# 23. Multi-modal Learning

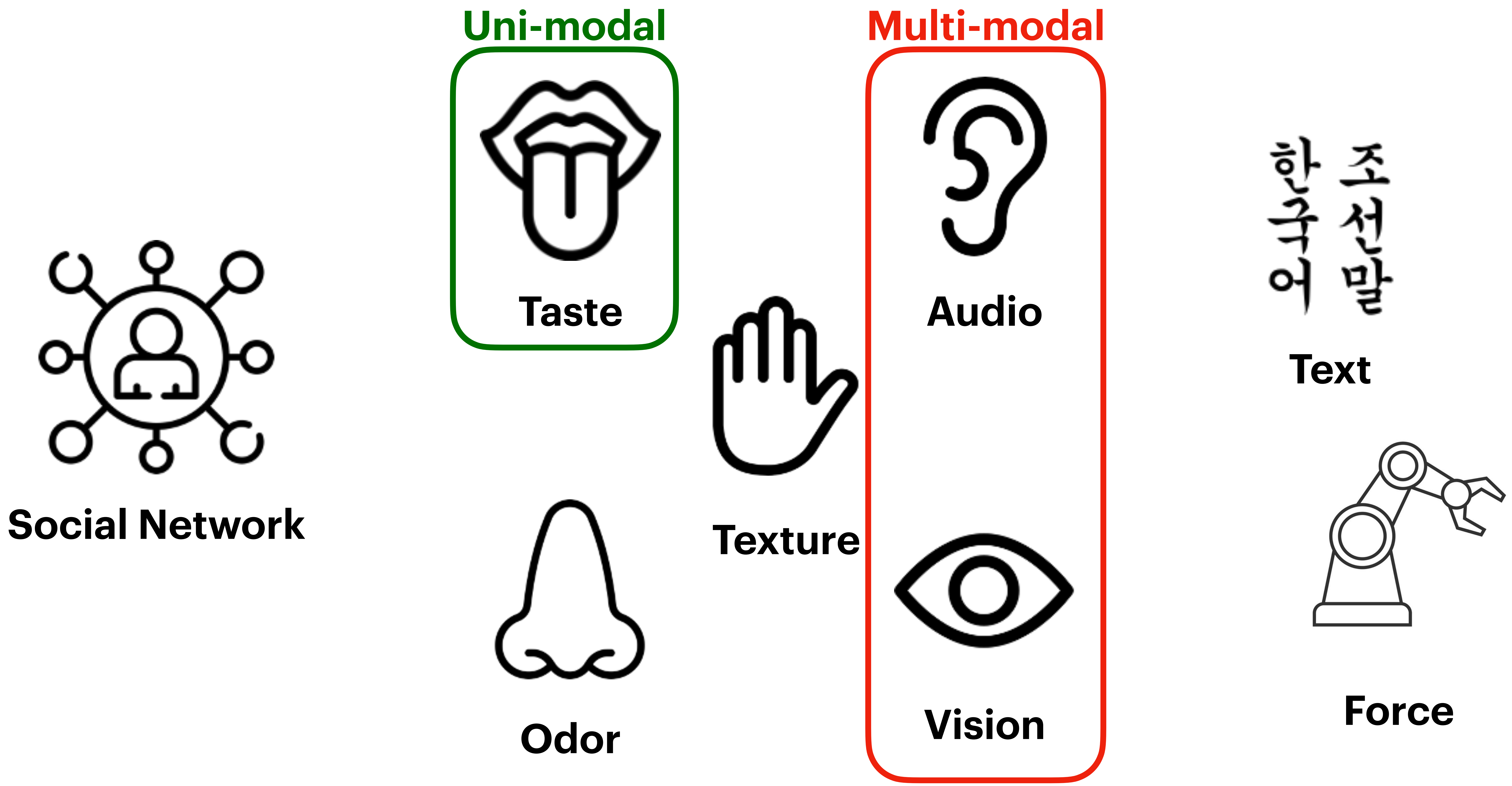
## EEEC454 Introduction to Machine Learning Systems

2023 Fall, Jaeho Lee

# Overview

# Multi-modality

- Modalities in multi-modal learning



# Challenges

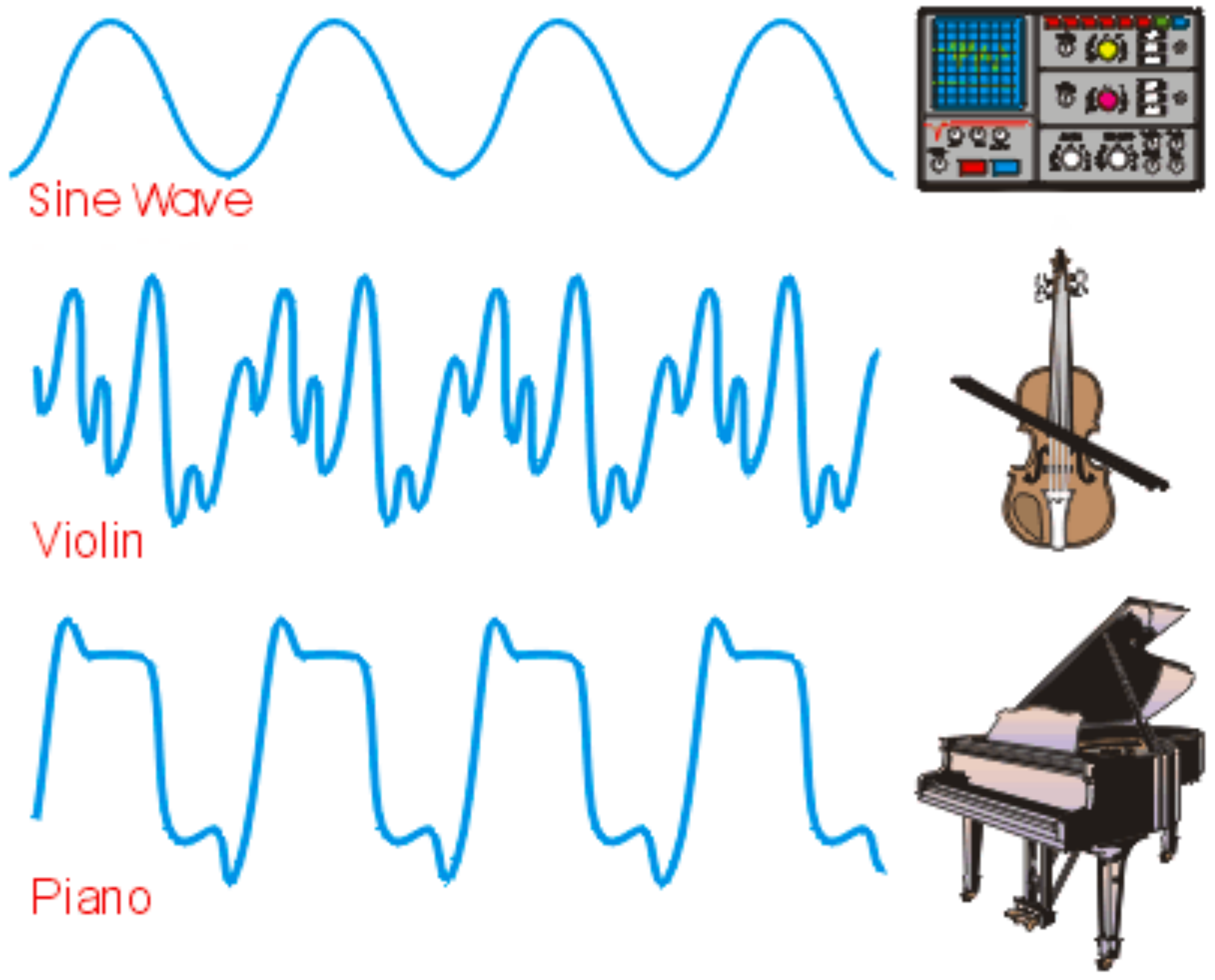
**1. Representation.** Data in each domain have different representations

