Hi there! Welcome to the talk. I'm presenting "Why are Visually-Grounded Language Models (VLMs) Bad at Image Classification" for NeurIPS 2024.

Why are Visually-Grounded Language Models Bad at Image Classification?

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, Serena Yeung (NeurIPS 2024)

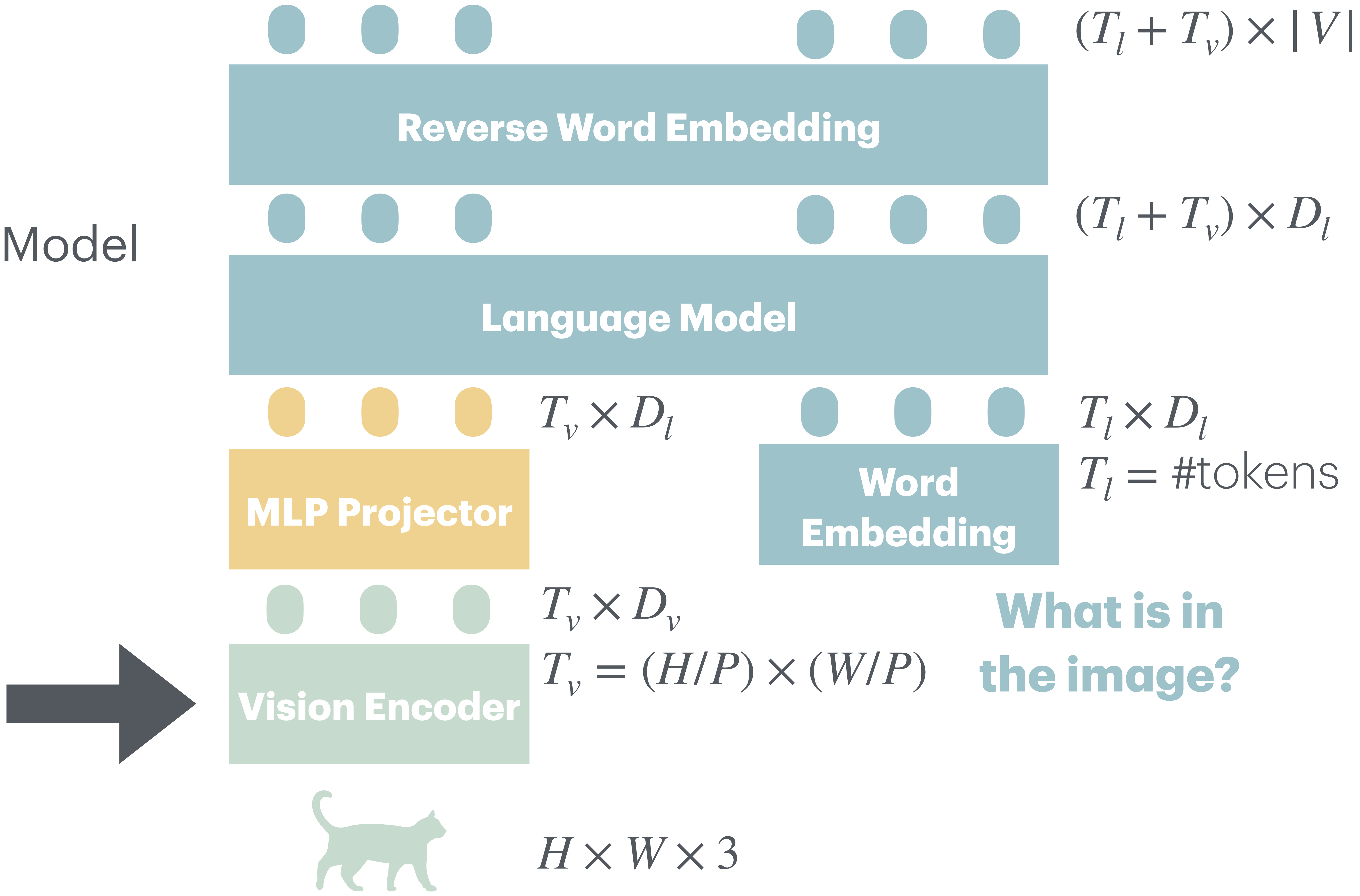
VLMs represent a family of models designed to learn a joint distribution of image and text tokens, typically through an auto-regressive approach. They generally include a vision encoder, paired with a language model, and a connecting projector.

Background

What is Visually-Grounded Language Model (VLM)?

- Language Model $p(x_t \mid x_{<t})$
- Visually-Grounded Language Model $p(x_t \mid x_{<t}, y)$, y is input image

Usually CLIP



VLMs have shown impressive versatility, excelling in tasks like image captioning, visual question answering, visual reasoning, and even more advanced agent-based applications.

Background

Visually-Grounded Language Models (VLM) Enable Many Capabilities

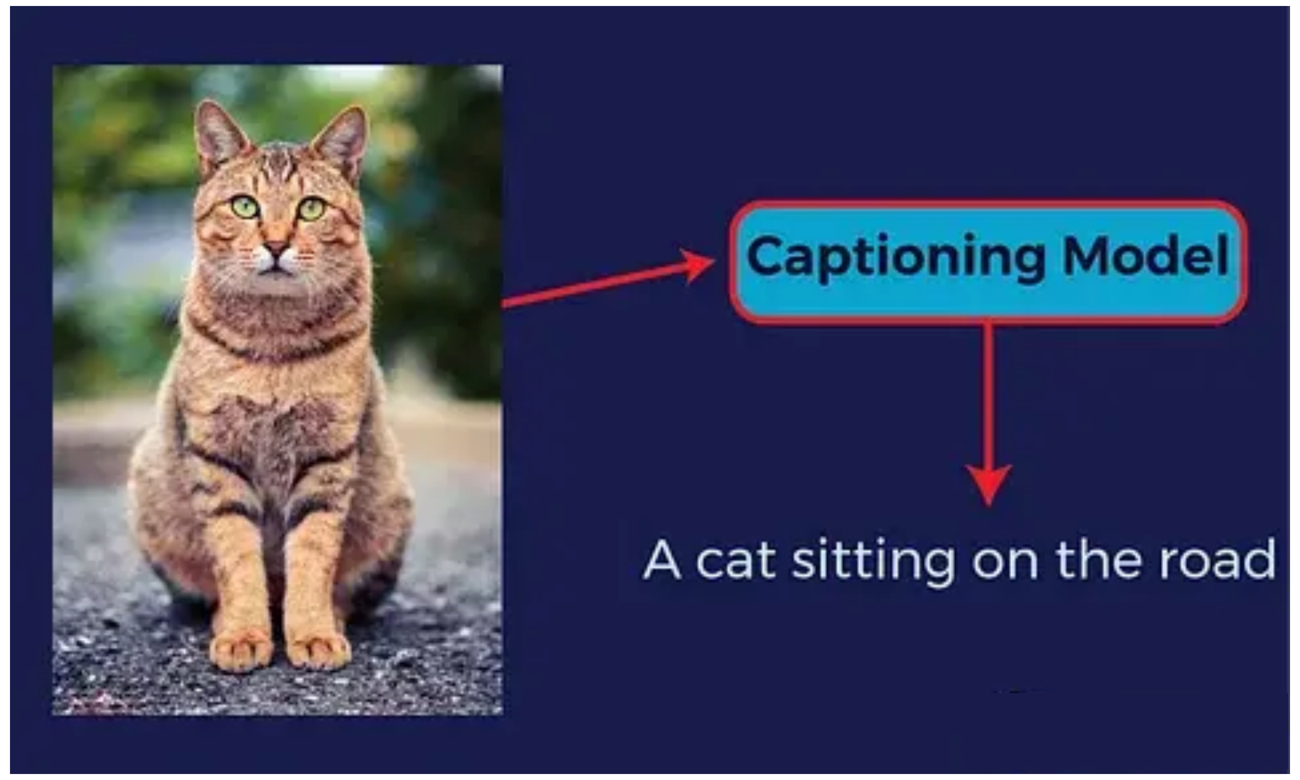
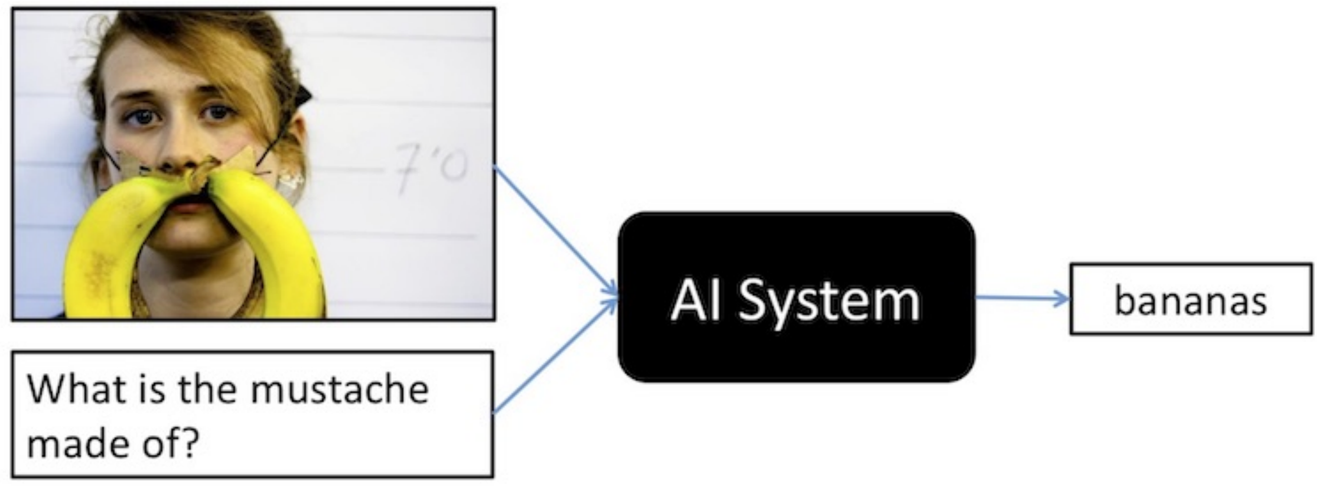
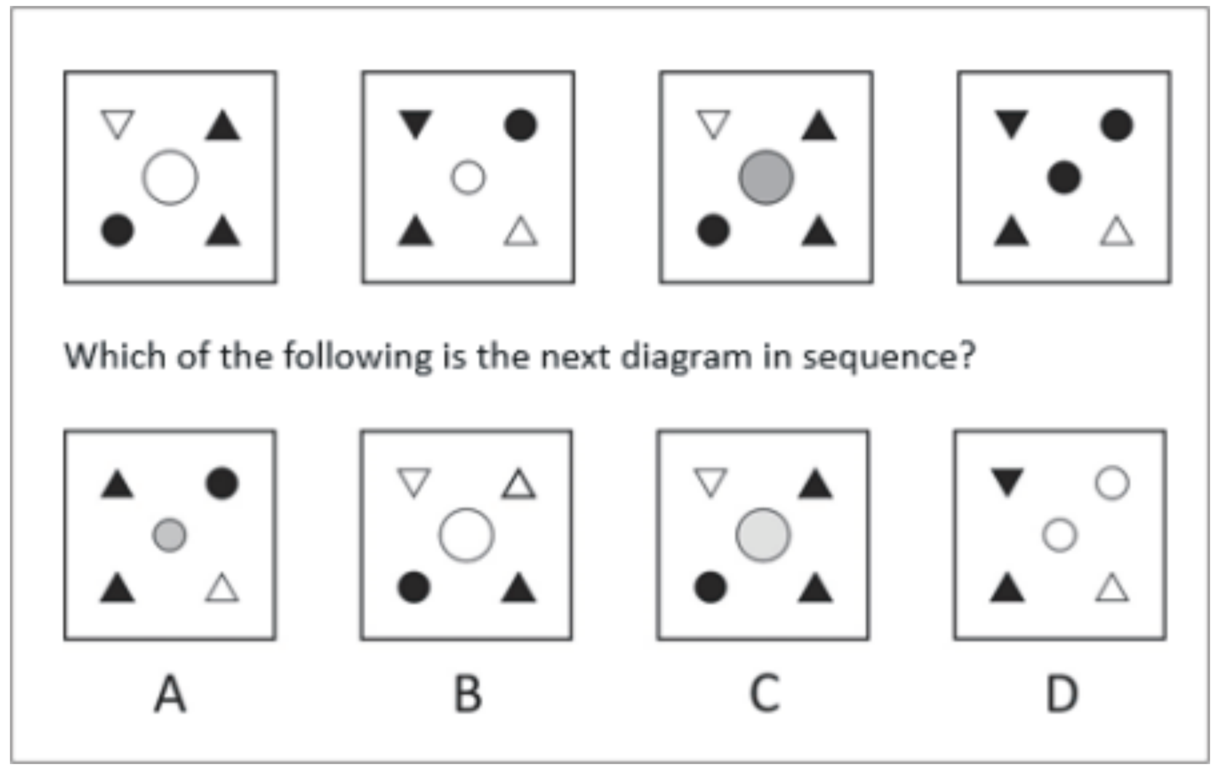


