediction.

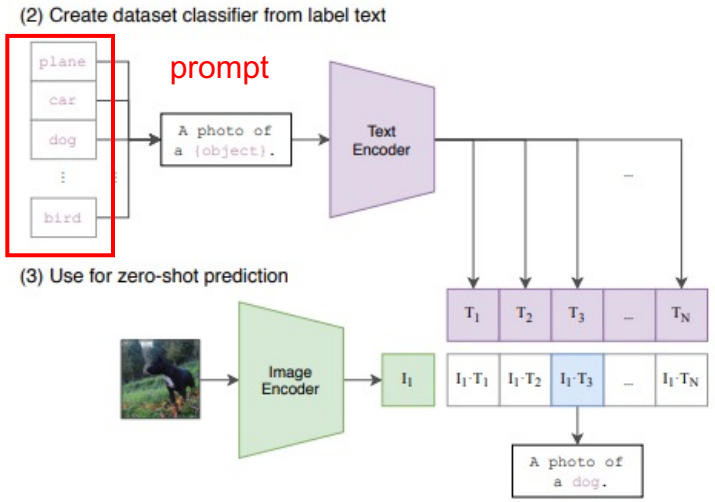
Radford, Alec et al. "Learning Transferable Visual Models From Natural Language Supervision." International Conference on Machine Learning (2021).

Prompting on VLM

[CLIP inference for downstream task]

- Thanks to the flexibility of contrastive learning, CLIP is able to handle **open-set** problem.

The downstream labels can be arbitrary.



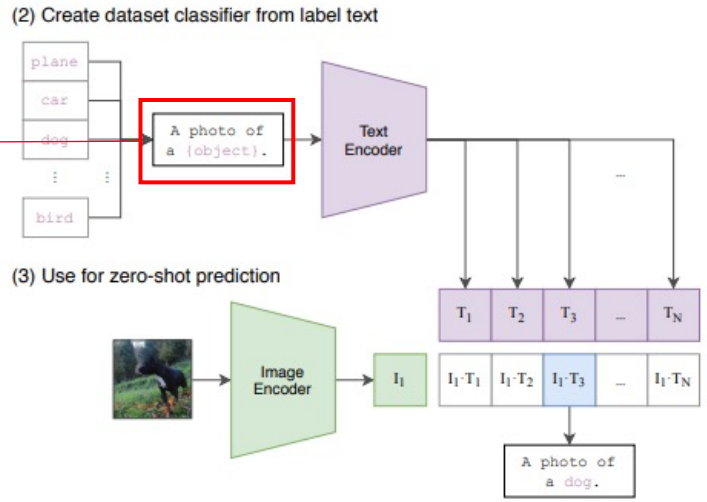
Radford, Alec et al. "Learning Transferable Visual Models From Natural Language Supervision." International Conference on Machine Learning (2021).

Prompting on VLM

[CLIP limitation]

- The prompt used to form new text inputs is manually designed and fixed for all downstream tasks.
- This kind of prompt does not contain task-specific information.

Manually designed and fixed!



Radford, Alec et al. "Learning Transferable Visual Models From Natural Language Supervision." International Conference on Machine Learning (2021).

Prompting on VLM

[CLIP limitation]

- While adapting CLIP to the downstream task without additional training, the performance is sensitive to the used prompt.

Flowers102	Prompt	Accuracy
	a photo of a [CLASS].	60.86
	a flower photo of a [CLASS].	65.81
	a photo of a [CLASS], a type of flower.	66.14
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

Adapt CLIP to classify flowers

EuroSAT	Prompt	Accuracy
	a photo of a [CLASS].	24.17
	a satellite photo of [CLASS].	37.46
	a centered satellite photo of [CLASS].	37.56
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

