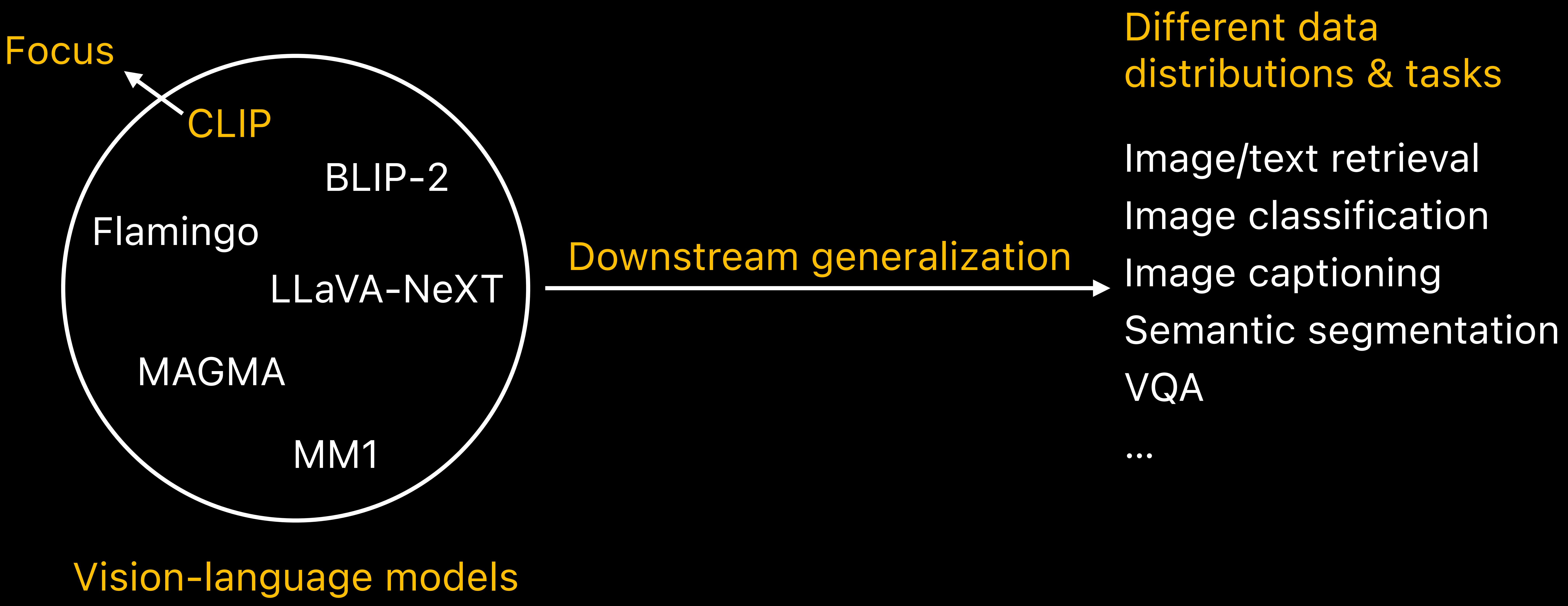
Aggregate-and-Adapt Natural Language Prompts for Downstream Generalization of CLIP

Chen Huang

NeurIPS | Apple | Dec 13, 2024

Goal

Improve downstream generalization of vision-language models



Focus

CLIP

BLIP-2

Flamingo

LLaVA-NeXT

MAGMA

MM1

Vision-language models

Downstream generalization

Different data distributions & tasks

Image/text retrieval

Image classification

Image captioning

Semantic segmentation

VQA

...

Challenges

- *Tail class concepts* are unseen or under-represented during model pretraining
- Limited downstream task data for model adaptation
- Domain gap between pretraining and downstream data

Jeep Compass SUV

Spyker C8 Convertible

Tail class (pretraining)

Downstream task →

?

Fine-grained car model

Idea 💡

Distill textual knowledge from natural language prompts
for downstream adaptation

Natural Language Prompts

LLM-generated prompts

For object-centric images

"The exterior of a Jeep Compass SUV 2012 is very sleek and modern."

⋮

"A Jeep Compass SUV 2012 has a boxy shape with a sloping roofline."