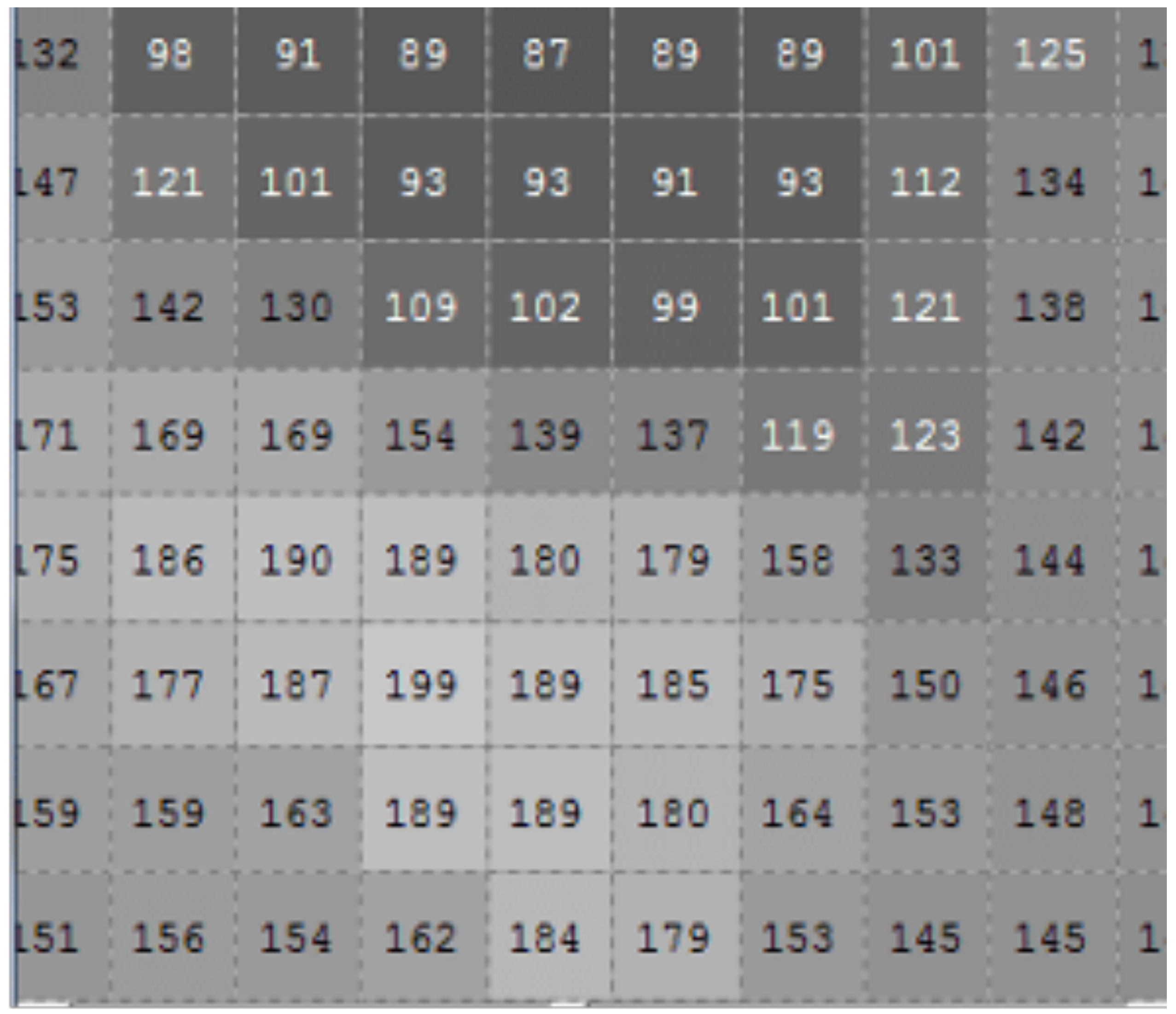
**Audio**

**Vision**

한국어 조선말

**Text**



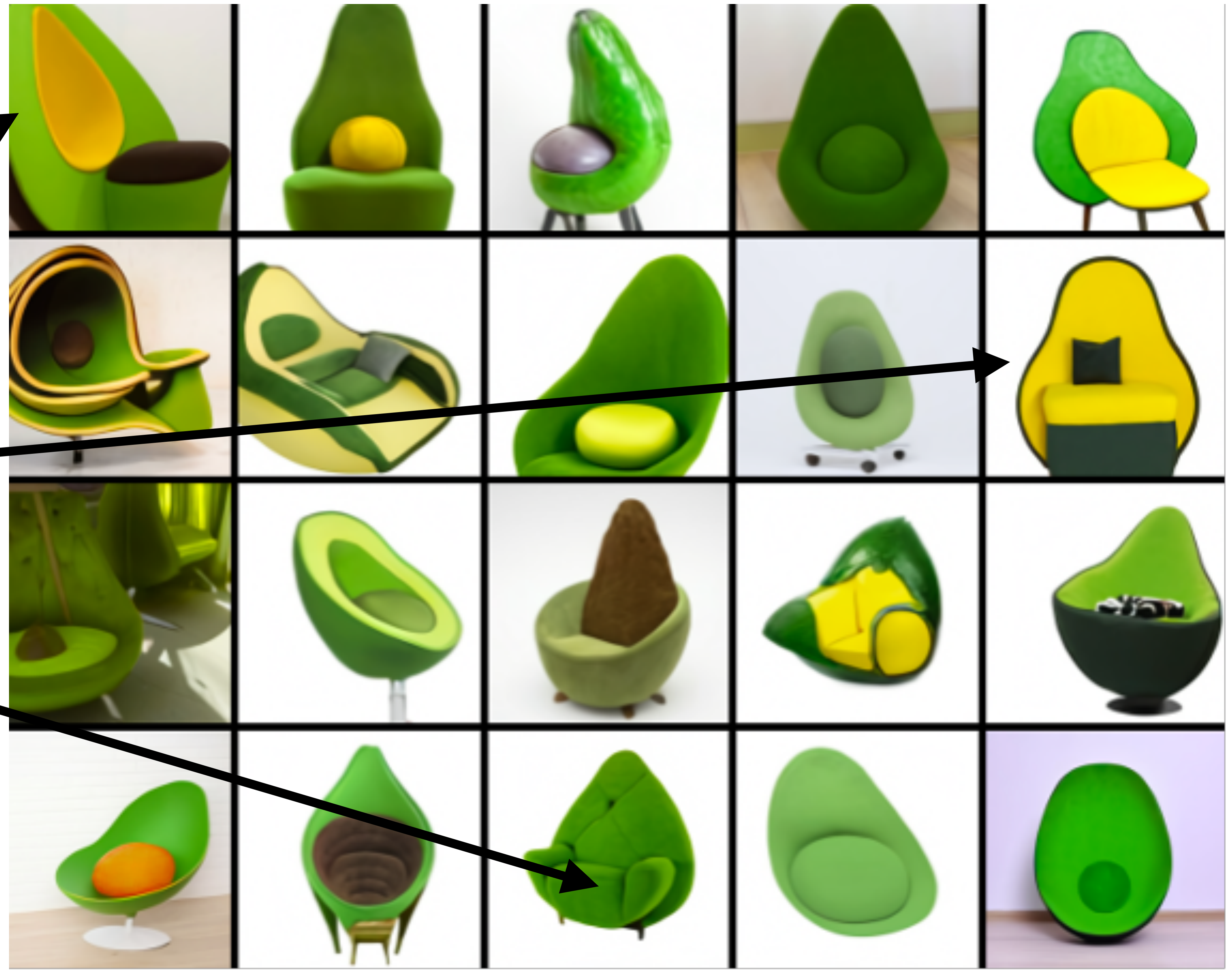

| | living being | feline | human | gender | royalty | verb | plural |
|---|---|---|---|---|---|---|---|
| *cat* → | 0.6 | 0.9 | 0.1 | 0.4 | −0.7 | −0.3 | −0.2 |
| *kitten* → | 0.5 | 0.8 | −0.1 | 0.2 | −0.6 | −0.5 | −0.1 |
| *dog* → | 0.7 | −0.1 | 0.4 | 0.3 | −0.4 | −0.1 | −0.3 |
| *houses* → | −0.8 | −0.4 | −0.5 | 0.1 | −0.9 | 0.3 | 0.8 |