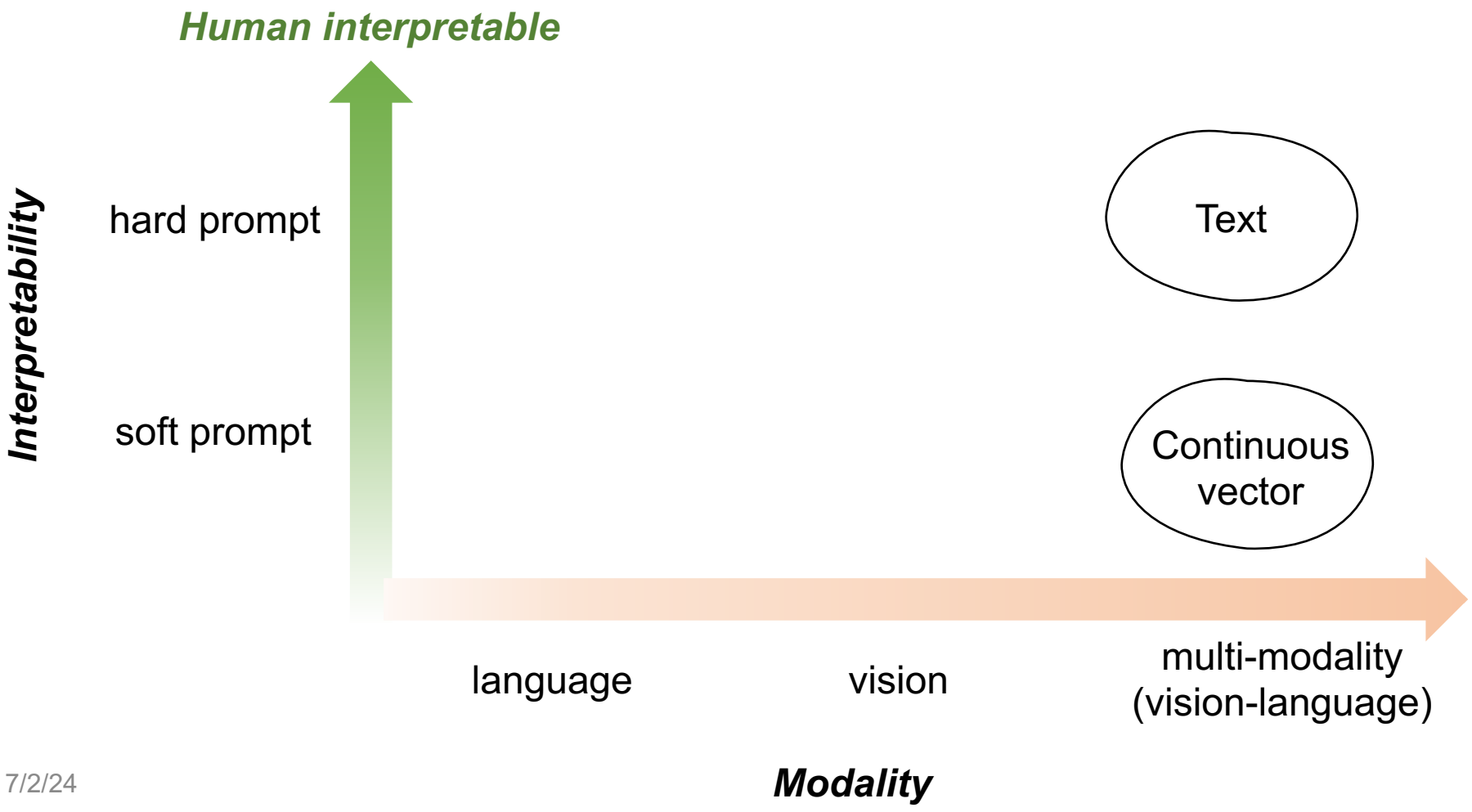
Adapt CLIP to classify satellite photos

Zhou, Kaiyang et al. "Learning to Prompt for Vision-Language Models." International Journal of Computer Vision 130 (2021).

Prompt engineering aids to adapt the pre-trained CLIP to various downstream tasks

to get better performance.

Prompt Engineering on VLM



Hard Prompt Engineering on VLM

Hard prompt engineering:

- is non-trivial and may require great manual effort.
- constraints prompt to be human interpretable text in a discrete space, impeding to find optimal prompt.

a bad photo of a {}.
a photo of many {}.
a sculpture of a {}.
a photo of the hard to see {}.
a low resolution photo of the {}.
a rendering of a {}.
graffiti of a {}.
a bad photo of the {}.
a cropped photo of the {}.
a tattoo of a {}.
the embroidered {}.
a photo of a hard to see {}.
a bright photo of a {}.
a photo of a clean {}.
a photo of a dirty {}.
a dark photo of the {}.
a drawing of a {}.
a photo of my {}.
the plastic {}.
a photo of the cool {}.
a close-up photo of a {}.
a black and white photo of the {}.
a painting of the {}.
a painting of a {}.

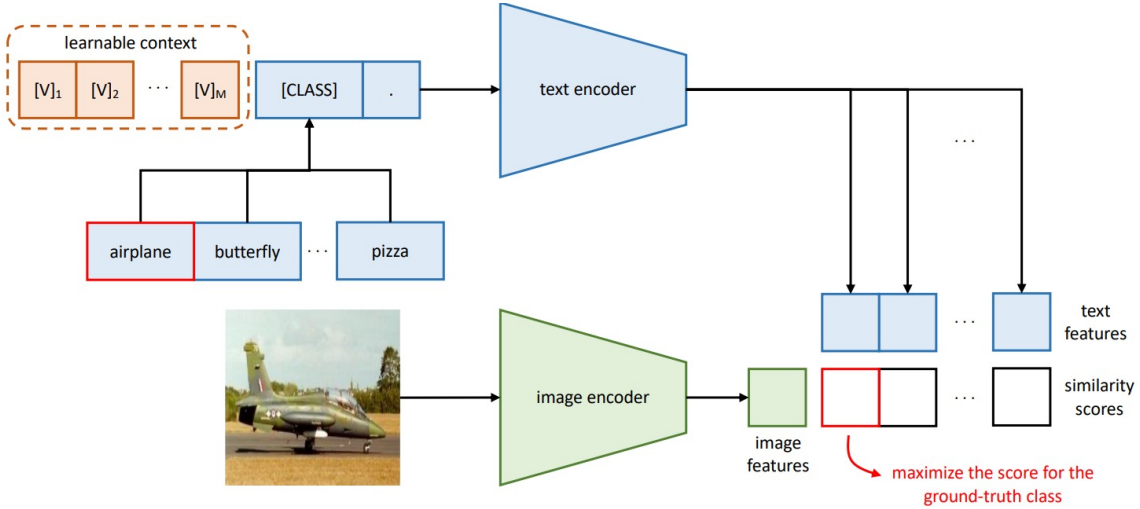
a pixelated photo of the {}.
a sculpture of the {}.
a bright photo of the {}.
a cropped photo of a {}.
a plastic {}.
a photo of the dirty {}.
a jpeg corrupted photo of a {}.
a blurry photo of the {}.
a photo of the {}.
a good photo of the {}.
a rendering of the {}.
a {} in a video game.
a photo of one {}.
a doodle of a {}.
a close-up photo of the {}.
a photo of a {}.
the origami {}.
the {} in a video game.
a sketch of a {}.
a doodle of the {}.
a origami {}.
a low resolution photo of a {}.
the toy {}.
a rendition of the {}.

