Prompting in Visual Intelligence

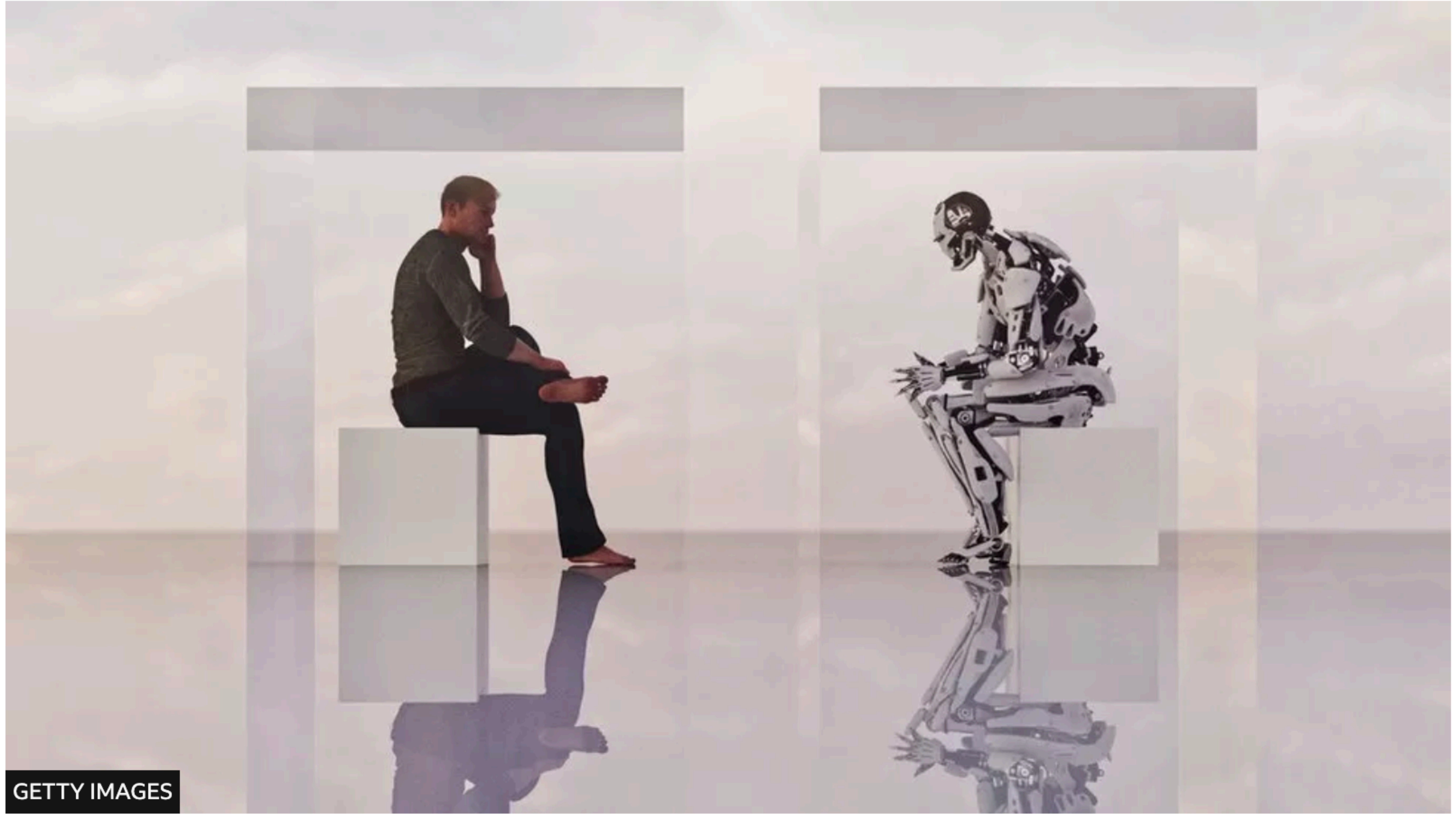
Kaiyang Zhou

@kaiyangzhou

Image: unsplash.com/@jean_vella

ChatGPT: New AI chatbot has everyone talking to it

7 December 2022



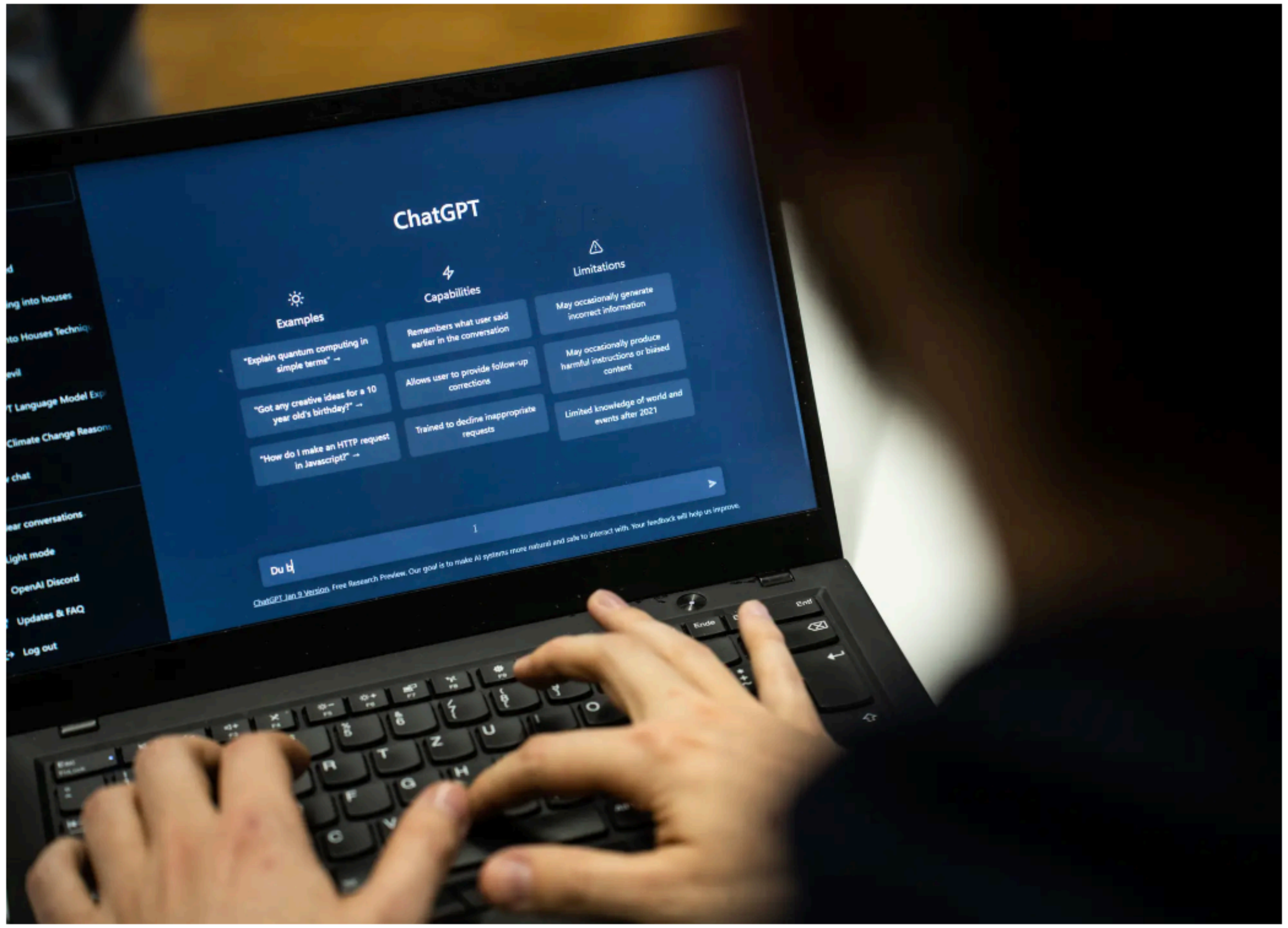
https://www.bbc.com/news/technology-63861322

Photographer: Gabby Jones/Bloomberg

AI 'prompt engineer' jobs can pay up to $375,000 a year and don't always require a background in tech

Britney Nguyen May 1, 2023, 11:34 PM GMT+8

Read in app



The rise of generative AI tools like ChatGPT is creating a hot market for "prompt engineers" who test and improve chatbot answers. Getty Images

USD 375,000
JPY 54,094,500
CNY 2,718,938

https://businessinsider.com

GPT-4

Illustration: Ruby Chen

We've created GPT-4, the latest milestone in OpenAI's effort in scaling up deep learning. GPT-4 is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks.

User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

https://openai.com/research/gpt-4

OpenAI @OpenAI

"A photo of an astronaut riding a horse" #dalle

Stable Diffusion @ai_diffusion

On stage 🎤

#CoolPope #PopeFrancis @Pontifex
#stablledifussion #detailedprompt #prompt #stablediffusionart
#digitalartwork #aigenerated

detailed prompt in image description

ARTIFICIAL INTELLIGENCE

Meta’s new AI can turn text prompts into videos

It’s a breakthrough in generative AI that raises some tough ethical questions.

By Melissa Heikkilä

September 29, 2022

META AI

"A teddy bear painting a portrait"

“A young couple walking in heavy rain”

Singer et al. Make-A-Video: Text-to-Video Generation without Text-Video Data. 2022.

Outline

- Prompting in natural language processing
  - Language model, hard prompt, and soft prompt
- Prompting in computer vision
  - White-box prompt learning
  - Black-box prompt learning

Outline

- Prompting in natural language processing
  - Language model, hard prompt, and soft prompt
- Prompting in computer vision
  - White-box prompt learning
  - Black-box prompt learning

Language model

Autoregressive training:
Predict the next word (token) based on previous words, e.g., GPT

“AI is the new” → Language model → “electricity”

A probability distribution over sequences of words

Training data is huge (gigabytes of text) and contains diverse sources such as Wikipedia, news articles, books, and so on

Language model

Prompting is used to elicit knowledge from pre-trained language models

"Dante was born in ___" → Language model → "Florence"

Fill-in-the-blank cloze test

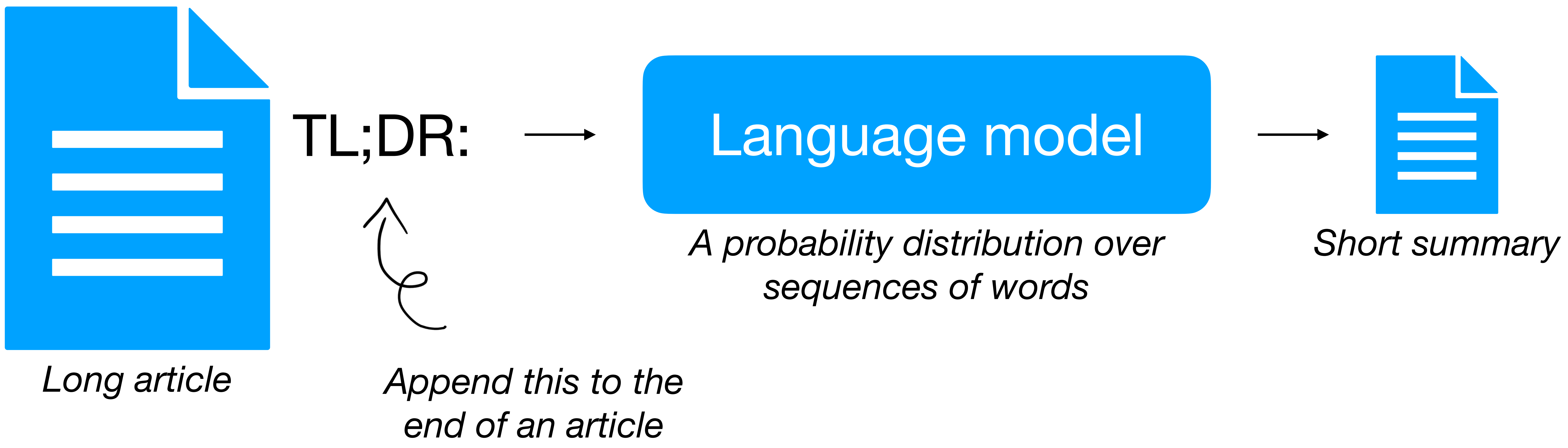
A probability distribution over sequences of words

Idea: Convert input into a language modeling format

Petroni et al. Language models as knowledge bases? 2019.

Language model

Prompting is used to elicit knowledge from pre-trained language models



Idea: Convert input into a language modeling format

Radford et al. Language Models are Unsupervised Multitask Learners. 2019.

Language model

Prompting is used to elicit knowledge from pre-trained language models

"Translate English into French:
sea otter => loutre de mer
…
cheese =>" → Language model → "fromage"

Task description + examples (in-context learning)

A probability distribution over sequences of words

Idea: Convert input into a language modeling format

Brown et al. Language Models are Few-Shot Learners. 2020

Prompting does not introduce large amounts of learnable parameters and can handle open-set queries

… but manually crafting a good prompt is non-trivial
(a bad prompt might fail to retrieve the correct knowledge)

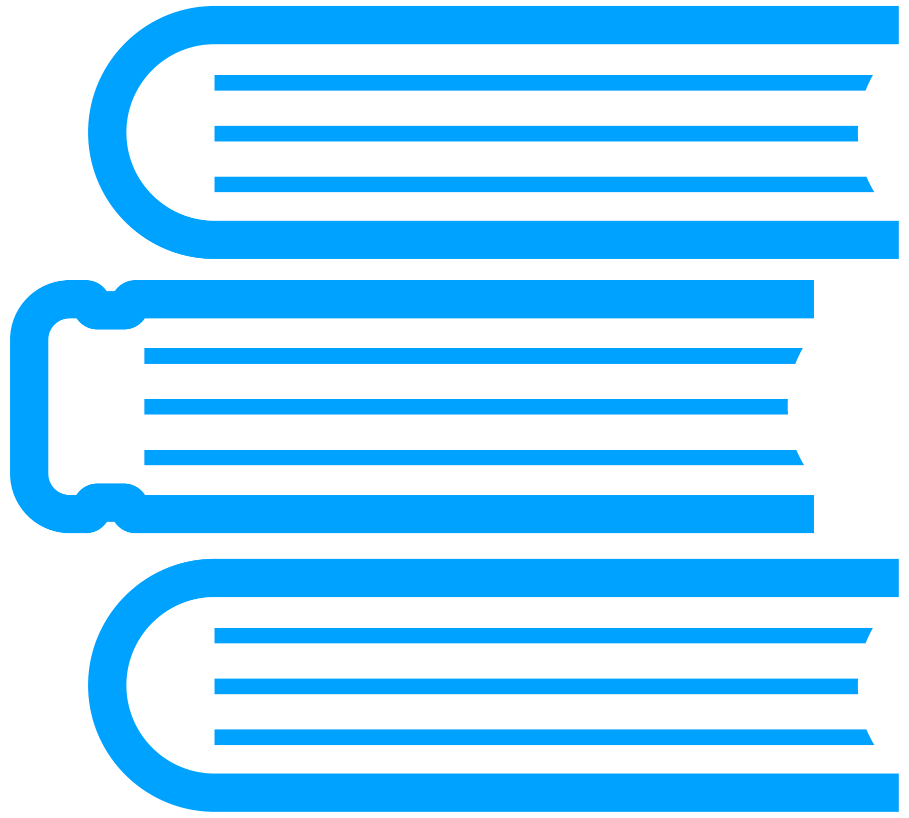
… so some kind of adaptation is needed for downstream tasks

Hard prompt

- Mining-based prompt generation

Relation: subclass of
Manual: x <u>is a subclass of</u> y
Mined: x <u>is a type of</u> 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 <u>is a type of</u> y (92.7%)
x <u>belongs to</u> y (80.2%)
…
x <u>is a subclass of</u> y (70.0%)

Jiang et al. How Can We Know What Language Models Know? 2020.

Hard prompt

- Paraphrasing-based prompt generation

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

A seed prompt (manual or mined)

Paraphrasing →

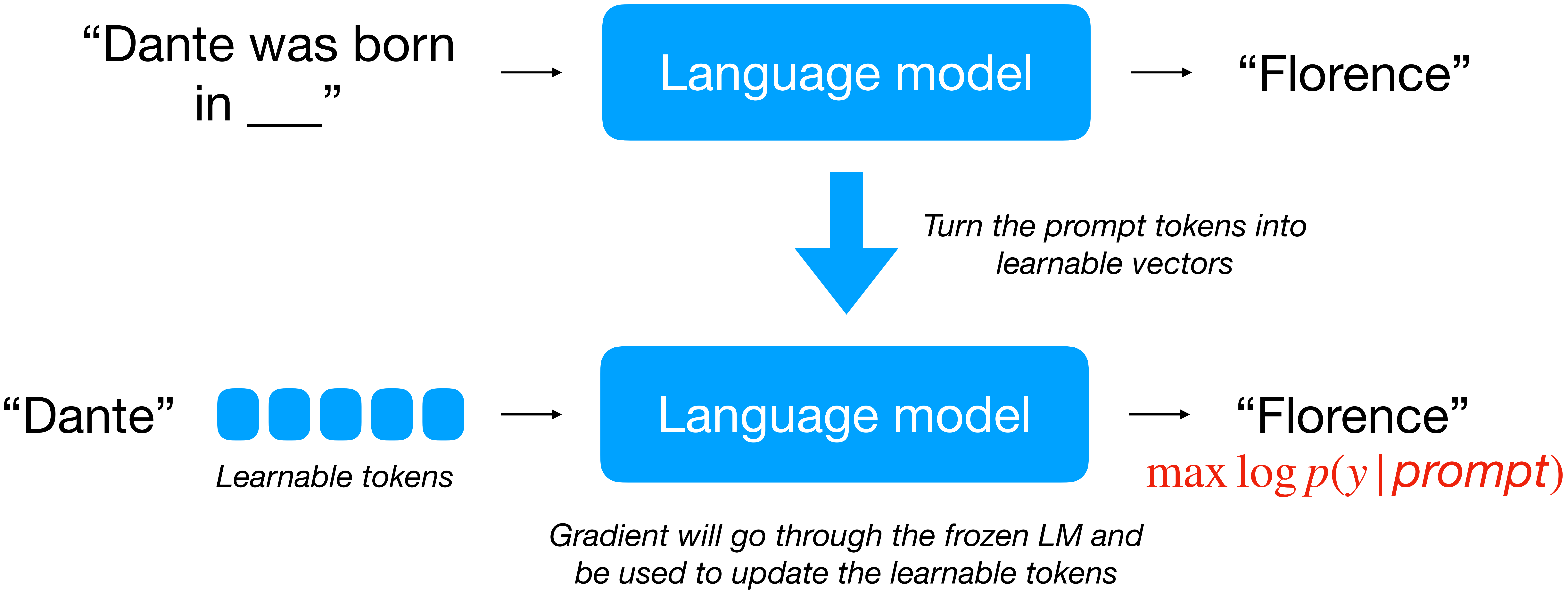
- *Back translation (Jiang et al., 2020)*
- *Neural prompt rewriter (Haviv et al., 2021)*
- *etc.*

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

Jiang et al. How Can We Know What Language Models Know? 2020.