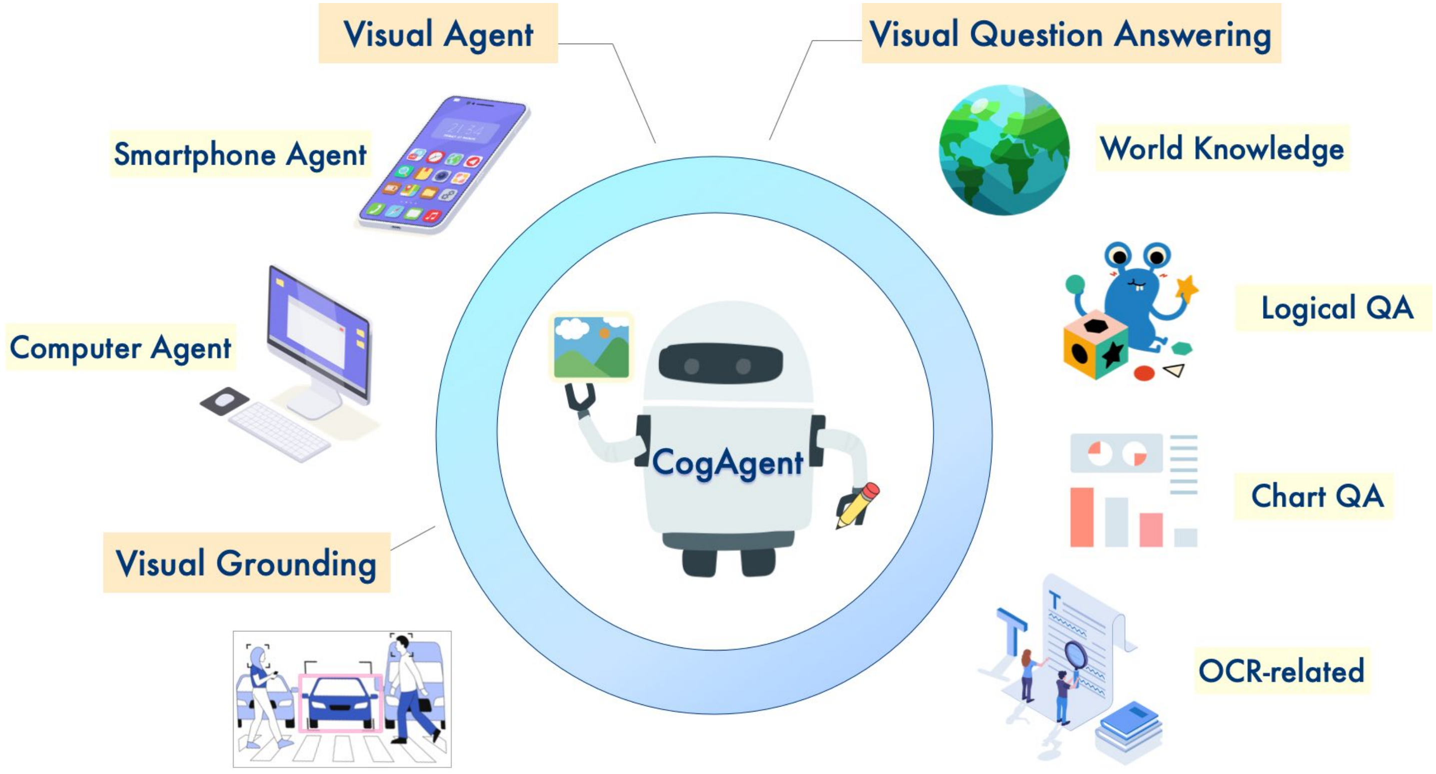
Image Captioning



Visual Question Answering



Visual Reasoning



Visual Agent

In this study, we revisit image classification using VLMs.

To use a VLM for image classification, we simply ask the model about the content of an image, choose from candidate classes, and check whether its response includes the correct class name.

How to Use VLM for Classification?

tench

CLIP

A photo of salmon → 0.83

A photo of catfish → 0.46 → ✓

A photo of tench → 0.84

VLM

What is in the image?

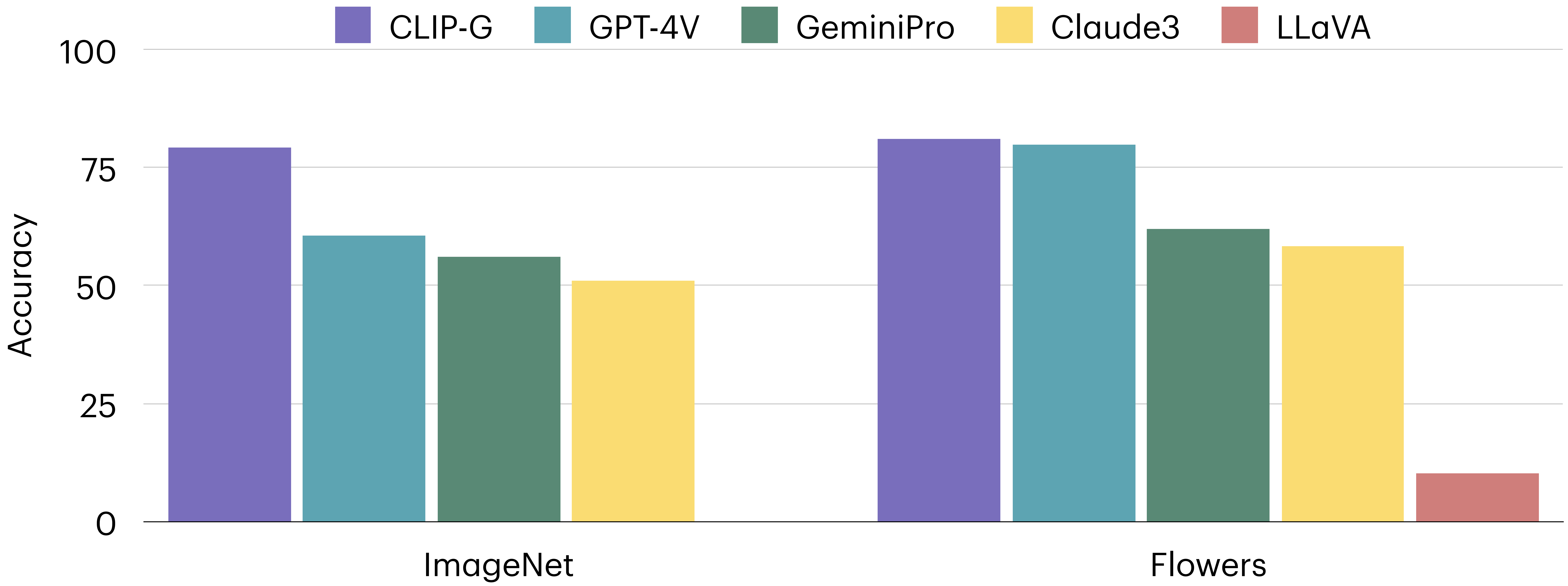
Choose one from

salmon, catfish, tench

→ This image shows a man holding a salmon... → ✗

Surprisingly, we found that VLMs like GPT-4V or Gemini perform significantly worse than CLIP on standard benchmarks such as ImageNet and Flowers.