a photo of the clean {}.
a photo of a large {}.
a rendition of a {}.
a photo of a nice {}.
a photo of a weird {}.
a blurry photo of a {}.
a cartoon {}.
art of a {}.
a sketch of the {}.
a embroidered {}.
a pixelated photo of a {}.
itap of the {}.
a jpeg corrupted photo of the {}.
a good photo of a {}.
a plushie {}.
a photo of the nice {}.
a photo of the small {}.
a photo of the weird {}.
the cartoon {}.
art of the {}.
a drawing of the {}.
a photo of the large {}.
a black and white photo of a {}.
the plushie {}.

Soft Prompt Engineering on VLM

Soft prompt tuning (Context Optimization (CoOp), Zhou et al.)

- **Continuous relaxation**: cast the prompt combined with text labels as continuous embedding vectors (soft prompt) instead of natural language.

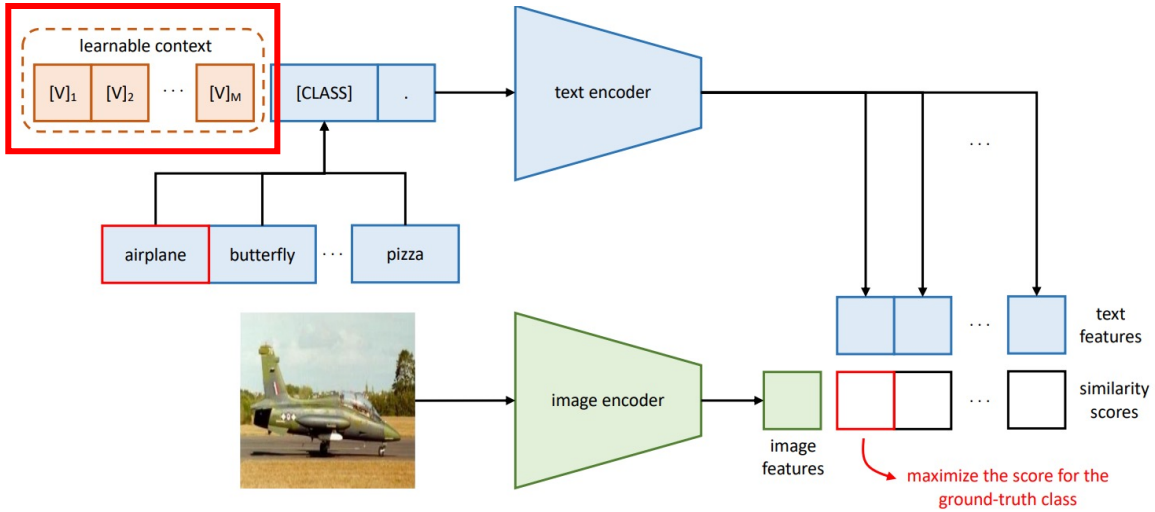


Zhou, Kaiyang et al. "Learning to Prompt for Vision-Language Models." International Journal of Computer Vision 130 (2021).

Soft Prompt Engineering on VLM

Soft prompt tuning (Context Optimization (CoOp), Zhou et al.)

- **Few-shot learning**: the soft prompt is learnable for each downstream task, requiring a few labeled data as supervision to optimize the prompt. This is different from the CLIP's zero-shot inference.