# Challenges

**2. Correspondence.** Heterogeneous feature spaces with potentially limited correspondence

*"An armchair in the shape of avocado"*

1-to-N mapping

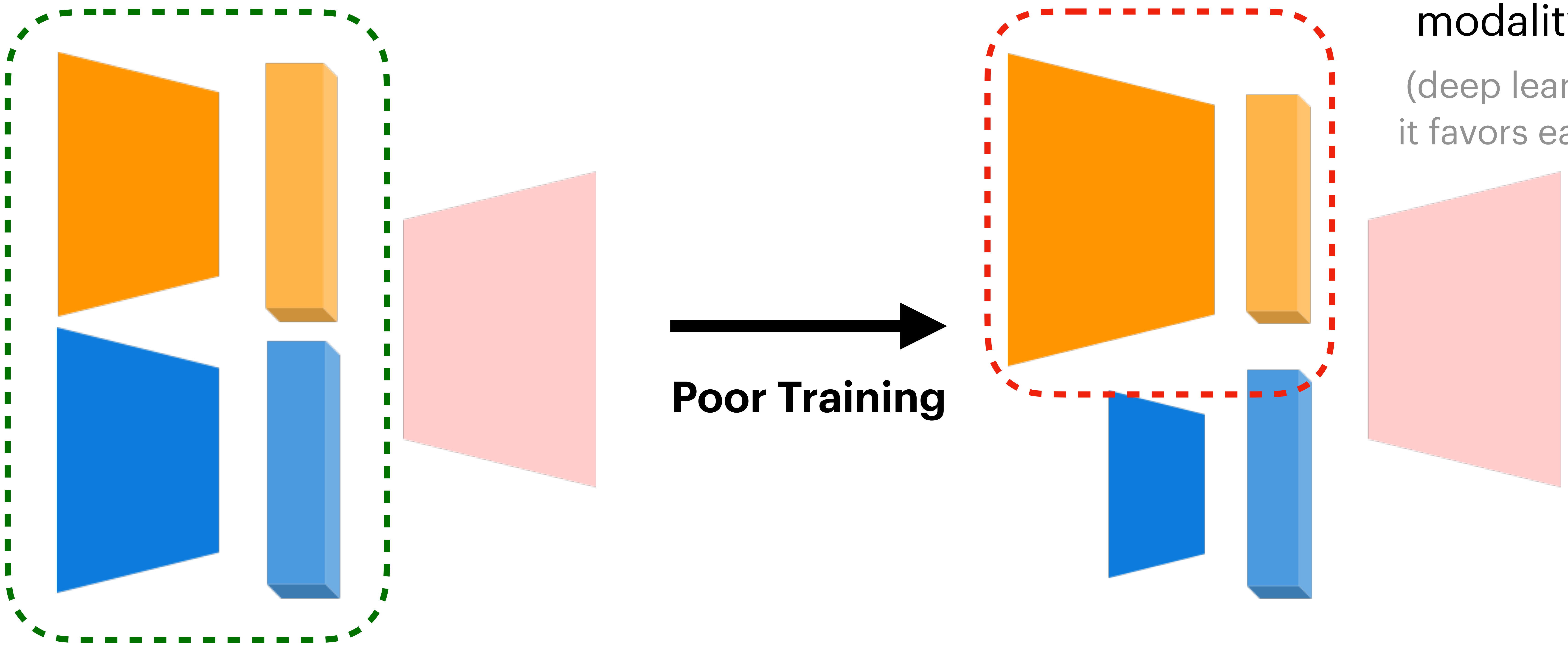


**Text Space**

**Image Space**

# Challenges

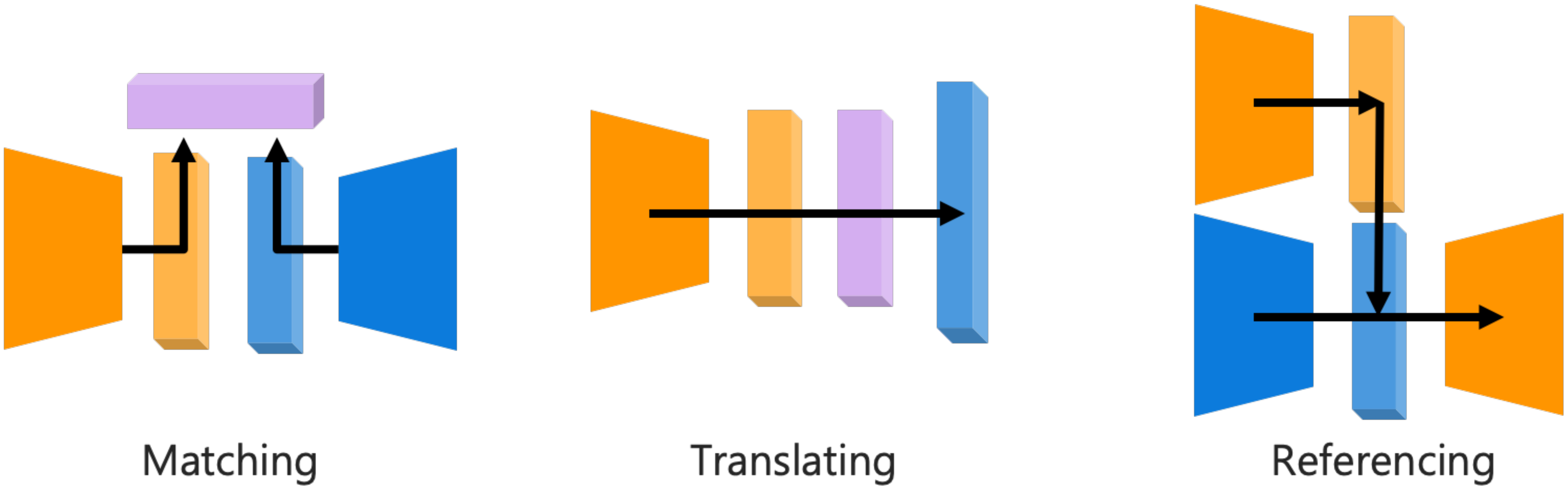
**3. Bias.** Imbalance between heterogeneous feature spaces



# Yet...

Despite the challenges, we expect much fruitful outcomes

- We look at the example of CLIP, which handles ***vision + text***



# Vision & Language

# Text Embedding

- Map each word / token to a continuous Euclidean space.
  - Discrete characters are difficult to use.

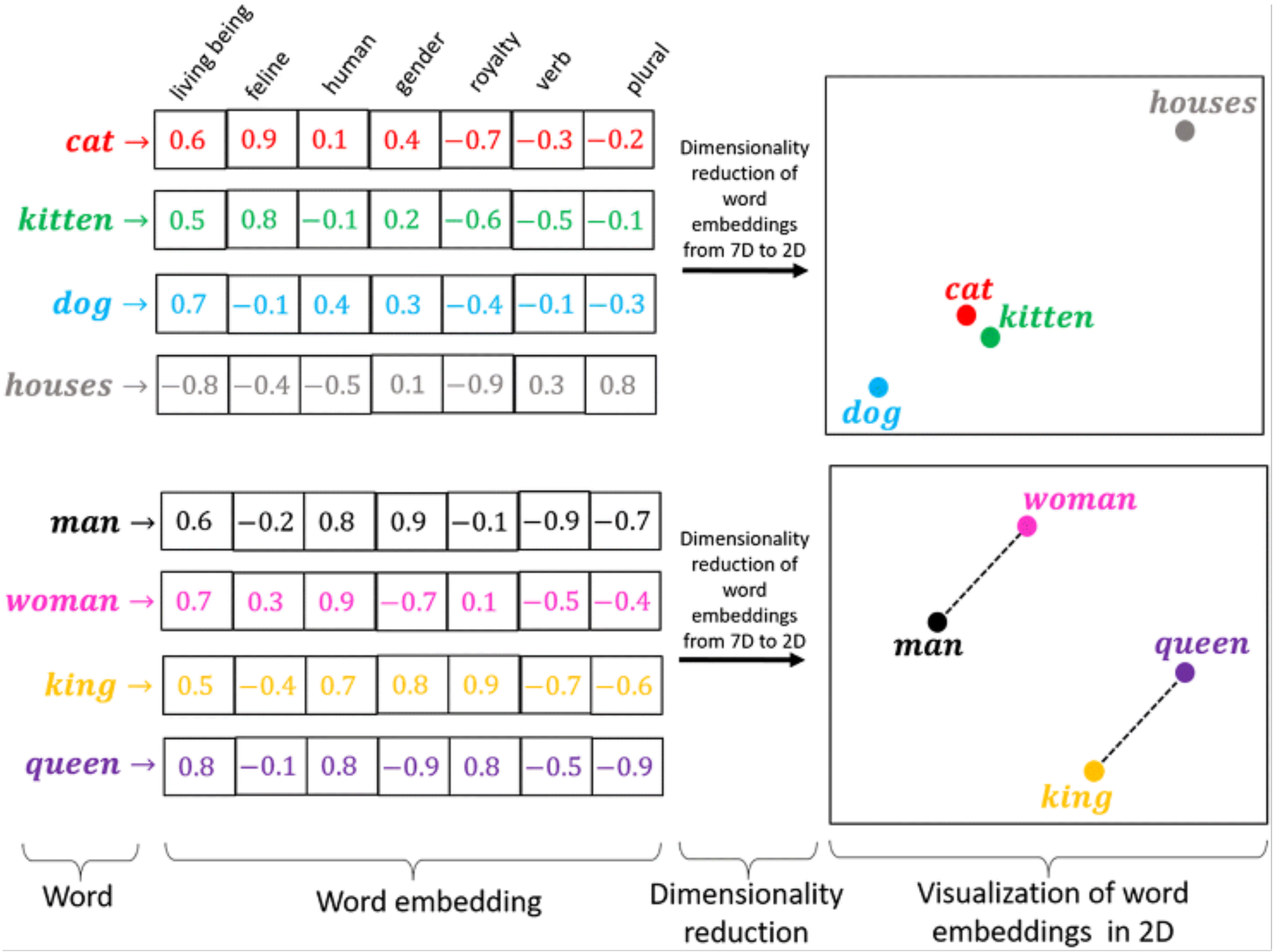
| #14 | living being | feline | human | gender | royalty | verb | plural |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *cat* → | 0.6 | 0.9 | 0.1 | 0.4 | −0.7 | −0.3 | −0.2 |
| *kitten* → | 0.5 | 0.8 | −0.1 | 0.2 | −0.6 | −0.5 | −0.1 |
| *dog* → | 0.7 | −0.1 | 0.4 | 0.3 | −0.4 | −0.1 | −0.3 |
| *houses* → | −0.8 | −0.4 | −0.5 | 0.1 | −0.9 | 0.3 | 0.8 |
| *man* → | 0.6 | −0.2 | 0.8 | 0.9 | −0.1 | −0.9 | −0.7 |
| *woman* → | 0.7 | 0.3 | 0.9 | −0.7 | 0.1 | −0.5 | −0.4 |
| *king* → | 0.5 | −0.4 | 0.7 | 0.8 | 0.9 | −0.7 | −0.6 |
| *queen* → | 0.8 | −0.1 | 0.8 | −0.9 | 0.8 | −0.5 | −0.9 |