Restricting the search space to existing vocabulary tokens is suboptimal

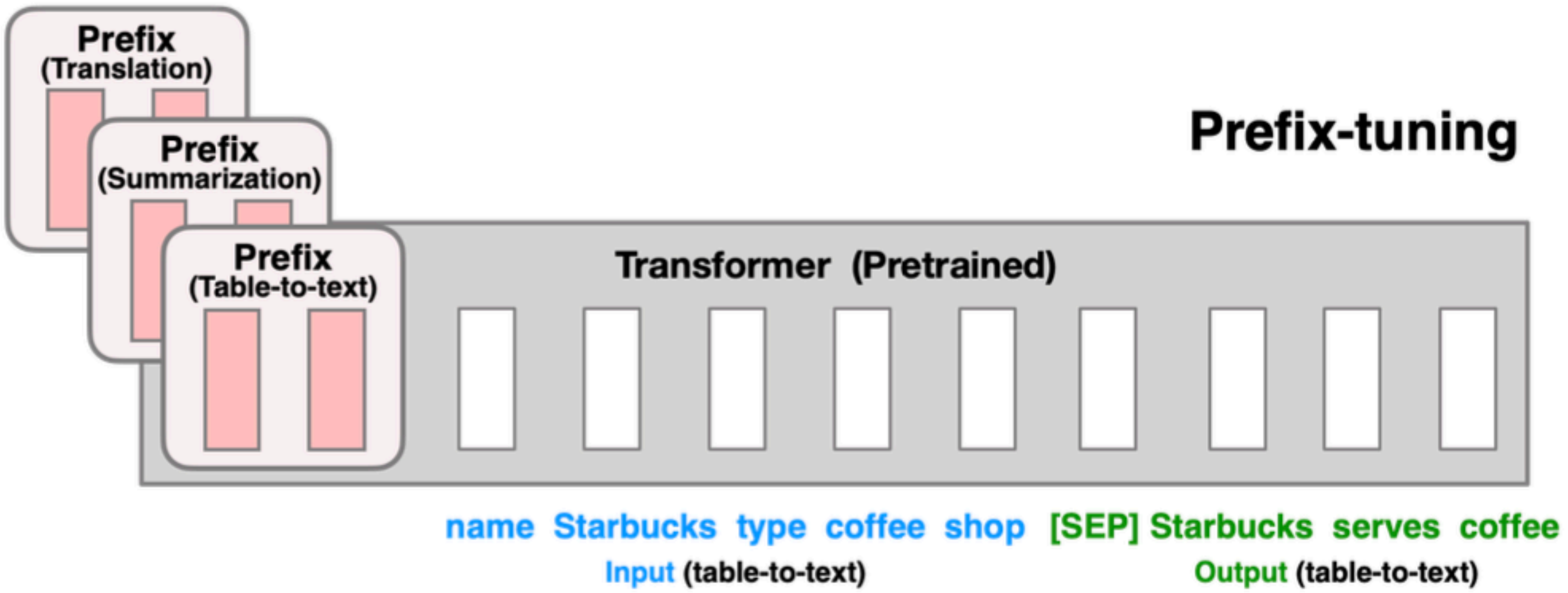
Soft prompt



Zhong et al. Factual Probing Is [MASK]: Learning vs. Learning to Recall. 2021.

Soft prompt

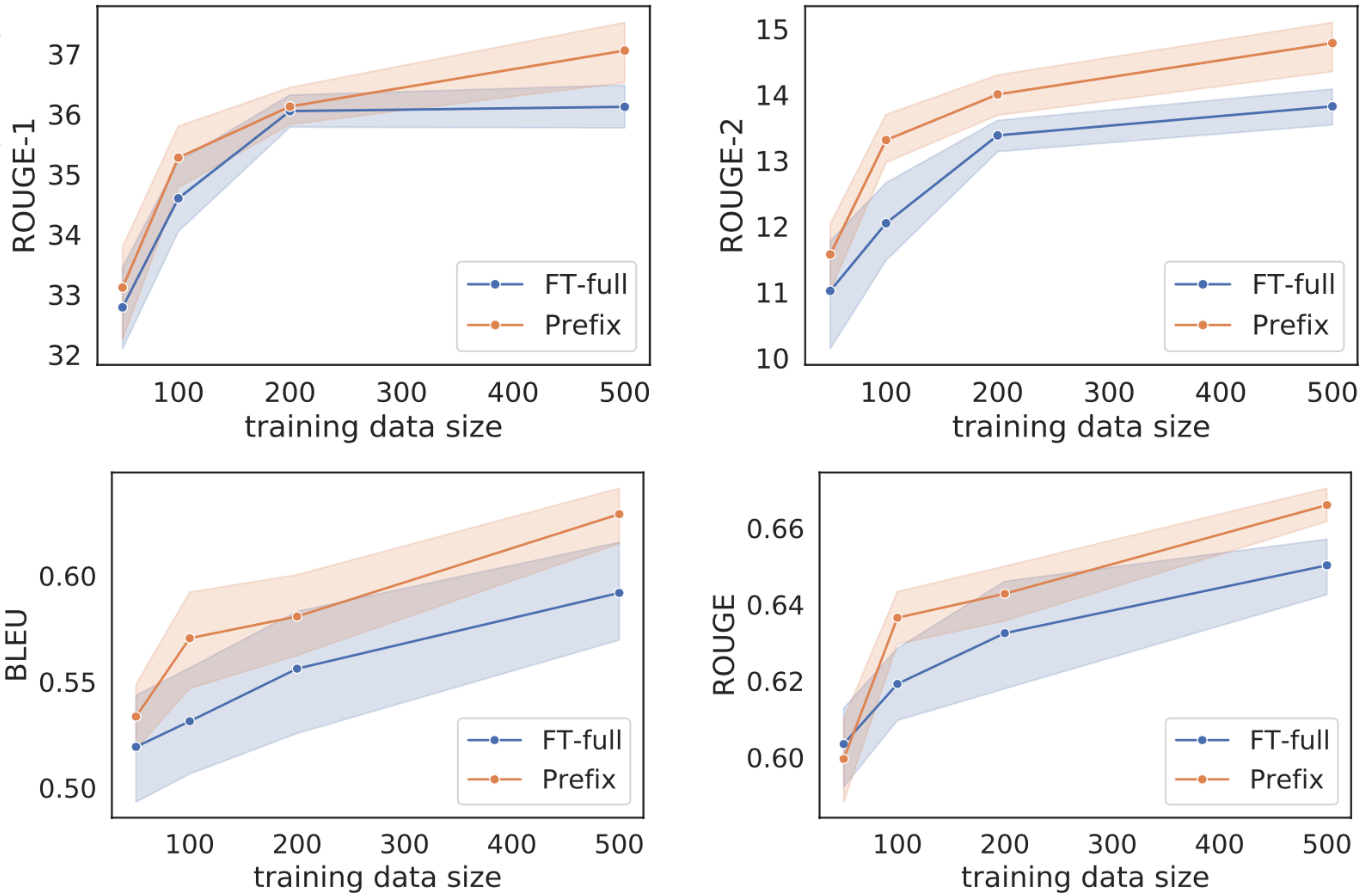
- Only learns task/user-specific prompt vectors
- Only needs to store these vectors for each task/user



Li and Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 2021.

Insights about soft prompt in NLP

- Can handle low-data regimes



Li and Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 2021.

Insights about soft prompt in NLP

- Is domain-generalizable

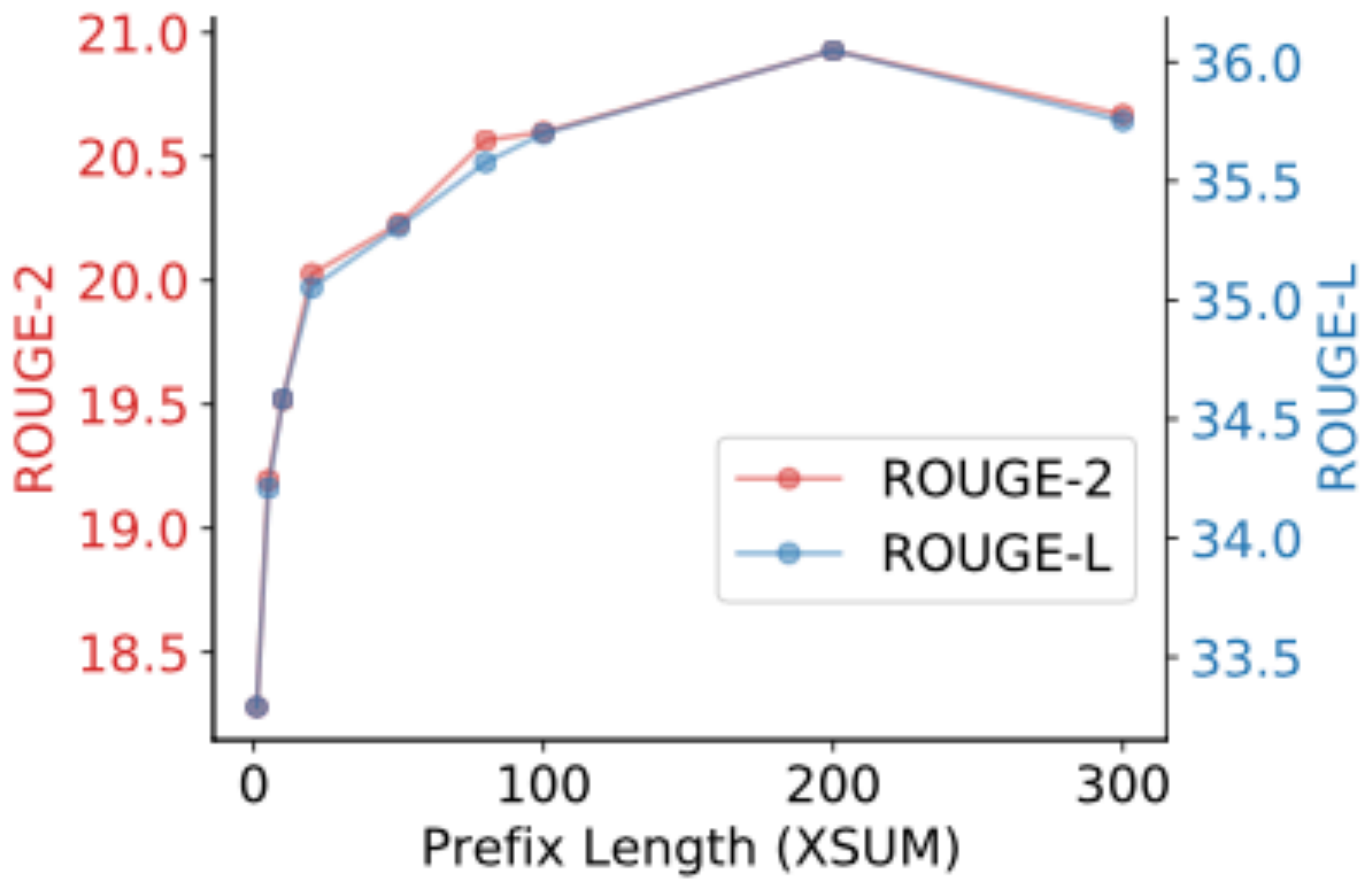
Dataset	Domain	Model	Prompt	Δ
SQuAD	Wiki	94.9 ±0.2	94.8 ±0.1	−0.1
TextbookQA	Book	54.3 ±3.7	**66.8** ±2.9	+12.5
BioASQ	Bio	77.9 ±0.4	**79.1** ±0.3	+1.2
RACE	Exam	59.8 ±0.6	**60.7** ±0.5	+0.9
RE	Wiki	88.4 ±0.1	**88.8** ±0.2	+0.4
DuoRC	Movie	**68.9** ±0.7	67.7 ±1.1	−1.2
DROP	Wiki	**68.9** ±1.7	67.1 ±1.9	−1.8

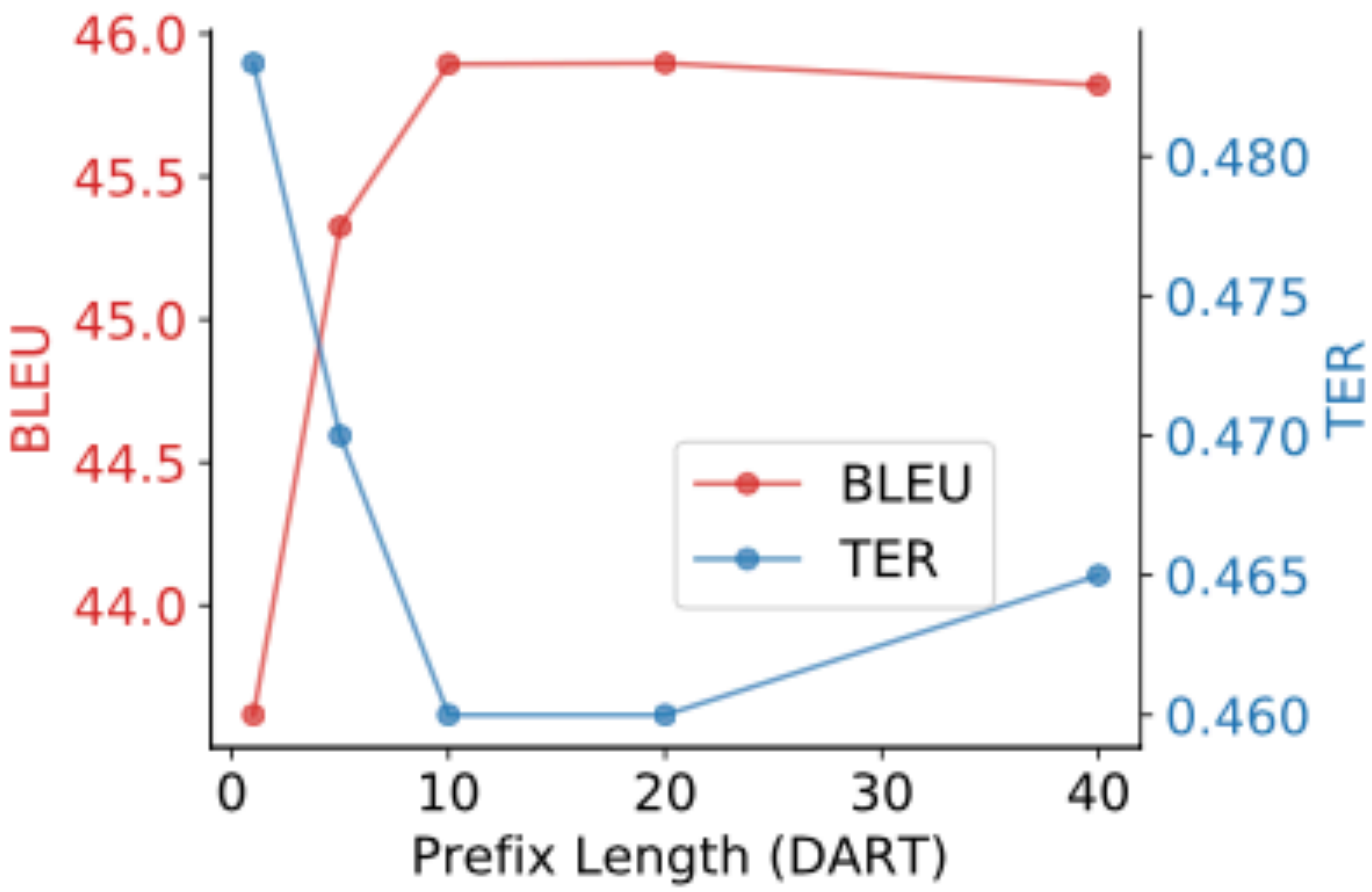
“Prompt tuning tends to give stronger zero-shot performance than model tuning, especially on datasets with large domain shifts like TextbookQA.”

Lester et al. The Power of Scale for Parameter-Efficient Prompt Tuning. 2021.

Insights about soft prompt in NLP

- Longer prompt works better but should not be too long

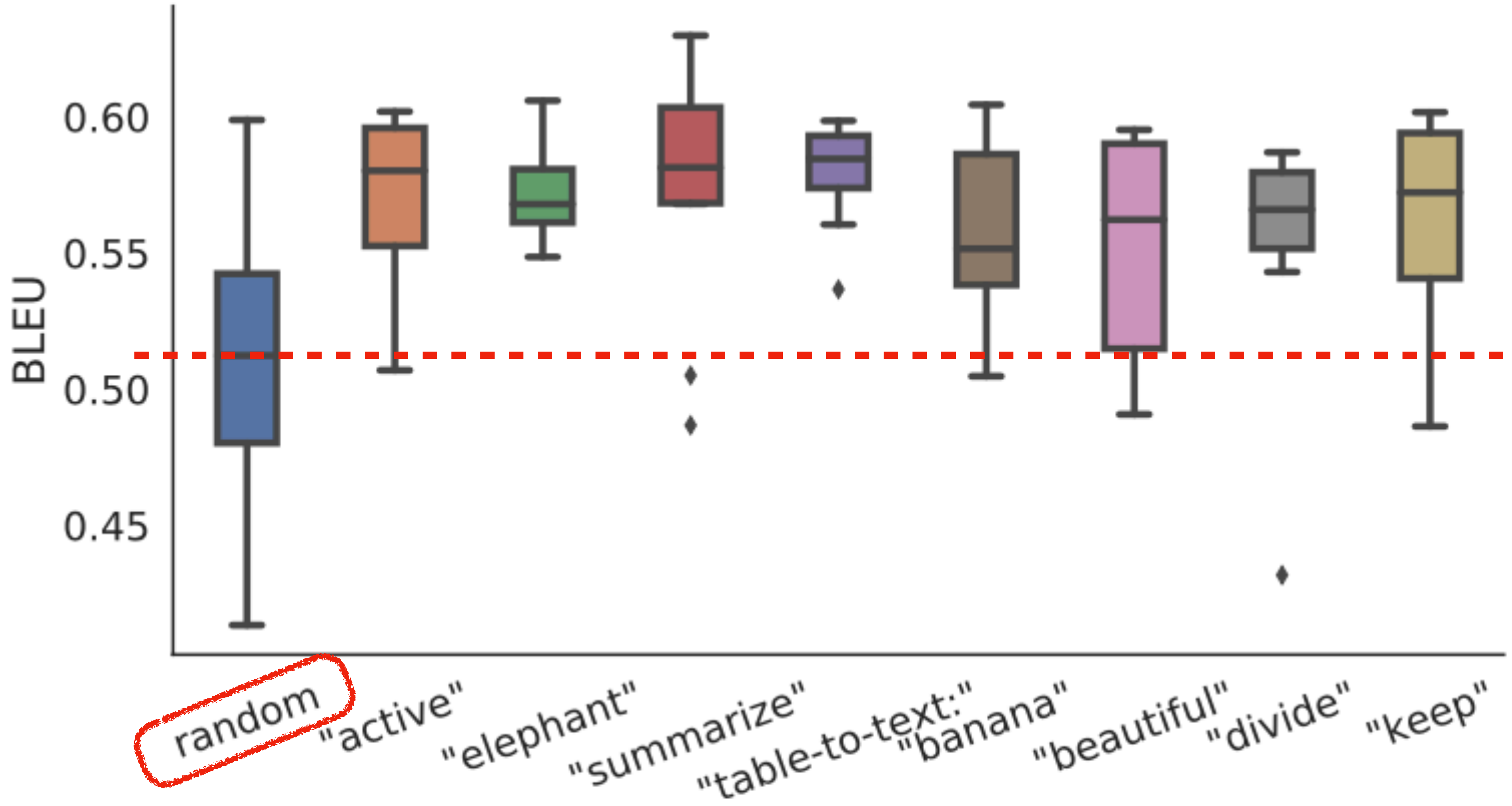




Li and Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 2021.

Insights about soft prompt in NLP

- Initialization matters a lot (word embeddings >> random)



Li and Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 2021.