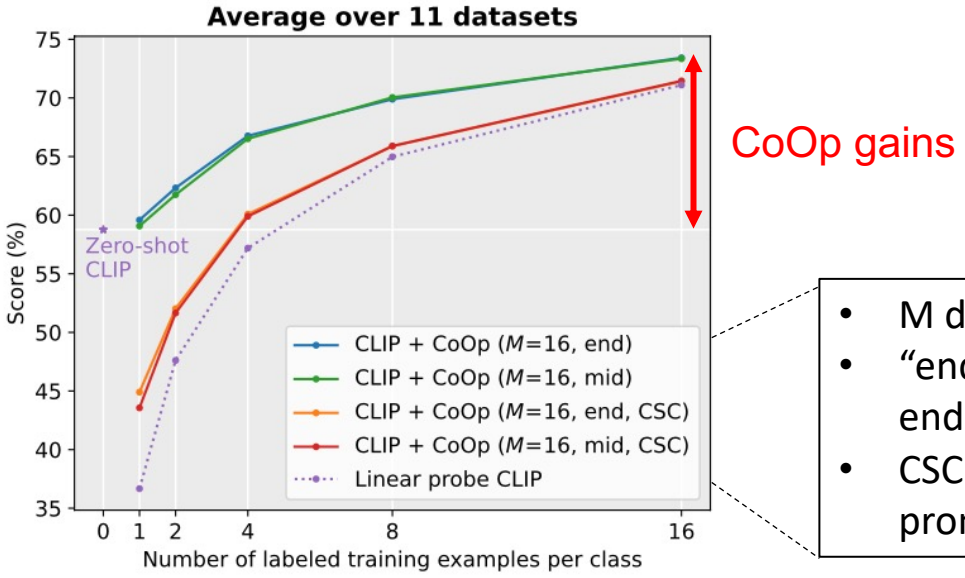


Zhou, Kaiyang et al. "Learning to Prompt for Vision-Language Models." International Journal of Computer Vision 130 (2021).

Soft Prompt Engineering on VLM

CoOp Performance

- Overall, CoOp in few-shot setting outperforms zero-shot CLIP, utilizing a few downstream labeled data to tune the prompt.



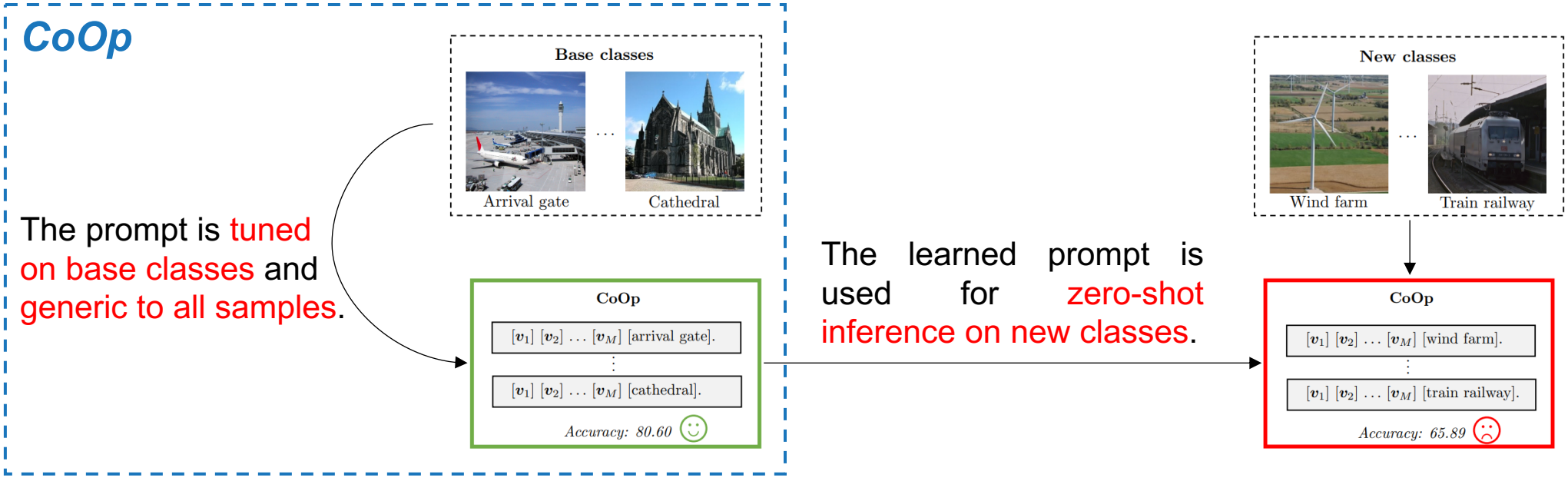
- M denotes the prompt length.
- "end" or "mid" means putting the class token in the end or middle of the prompt.
- CSC means class-specific prompt, i.e., distinct prompt for each class in the downstream task.

Zhou, Kaiyang et al. "Learning to Prompt for Vision-Language Models." International Journal of Computer Vision 130 (2021).

Soft Prompt Engineering on VLM

CoOp limitation

- The learned prompt via CoOp is not generalizable to wilder new classes, which are unseen during prompt tuning process.



Zhou, Kaiyang et al. "Conditional Prompt Learning for Vision-Language Models." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022).