↑

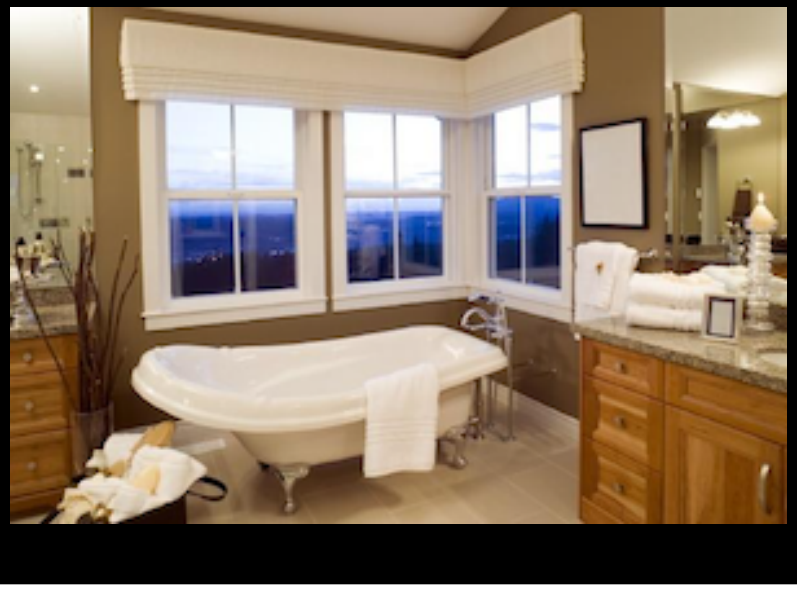
GPT or any LLM

↑

How can you identify a(n) Jeep Compass SUV 2012?

Human-generated prompts

For multi-object images



"a bathroom with a bath tub near windows"

⋮

"A bathroom scene is shown with a tub and counter."

↑

COCO image captions

Learn to Distill Task-Relevant Text Knowledge

AAPE (Aggregate-and-Adapted Prompt Embedding)

- $\mathscr{L}_{\text{distill}}$ — distill image-aligned, aggregated textual knowledge
- $\mathscr{L}_{\text{task}}$ — downstream task adaptation