Insights about soft prompt in NLP

- Interpretable? … sort of

The Power of Scale for Parameter-Efficient Prompt Tuning

Brian Lester* **Rami Al-Rfou** **Noah Constant**
Google Research
{brianlester,rmyeid,nconstant}@google.com

Finding 1:
Top-5 nearest words form clusters,
e.g., lexically similar cluster {Technology, Technologies, technological},
or diverse but related cluster {entirely, totally, completely, 100%}

Finding 2:
Init words tend to persist through training

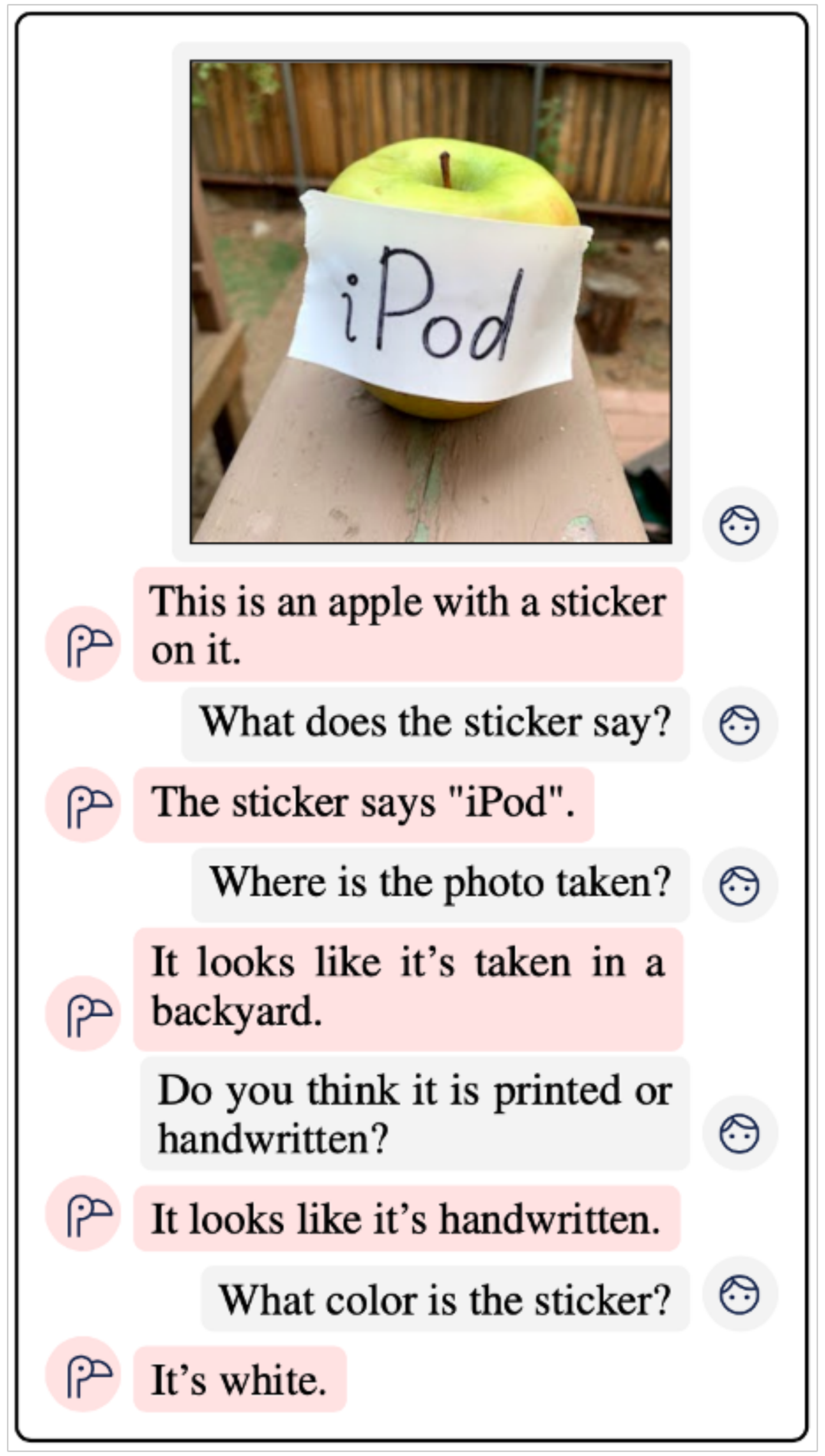
Lester et al. The Power of Scale for Parameter-Efficient Prompt Tuning. 2021.

Outline

- Prompting in natural language processing
  - Language model, hard prompt, and soft prompt
- Prompting in computer vision
  - White-box prompt learning
  - Black-box prompt learning

2023: Prompting is everywhere

DeepMind's Flamingo



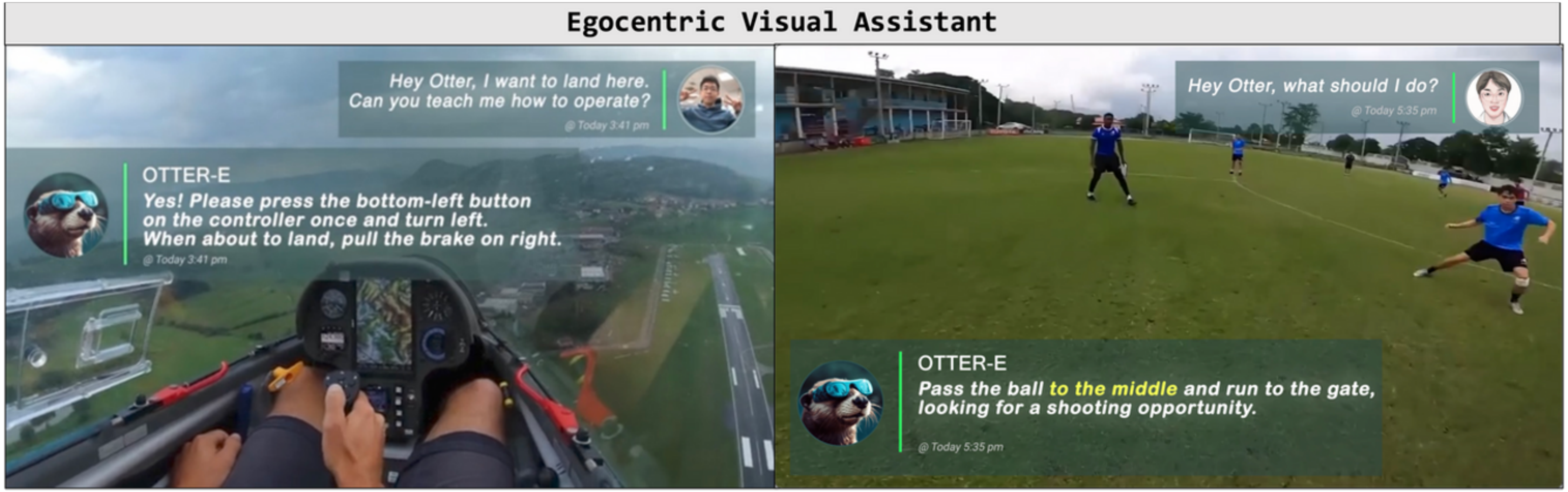
- Image/video recognition
- Captioning
- Question answering
- Visual dialogue
- etc.

Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. 2022.

2023: Prompting is everywhere

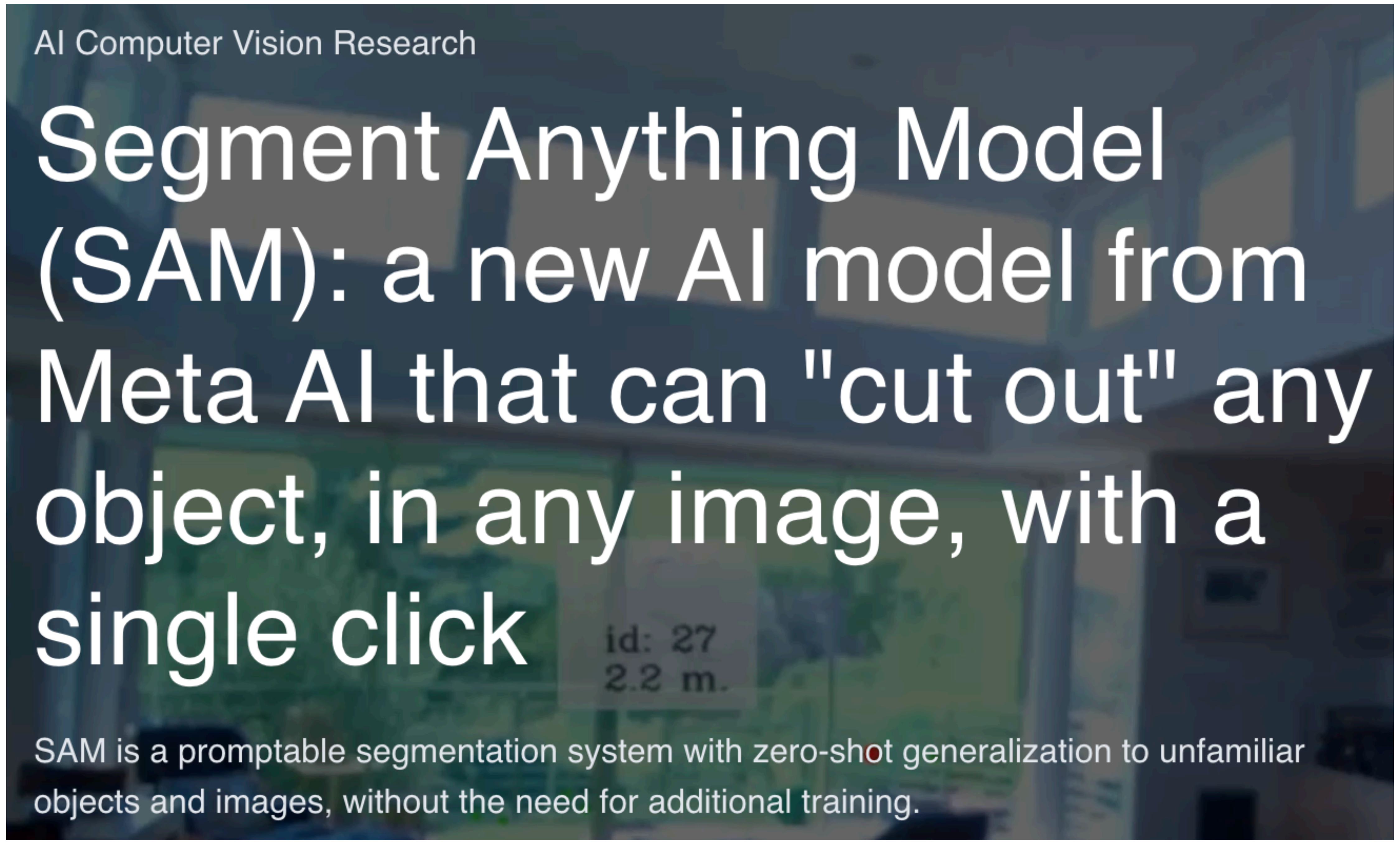
OTTER

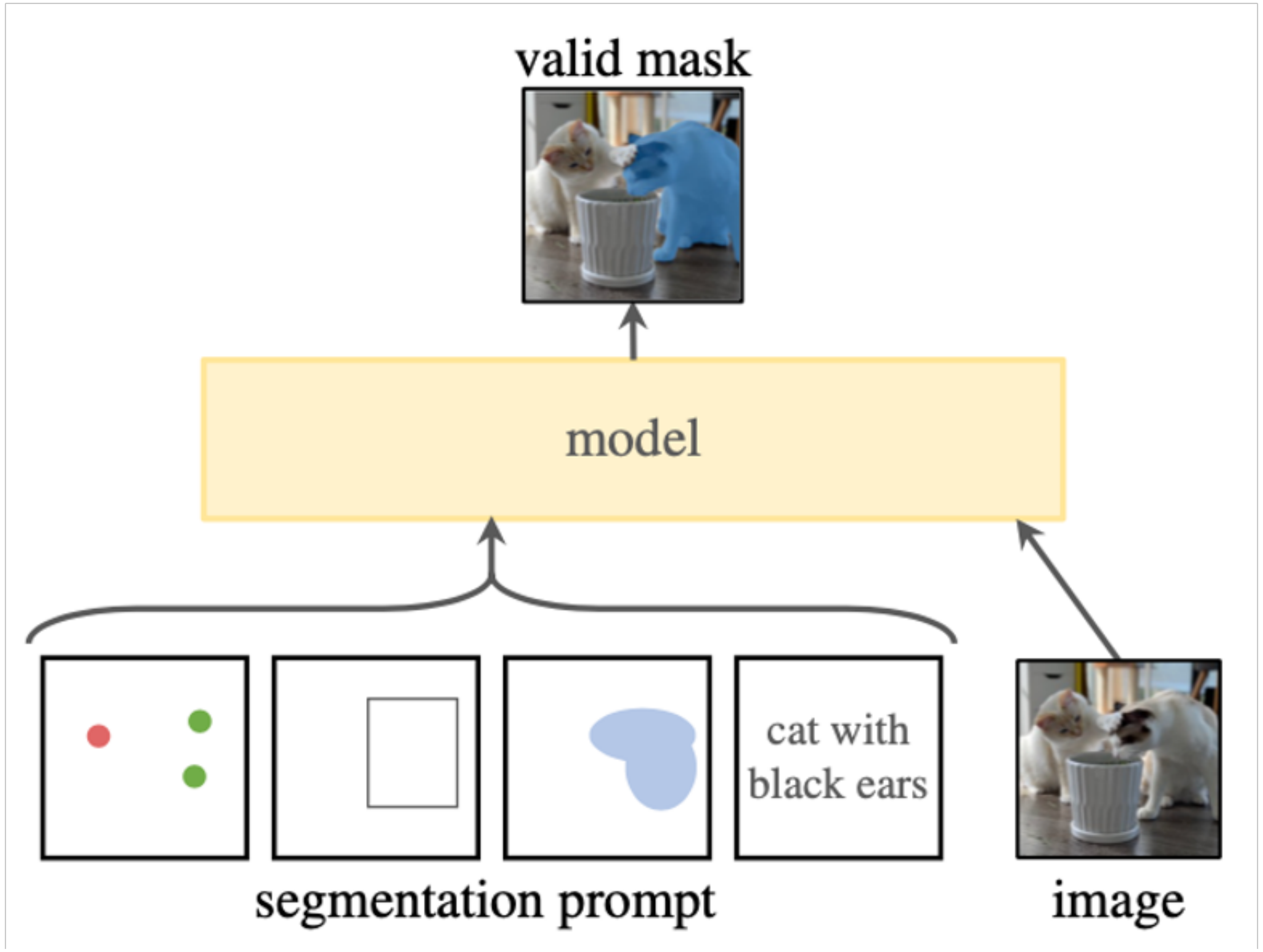
A Multi-Modal Model with
In-Context Instruction Tuning



Li et al. Otter: A Multi-Modal Model with In-Context Instruction Tuning. 2023.

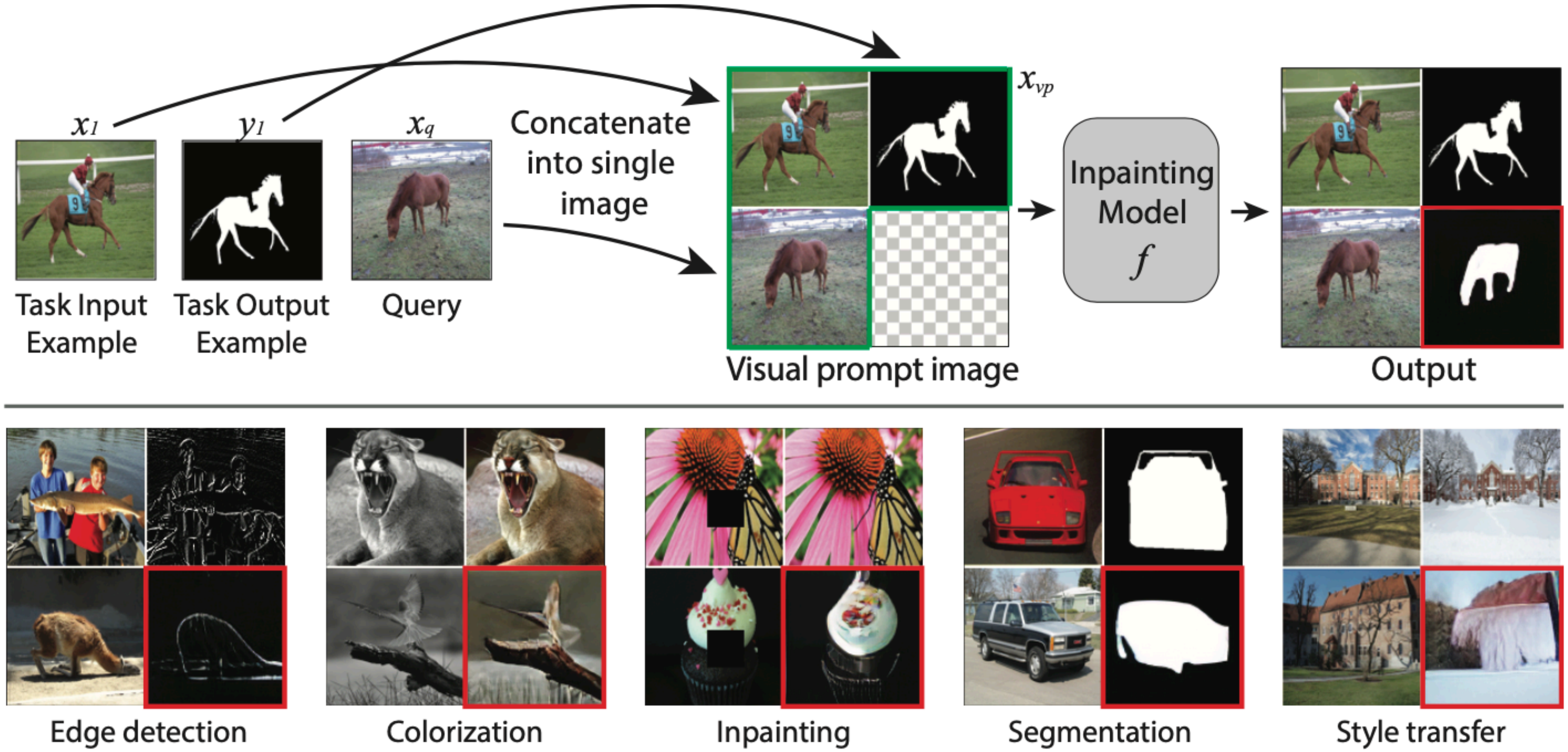
2023: Prompting is everywhere





Kirillov et al. Segment Anything. 2023.