Word

Word embedding

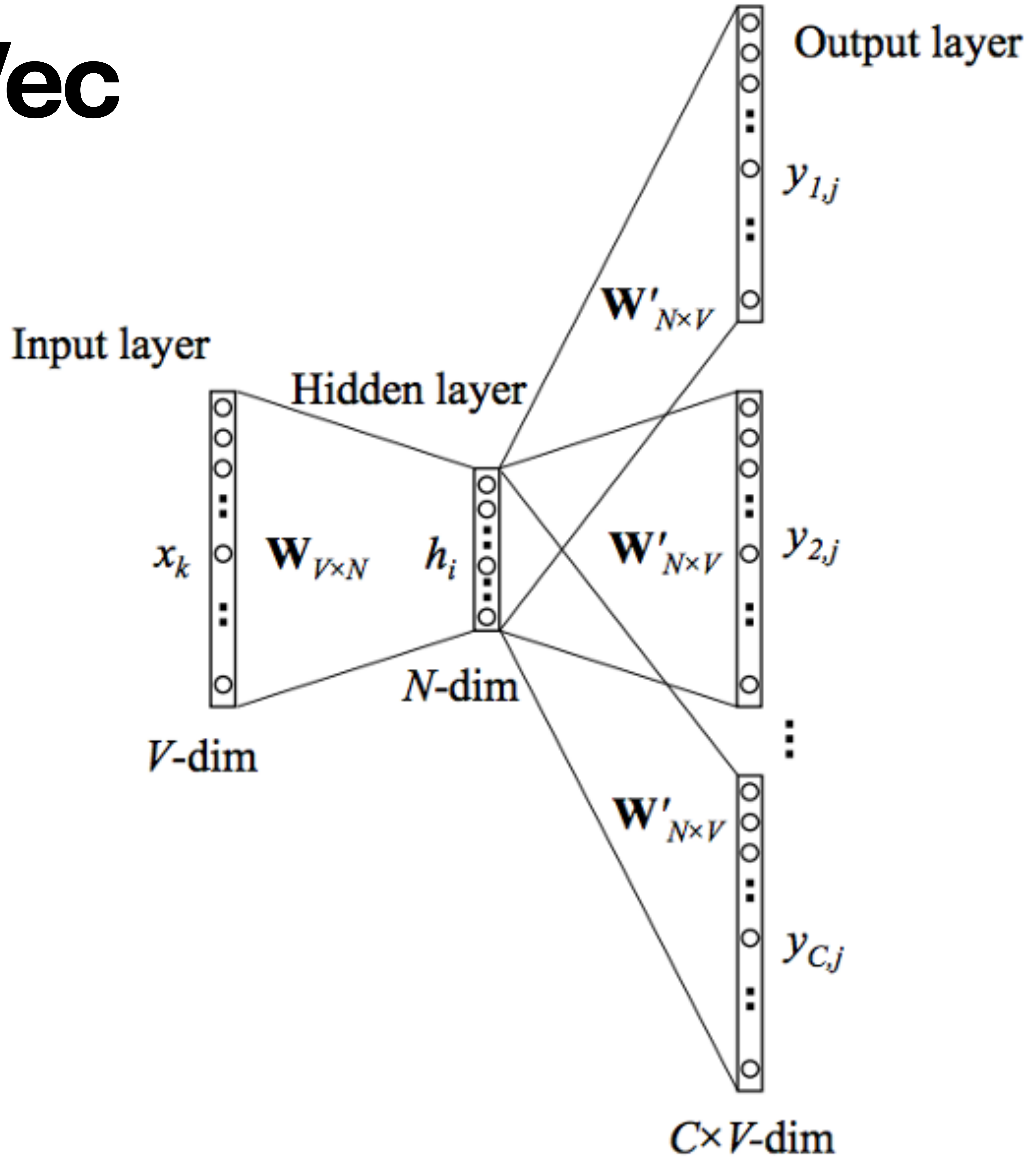
# Text Embedding

- Map each word / token to a continuous Euclidean space.
  - Discrete characters are difficult to use.
- Surprisingly, learned embeddings are rich in semantics (e.g., cat & kitten)



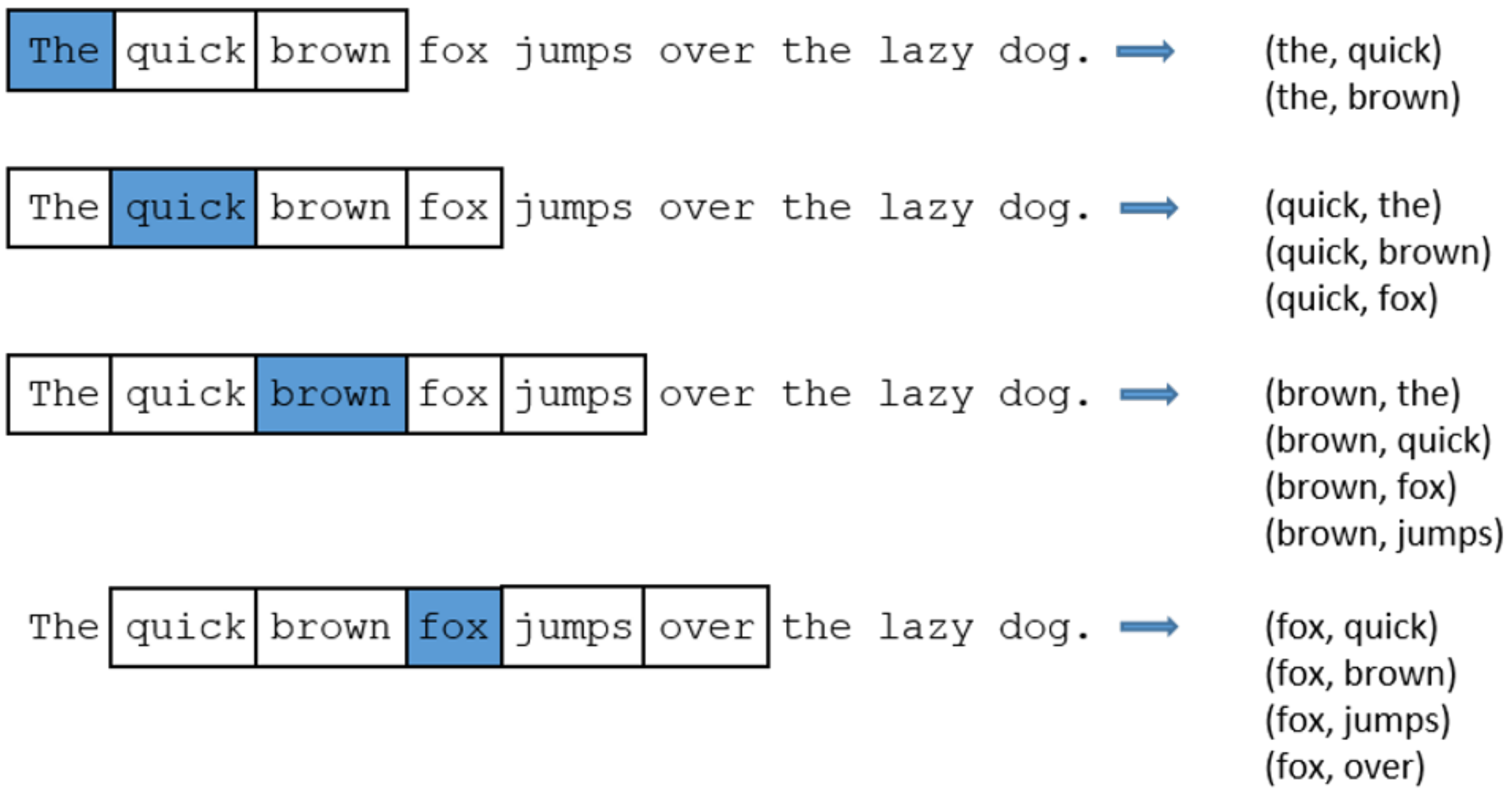
# Word2Vec

- One way to train text embeddings.
  - A *skip-gram* model



# Word2Vec

- One way to train text embeddings.
  - A *skip-gram* model
- **Idea.** Predict the ***surrounding words*** from the center word.



# Word2Vec

- One way to train text embeddings.
    - A ***skip-gram*** model
- **Idea.** Predict the ***surrounding words*** from the center word.

- **Question.** Can we use similar idea to train the ***joint embedding*** of image and text data?

Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML 2021

# CLIP

- Trains such joint embedding using the transformer, and a lot of data
- **Scale matters.** Not the first attempt;
  but the first to use very large dataset
  - Used 400 million image-text pairs.

french cat

french cat

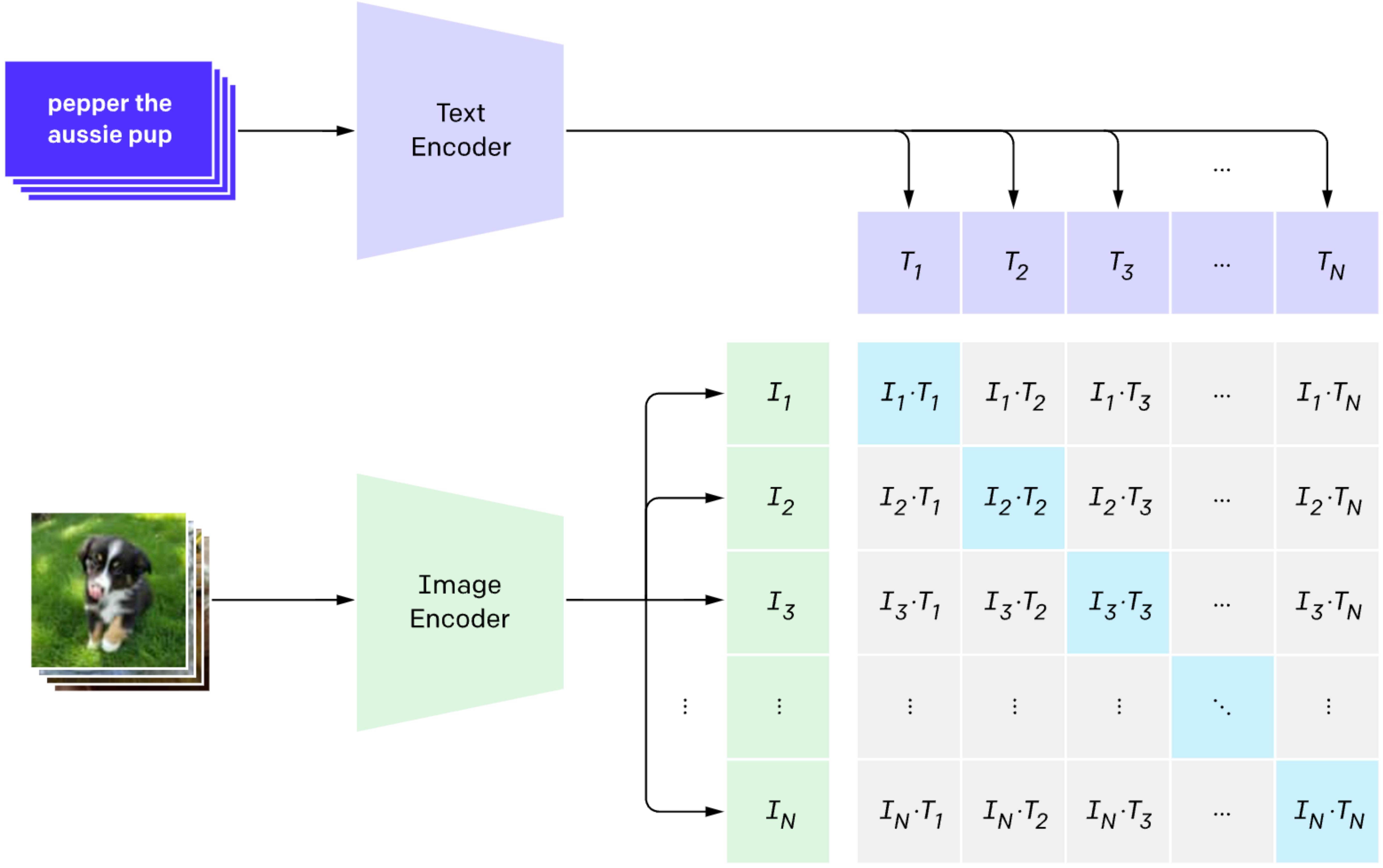
How to tell if your feline is french. He wears a b...