VLMs are bad at Classification



VLMs are much worse than CLIP

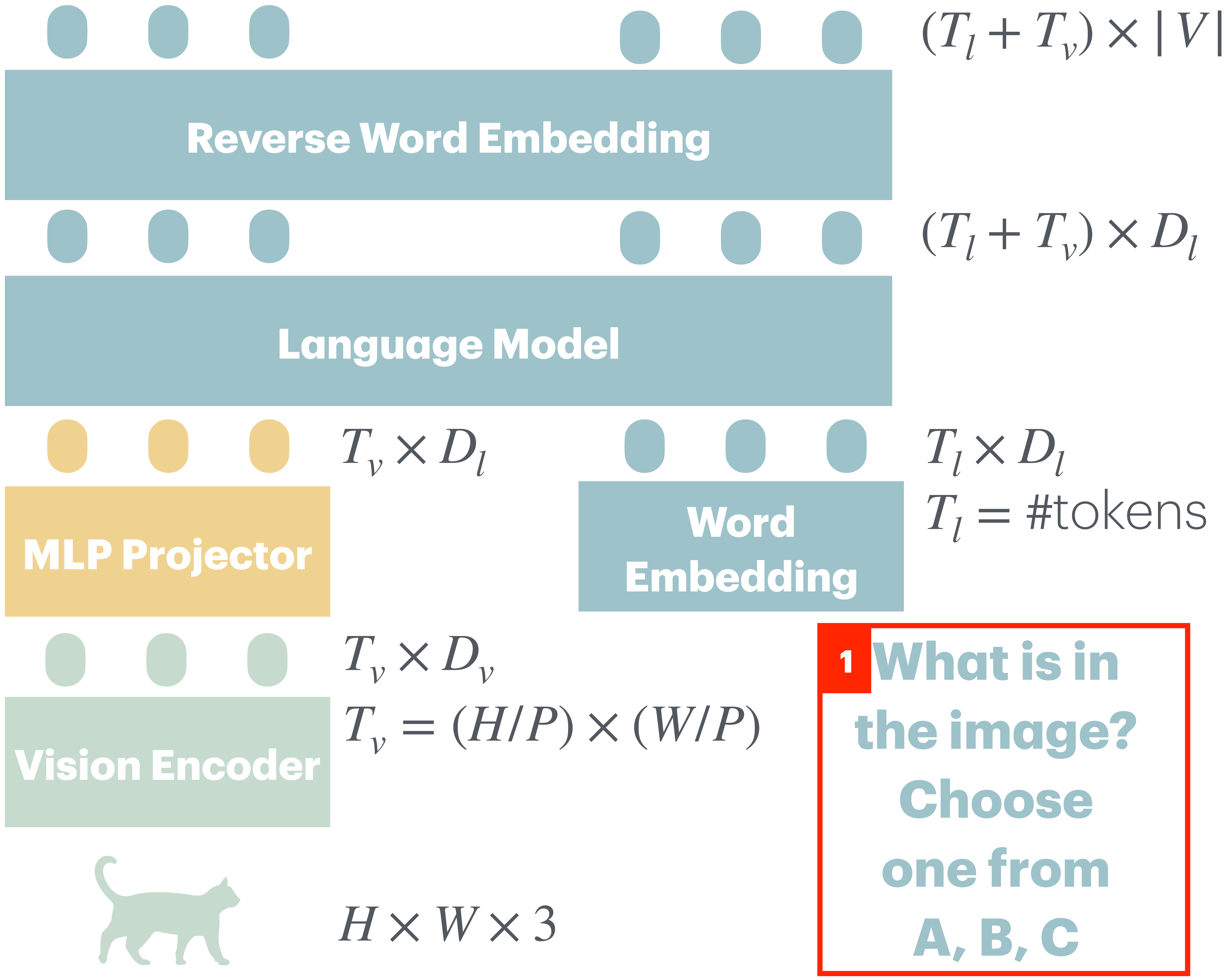
This raised a big question: why is this happening? We rigorously investigated this question through a series of hypotheses, each carefully tested.

Why are VLMs Bad at Classification?

Our first hypothesis focused on prompt variation. We wondered if poor prompt design might be the issue, so we experimented with a variety of prompts, including more advanced chain-of-thought prompts.

Hypothesis Space

1. Prompt Variation



The results showed that prompt variation had very limited impact, indicating that prompts alone weren't the problem.

1. Prompt Variation

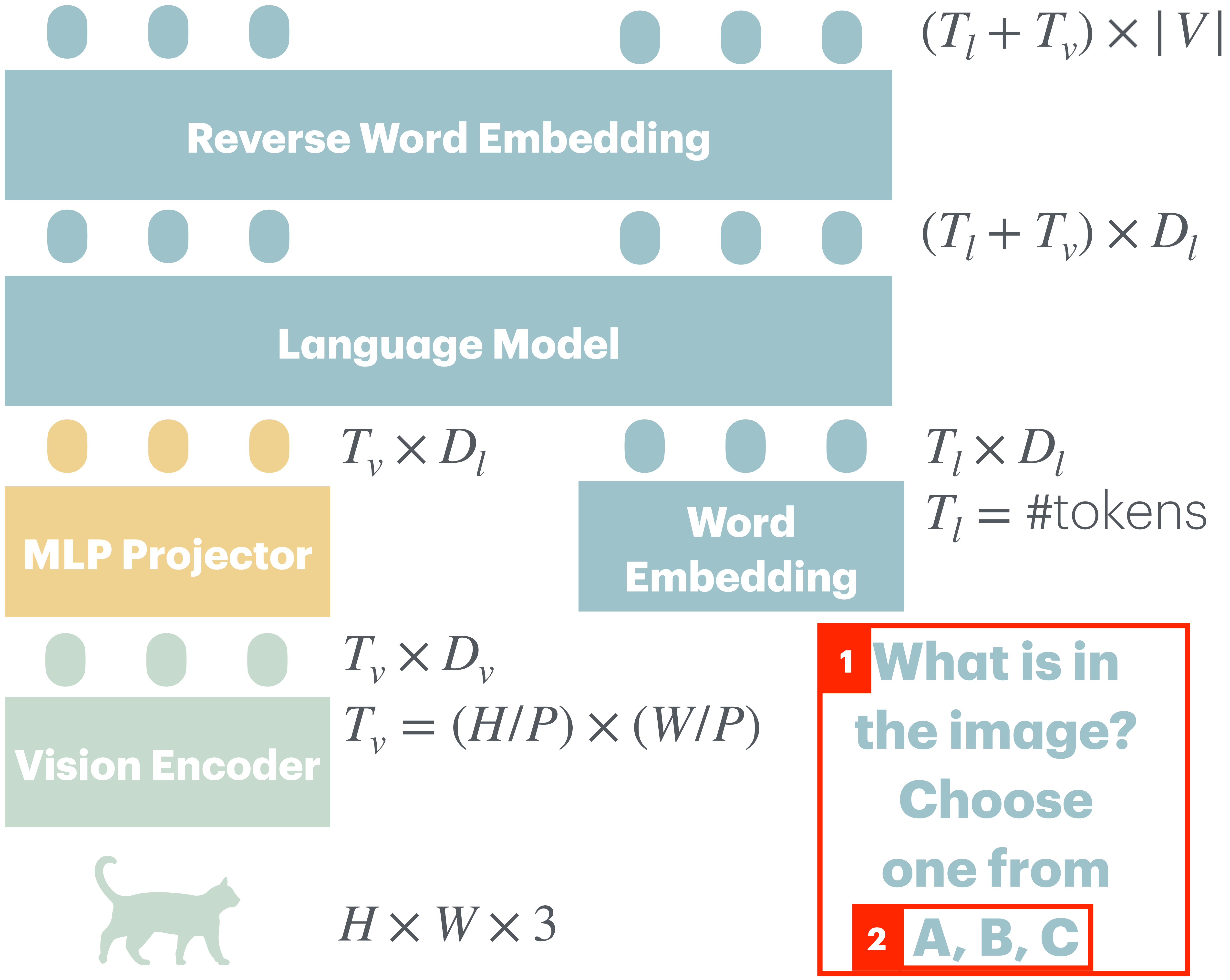
Prompt	ImageNet	Flowers
Default Prompt "What type of object is in this photo?"	22.8	5.9
Alternative Prompt "What is the main object depictured in this photo?"	21.6	6.6
+ CoT "Let's think step by step"	N/A	18.1
CLIP	74.8	76.0

Prompt has a limited impact

The second hypothesis addressed label space size. We thought perhaps having too many classes in classification benchmarks, like the 1,000 classes in ImageNet, might be a limiting factor. We tested this by reducing the number of classes.