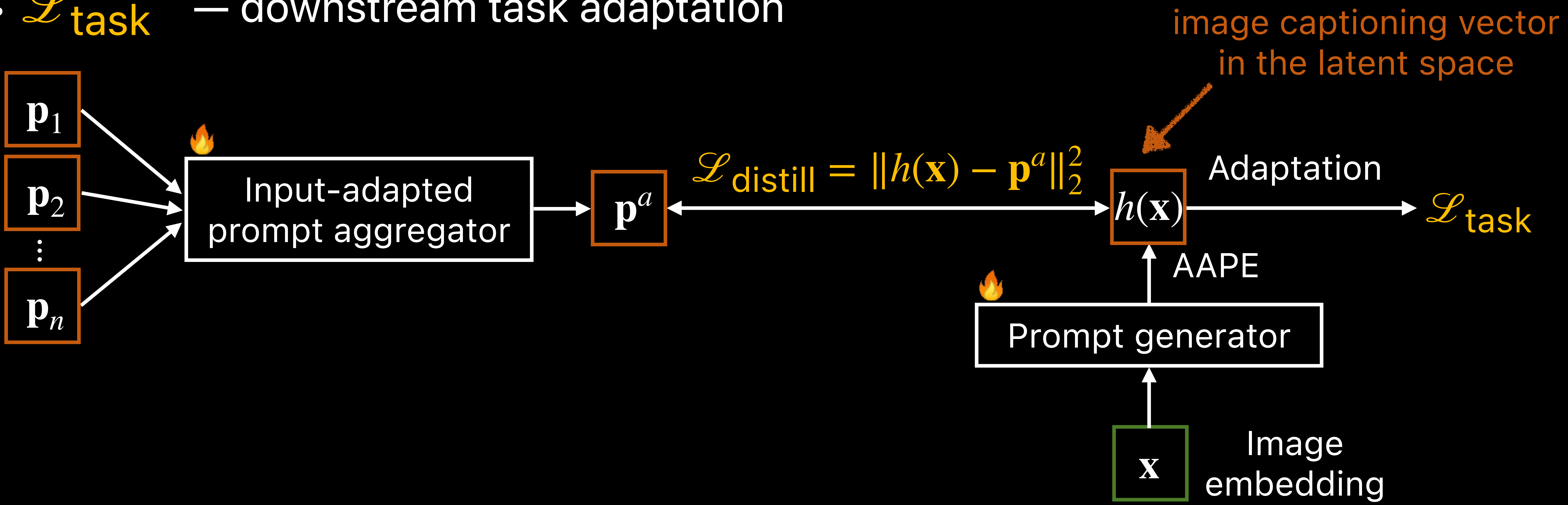
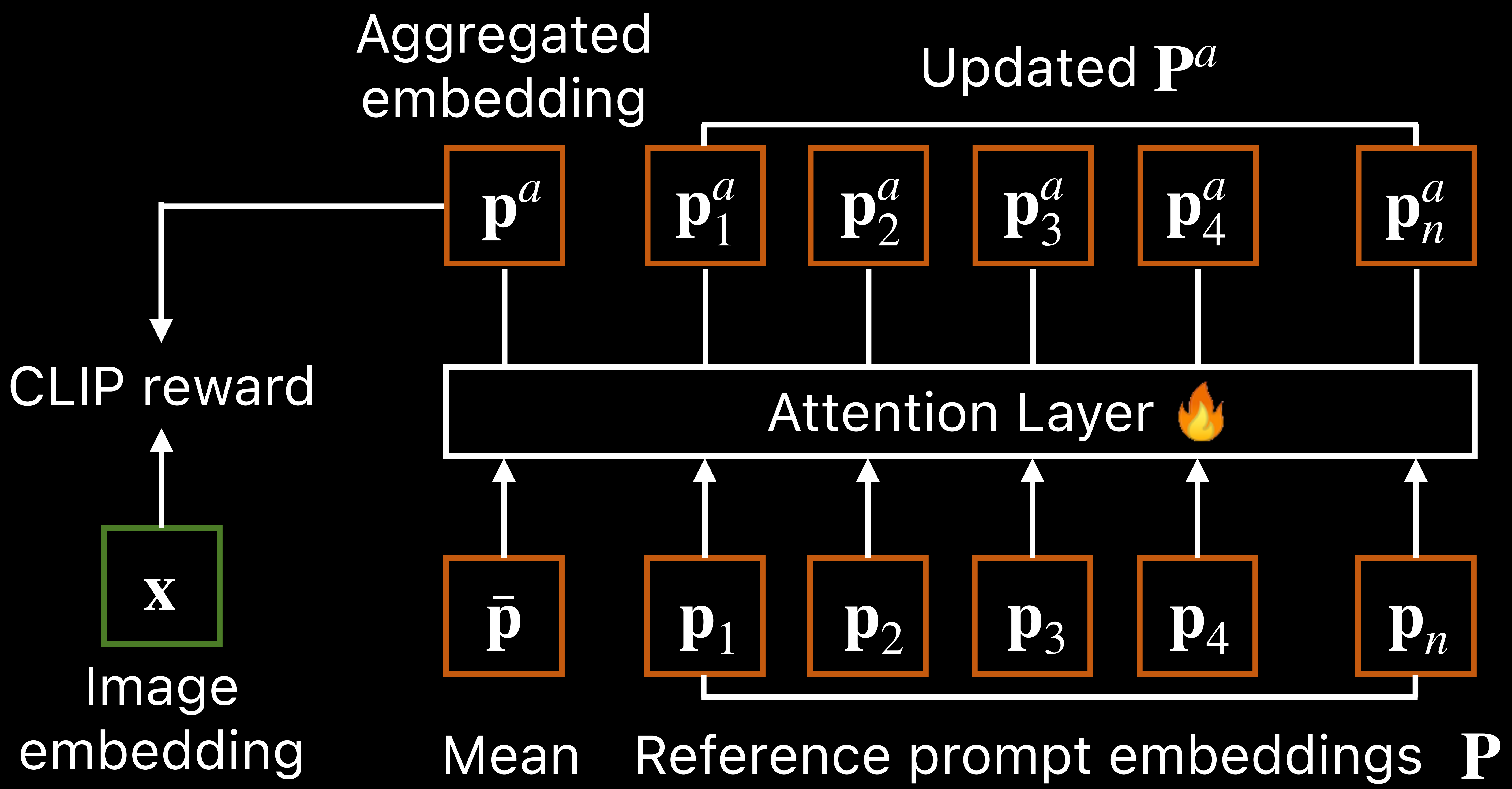