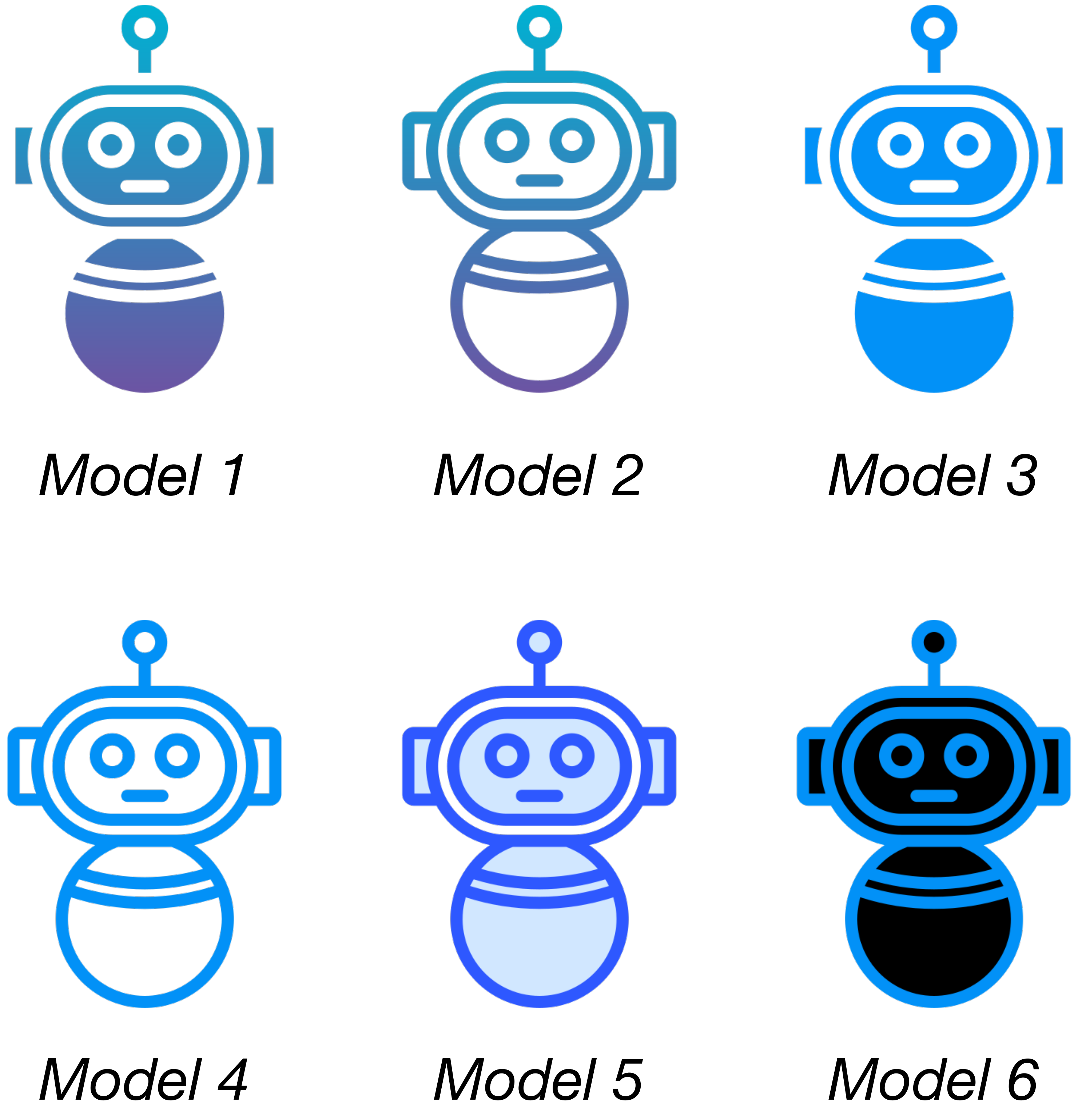
2023: Prompting is everywhere



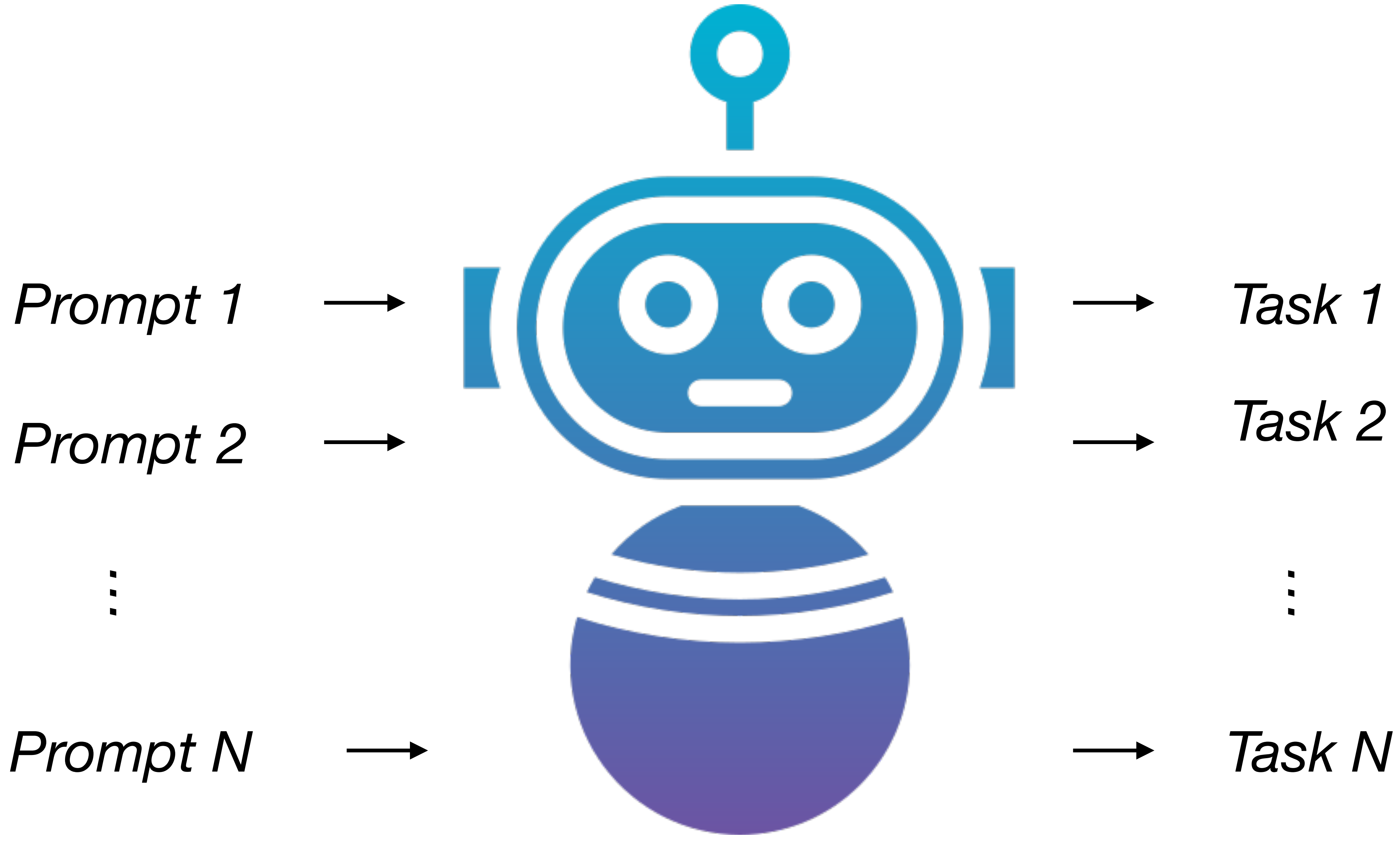
Bar et al. Visual Prompting via Image Inpainting. 2022.

2013 vs. 2023

Old days: one model for one purpose



Now: one model for multiple purposes

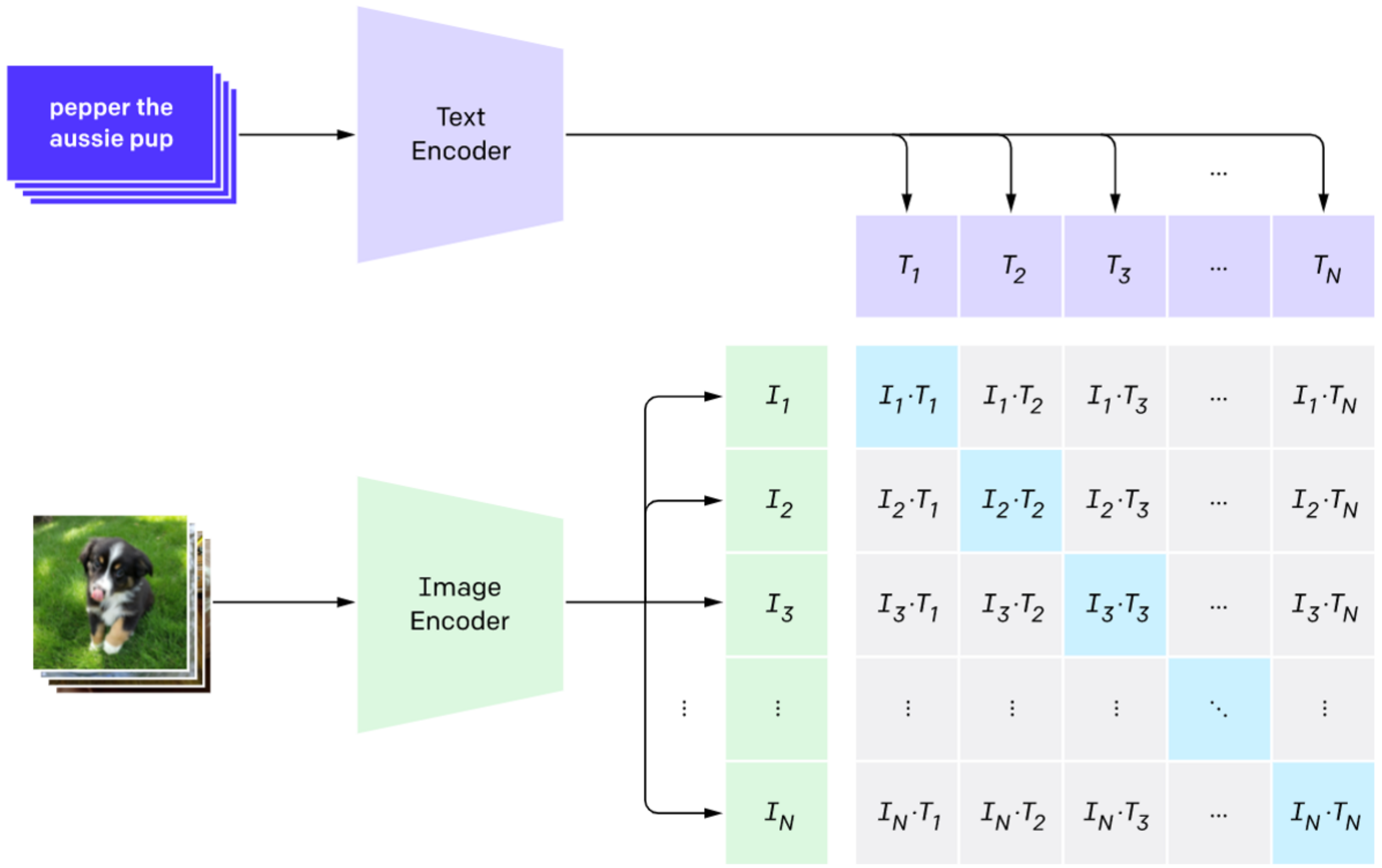


“Foundation model”

Credit: Visual Prompting Livestream With Andrew Ng, Landing AI

Prompt learning for visual language models

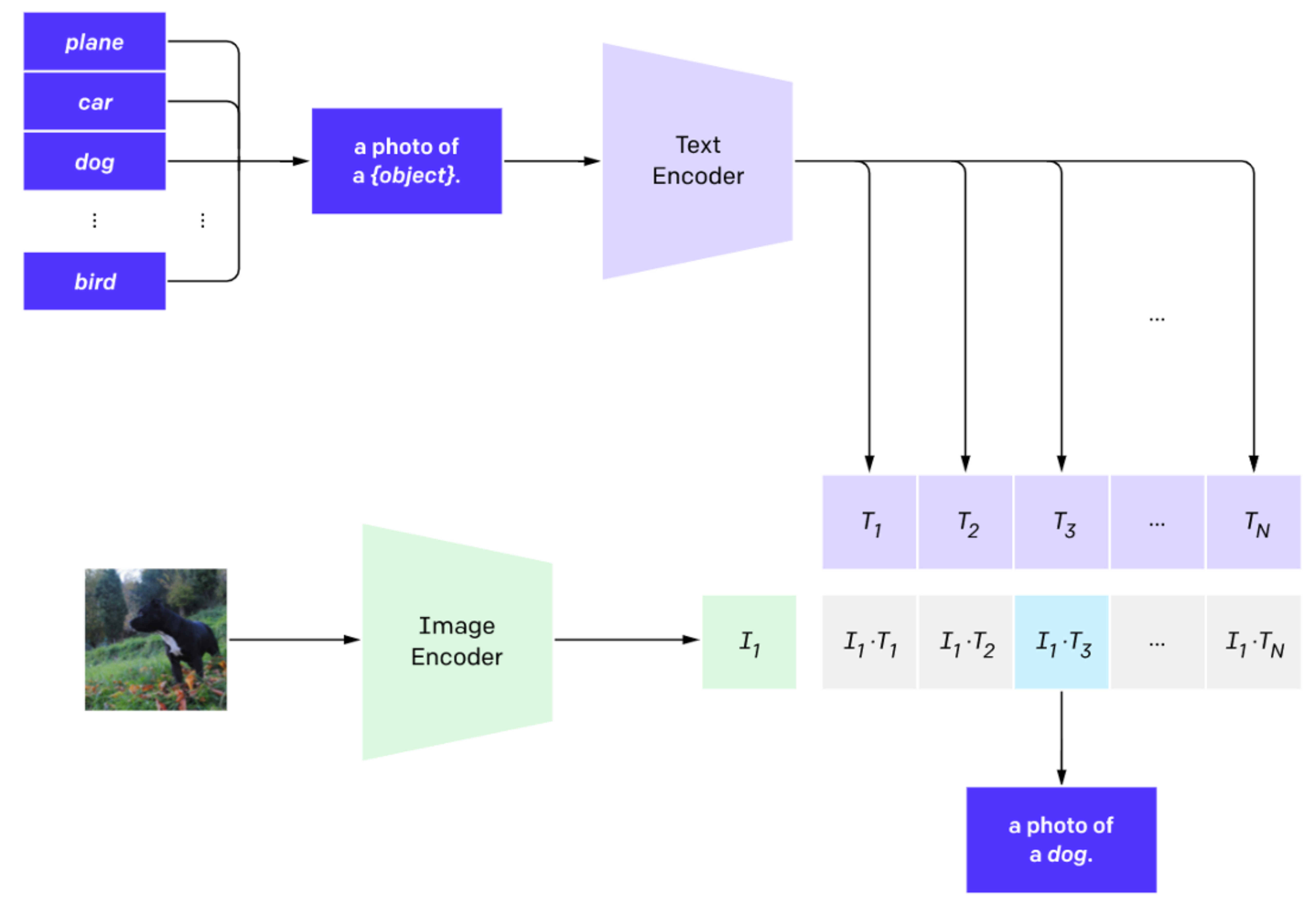
CLIP: Contrastive Language-Image Pre-training

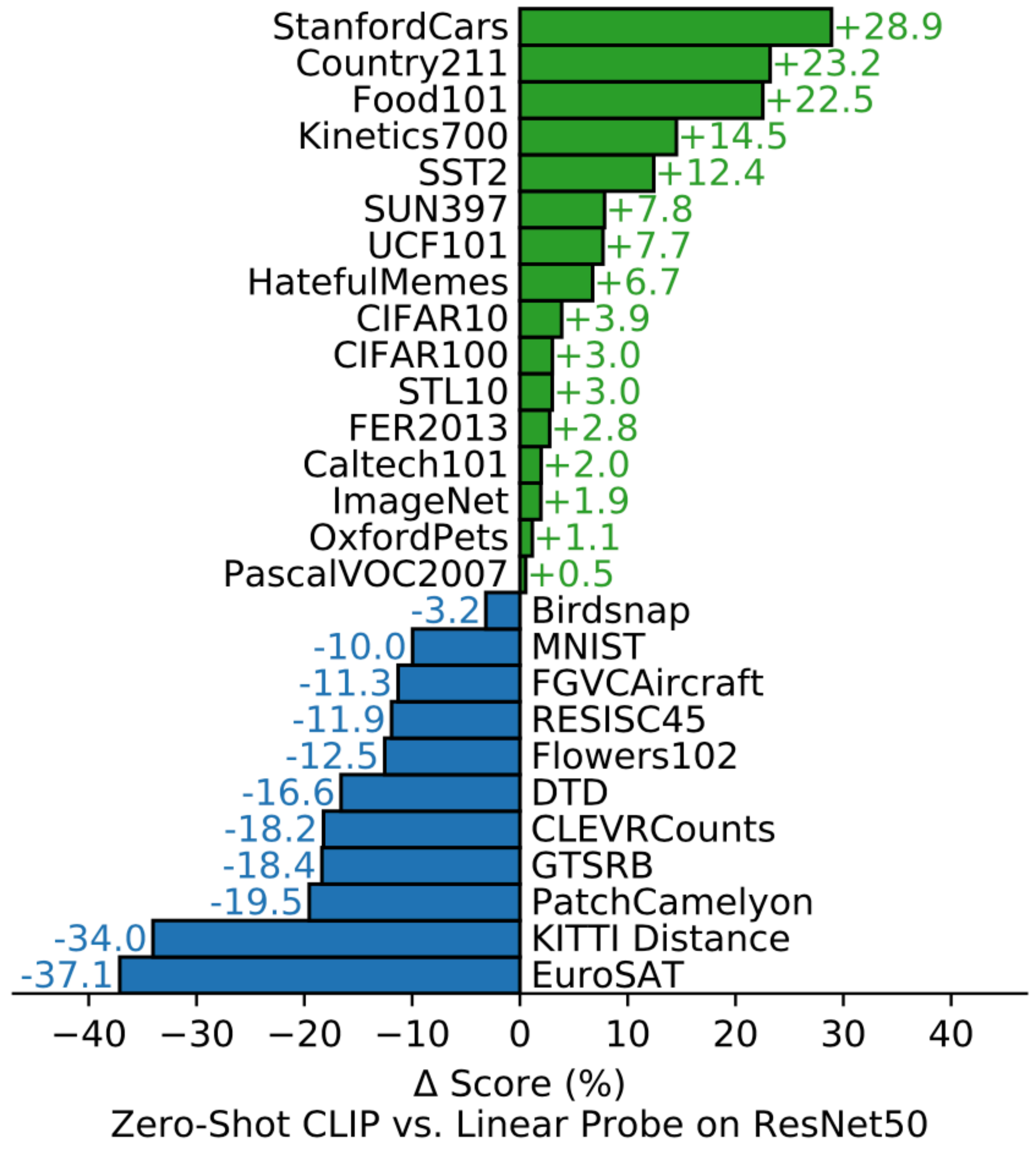


- Data: 400M image-text pairs
- Compute: 250-600 GPUs
- Training time: up to 18 days

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. 2021.

Zero-shot recognition via prompting

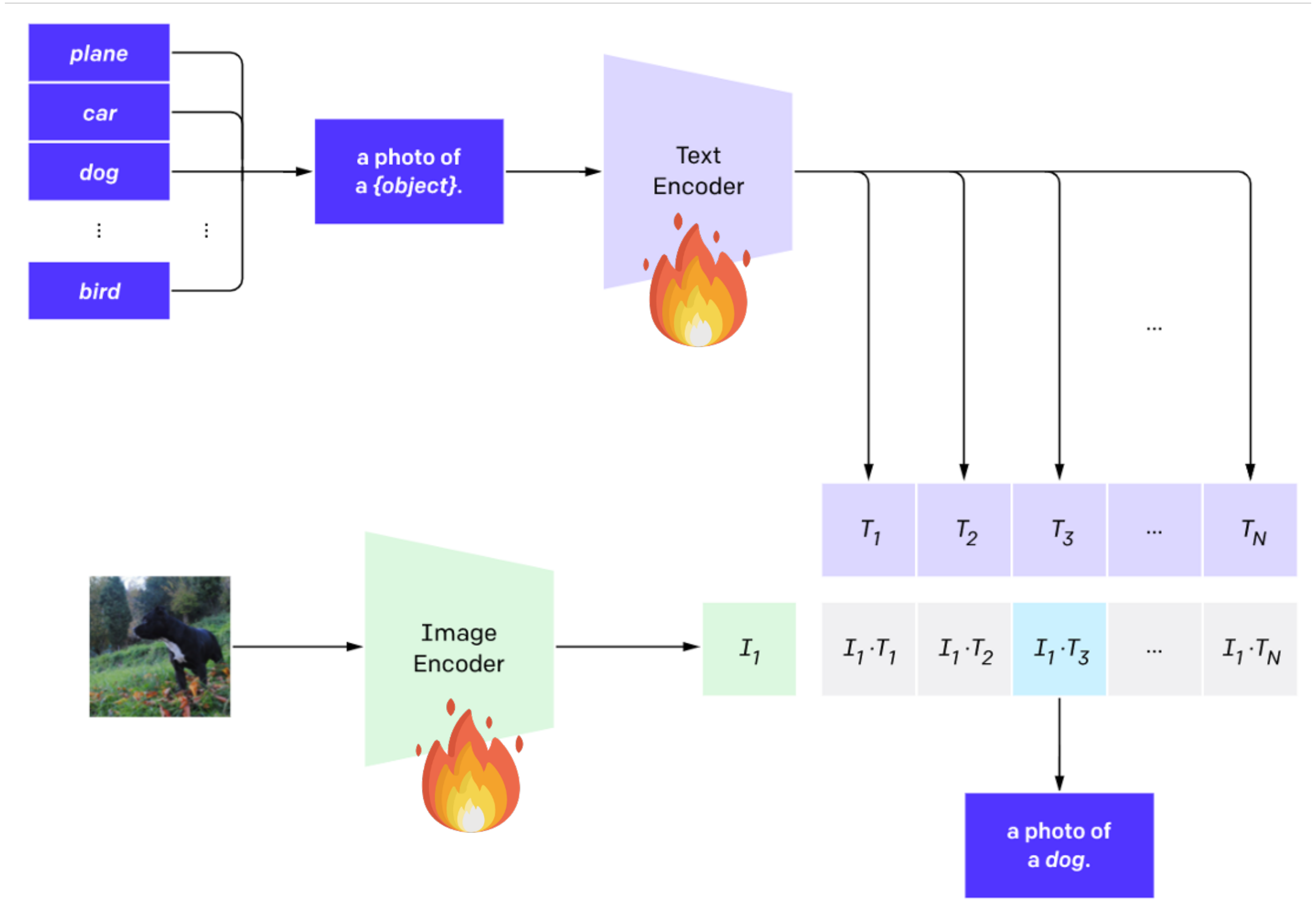




Radford et al. Learning Transferable Visual Models From Natural Language Supervision. 2021.