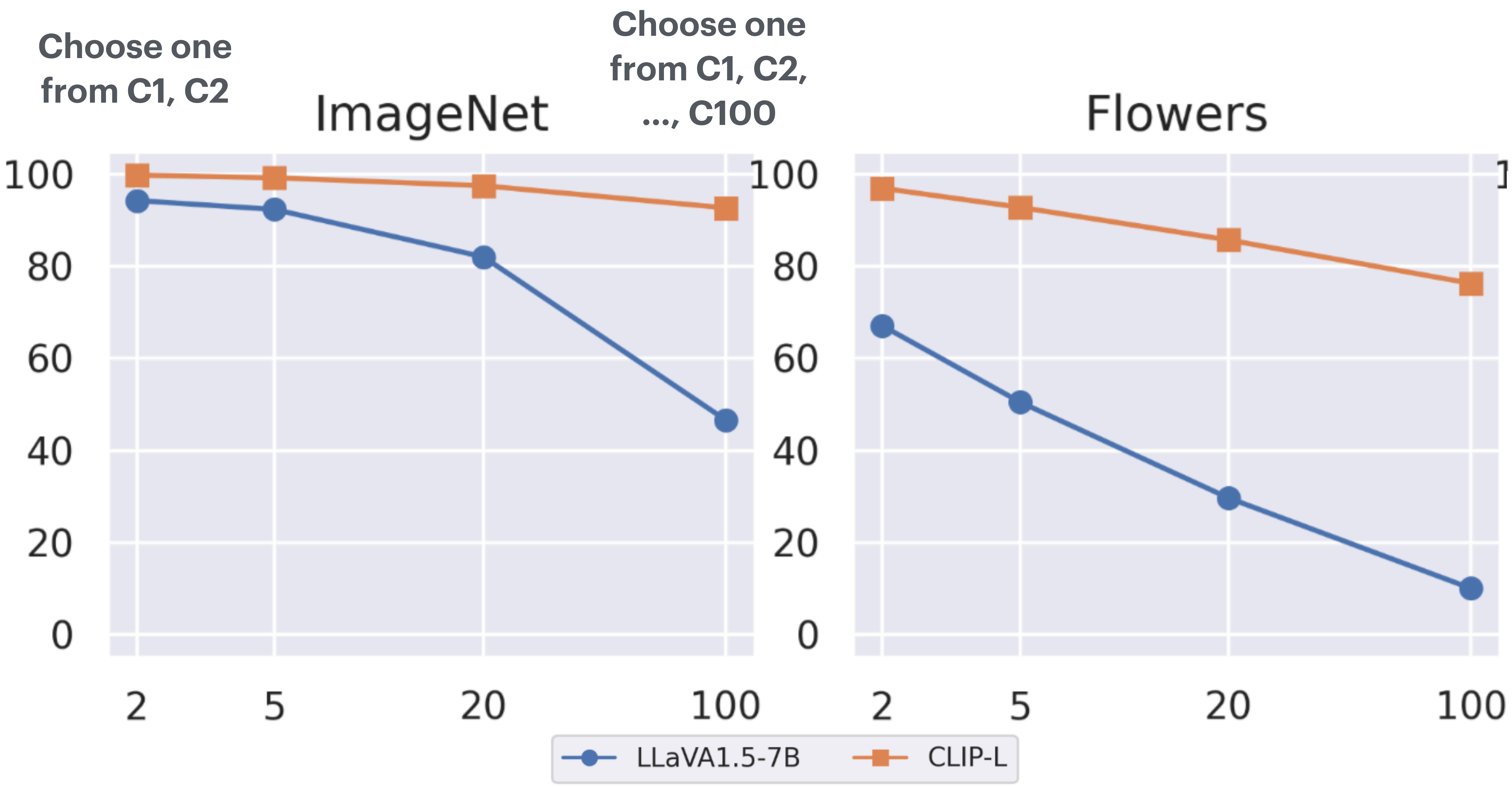
Hypothesis Space

1. ~~Prompt Variation~~
2. Label Space



When reducing from 100 to just 2 classes, we saw the gap between LLaVA and CLIP narrow slightly, but it still persisted—even in binary classification.

2. Label Space

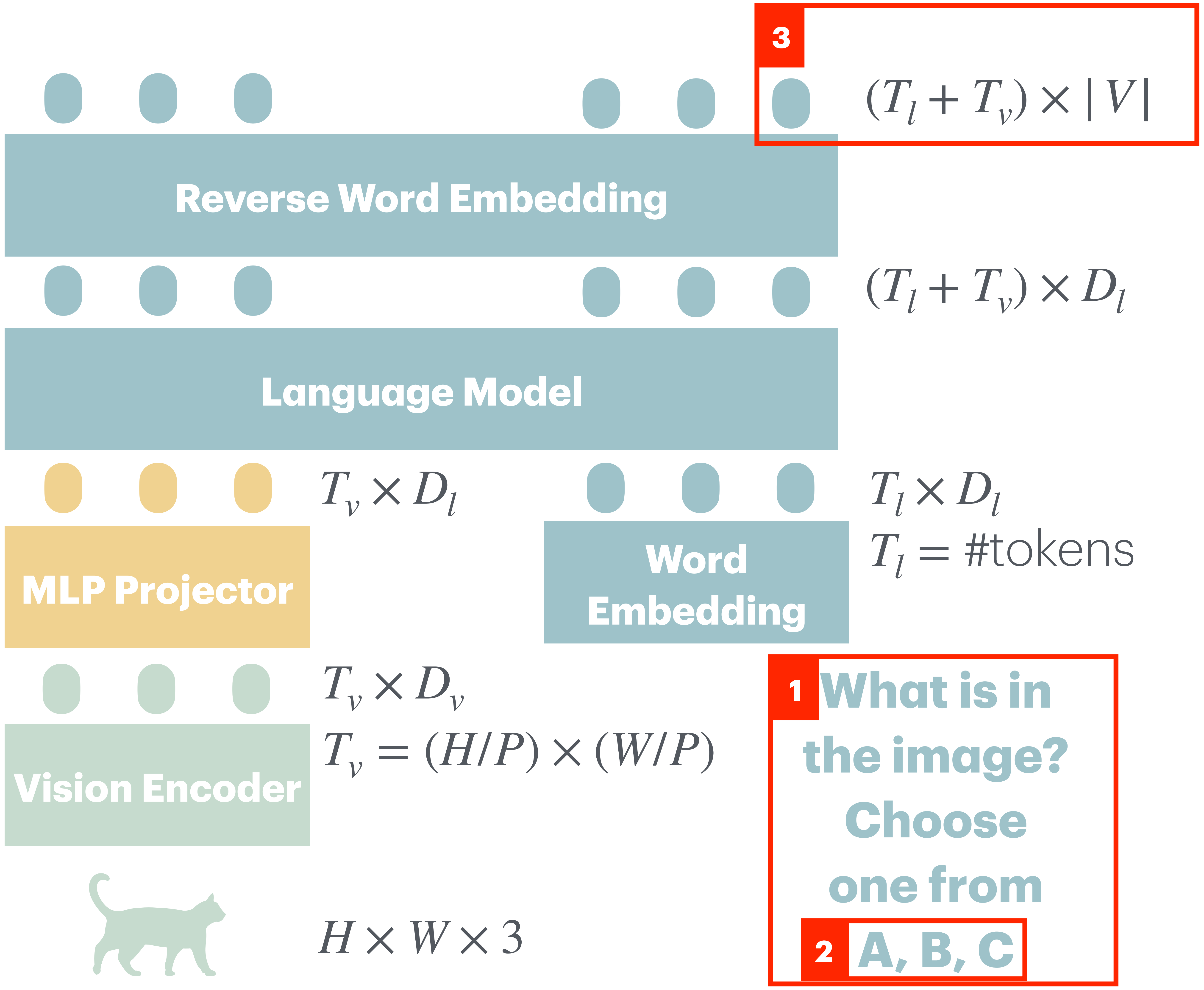


Gap narrows but always exists (even N=2)

Our third hypothesis examined the inference method. Asking the VLM to generate exact class names might be too challenging, as it might produce synonyms. We adjusted our approach to measure the probability of each class name.

Hypothesis Space

1. ~~Prompt Variation~~
2. ~~Label Space~~
3. Inference Algorithm



We used the probability inference technique to select the highest probability of all the class names. While this probability-based approach improved accuracy, the gap between LLaVA and CLIP remained.