イケメン猫モデル「トキ・ナンタケット」がかっこいい - NAVER まとめ

Hilarious pics of funny cats! funnycatsgif.com

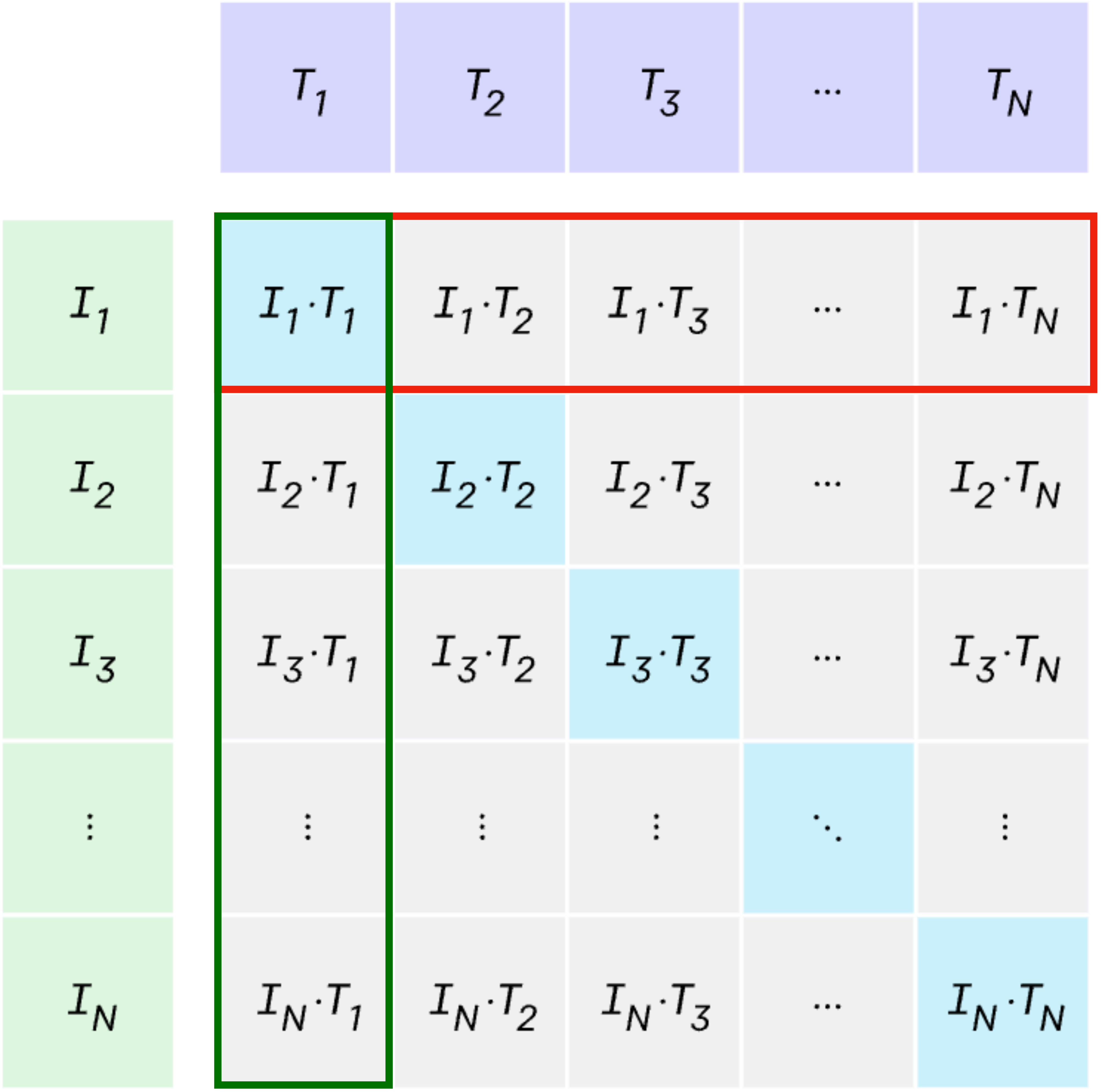
# CLIP

- **Algorithm.**
  Contrastive pre-training
  - Draws $N$ image-text pairs as a batch.
  - *Increase* the similarity between $(I_i, T_i)$
  - *Decrease* the similarity between $(I_i, T_j)$



Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML 2021

# CLIP




- **Concretely...**
  Minimize the mixture of two losses.
  - Image-to-text loss

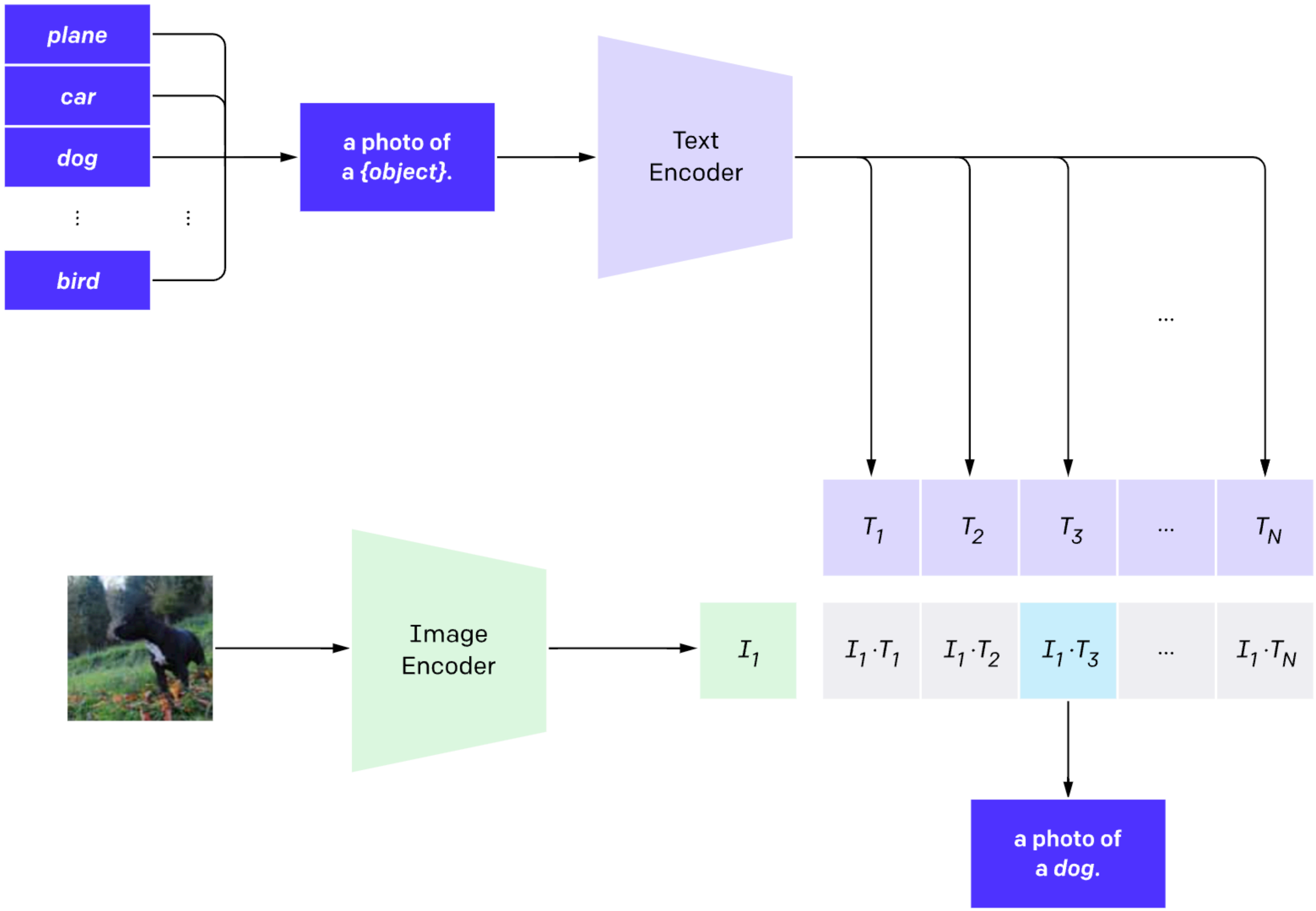
$$L_{i\to t} = -\sum_{i=1}^{N} \log \boxed{\frac{\exp(I_i \cdot T_i/\tau)}{\sum_j \exp(I_i \cdot T_j/\tau)}}$$

  - Text-to-image loss

$$L_{i\to t} = -\sum_{j=1}^{N} \log \boxed{\frac{\exp(I_j \cdot T_j/\tau)}{\sum_i \exp(I_i \cdot T_j/\tau)}}$$