How to adapt such gigantic models to downstream tasks
to get better performance?

Fine-tuning?



- Fine-tune image encoder: -40%
- Fine-tune both: collapse

The model is too large so it needs a lot of data to avoid overfitting

Prompt engineering?

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 {}.

A slight change in wording could lead to big changes in performance

https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb

Prompt engineering is also hard

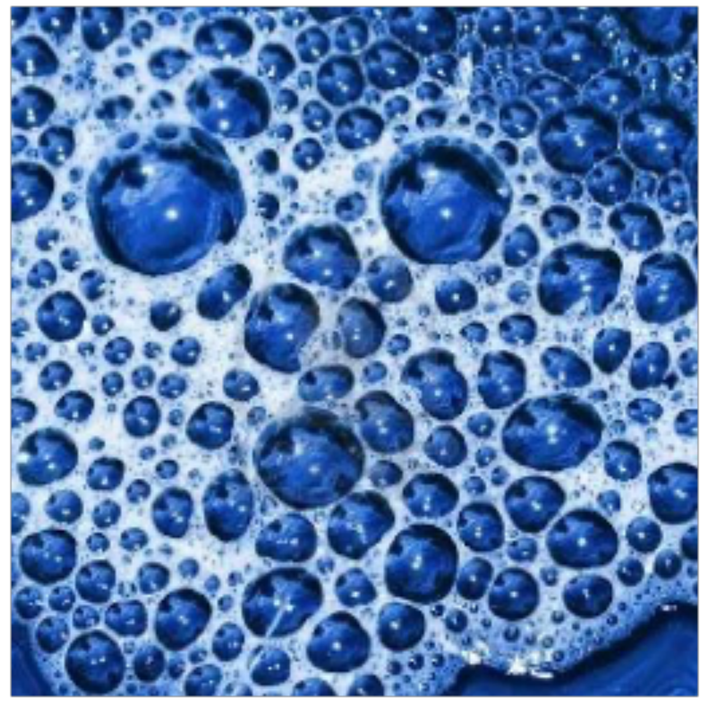
Caltech101

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**

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**

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**

EuroSAT

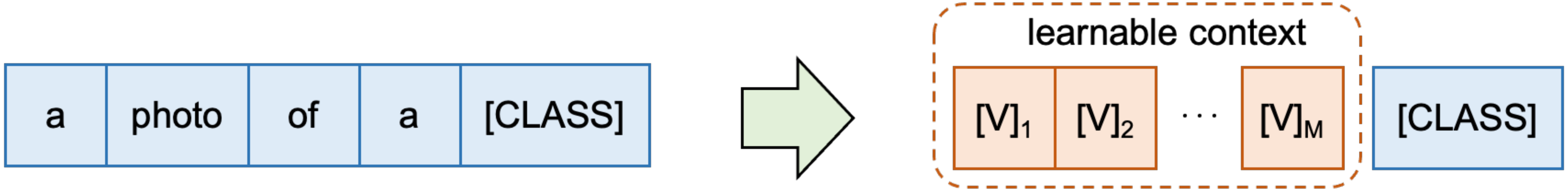
Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**83.53**

A slight change in wording could lead to big changes in performance

Zhou et al. Learning to Prompt for Vision-Language Models. 2022.

Context Optimization (CoOp)

/ku:p/



$$p(y=i|\boldsymbol{x}) = \frac{\exp(\cos(\boldsymbol{w_i},\boldsymbol{f})/\tau)}{\sum_{j=1}^{K}\exp(\cos(\underline{\boldsymbol{w_j}},\boldsymbol{f})/\tau)}$$

Hand-designed prompt

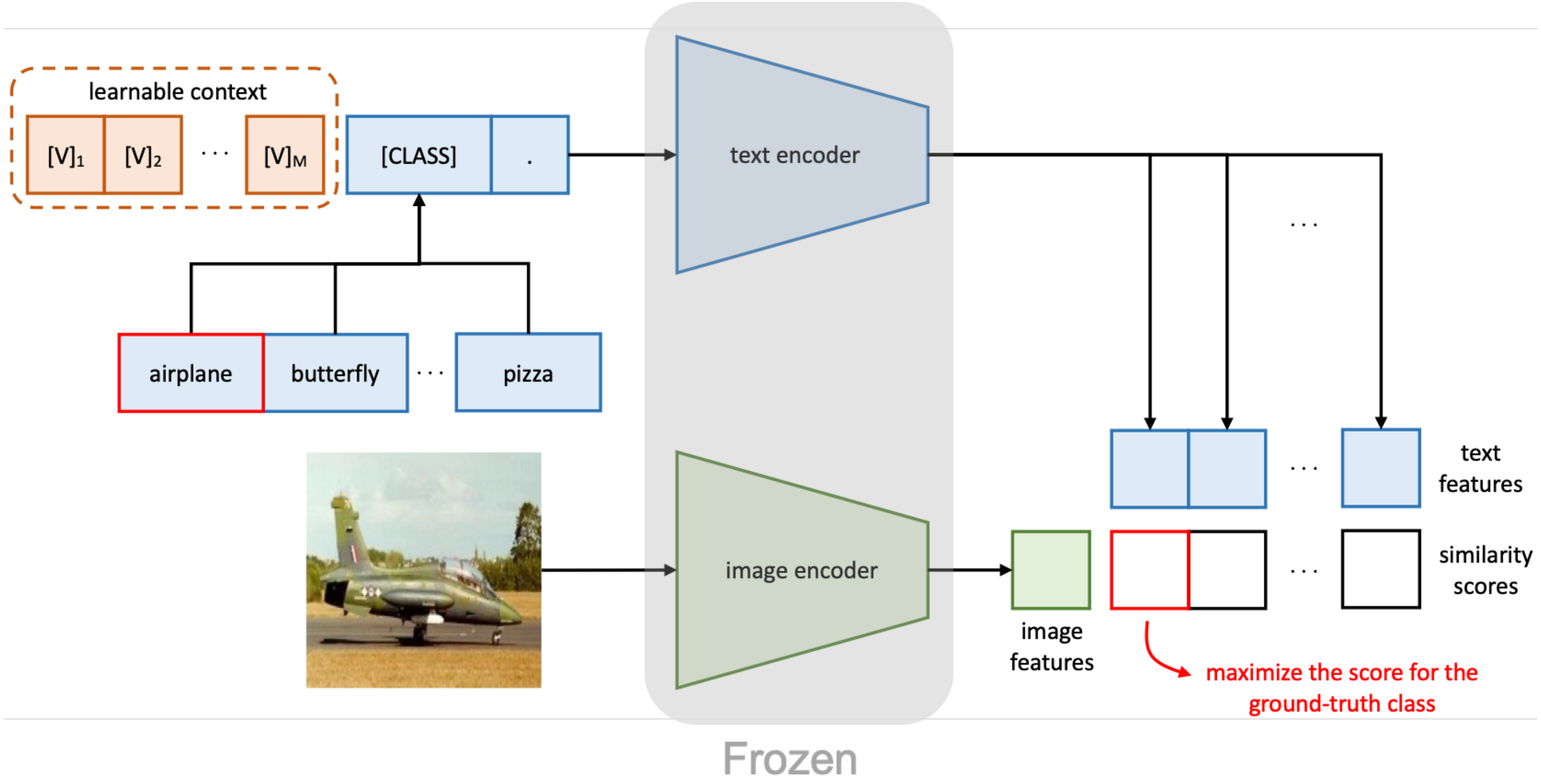
$$p(y=i|\boldsymbol{x}) = \frac{\exp(\cos(g(\boldsymbol{t}_i),\boldsymbol{f})/\tau)}{\sum_{j=1}^{K}\exp(\cos(\underline{g(\boldsymbol{t}_j)},\boldsymbol{f})/\tau)}$$

Learnable prompt

Zhou et al. Learning to Prompt for Vision-Language Models. 2022.

Context Optimization (CoOp)

/ku:p/



Zhou et al. Learning to Prompt for Vision-Language Models. 2022.

CoOp is a few-shot learner

11 datasets covering diverse classification problems:
generic/fine-grained objects, scenes, actions, etc.

Zhou et al. Learning to Prompt for Vision-Language Models. 2022.

CoOp is domain-generalizable

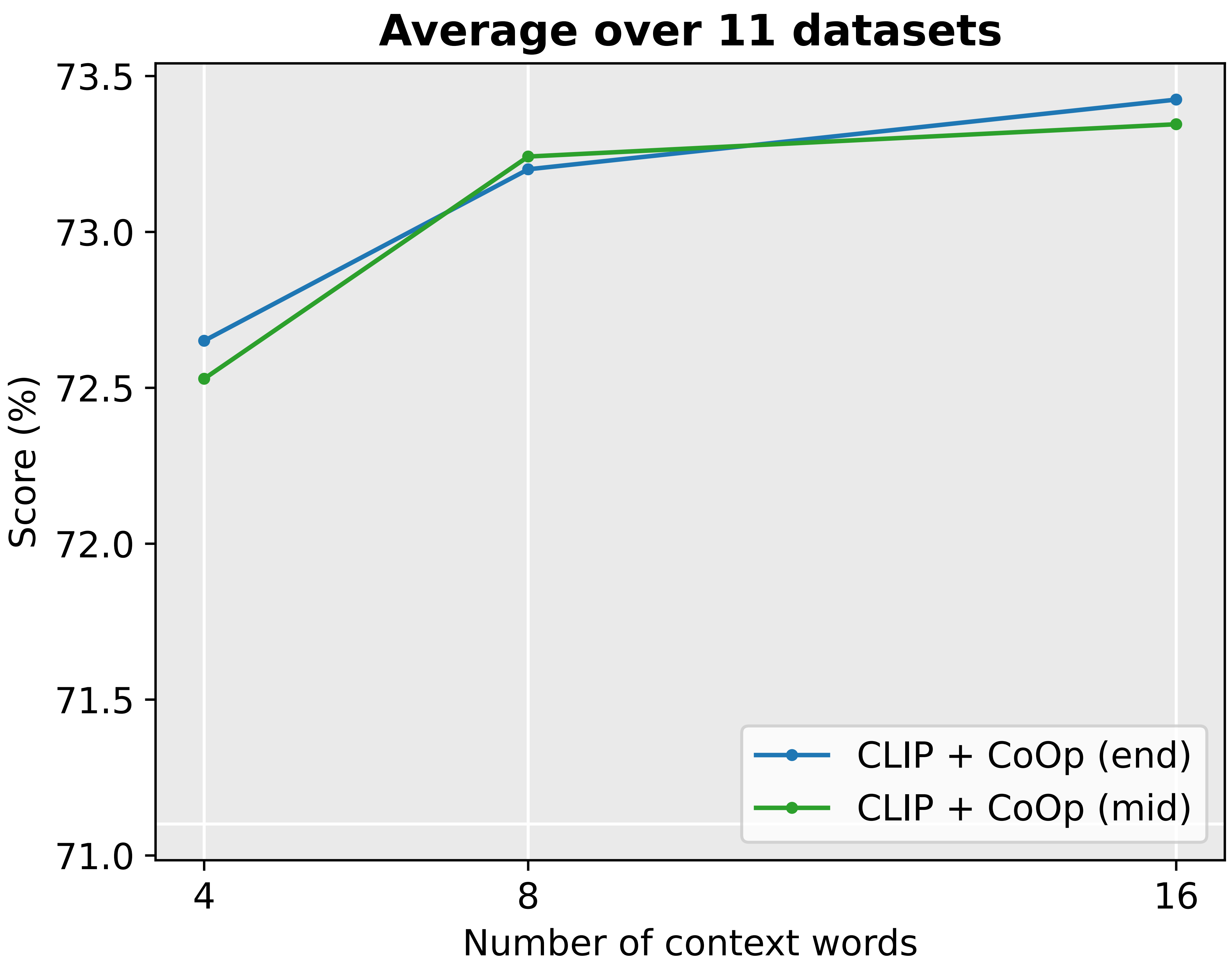
	CLIP	Ours
ImageNet (source)	58.18	**63.33**
V2 (target)	51.34	**55.40**
Sketch (target)	33.32	**34.67**
Adversarial (target)	21.65	**23.06**
Rendition (target)	56.00	**56.60**

Method	Source ImageNet	Target -V2	-Sketch	-A	-R
ResNet-50					
Zero-Shot CLIP	58.18	51.34	33.32	21.65	56.00
Linear Probe CLIP	55.87	45.97	19.07	12.74	34.86
CLIP + CoOp ($M=16$)	62.95	55.11	32.74	22.12	54.96
CLIP + CoOp ($M=4$)	**63.33**	**55.40**	**34.67**	**23.06**	**56.60**
ResNet-101					
Zero-Shot CLIP	61.62	54.81	38.71	28.05	64.38
Linear Probe CLIP	59.75	50.05	26.80	19.44	47.19
CLIP + CoOp ($M=16$)	**66.60**	**58.66**	39.08	28.89	63.00
CLIP + CoOp ($M=4$)	65.98	58.60	**40.40**	**29.60**	**64.98**
ViT-B/32					
Zero-Shot CLIP	62.05	54.79	40.82	29.57	**65.99**
Linear Probe CLIP	59.58	49.73	28.06	19.67	47.20
CLIP + CoOp ($M=16$)	**66.85**	58.08	40.44	30.62	64.45
CLIP + CoOp ($M=4$)	66.34	**58.24**	**41.48**	**31.34**	65.78
ViT-B/16					
Zero-Shot CLIP	66.73	60.83	46.15	47.77	73.96
Linear Probe CLIP	65.85	56.26	34.77	35.68	58.43
CLIP + CoOp ($M=16$)	**71.92**	64.18	46.71	48.41	74.32
CLIP + CoOp ($M=4$)	71.73	**64.56**	**47.89**	**49.93**	**75.14**

Zhou et al. Learning to Prompt for Vision-Language Models. 2022.

More insights about CoOp

- Longer prompt works better but there is diminishing return



Zhou et al. Learning to Prompt for Vision-Language Models. 2022.

More insights about CoOp

- Initialization does not matter

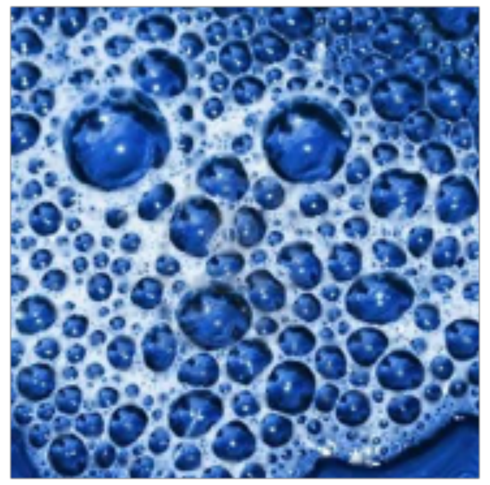
	Avg %
$[V]_1[V]_2[V]_3[V]_4$	72.65
"a photo of a"	72.65

Zhou et al. Learning to Prompt for Vision-Language Models. 2022.

More insights about CoOp

- Interpretable? … not really




Finding 1:
Few are somewhat relevant, e.g., “fluffy” and “paw”

Finding 2:
The whole prompt does not make much sense

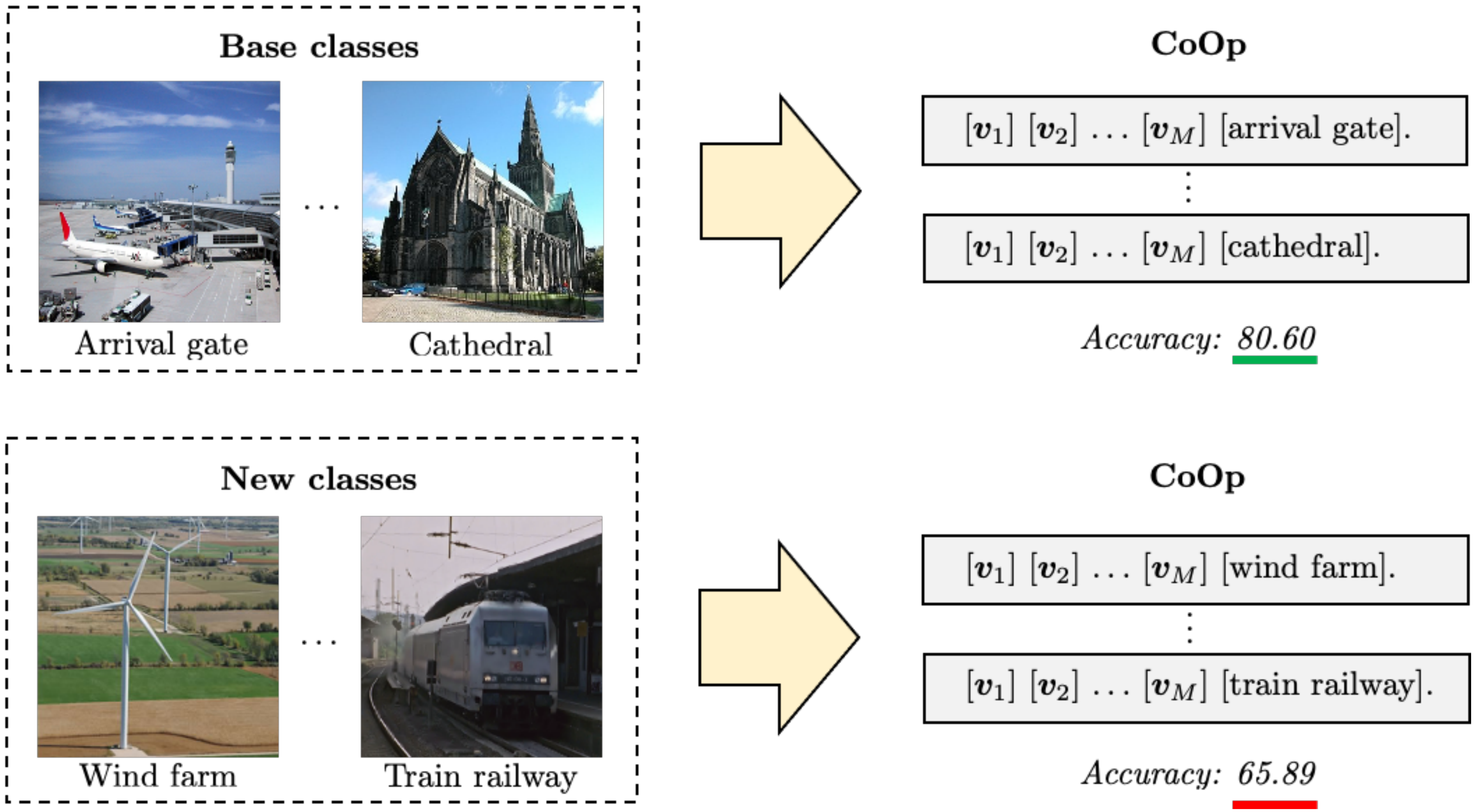
#	ImageNet	Food101	OxfordPets	DTD	UCF101
1	Potd (1.7136)	Lc (0.6752)	Tosc (2.5952)	Boxed (0.9433)	Meteorologist (1.5377)
2	That (1.4015)	Enjoyed (0.5305)	Judge (1.2635)	Seed (1.0498)	Exe (0.9807)
3	Filmed (1.2275)	Beh (0.5390)	Fluffy (1.6099)	Anna (0.8127)	Parents (1.0654)
4	Fruit (1.4864)	Matches (0.5646)	Cart (1.3958)	Mountain (0.9509)	Masterful (0.9528)
5	,... (1.5863)	Nytimes (0.6993)	Harlan (2.2948)	Eldest (0.7111)	Fe (1.3574)
6	°(1.7502)	Prou (0.5905)	Paw (1.3055)	Pretty (0.8762)	Thof (1.2841)
7	Excluded (1.2355)	Lower (0.5390)	Incase (1.2215)	Faces (0.7872)	Where (0.9705)
8	Cold (1.4654)	N/A	Bie (1.5454)	Honey (1.8414)	Kristen (1.1921)
9	Stery (1.6085)	Minute (0.5672)	Snuggle (1.1578)	Series (1.6680)	Imam (1.1297)
10	Warri (1.3055)	~ (0.5529)	Along (1.8298)	Coca (1.5571)	Near (0.8942)
11	Marvelcomics (1.5638)	Well (0.5659)	Enjoyment (2.3495)	Moon (1.2775)	Tummy (1.4303)
12	.: (1.7387)	Ends (0.6113)	Jt (1.3726)	Ih (1.0382)	Hel (0.7644)
13	N/A	Mis (0.5826)	Improving (1.3198)	Won (0.9314)	Boop (1.0491)
14	Lation (1.5015)	Somethin (0.6041)	Srsly (1.6759)	Replied (1.1429)	N/A
15	Muh (1.4985)	Seminar (0.5274)	Asteroid (1.3395)	Sent (1.3173)	Facial (1.4452)
16	.# (1.9340)	N/A	N/A	Piedmont (1.5198)	During (1.1755)

Zhou et al. Learning to Prompt for Vision-Language Models. 2022.