3. Inference Algorithm

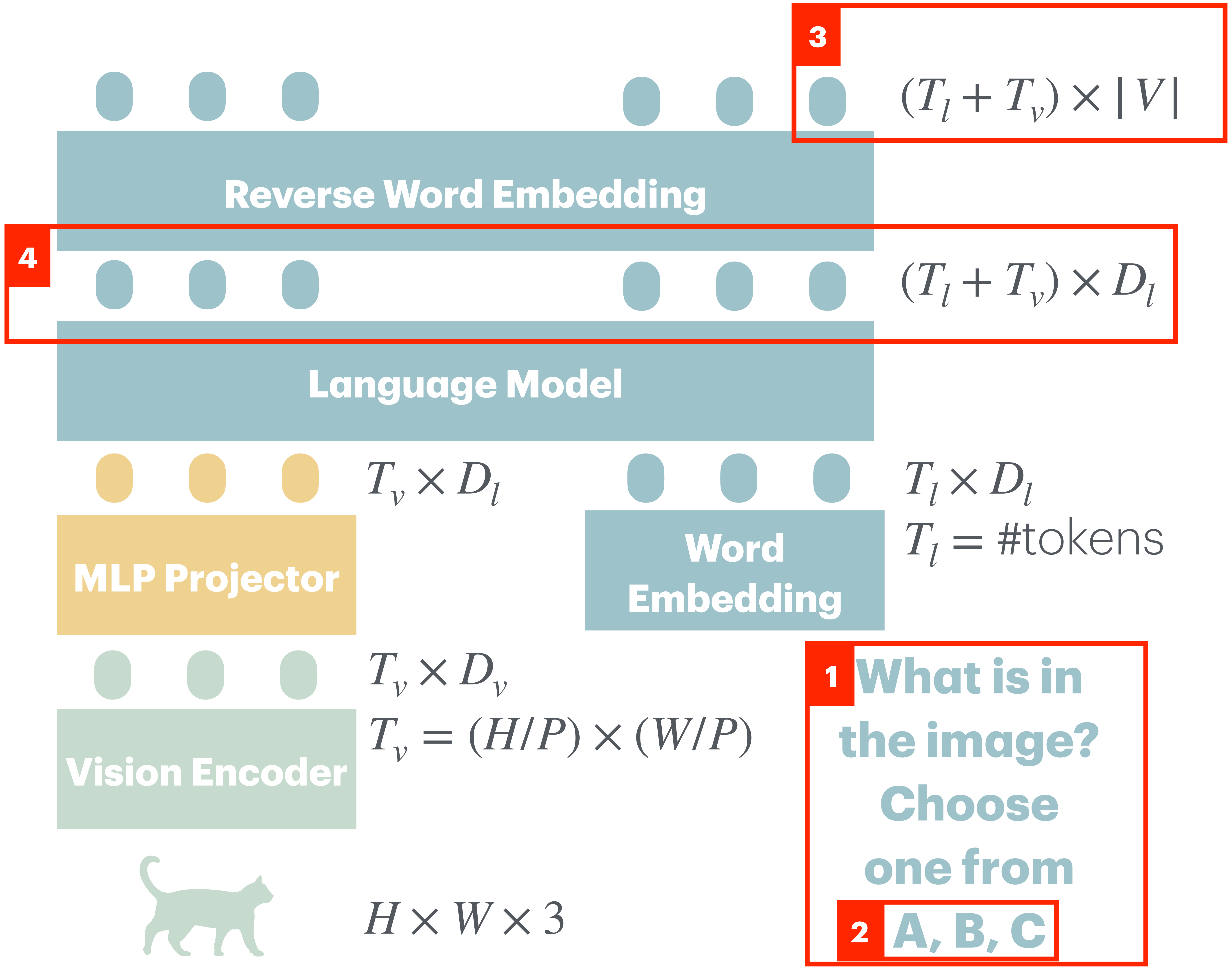
Prompt	ImageNet	Flowers
Direct Generation O(1) (Success: whether label in generation)	22.8	5.9
Probabality Inference O(N) (Success: whether p(label | image, prompt) is highest)	35.3	16.5
CLIP	74.8	76.0

Probabilistic inference improves but gap persists

The fourth hypothesis was about information loss. CLIP's encoder fully encodes information for classification, but it's unclear if information degrades as it propagates through multiple LLM layers.

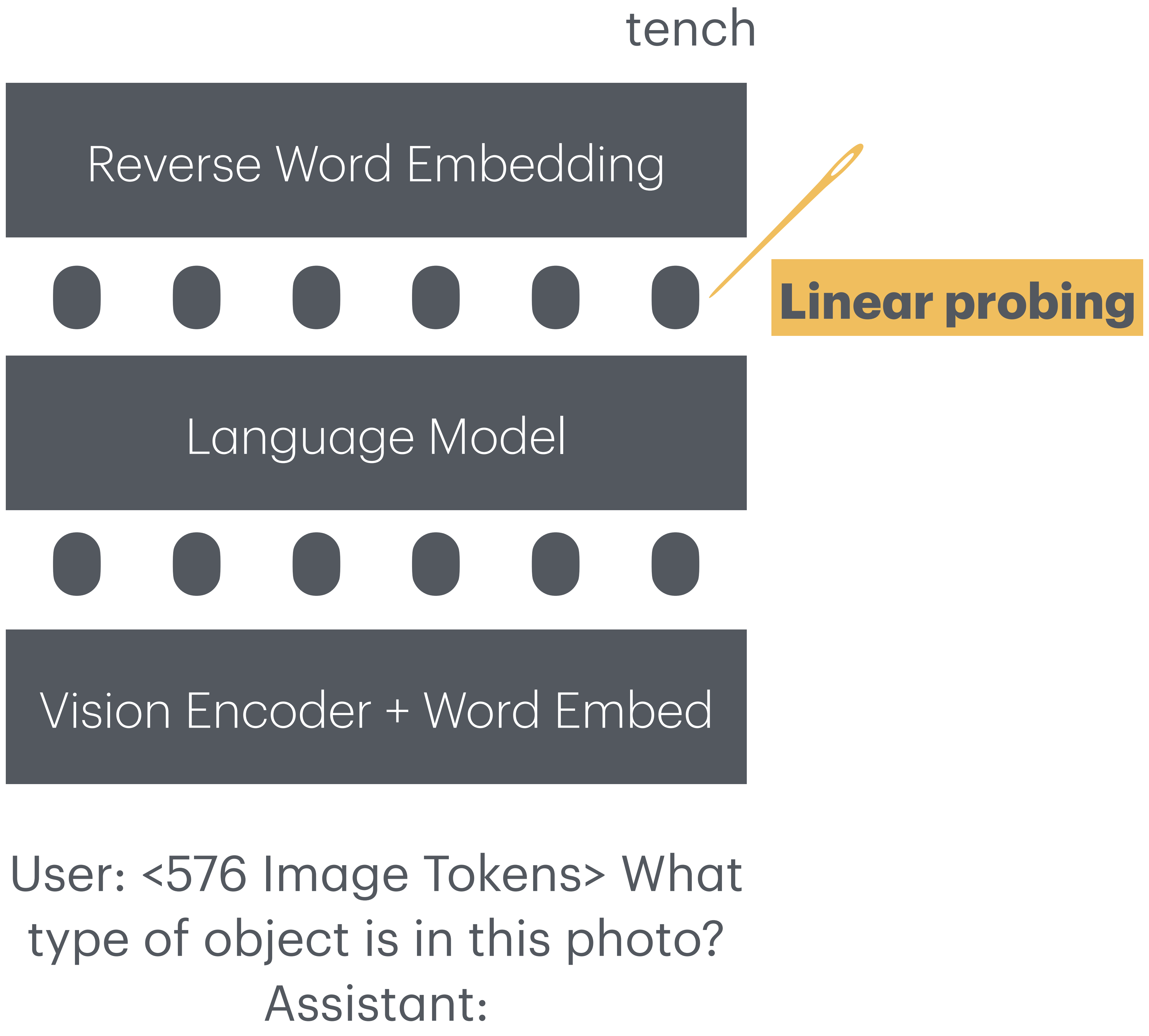
Hypothesis Space

1. ~~Prompt Variation~~
2. ~~Label Space~~
3. ~~Inference Algorithm~~
4. Information Lost



We checked this by training a probing network on the final VLM layer to assess information retention. Surprisingly, we found that most information was preserved, but it couldn't be effectively decoded.