image captioning vector in the latent space
$\mathbf{p}_1$
$\mathbf{p}_2$
$\mathbf{p}_n$
Input-adapted prompt aggregator
$\mathbf{p}^a$
$\mathscr{L}_{\text{distill}} = \|h(\mathbf{x}) - \mathbf{p}^a\|_2^2$
$h(\mathbf{x})$
Adaptation
$\mathscr{L}_{\text{task}}$
AAPE
Prompt generator
$\mathbf{x}$
Image embedding

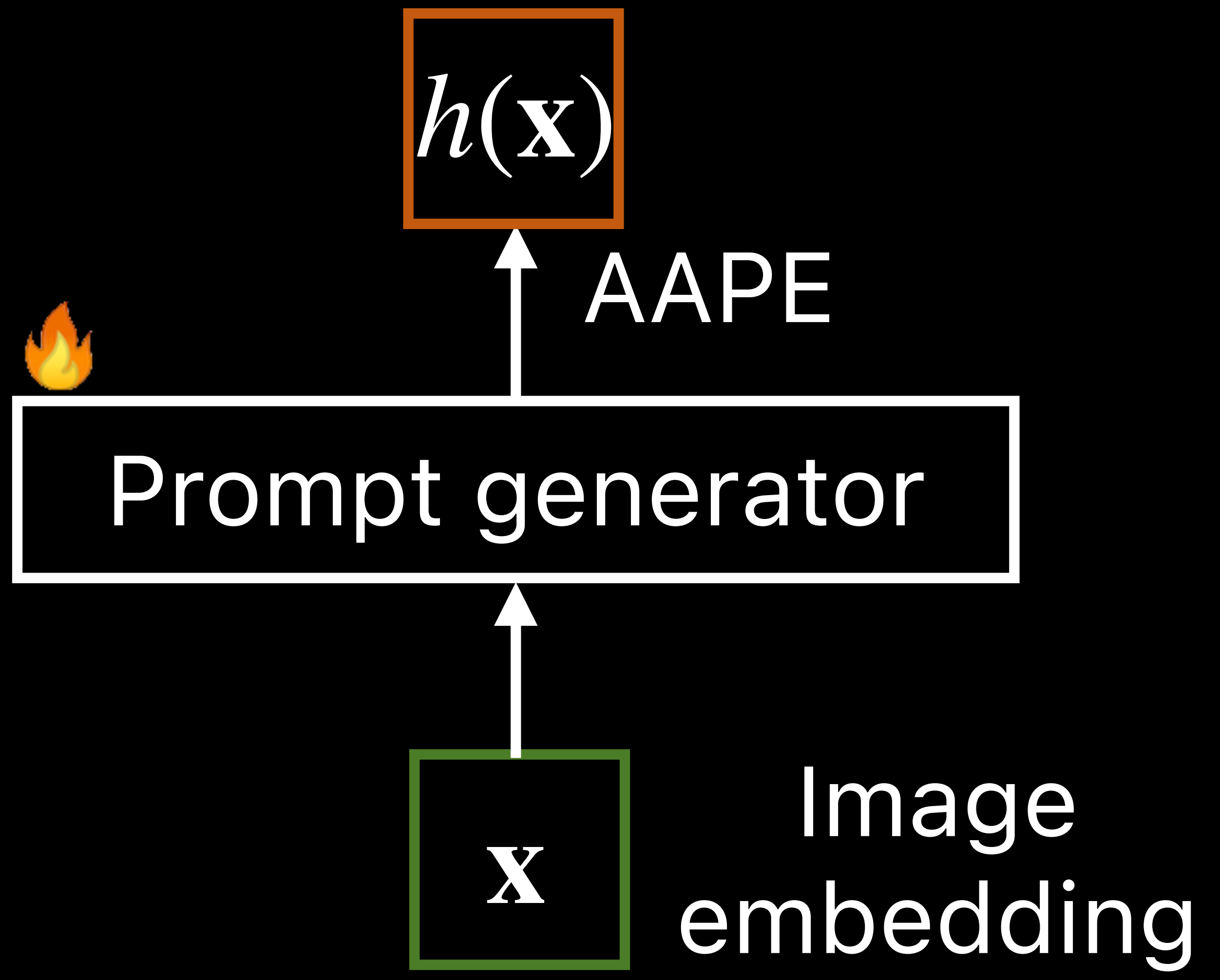
Input-Adapted Prompt Aggregator

Produce an image-aligned, condensed prompt summary



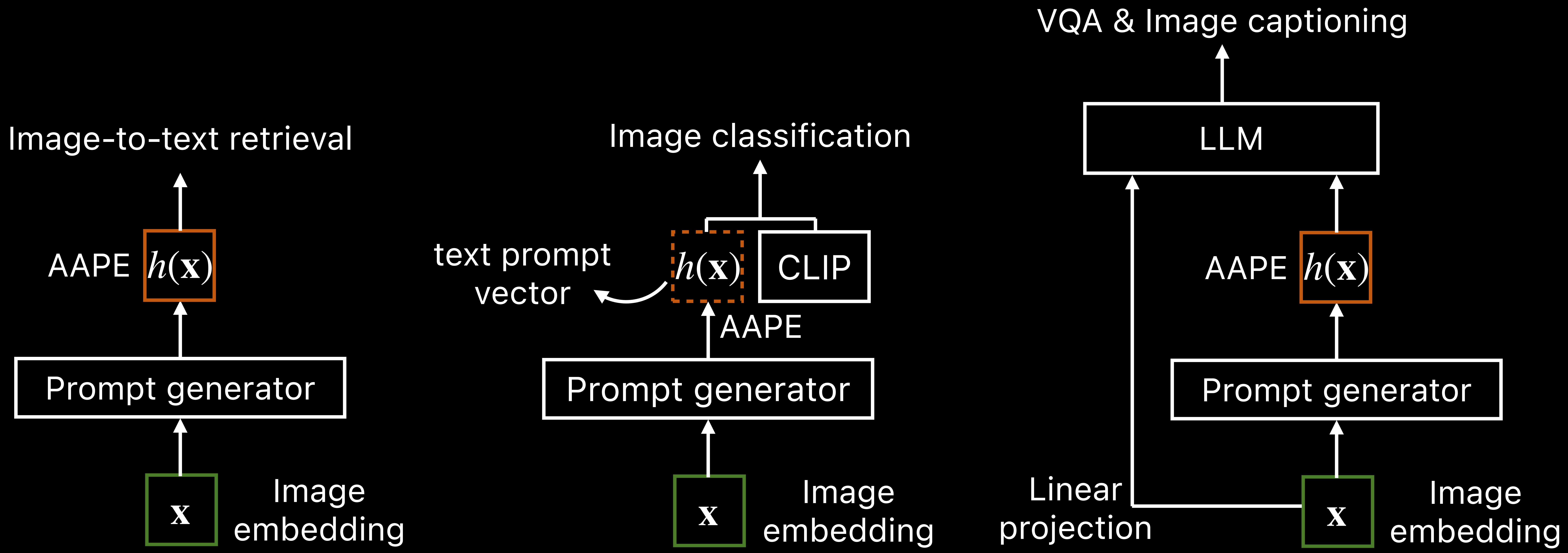
Inference

Keep the prompt generator (prompt aggregator discarded)



Inference

AAPE is applicable to different vision-language tasks



Results

Compelling performance on 4 vision-language tasks

- Example results on few-shot classification
  - Accuracy on seen classes during fine-tuning: downstream performance
  - Accuracy on held-out unseen classes: generalization performance

Average Accuracy across 11 downstream datasets	**Seen**	**Unseen**
State-of-the-art OGEN [ICLR'24]	84.17	76.86
AAPE	84.72	77.54

Conclusions

- Language priors are useful to improve the downstream generalization of CLIP
- AAPE achieves compelling performance on various downstream tasks, especially in few-shot and OOD tasks
- Future plans: apply AAPE learning to model pretraining and to more vision-language models