Soft prompt in CV vs. NLP

	CV	NLP
Allows few-shot learning?	Yes	Yes
Is domain-generalizable?	Yes	Yes
Longer prompt works better?	Yes but has diminishing return	Yes but not too long
Initialization matters?	No	Yes
Is interpretable?	Not really	Sort of

Can CoOp generalize to broader (related) concepts within the same dataset?



The prompt only works for a subset of classes (i.e., overfitting)

Zhou et al. Conditional Prompt Learning for Vision-Language Models. 2022.

More failure cases of CoOp on unseen classes (same dataset)

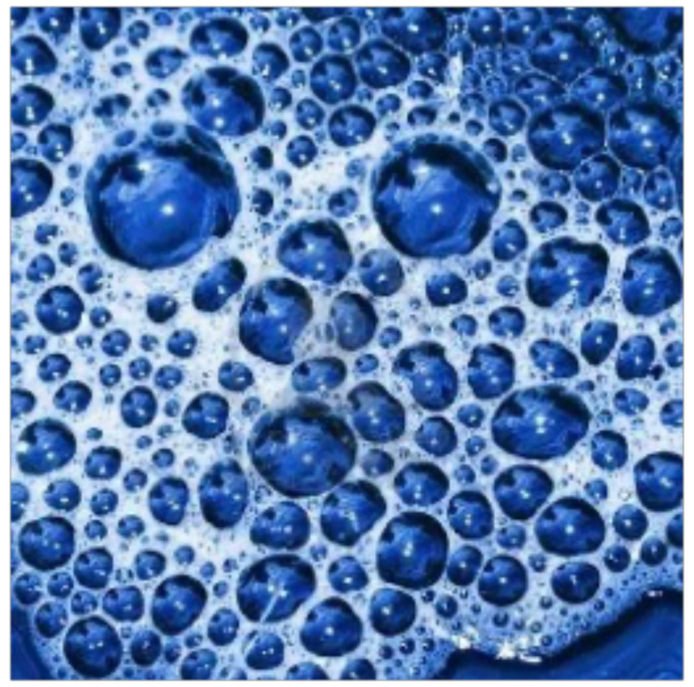
ImageNet
↓8.86%

Caltech101
↓8.19%

Flowers102
↓37.93%

StanfordCars
↓17.72%

FGVCAircraft
↓18.14%



DTD
↓38.26%

EuroSAT
↓37.45%

UCF101
↓28.64%

Zhou et al. Conditional Prompt Learning for Vision-Language Models. 2022.

What is a good prompt?

A good prompt should characterize each instance with some specific context words

Zhou et al. Conditional Prompt Learning for Vision-Language Models. 2022.

Conditional Context Optimization (CoCoOp)

/kəʊˌkuːp/

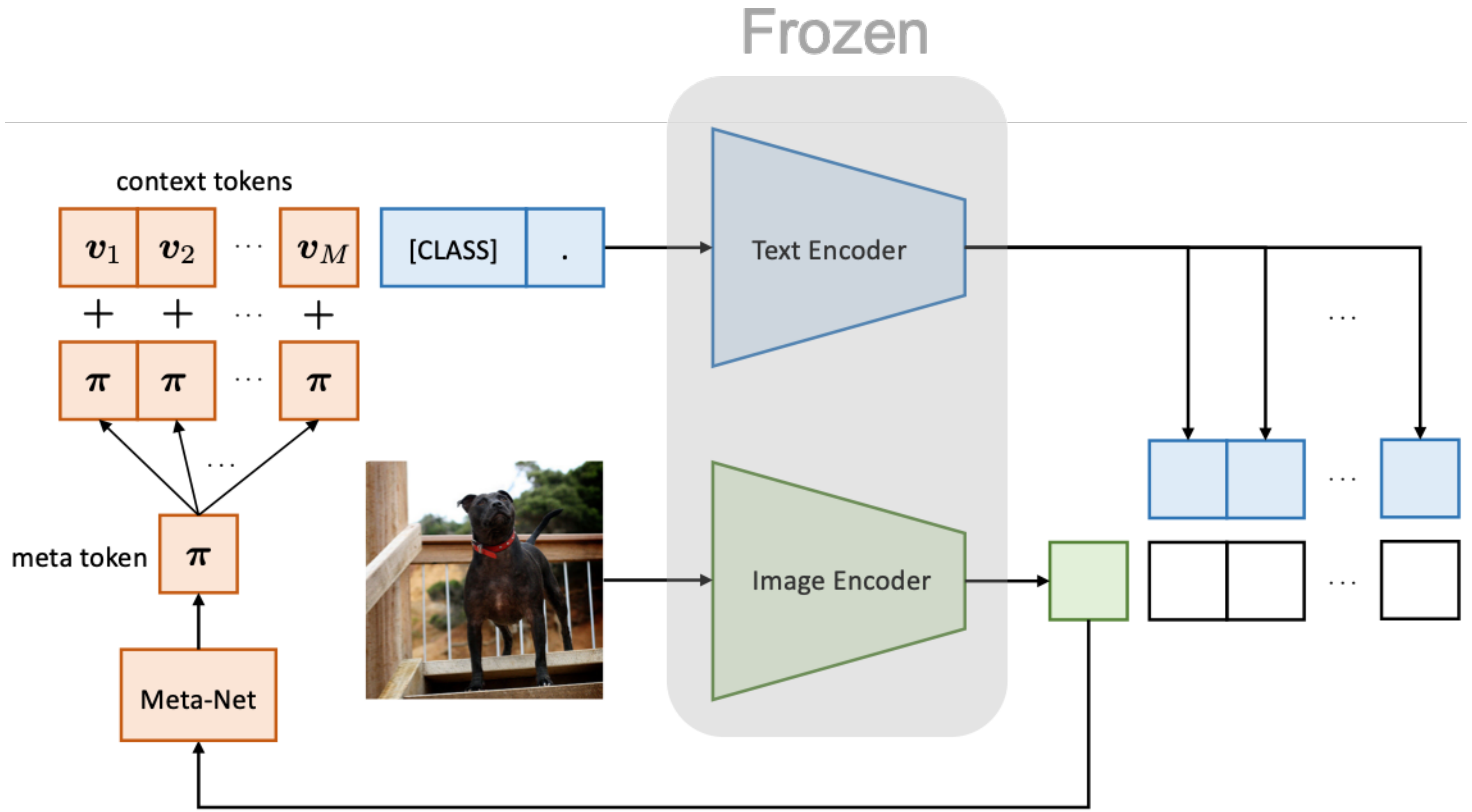
$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_y(\boldsymbol{x})))/\tau)}{\sum_{i=1}^{K} \exp(\text{sim}(\boldsymbol{x}, g(\underline{\boldsymbol{t}_i(\boldsymbol{x})}))/\tau)}$$

Conditioned on image

A parameter-efficient design:
Learn a single mini-network h_θ

$$\boldsymbol{t}_i(\boldsymbol{x}) = \{\boldsymbol{v}_1(\boldsymbol{x}), \ldots, \boldsymbol{v}_M(\boldsymbol{x})\}$$

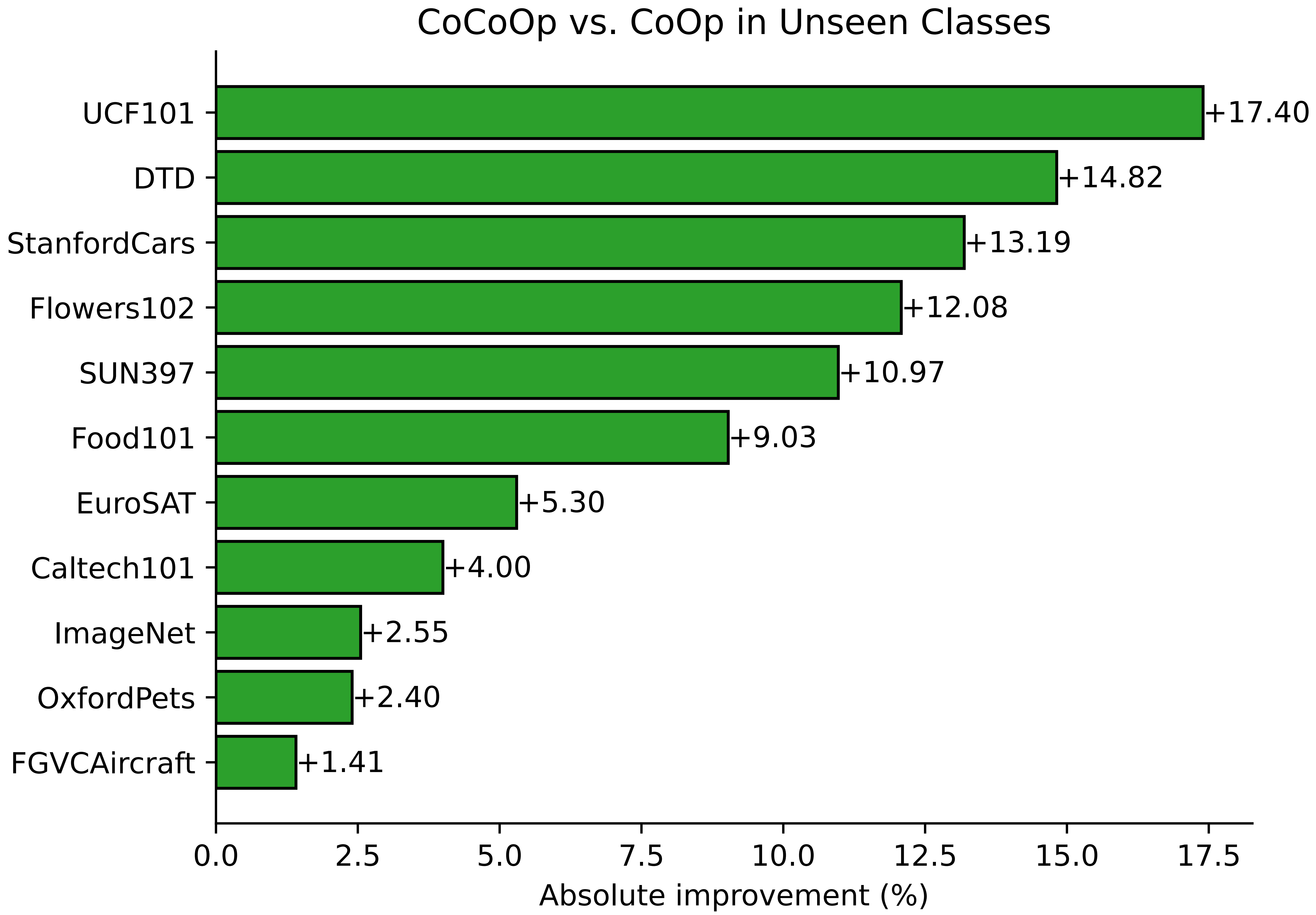
$$\boldsymbol{v}_m(\boldsymbol{x}) = \boldsymbol{v}_m + h_{\boldsymbol{\theta}}(\boldsymbol{x})$$



Zhou et al. Conditional Prompt Learning for Vision-Language Models. 2022.

Findings of conditional prompt learning

- Is more generalizable



Zhou et al. Conditional Prompt Learning for Vision-Language Models. 2022.

Findings of conditional prompt learning

- Is more transferable

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp [62]	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	**46.39**	66.55	63.88
CoCoOp	71.02	**94.43**	**90.14**	**65.32**	**71.88**	**86.06**	**22.94**	**67.36**	**45.73**	45.37	**68.21**	**65.74**
Δ	**-0.49**	**+0.73**	**+1.00**	**+0.81**	**+3.17**	**+0.76**	**+4.47**	**+3.21**	**+3.81**	**-1.02**	**+1.66**	**+1.86**

Zhou et al. Conditional Prompt Learning for Vision-Language Models. 2022.

Findings of conditional prompt learning

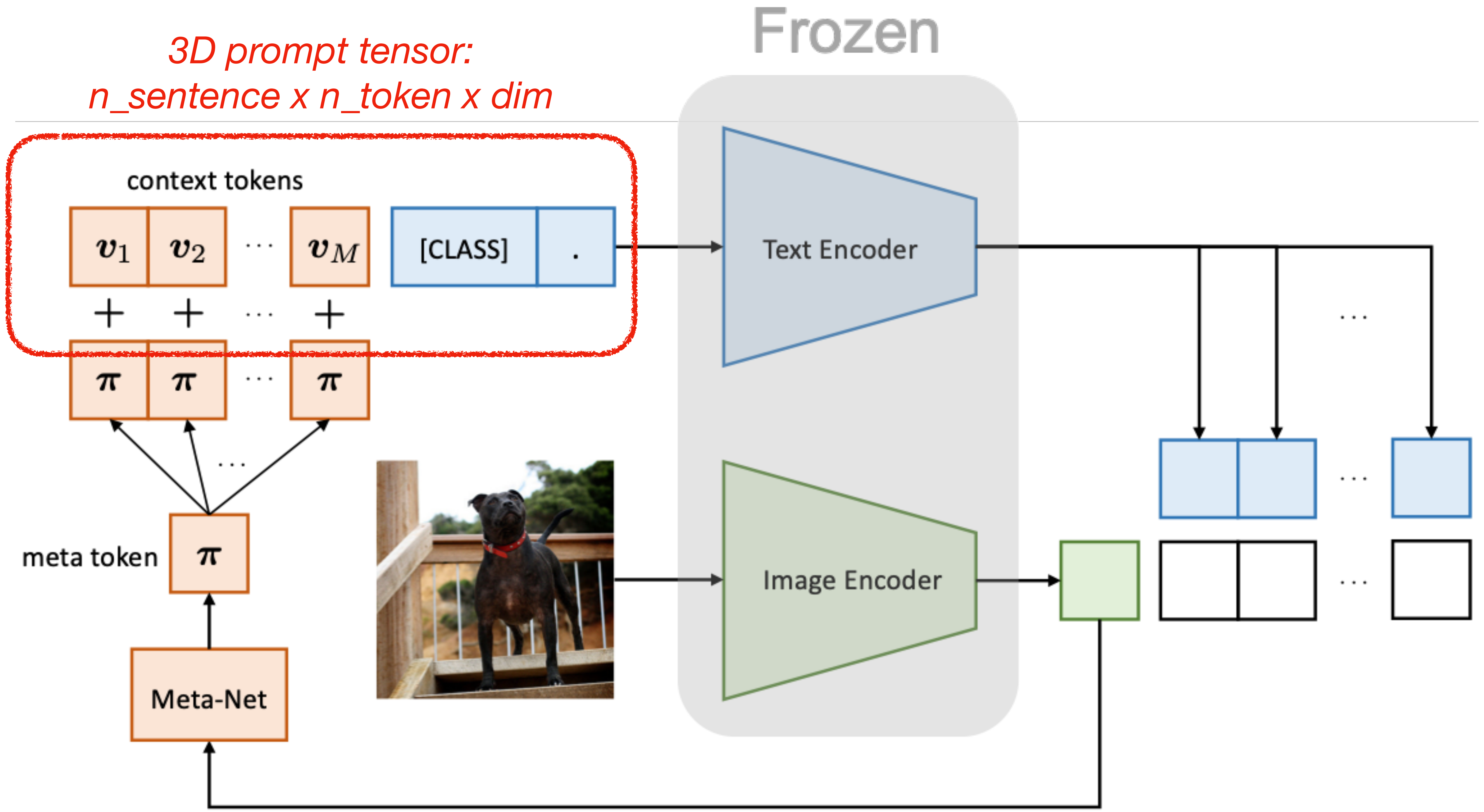
- Is more robust to distribution shifts

		Source	Target			
	Learnable?	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP [40]		66.73	60.83	46.15	47.77	73.96
CoOp [62]	✓	**71.51**	**64.20**	47.99	49.71	75.21
CoCoOp	✓	71.02	64.07	**48.75**	**50.63**	**76.18**

Zhou et al. Conditional Prompt Learning for Vision-Language Models. 2022.

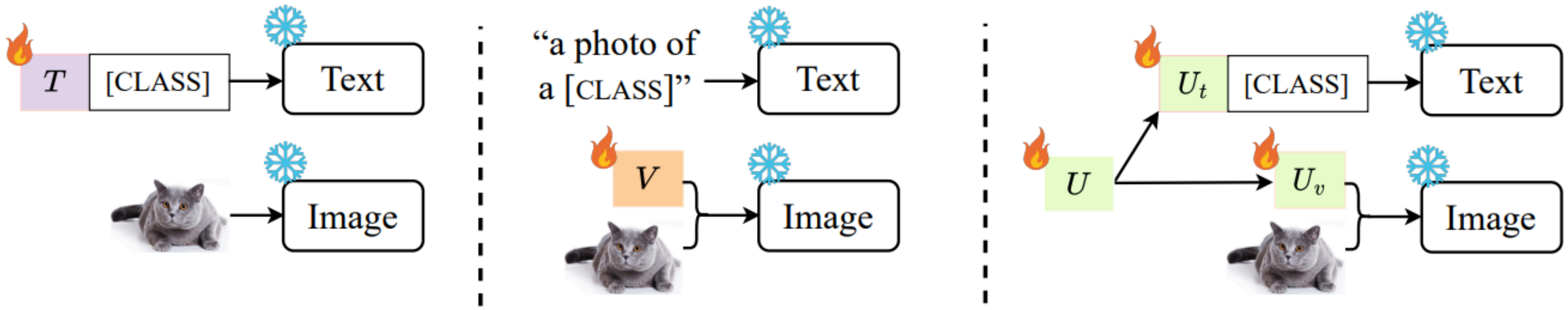
Findings of conditional prompt learning

- Is very slow to train



Zhou et al. Conditional Prompt Learning for Vision-Language Models. 2022.