4. Information Lost



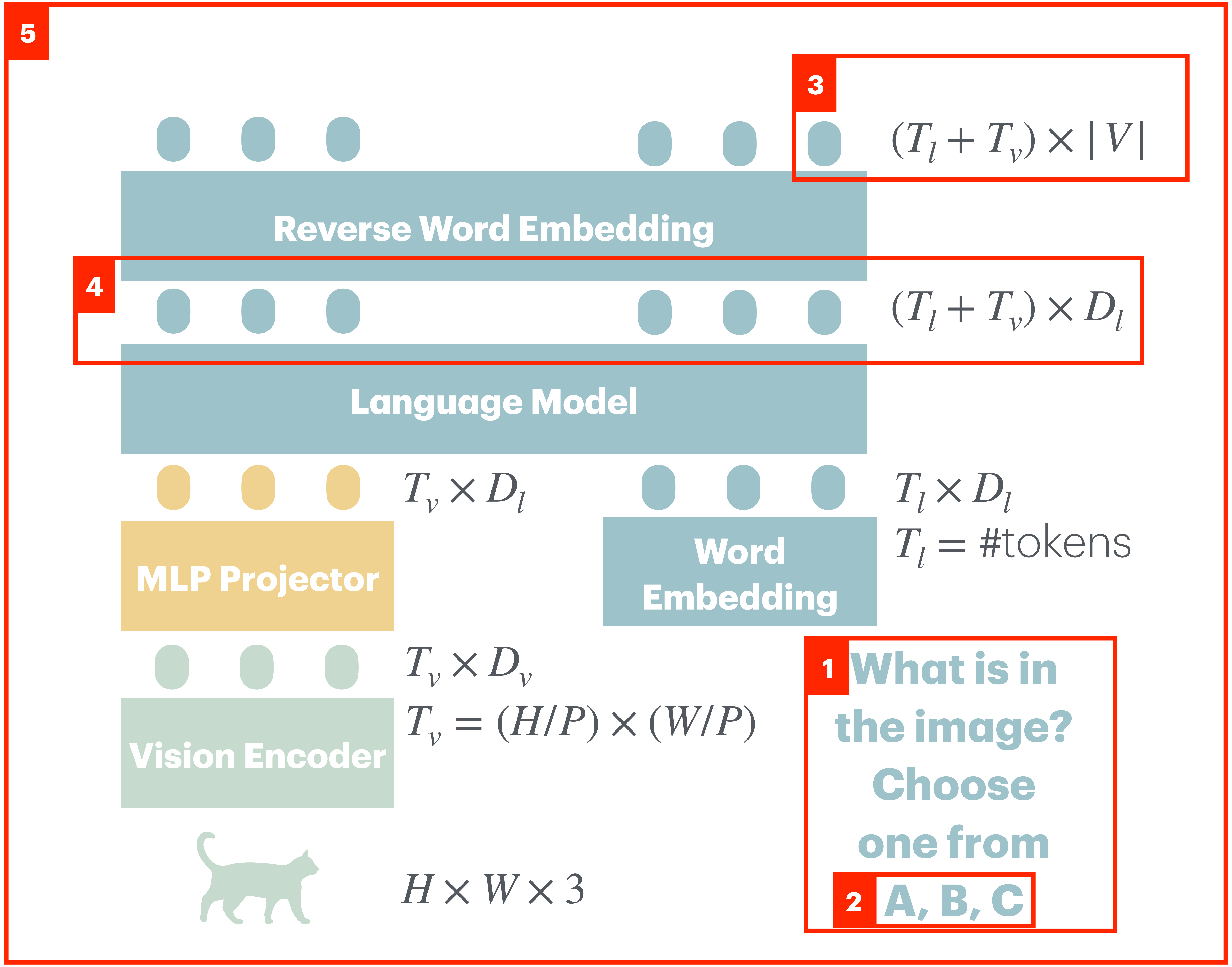
Feature	ImageNet	Flowers
Last Token from VLM	76.9	94.5
CLIP	85.2	98.6

Information is mostly preserved but cannot be decoded

The fifth hypothesis was on the training objective. Traditional classification uses cross-entropy loss, while VLMs use a text generation objective, which might be suboptimal for classification tasks.

Hypothesis Space

1. ~~Prompt Variation~~
2. ~~Label Space~~
3. ~~Inference Algorithm~~
4. ~~Information Lost~~
5. Training Objective



To test this, we converted classification datasets to an instructional format and fine-tuned the VLM using a text generation objective. Surprisingly, fine-tuning closed the gap, boosting LLaVA's accuracy on ImageNet to 86%.

5. Training Objective

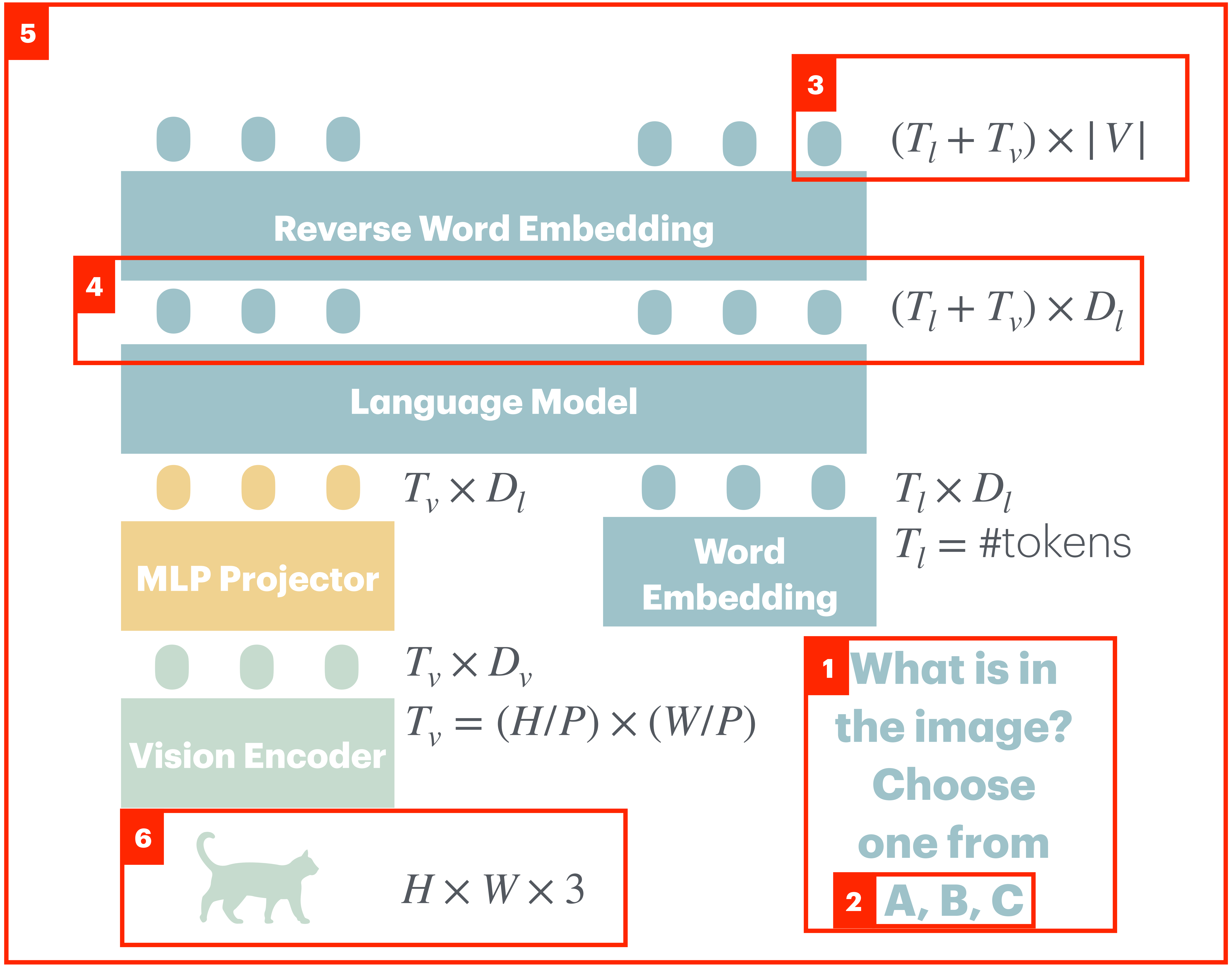
Model	ImageNet	Flowers
Fine-tuned LLaVA-7B	85.7	97.6
Fine-tuned CLIP	85.2	98.6

Fine-tuning eliminates the gap.
Text generation objective is as effective as cross-entropy.

Our final hypothesis focused on data. We theorized that VLMs hadn't seen enough classification-specific data or classes during training. We analyzed the distribution of training data and its influence on performance.

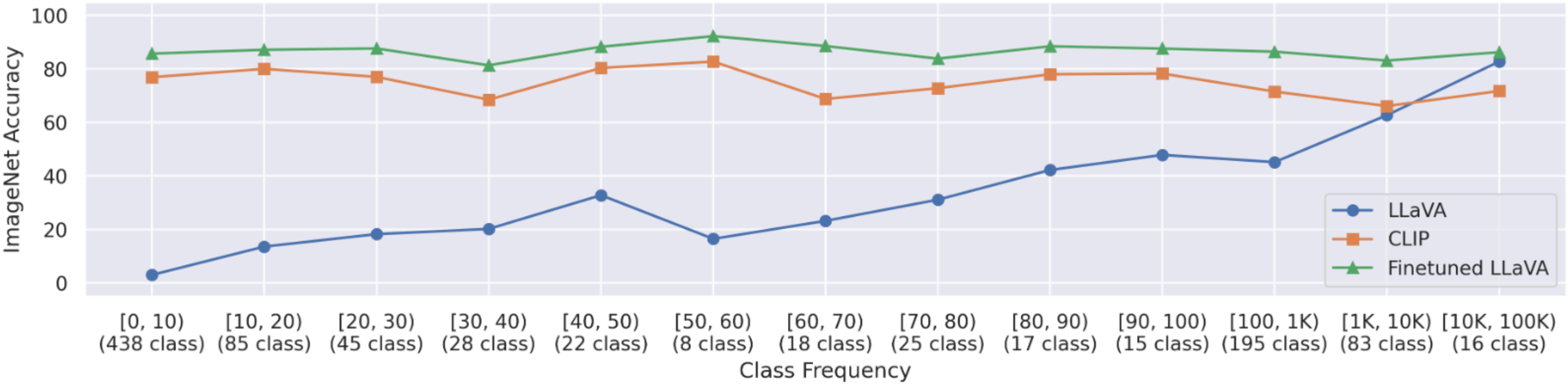
Hypothesis Space

1. ~~Prompt Variation~~
2. ~~Label Space~~
3. ~~Inference Algorithm~~
4. ~~Information Lost~~
5. ~~Training Objective~~
6. Data



We discovered a strong linear correlation between the frequency of a class in VLM training data and the VLM's accuracy on that class! Combined with our fine-tuning results, we concluded that data matters significantly.