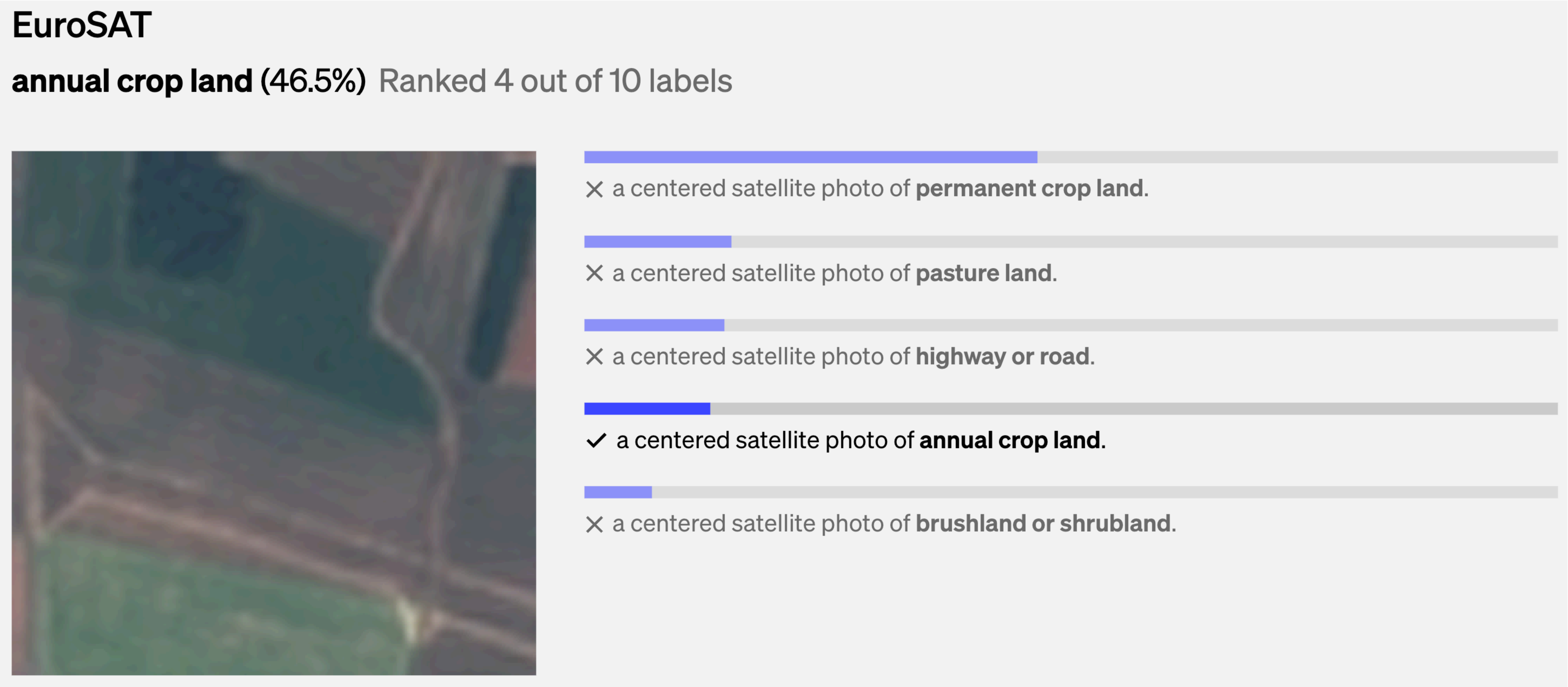
# Use cases

- Given a good joint embedding, one can use it for *classification*.



- Enables an effective *zero-shot classification*.

- Enables an effective ***zero-shot classification***.
  - Especially when we have ***good prompts***.

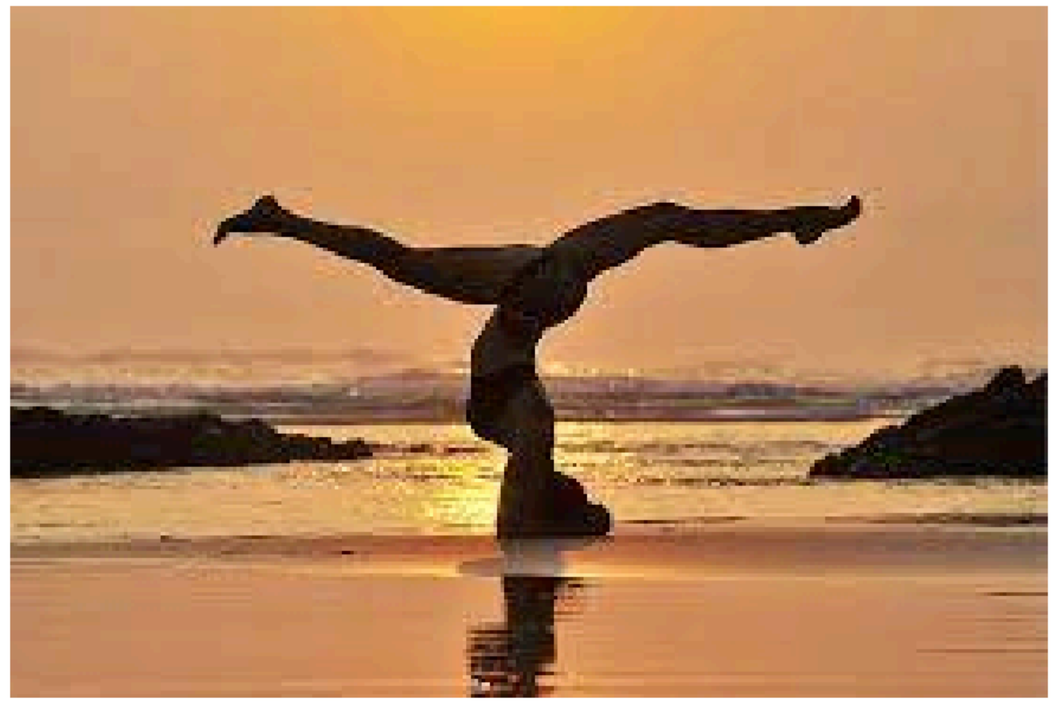


# Other use cases

- CLIP + LLMs = Captioning Models

A politician receives a gift from politician.

A collage of different colored ties on a white background.

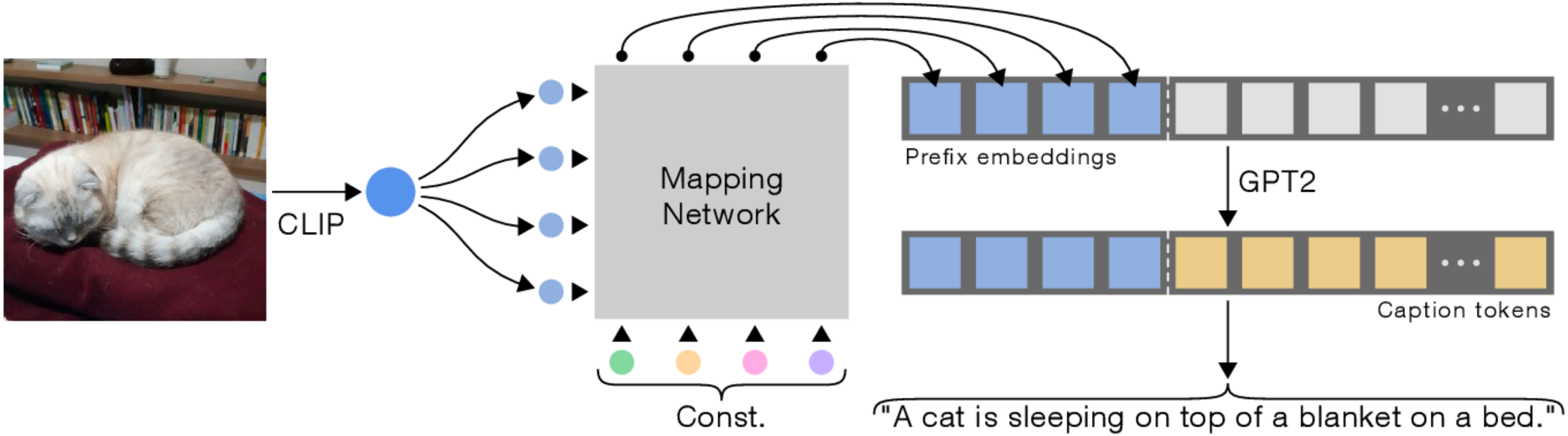


Silhouette of a woman practicing yoga on the beach at sunset.

Aerial view of a road in autumn.

# Other use cases

- CLIP + LLMs = Captioning Models

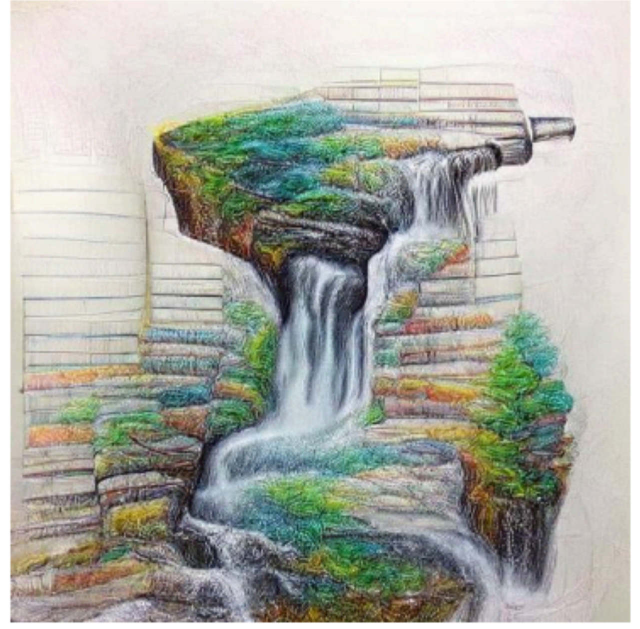


Crowson et al., “VQGAN-CLIP: Open domain image generation and editing with natural language guidance,” 2022