Want faster training? Try multimodal prompt learning

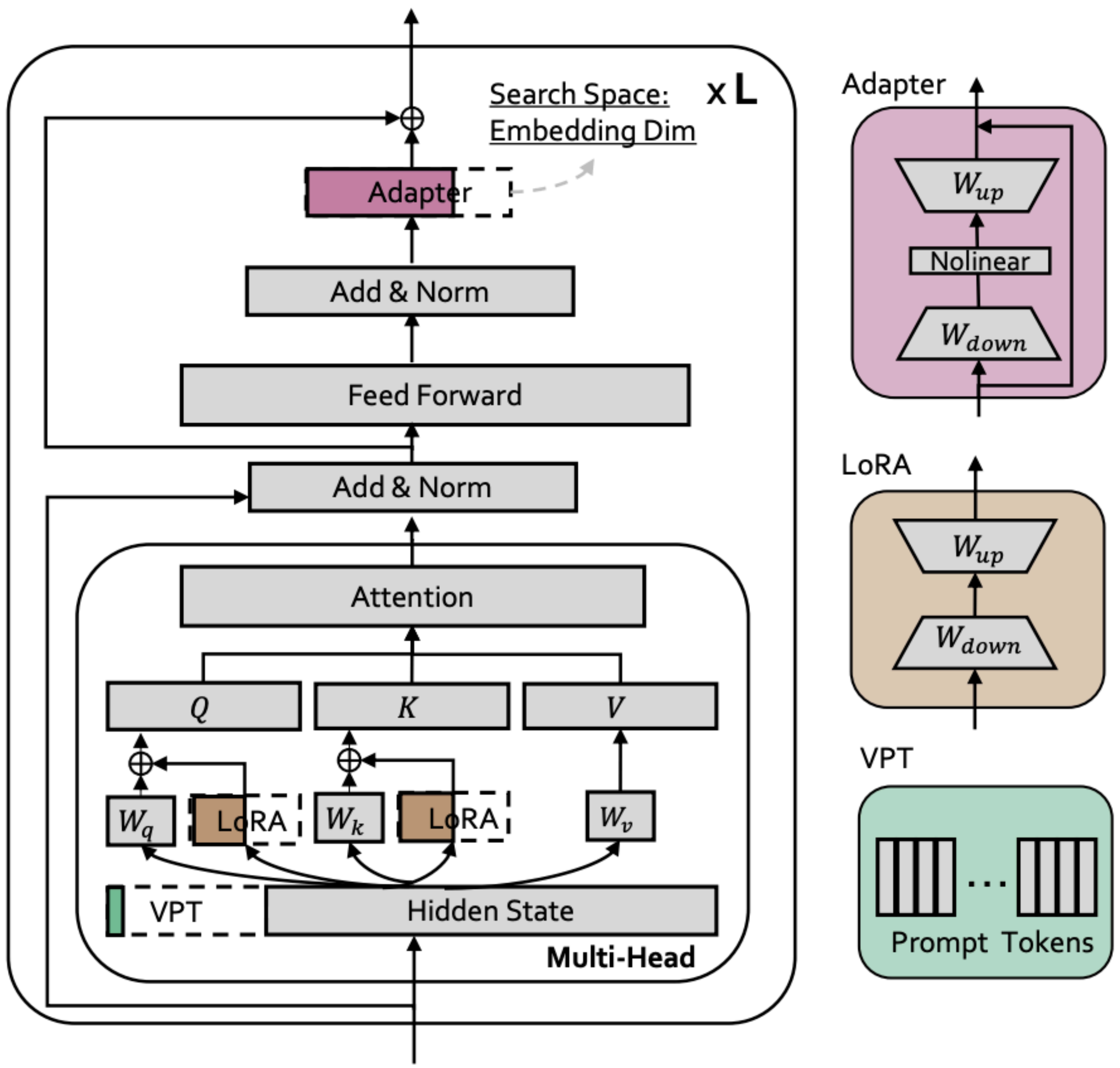


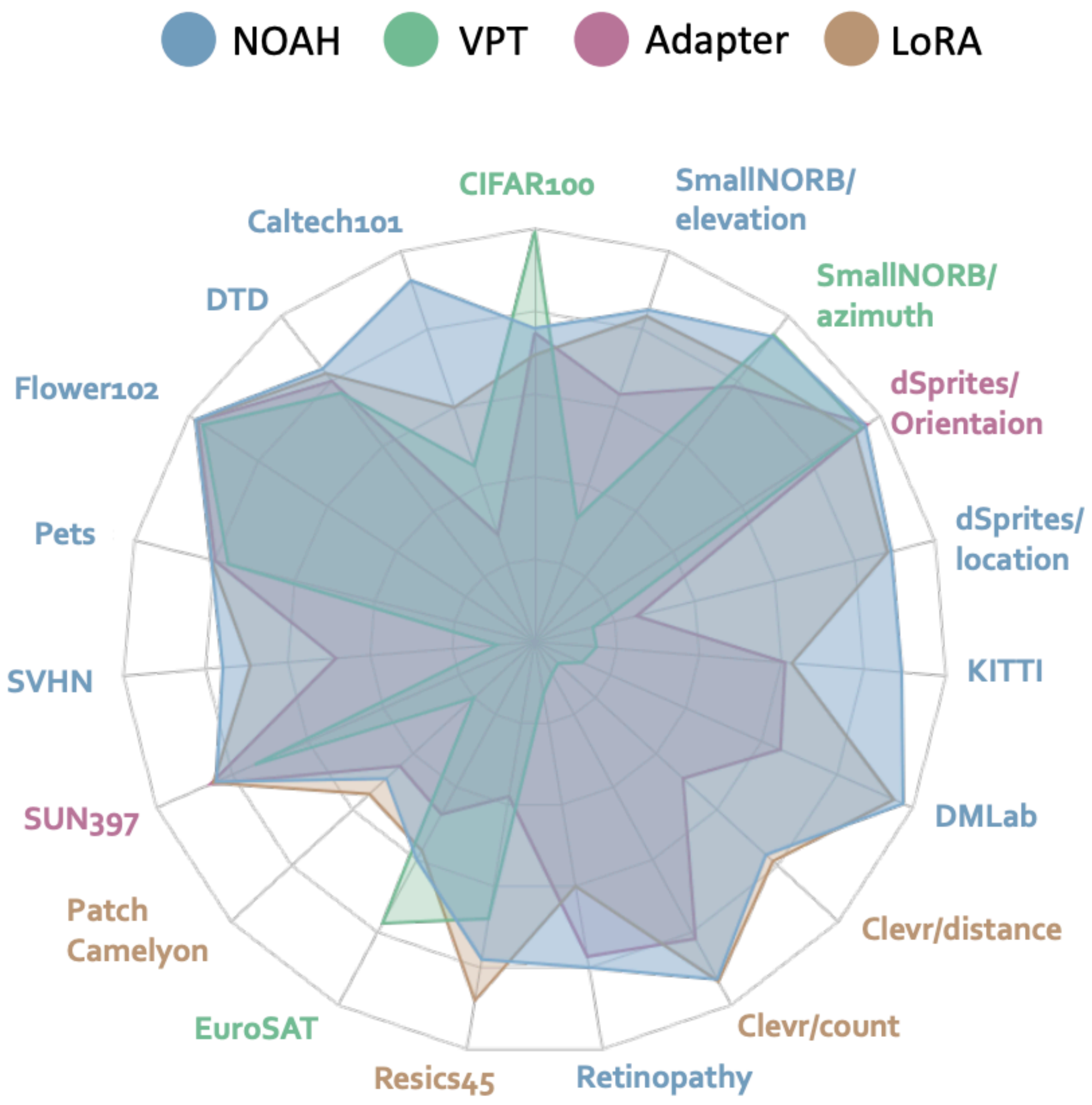
(a) Text Prompt - CoOp **(b)** Visual Prompt - VPT **(c)** Unified Prompt - Ours

#	Method	Source	Target				Average	*OOD* Average
		ImageNet	-V2	-S	-A	-R		
1	CoOp	71.51	64.20	47.99	49.71	75.21	61.72	59.28
2	CoCoOp	71.02	64.07	**48.75**	50.63	76.18	62.13	59.91
3	VPT-shallow	68.98	62.10	47.68	47.19	76.10	60.38	58.27
4	VPT-deep	70.57	63.67	47.66	43.85	74.42	60.04	57.40
5	UPT	**72.63**	**64.35**	48.66	**50.66**	**76.24**	**62.51**	**59.98**

Zang et al. Unified Vision and Language Prompt Learning. 2022.

Have more compute? Try neural prompt search





Zhang et al. Neural Prompt Search. 2022.

What if we only have access to model APIs?

Model-as-a-Service (MaaS)

- APIs are provided to users instead of model weights
- Reasons: model size, accessibility, maintenance, monetization, security, etc.

ChatGPT

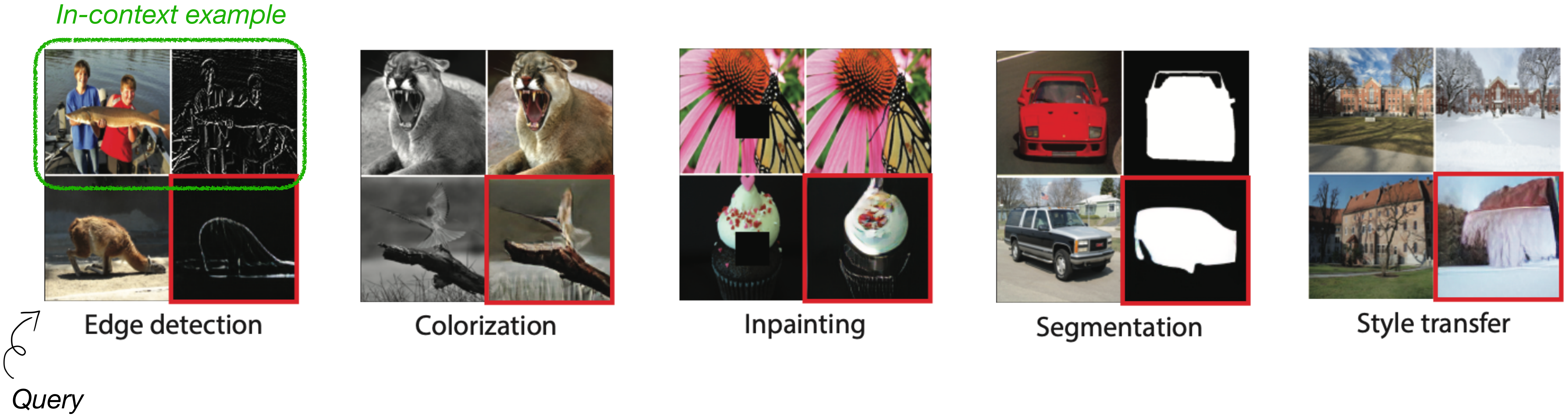
OpenAI DALL·E 2

PaLM 2

Midjourney

Visual in-context learning

User can customize models by "tuning" the in-context example(s)



Bar et al. Visual Prompting via Image Inpainting. 2022.

Visual in-context learning

- Train: Masked image modeling

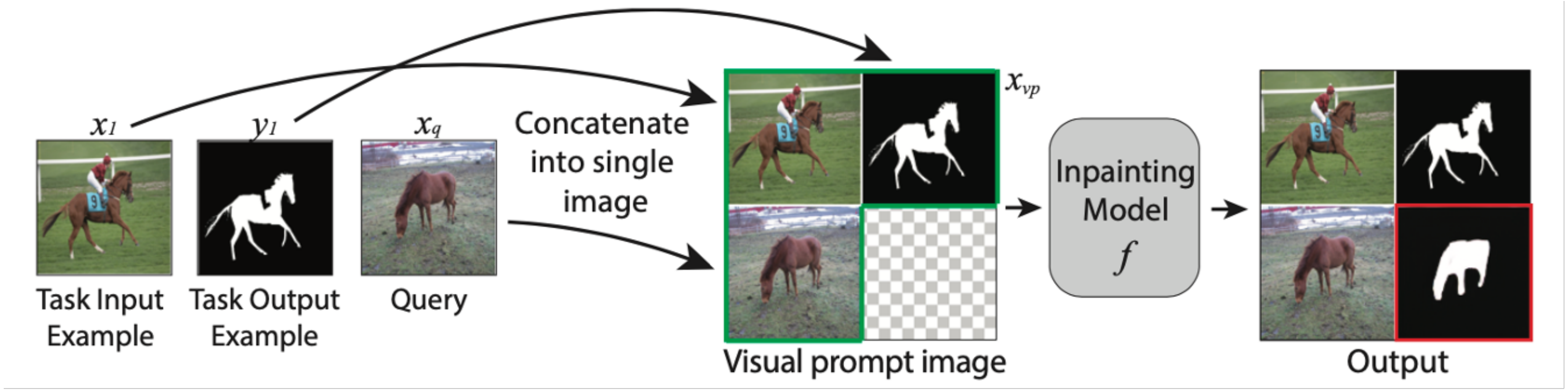
Key idea: Train the model to fill missing patches

*Training dataset: Computer Vision Figures, with 88k unlabeled **grid-like** images collected from computer vision papers*

Bar et al. Visual Prompting via Image Inpainting. 2022.

Visual in-context learning

- Test: In-context learning

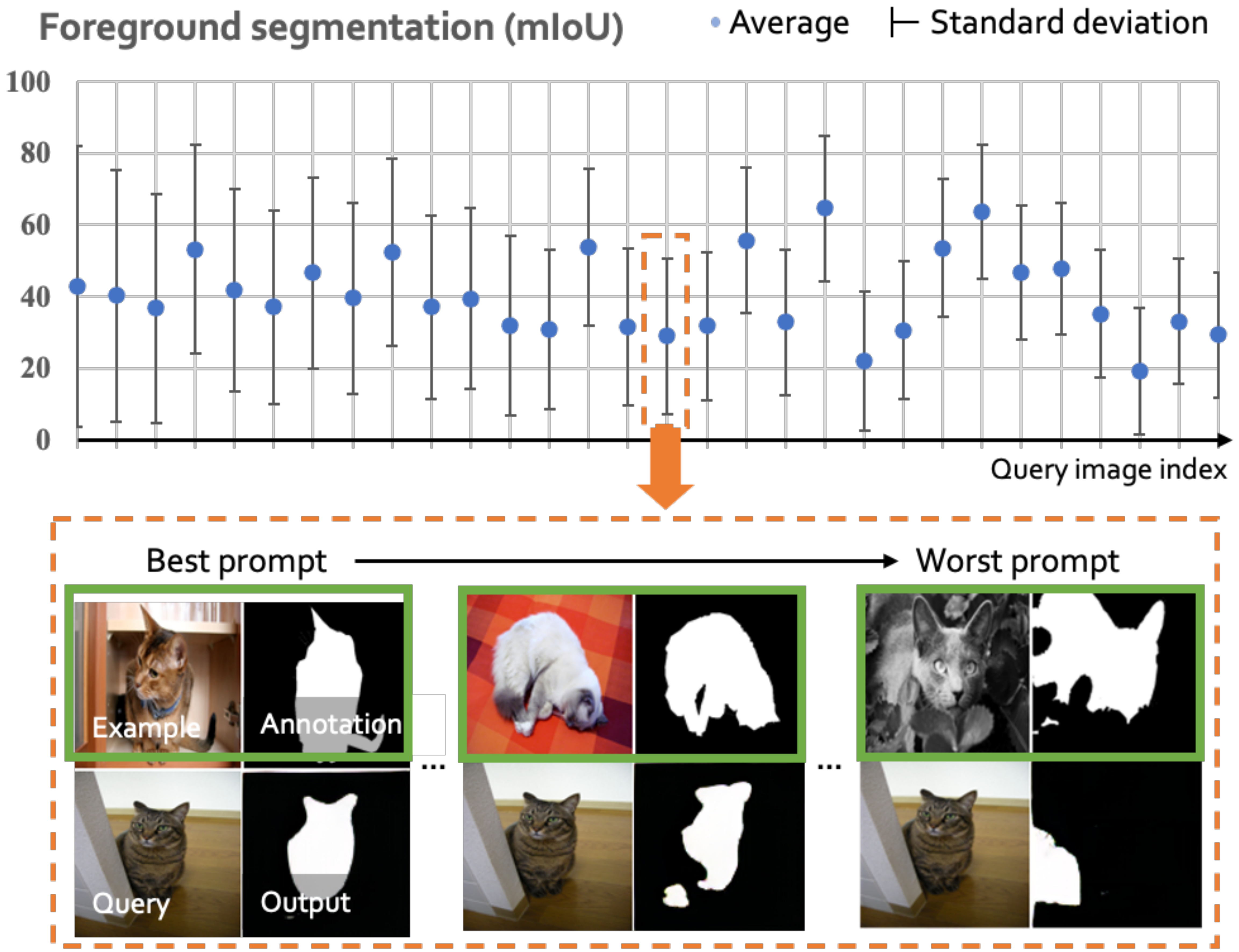


No parameter update!

Bar et al. Visual Prompting via Image Inpainting. 2022.

Visual in-context learning

- The choice of in-context examples matters a lot

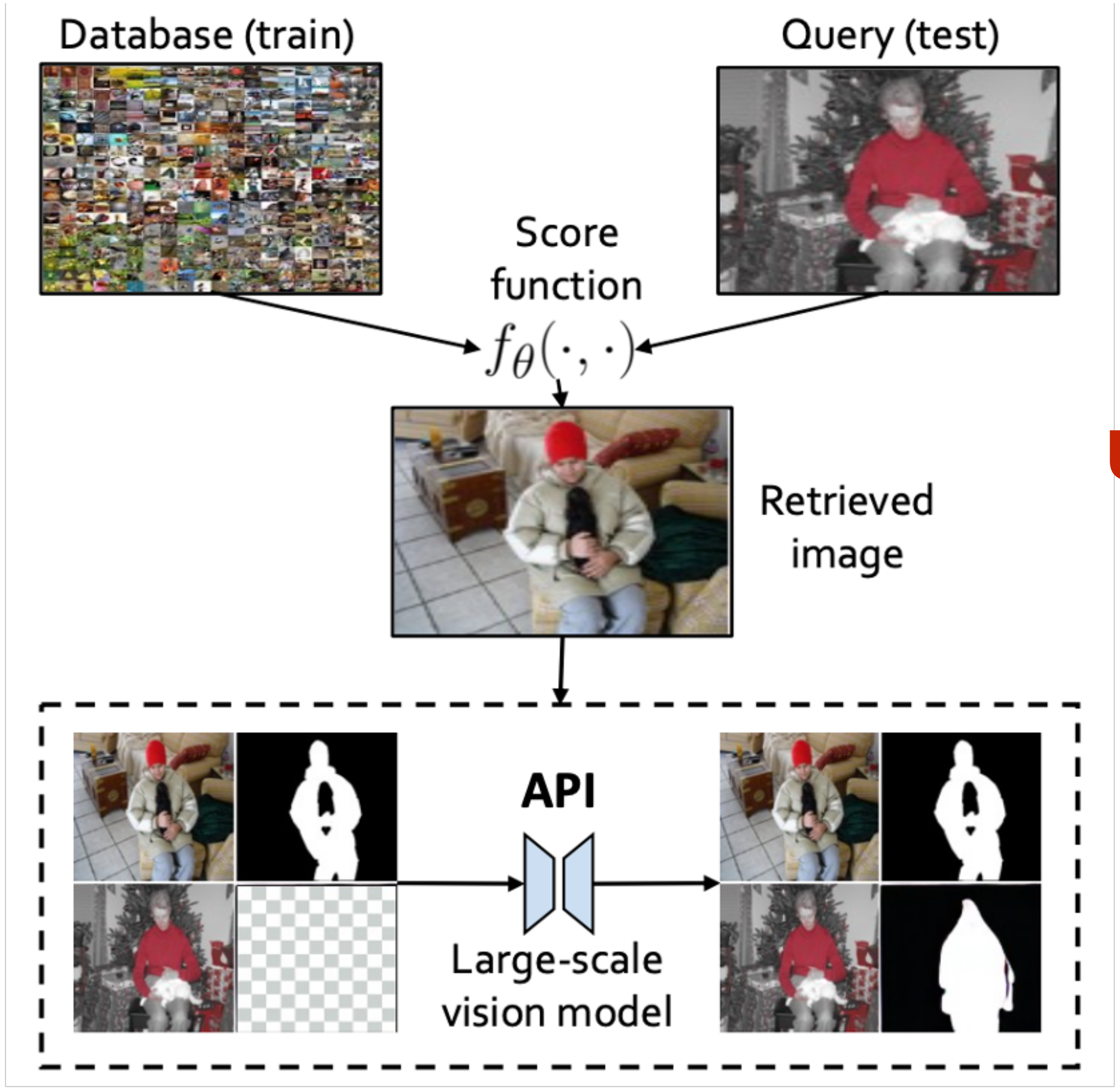


Random selection => large variances

Manual selection is time-consuming

Zhang et al. What Makes Good Examples for Visual In-Context Learning? 2023.