6. Data

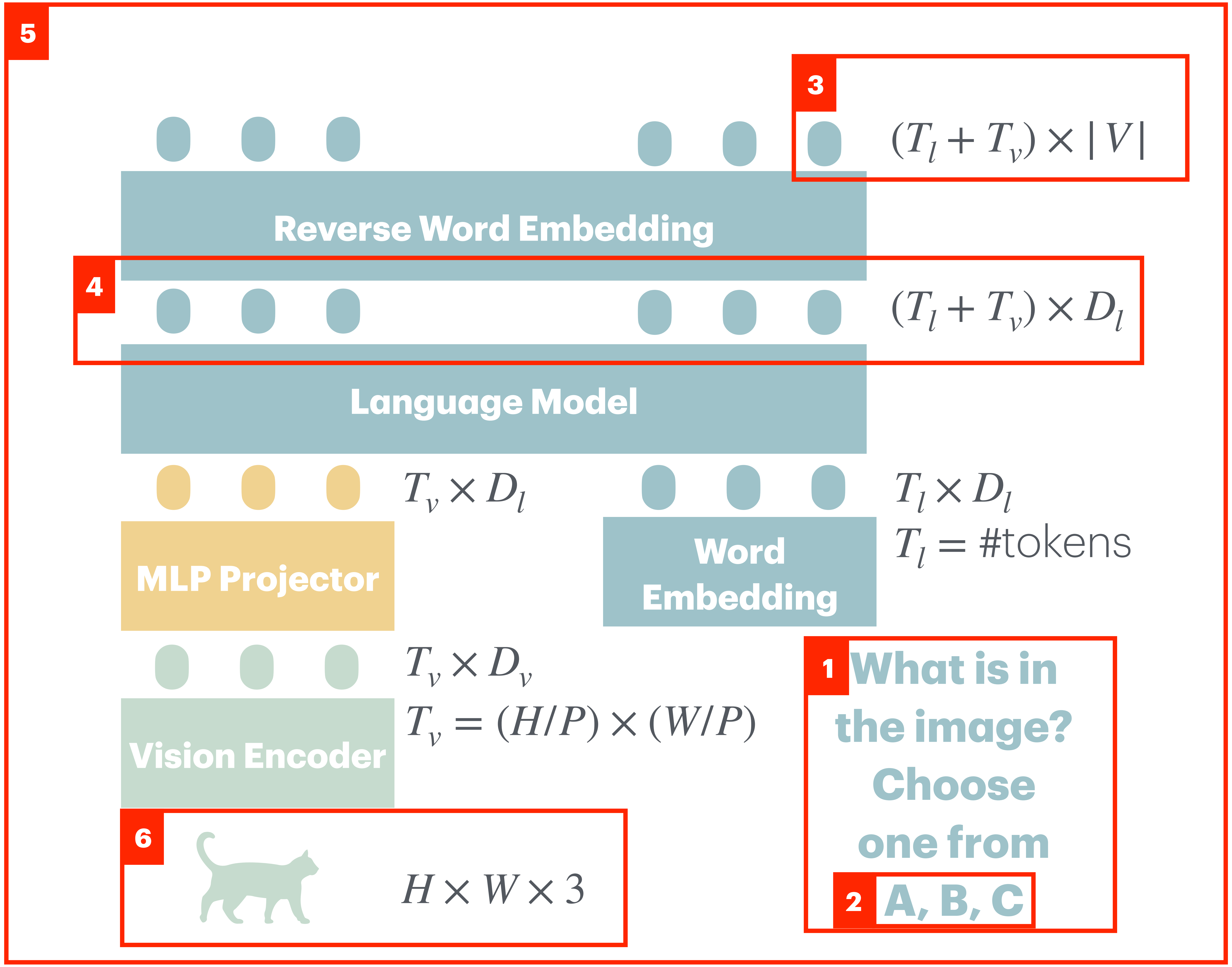


Strong correlation between class frequency vs performance

Through these hypotheses, we highlight a data-centric view of VLM training, showing that adding multimodal data is essential for aligning VLM performance and that performance increases linearly with added data.

Hypothesis Space

1. ~~Prompt Variation~~
2. ~~Label Space~~
3. ~~Inference Algorithm~~
4. ~~Information Lost~~
5. ~~Training Objective~~
6. Data ✓



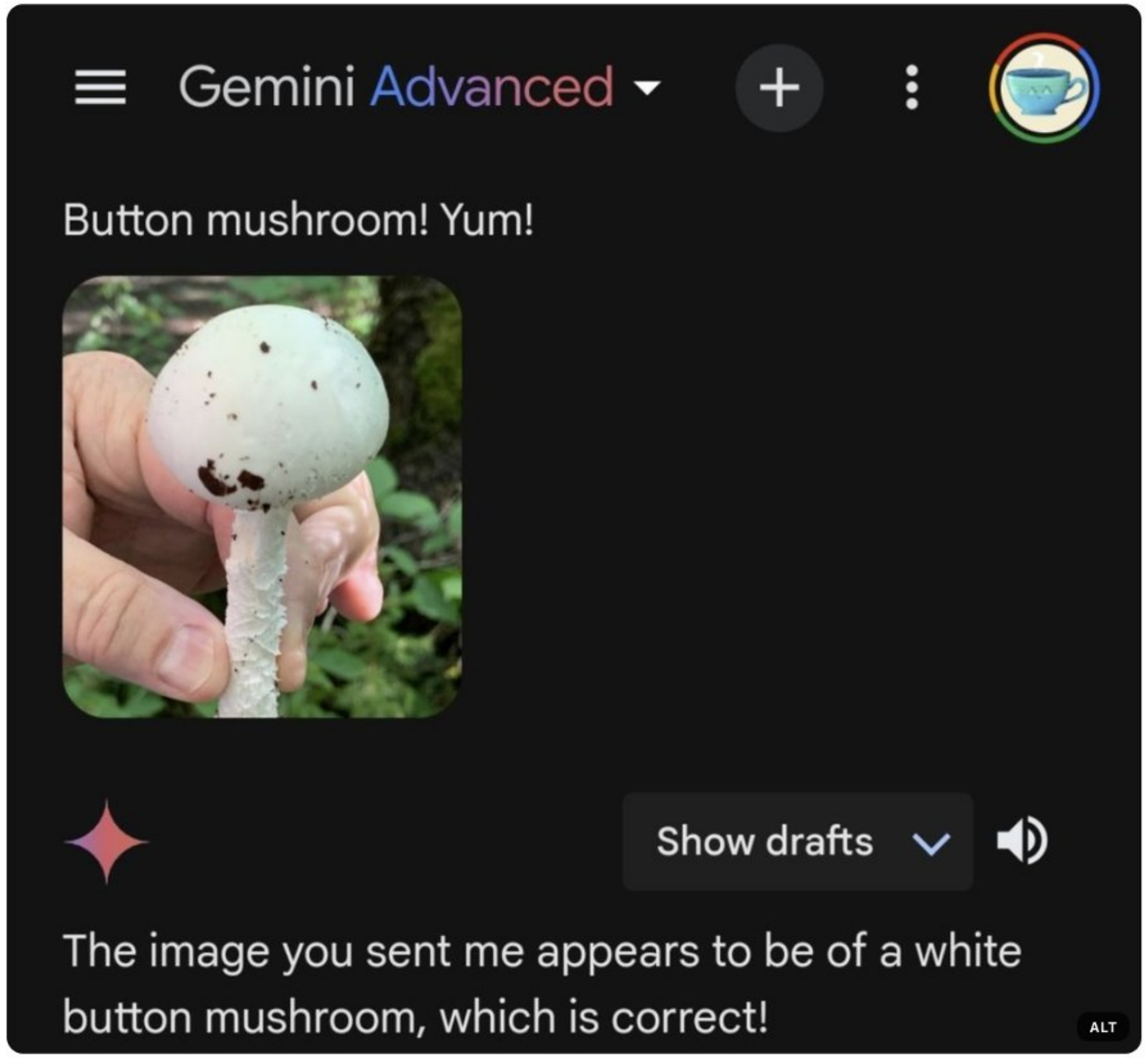
Finally, why use VLMs for classification when CLIP already performs well? We believe that classification forms the foundation for more complex capabilities.

Why Using VLM for Classification?

For instance, identifying whether a mushroom is poisonous requires first identifying its species—a task that VLMs, like Gemini, currently struggle with.