Soft Prompt Engineering on VLM

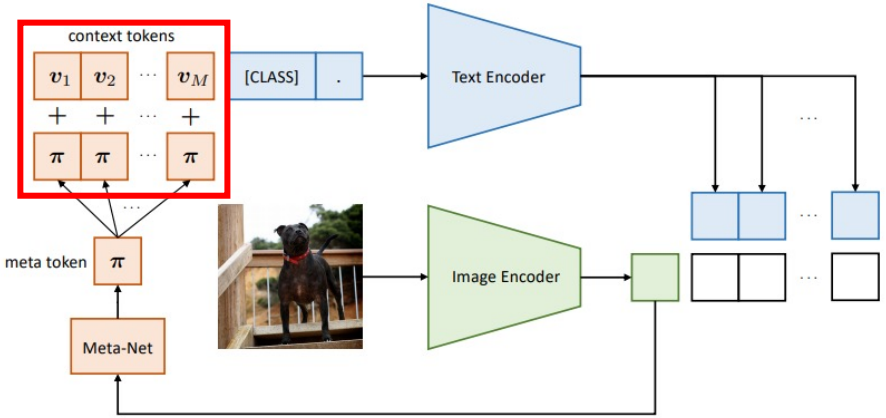
Generalizable soft prompt tuning

(Conditional Context Optimization (CoCoOp), Zhou et al.)

- Turn the prompt from generic to instance-dependent.

prompt = **generic** context tokens + **instance-dependent** meta token π

Instance-dependent prompt in CoCoOp consists of **two parts**: a generic part for all instances and an instance-dependent part determined by the input image.



Zhou, Kaiyang et al. "Conditional Prompt Learning for Vision-Language Models." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022).

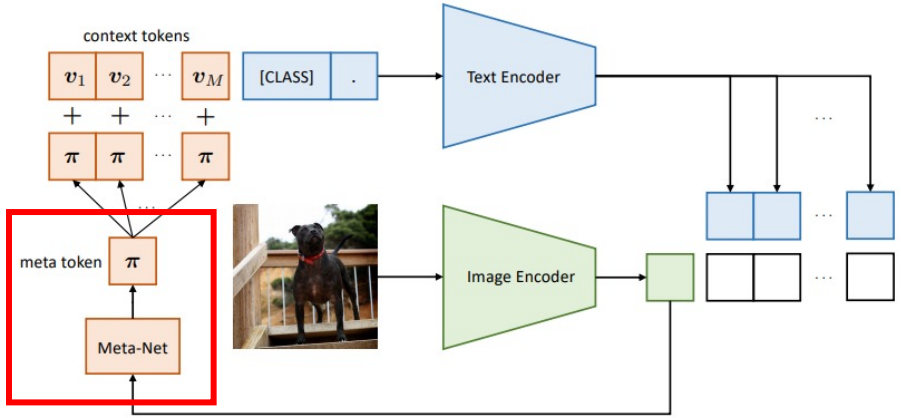
Soft Prompt Engineering on VLM

Generalizable soft prompt tuning

(Conditional Context Optimization (CoCoOp), Zhou et al.)

- Meta token π is predicted by a meta-net, which incorporates the image embedding as input and is learned.

Meta-net is shared for all image instances, it is the input image that makes the meta token instance-dependent.



Zhou, Kaiyang et al. "Conditional Prompt Learning for Vision-Language Models." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022).

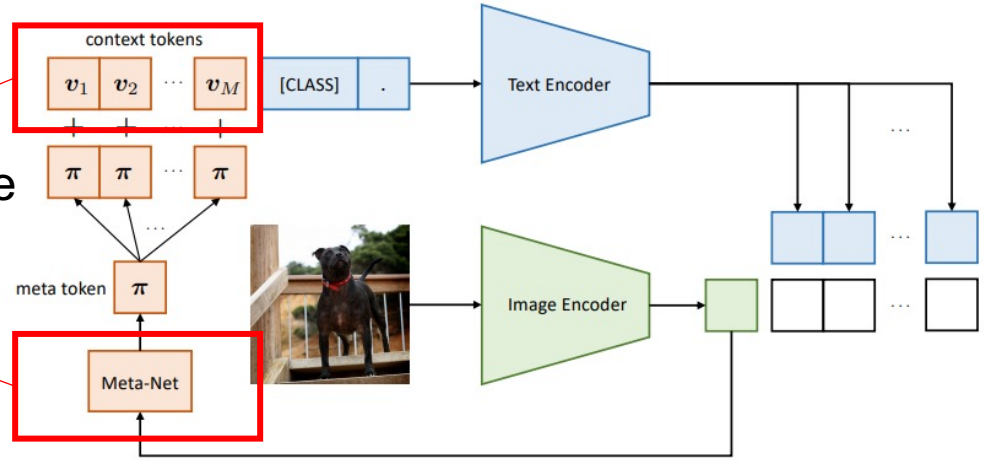
Soft Prompt Engineering on VLM

Generalizable soft prompt tuning

(Conditional Context Optimization (CoCoOp), Zhou et al.)

- During training on base classes, the generic context tokens and the meta-net are learned.

The generic part of prompt and the meta-net are learned.