Prompt retrieval



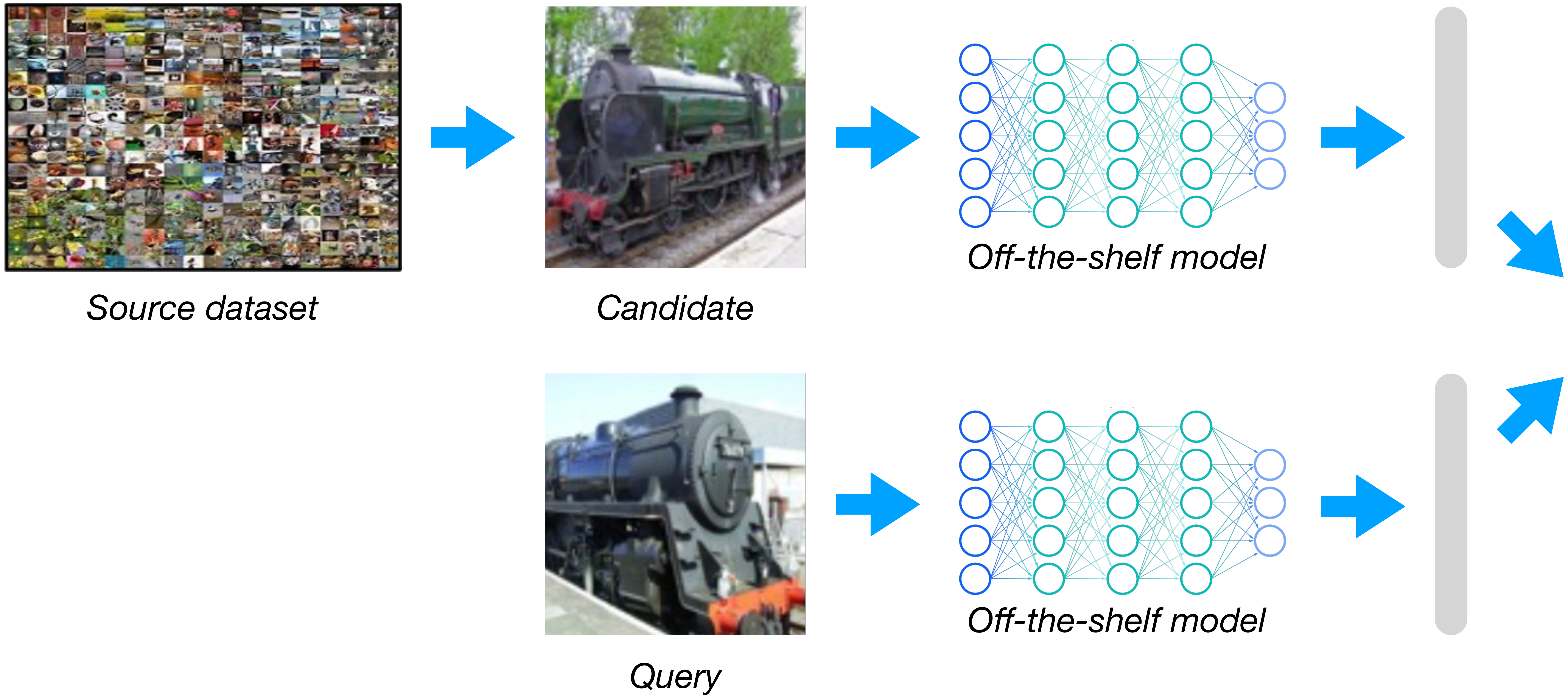
Use the API's output as supervision

$$x^* = \arg\max_{x_n \in \mathcal{D}} f_\theta(x_n, x_q)$$

Zhang et al. What Makes Good Examples for Visual In-Context Learning? 2023.

Unsupervised prompt retrieval

Semantic closeness



$$x^* = \arg\max_{x_n \in \mathcal{D}} f_\theta(x_n, x_q)$$

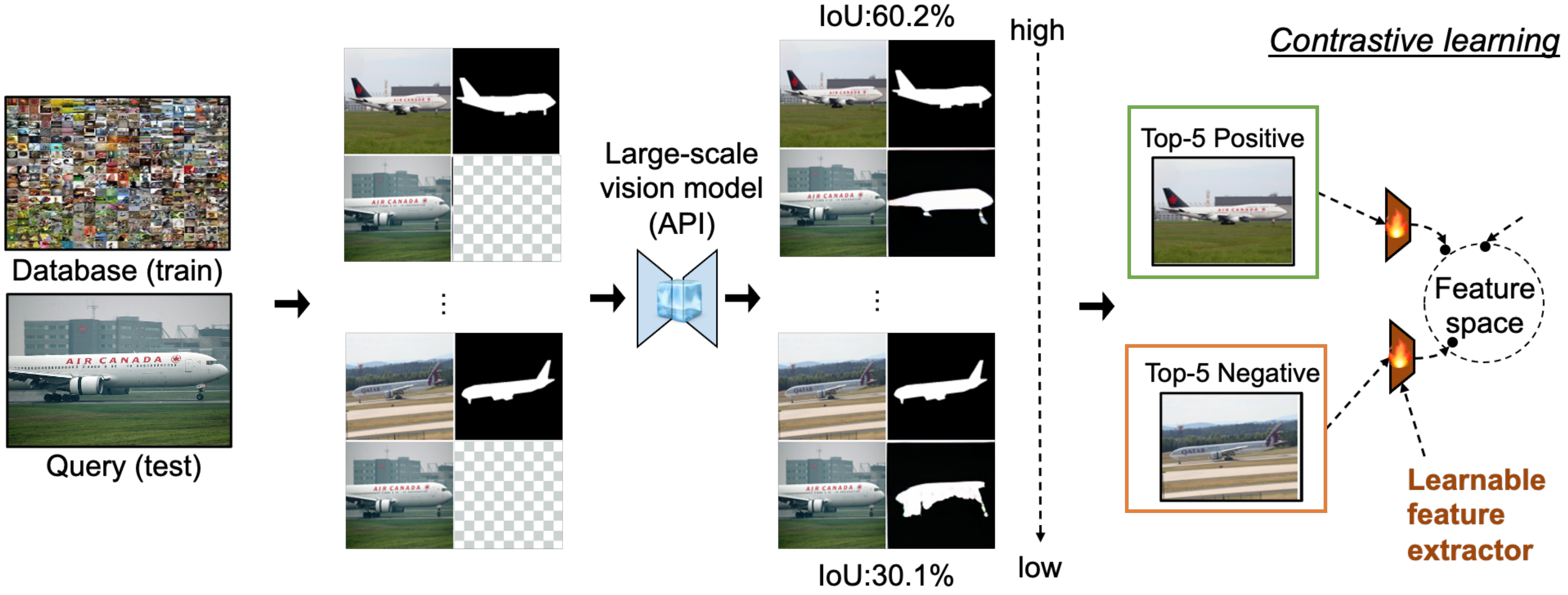
Find the closest

Zhang et al. What Makes Good Examples for Visual In-Context Learning? 2023.

Supervised prompt retrieval

Directly optimize in-context learning using a surrogate loss

$$\max_{\mathcal{P}} \quad \log p(y_q|\mathcal{P}, x_q)$$

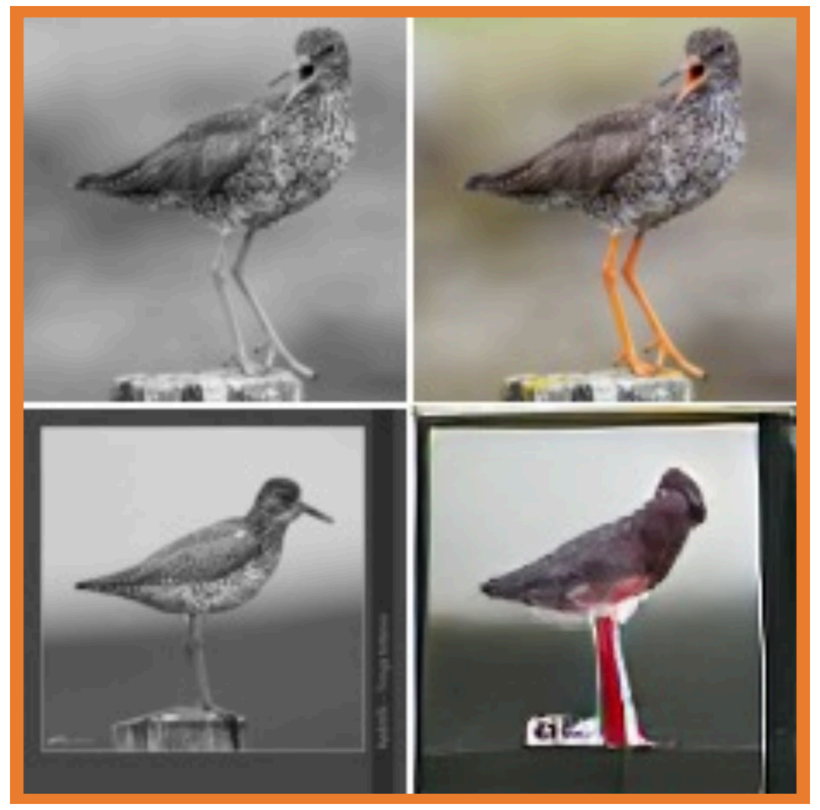


Zhang et al. What Makes Good Examples for Visual In-Context Learning? 2023.

Prompt retrieval vs. random selection

Foreground segmentation

Single object detection

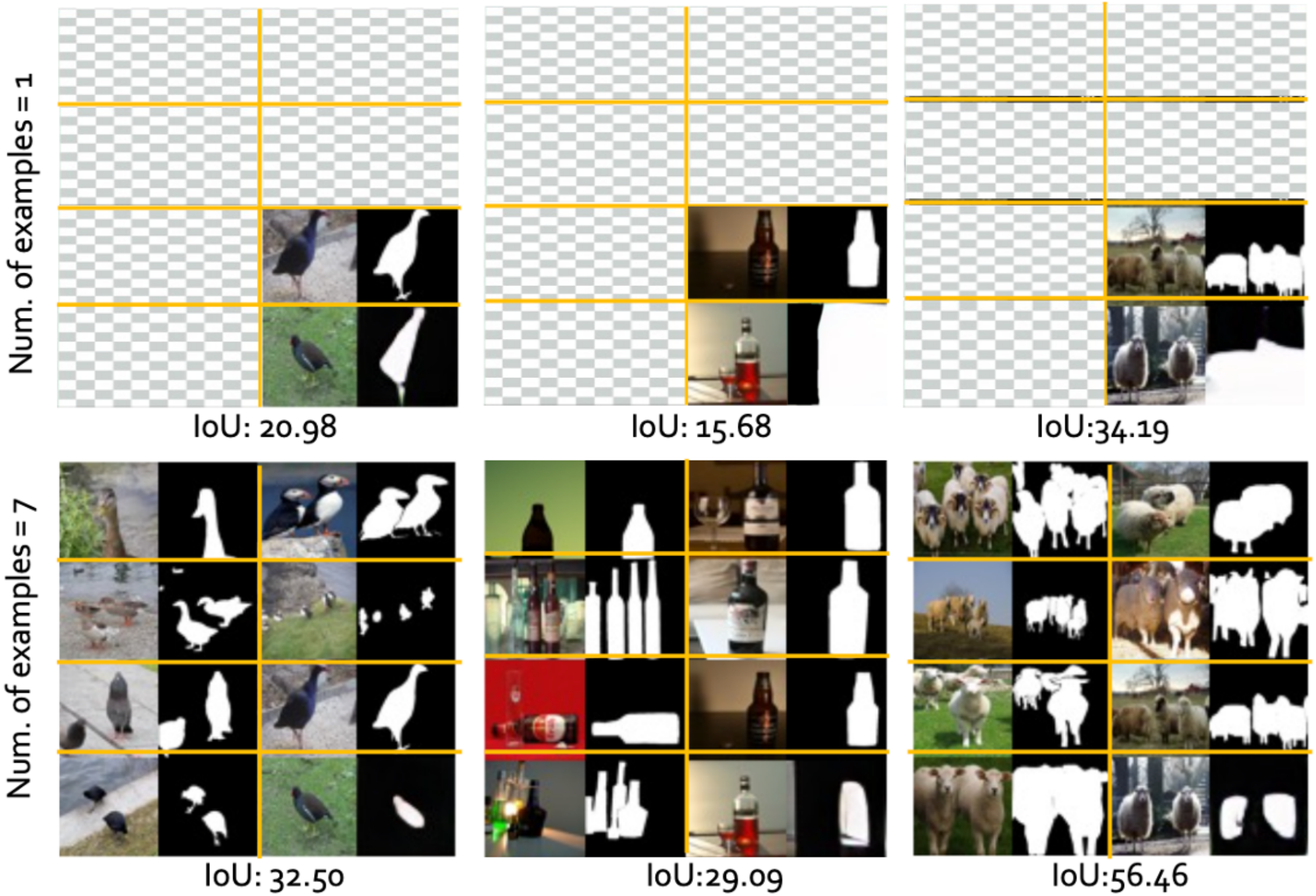


Colorization

	Seg. (mIOU) ↑	**Det.** (mIOU) ↑	**Color.** (mse) ↓
Random	27.56	25.45	0.67
UnsupPR	33.56	26.84	**0.63**
SupPR	**35.56**	**28.22**	**0.63**

Zhang et al. What Makes Good Examples for Visual In-Context Learning? 2023.

in-context examples: more is better



Zhang et al. What Makes Good Examples for Visual In-Context Learning? 2023.

Order of in-context examples: does not matter

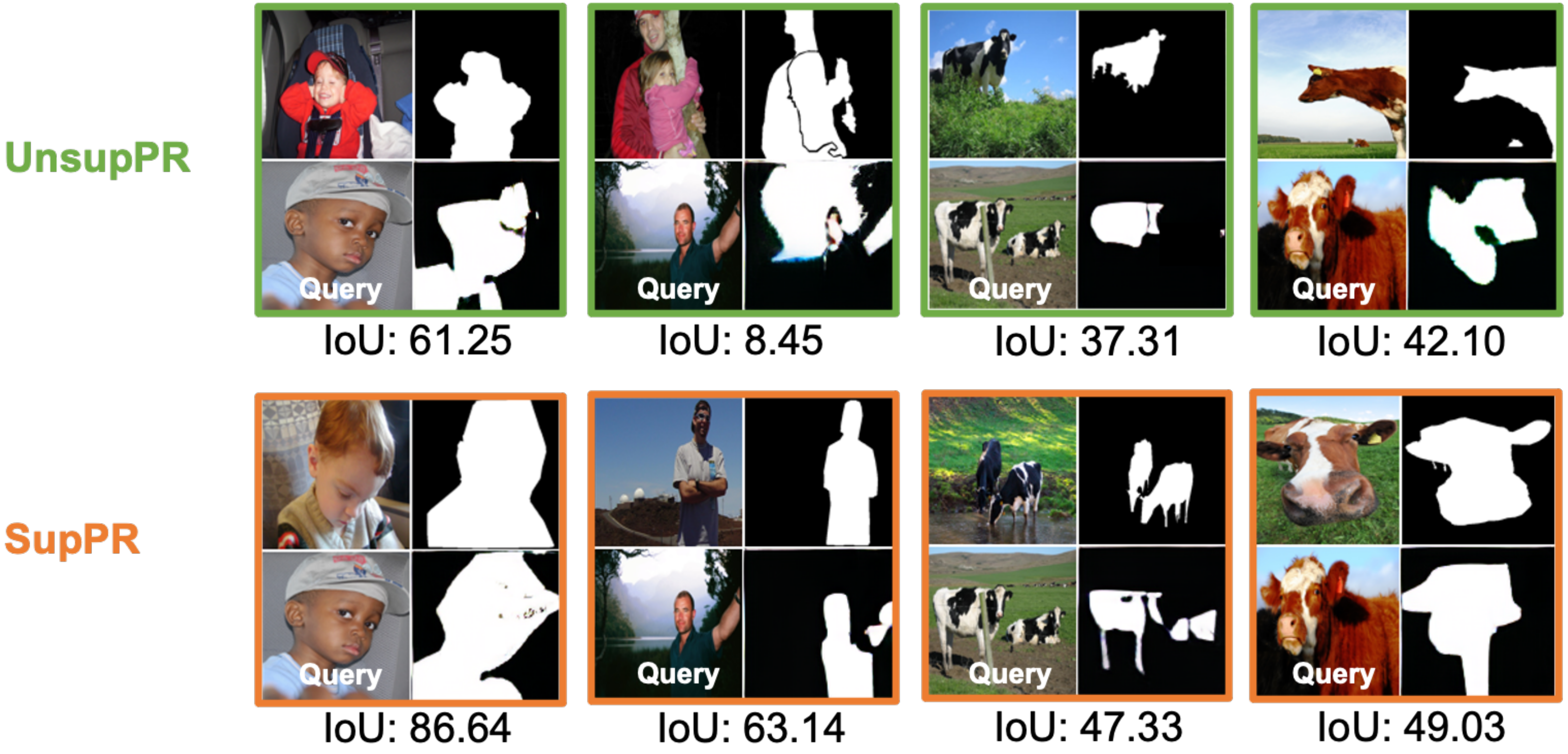
	Seg. (mIoU) ↑				
	Split-0	Split-1	Split-2	Split-3	Avg
Random	17.93 ± 0.20	25.48 ± 0.27	21.34 ± 0.73	21.12 ± 0.53	21.46 ± 0.43
UnsupPR	20.22 ± 0.31	27.58 ± 0.40	22.42 ± 0.38	23.36 ± 0.42	23.39 ± 0.37
SupPR	**20.74** ± 0.40	**28.19** ± 0.37	**23.09** ± 0.34	**24.22** ± 0.48	**24.06** ± 0.40

Variances of different orders

Zhang et al. What Makes Good Examples for Visual In-Context Learning? 2023.

What are good in-context examples?

Closeness in semantics, background, pose, appearance, view point, etc.



Zhang et al. What Makes Good Examples for Visual In-Context Learning? 2023.

Key takeaways

- Prompting has become a dominating paradigm in both NLP & CV
- Soft prompt learning in NLP & CV:
  - is data-efficient
  - is domain-generalizable
  - is difficult to interpret
- Conditional prompt learning works better but is slow to train
- Multimodal prompt learning offers better trade-offs
- Do neural prompt search if more compute is available
- Only APIs are available? Use their output as supervision

Prompting => conversational visual intelligence

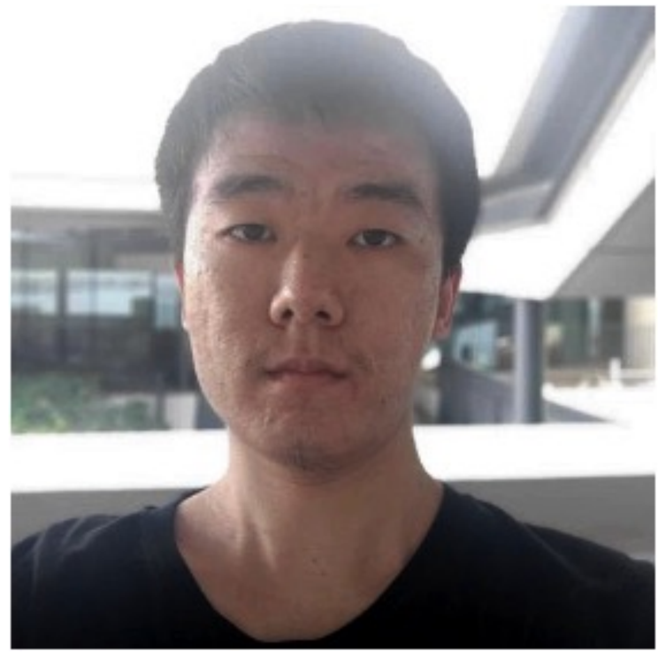
References

- Learning to Prompt for Vision-Language Models.
- Conditional Prompt Learning for Vision-Language Models.
- Unified Vision and Language Prompt Learning.
- Neural Prompt Search.
- What Makes Good Examples for Visual In-Context Learning?

Code: https://github.com/KaiyangZhou
Paper pdfs: https://kaiyangzhou.github.io/

Acknowledgement

Jingkang Yang



Yuhang Zang

Yuanhan Zhang

Ziwei Liu

Chen Change Loy

Thanks! Any question?