# Other use cases

- CLIP + GAN = Text-based Image Generation

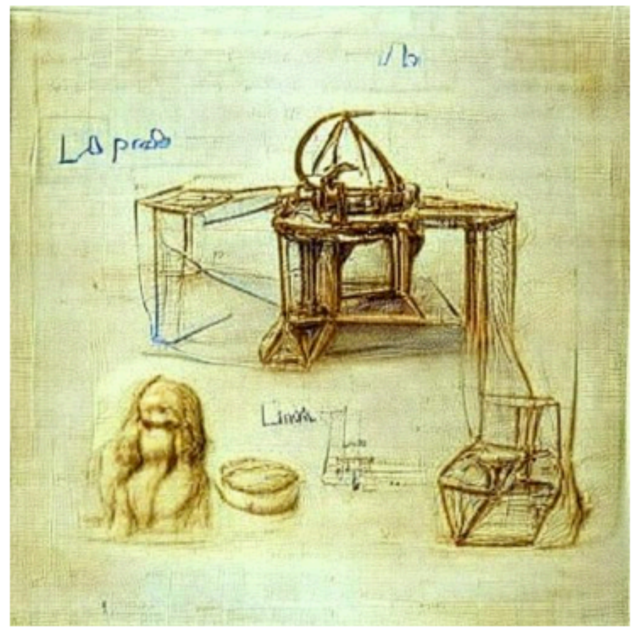
(a) Oil painting of a candy dish of glass candies, mints, and other assorted sweets



(b) A colored pencil drawing of a waterfall

(c) A fantasy painting of a city in a deep valley by Ivan Aivazovsky

(d) A beautiful painting of a building in a serene landscape

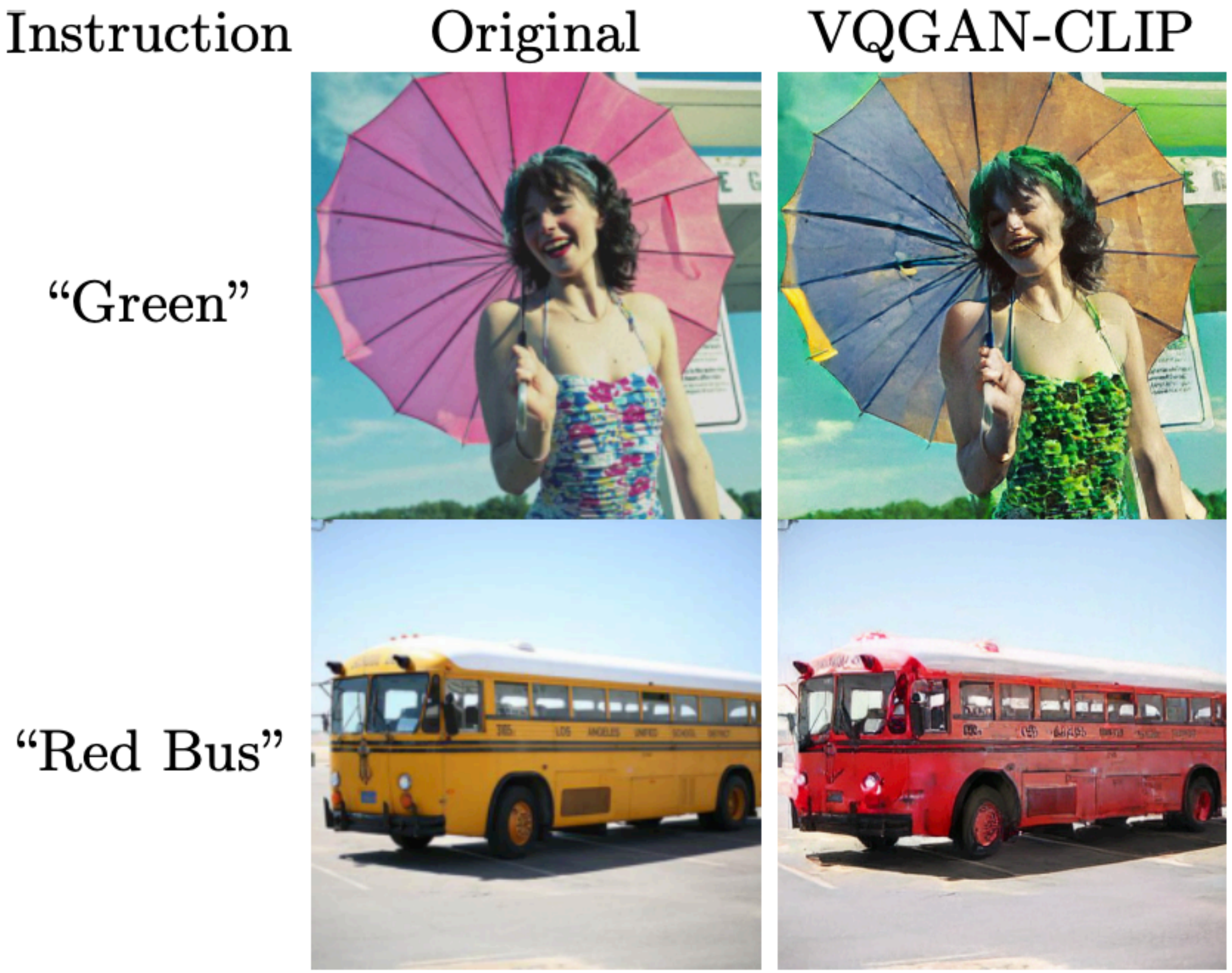


(e) sketch of a 3D printer by Leonardo da Vinci

(f) an autogyro flying car, trending on artstation

# Other use cases

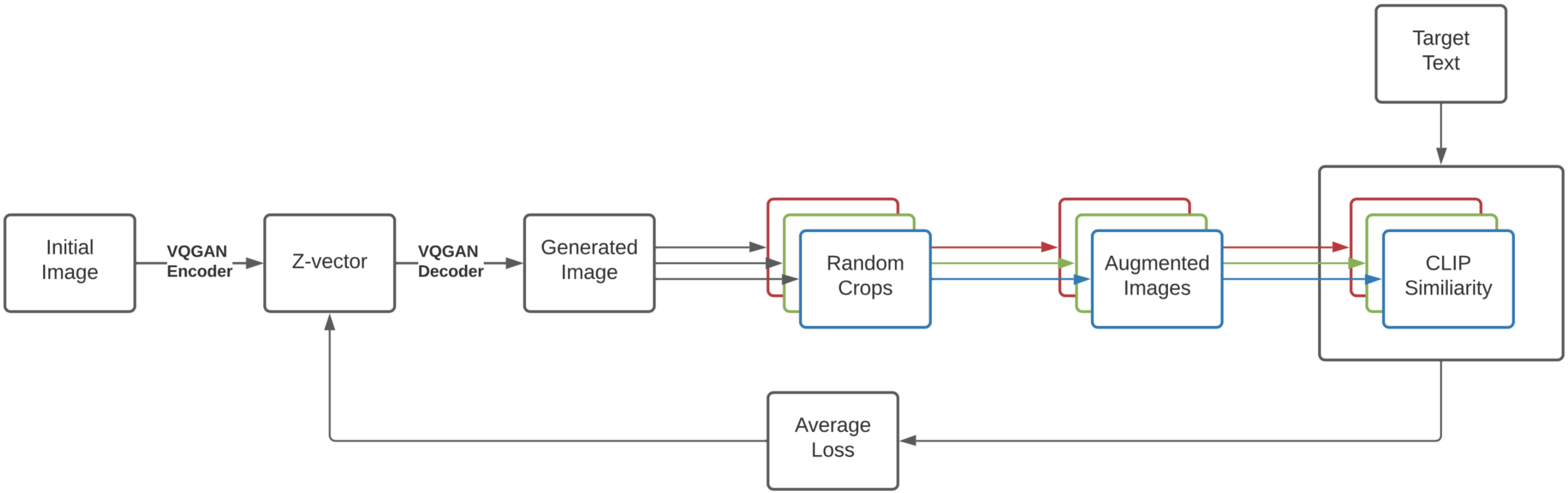
- CLIP + GAN = Text-based Image Generation and editing



Crowson et al., “VQGAN-CLIP: Open domain image generation and editing with natural language guidance,” 2022