Zhou, Kaiyang et al. "Conditional Prompt Learning for Vision-Language Models." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022).

Soft Prompt Engineering on VLM

CoCoOp performance

- CoCoOp strikes the best balance between performance on base classes and generalization ability on new classes compared to CLIP and CoOp.

(a) **Average over 11 datasets.**

	Base	New	H
CLIP	69.34	**74.22**	71.70
CoOp	**82.69**	63.22	71.66
CoCoOp	80.47	71.69	**75.83**

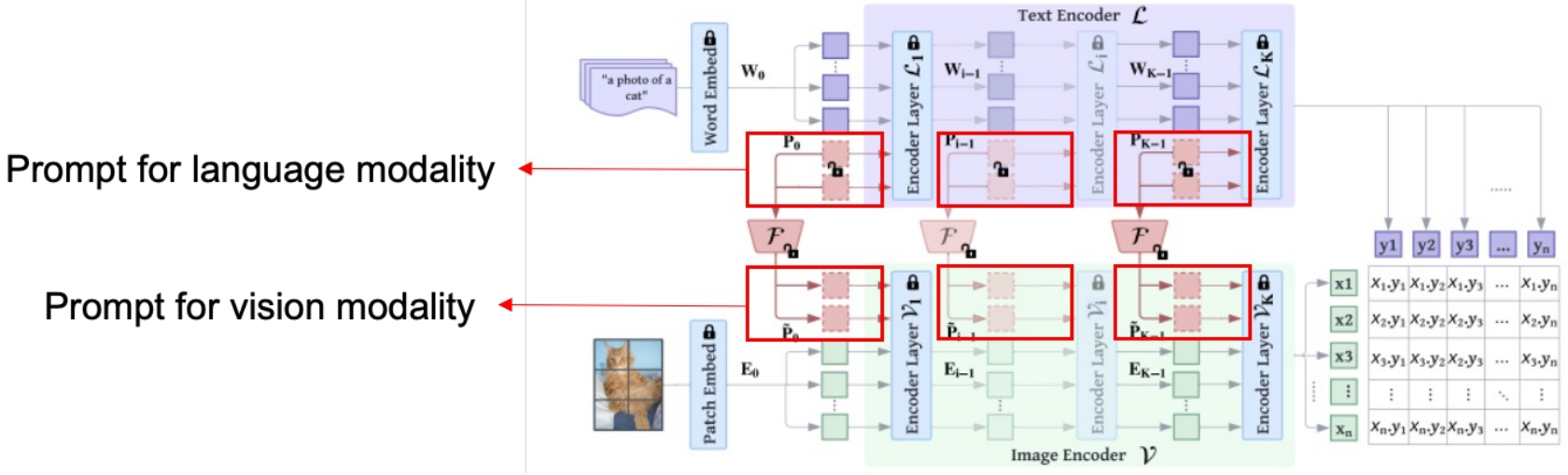
- CLIP performs zero-shot inference on both base and new classes.
- CoOp and CoCoOp are trained on based classes under 16-shot setting, and perform zero-shot inference on new classes.
- H means harmonic mean to highlight the generalization trade-off.

Zhou, Kaiyang et al. "Conditional Prompt Learning for Vision-Language Models." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022).

Soft Prompt Engineering on VLM

Multi-modal soft prompt tuning (MaPLe, Khattak et al.)

- Learning deep embedding-level **prompts for both language and vision modalities** to adapt the pre-trained CLIP for downstream tasks more completely.



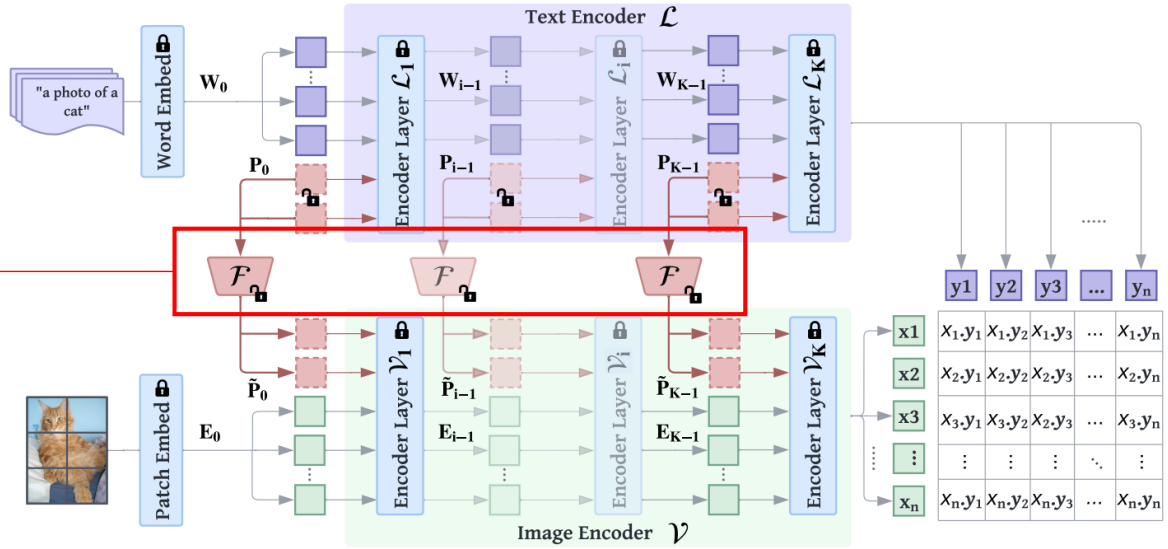
Khattak, Muhammad Uzair, et al. "Maple: Multi-modal prompt learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023).