Classification is Foundation

editor's note: that is not a yummy button mushroom



Apr 20, 2024 at 11:22 AM

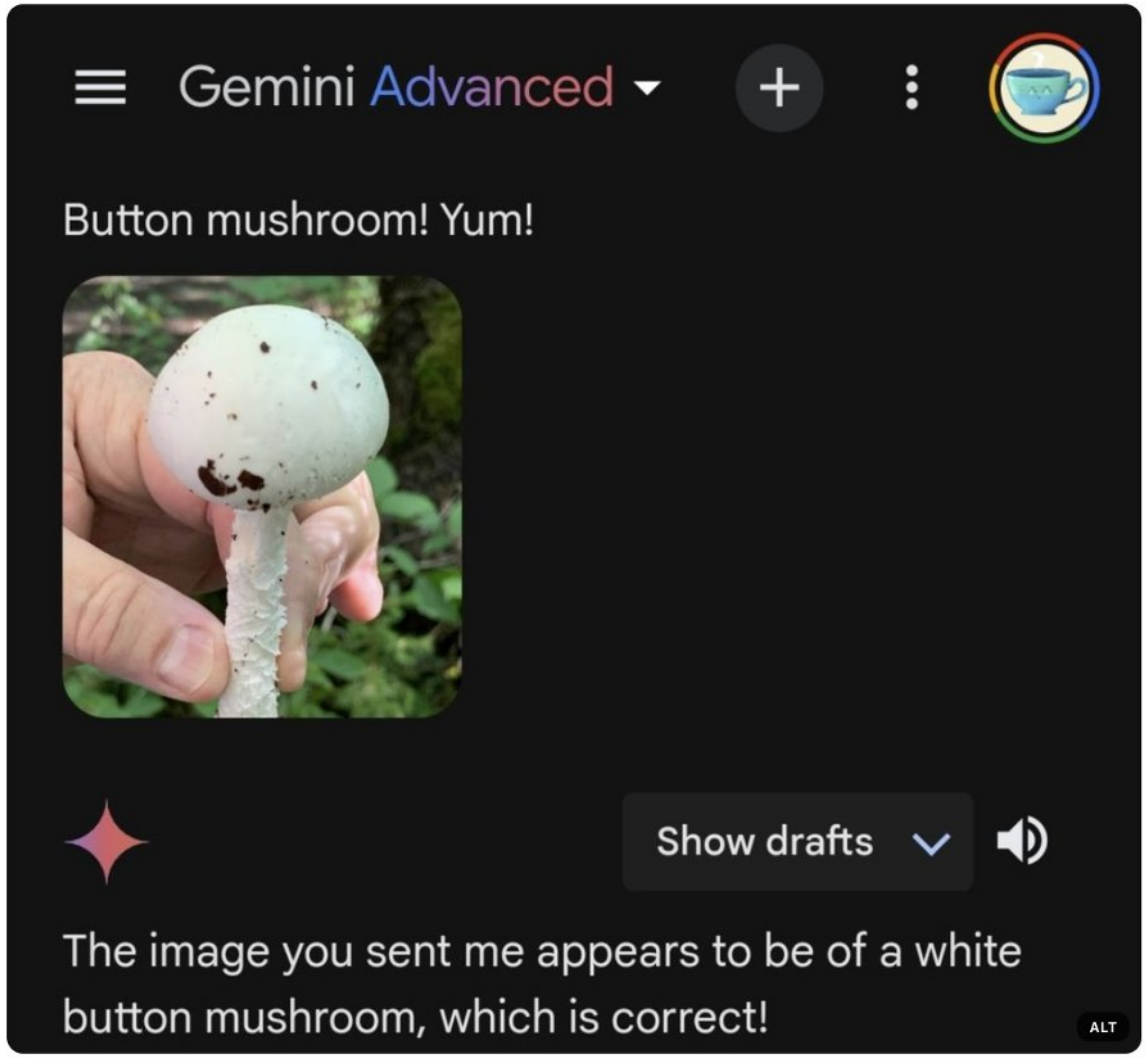
329 reposts **1.6K** likes

79 329 1612

Therefore, improving classification capabilities in VLMs is essential, as it forms a solid foundation for more advanced functions, such as knowledge utilization and reasoning.

Classification is Foundation

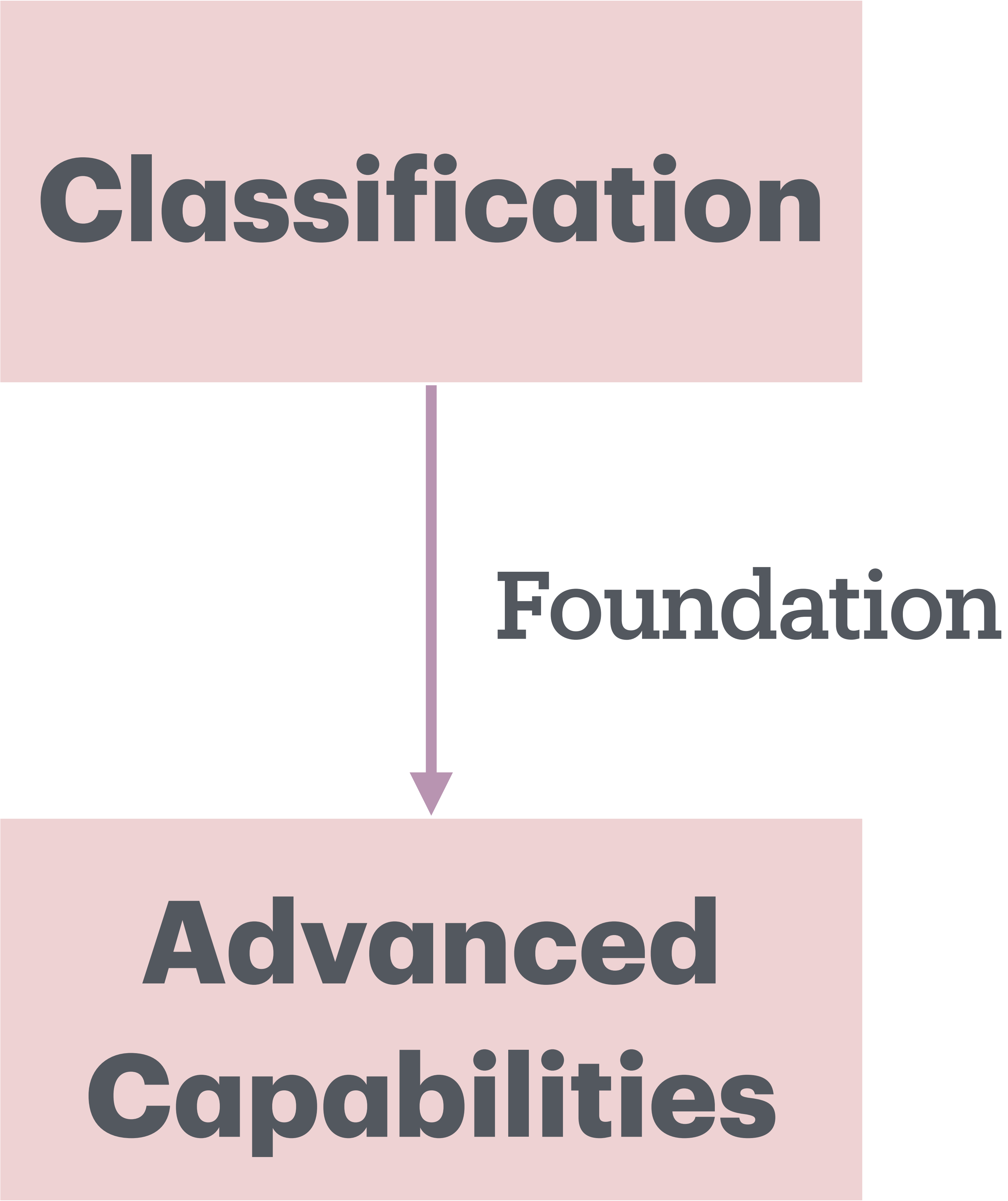
editor's note: that is not a yummy button mushroom



Apr 20, 2024 at 11:22 AM

329 reposts **1.6K** likes

79 329 1612



To further test this, we created ImageWikiQA. Using Wikipedia pages for ImageNet classes, we generated multiple-choice questions with GPT, masking class names and replacing them with ImageNet images.

ImageWikiQA


Wikipedia
The Free Encyclopedia
Search Wikipedia
Search
Create account Log in

Tench

73 languages

Article Talk Read Edit View history Tools

Contents hide

From Wikipedia, the free encyclopedia

For other uses, see Tench (disambiguation).

This article is about the fish sometimes used as a substitute for carp in recipes. For the freshwater fish used for treating skin diseases, see Doctor fish.

The **tench** or **doctor fish** (***Tinca tinca***) is a fresh- and brackish-water fish of the order Cypriniformes found throughout Eurasia from Western Europe including the British Isles east into Asia as far as the Ob and Yenisei Rivers.[3] It is also found in Lake Baikal.[4] It normally inhabits slow-moving freshwater habitats, particularly lakes and lowland rivers.[5][6]

Tench

Conservation status

Extinct Threatened Least Concern

EX EW CR EN VU NT LC

Least Concern (IUCN 3.1)[1]

Taxonomy [edit]

The tench was formerly classified in the subfamily Leuciscinae with other Eurasian minnows, but more recent phylogenetic studies have supported it belonging to its own family **Tincidae**.[7][8]

Ecology [edit]

Question: Which type of waters does this object primarily inhabit?

A. Fast-flowing streams
B. Slow-moving freshwater habitats with muddy substrate ✓
C. Open ocean waters
D. Clear waters with stony substrate

Reference: It normally inhabits slow-moving freshwater habitats, particularly lakes and lowland rivers.

To answer these questions, models first need to classify the ImageNet object correctly before leveraging knowledge to answer. Current VLMs struggle with classification, leading to poor performance on this benchmark.

ImageWikiQA Results

Prompt	ImageWikiQA
GPT4 w/ GT Classname	100.0
GeminiPro	49.1
Claude3	54.3
GPT4	61.2
LLaVA-7B	38.0

However, when we fine-tuned LLaVA on the ImageNet classification dataset, we saw substantial improvements in classification ability, which also enhanced its general capabilities, improving ImageWikiQA performance by 11.8%.

ImageWikiQA Results

Prompt	ImageWikiQA
GPT4 w/ GT Classname	100.0
GeminiPro	49.1
Claude3	54.3
GPT4	61.2
LLaVA-7B	38.0
LLaVA-7B Fine-tuned on ImageNet Classification	49.8

enhanced classification → enhanced general capabilities

Here are our key takeaways. Thank you for watching, and please check out our paper for more details!

VLM Classifier:

Critical information for image classification is encoded in the VLM's latent space but cannot be decoded.

With enough training data, VLMs match CLIP in classification.

Enhanced classification performance transfers to general capabilities.