# Other use cases

- CLIP + GAN = Text-based Image Generation and editing



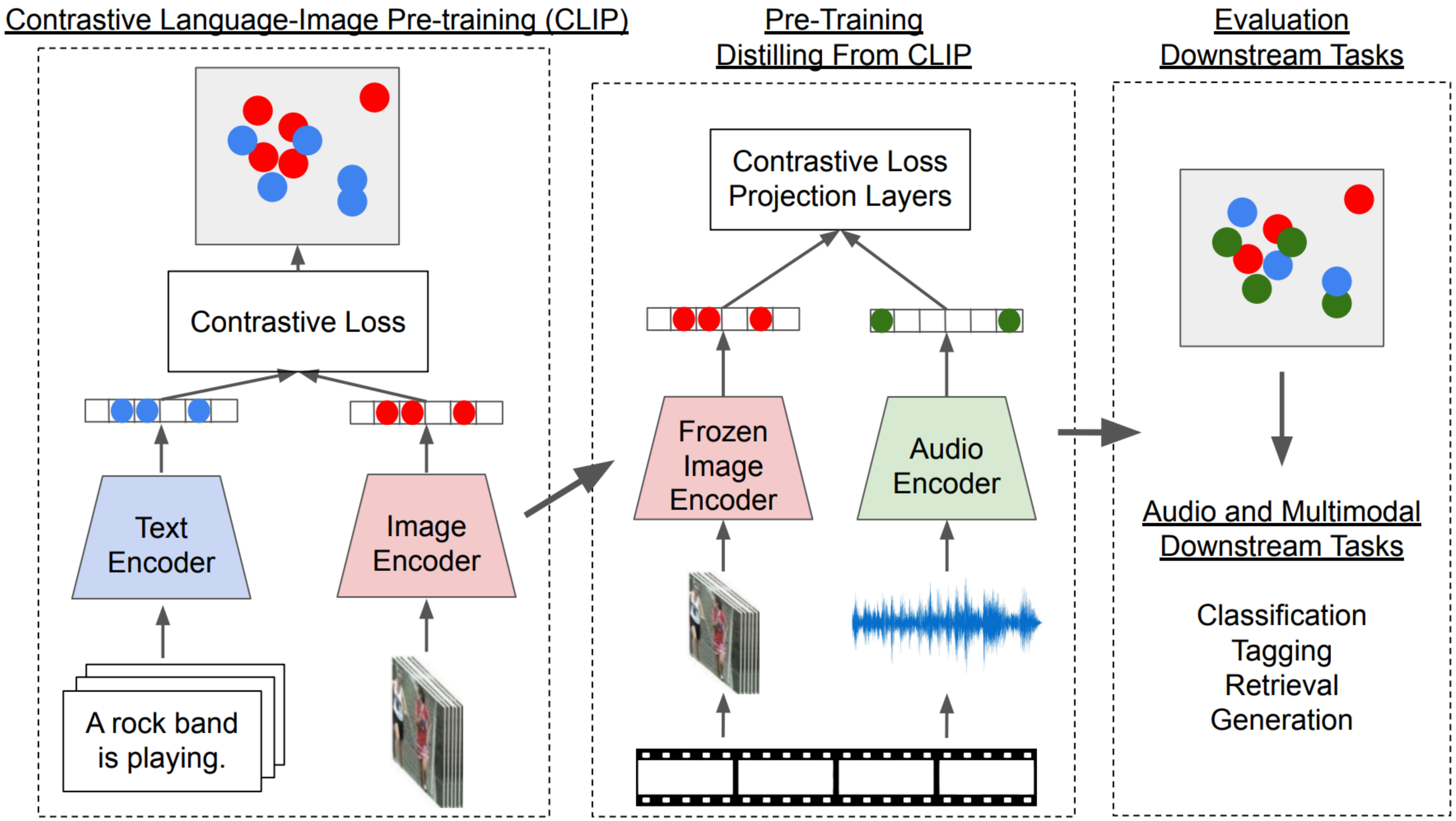
# Other use cases

- CLIP + GAN + Audio data

Text/Audio to Image Generation with VQGAN-CLIP

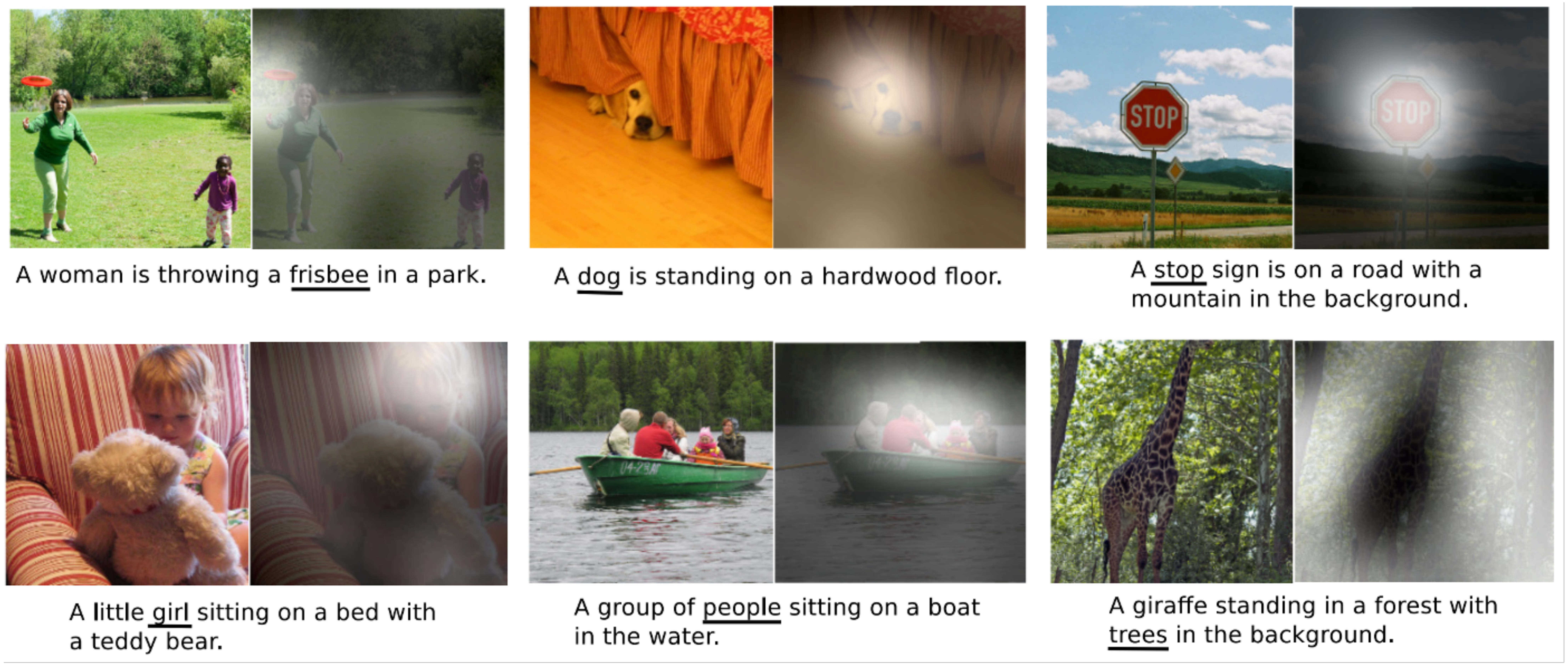
# Other use cases

- CLIP + GAN + Audio data



# Even older examples...

- Cross-modal reasoning



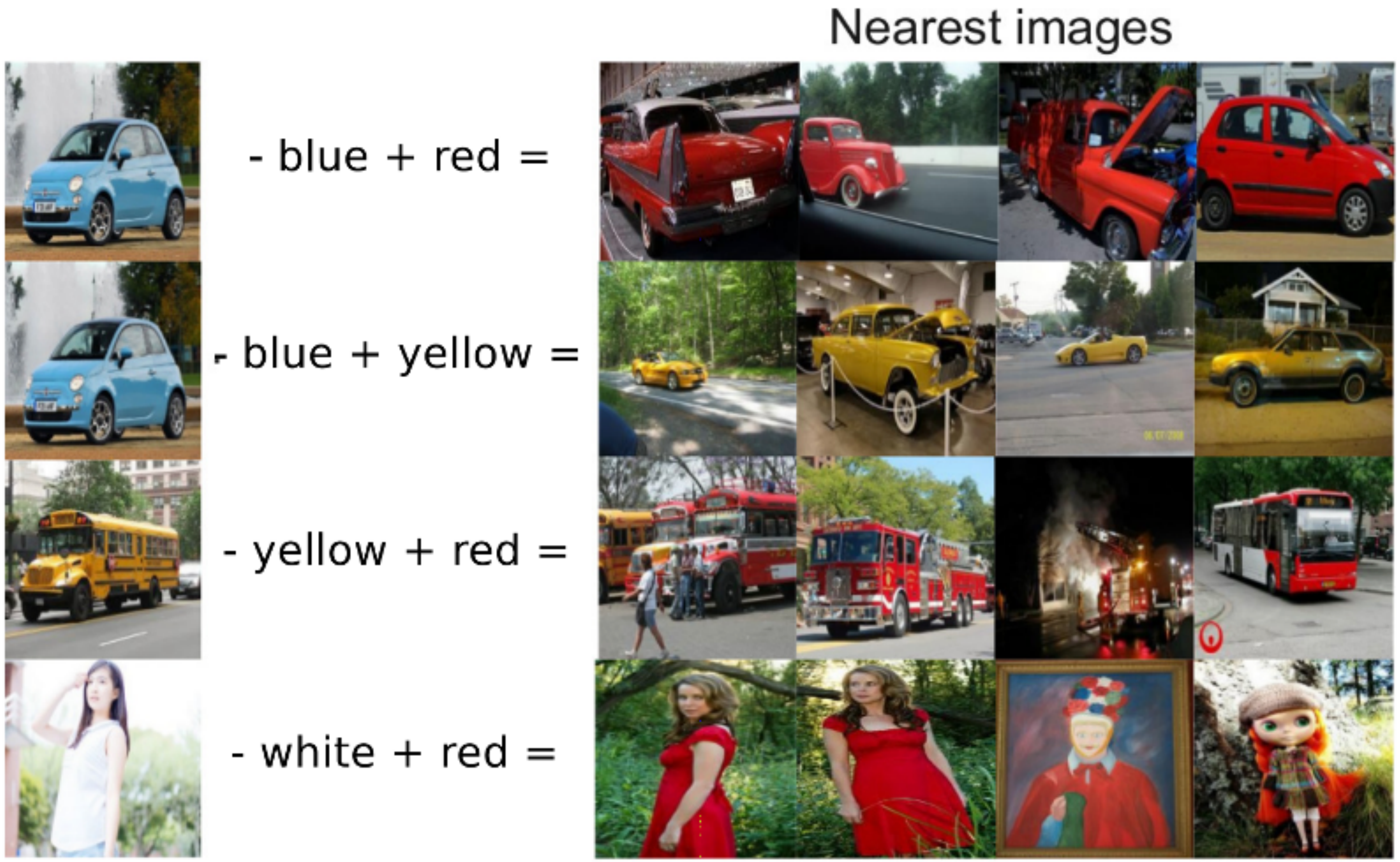
# Even older examples...

- Food recipe retrieval

Kiros et al., “Unifying visual-semantic embeddings with multimodal neural language models,” ICML 2014

# Even older examples...

- Image retrieval, with analogies



Cheers