Soft Prompt Engineering on VLM

Multi-modal soft prompt tuning (MaPLe, Khattak et al.)

- **Coupling the prompts** for both modalities encourages cross-modal synergy.

Projection function to couple the prompts for both modalities.



Khattak, Muhammad Uzair, et al. "Maple: Multi-modal prompt learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023).

Soft Prompt Engineering on VLM

Multi-modal soft prompt tuning (MaPLe, Khattak et al.)

- MaPLe shows improved performance on both base and new unseen classes.

(a) **Average over 11 datasets**

	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	**82.69**	63.22	71.66
Co-CoOp	80.47	71.69	75.83
MaPLe	82.28	**75.14**	**78.55**
MaPLe compared to CoCoOp ←	+1.81	+3.45	+2.72

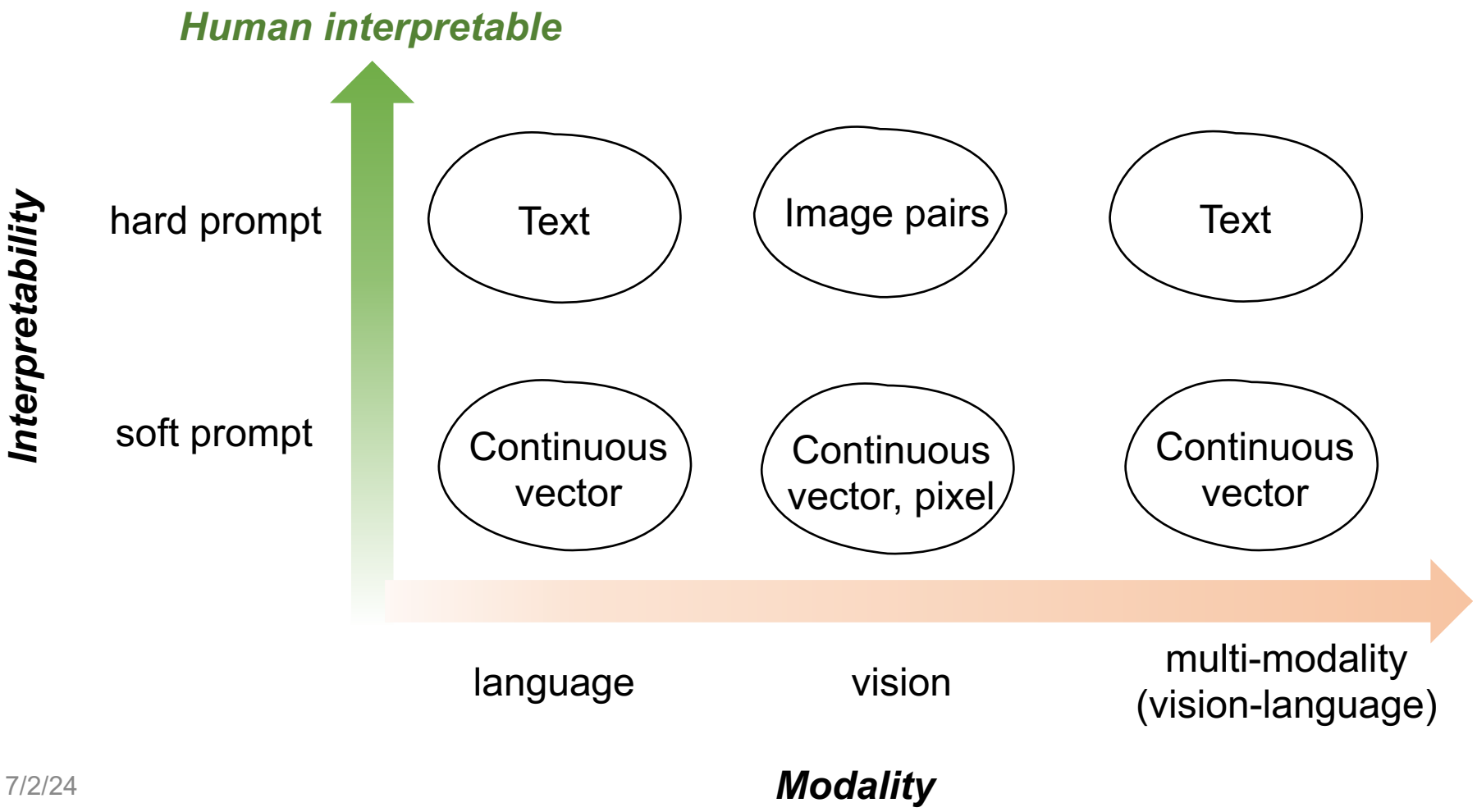
- CLIP performs zero-shot inference on both base and new classes.
- CoOp, CoCoOp and MaPLe are trained on based classes under 16-shot setting, and perform zero-shot inference on new classes.
- HM means harmonic mean to highlight the generalization trade-off.

Khattak, Muhammad Uzair, et al. "Maple: Multi-modal prompt learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023).

Outline

I. Motivation for Prompting
II. Development of Prompting
III. Prompting & Prompt Engineering in NLP
IV. Prompting & Prompt Engineering in CV
V. Prompting & Prompt Engineering on VLM
VI. Summary

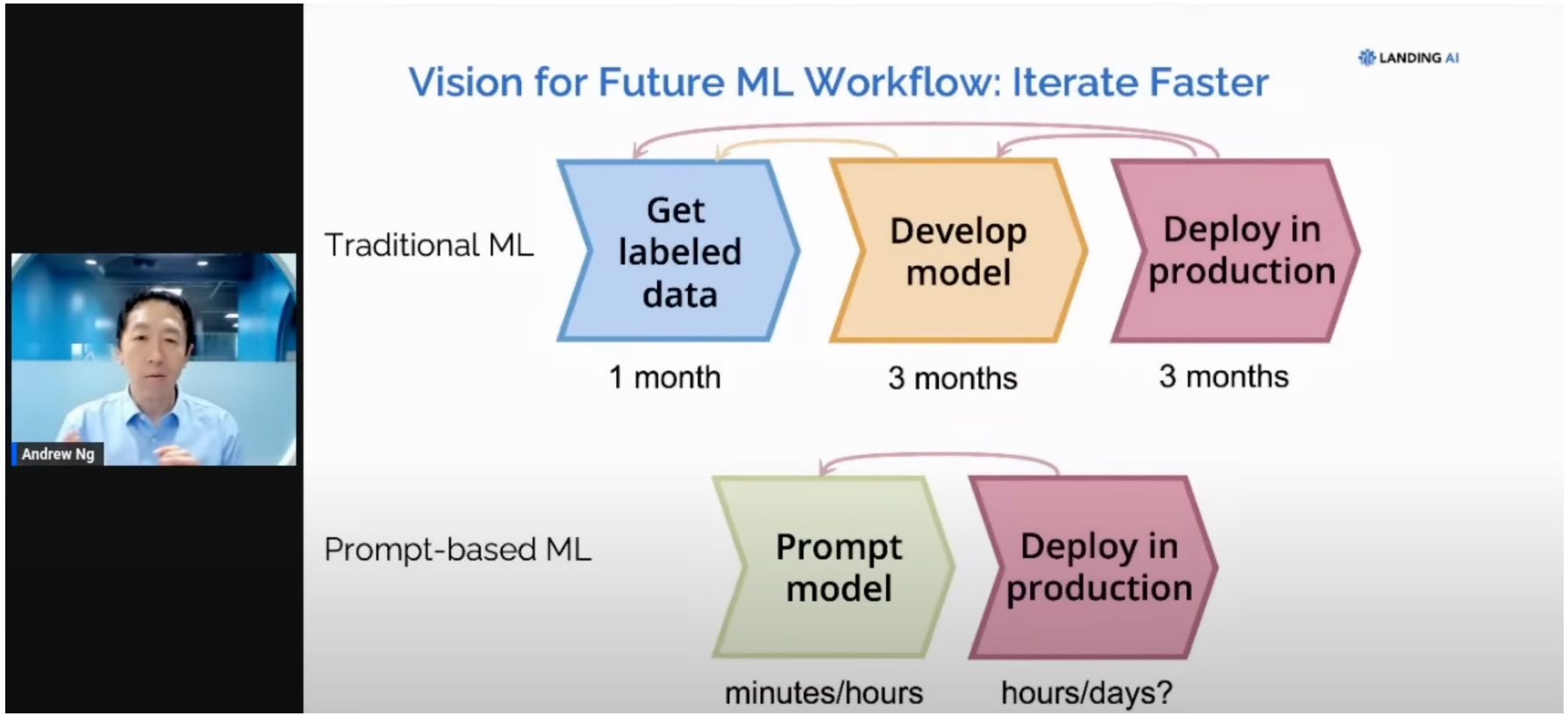
Prompt Engineering in Different Cases



Summary

	Pre-train & Fine-tune Paradigm	Pre-train & Prompt Paradigm
Saving cost	High	Low
Computational cost	High	Low
Downstream training	Yes	No – directly prompting Yes – prompt tuning
Data demand	High	Low
Accessible for everyone?	No	Yes
Towards general foundation model?	No	Yes

Traditional ML vs Prompt-Based ML



Landing AI, "Visual Prompting Livestream With Andrew Ng". https://www.youtube.com/watch?v=FE88OOUBonQ&t=577s

All of today’s class