# VLM for text: Language generation from images

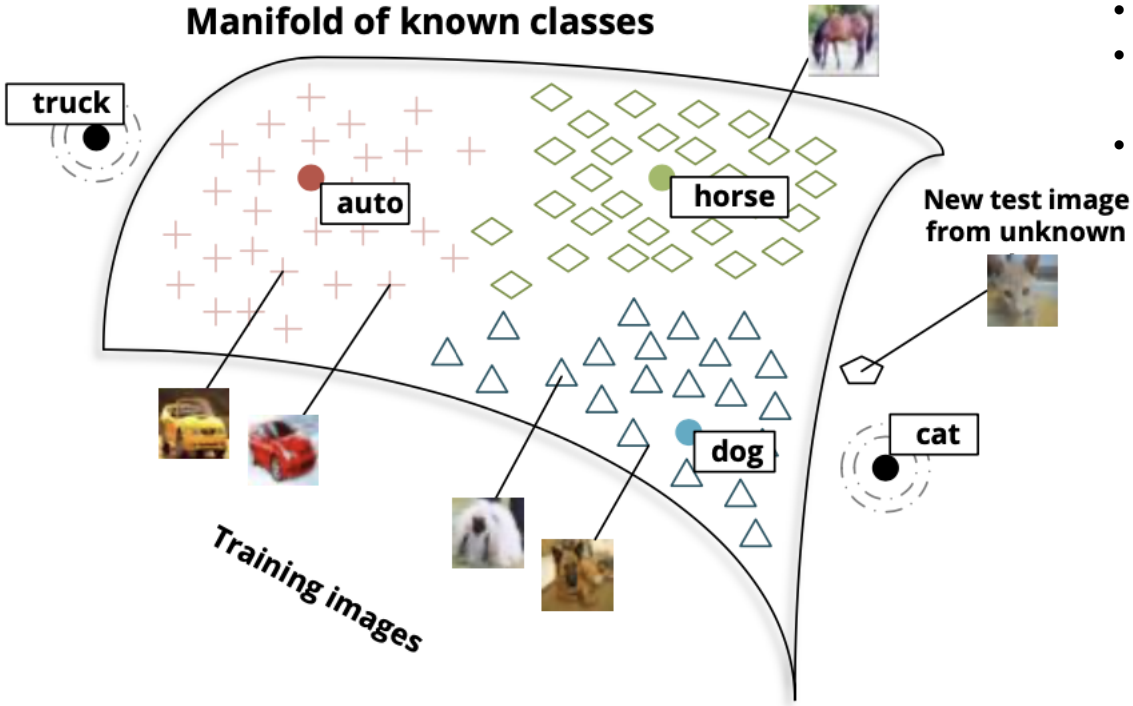
Tanveer Syeda-Mahmood

Stanford MEDICINE

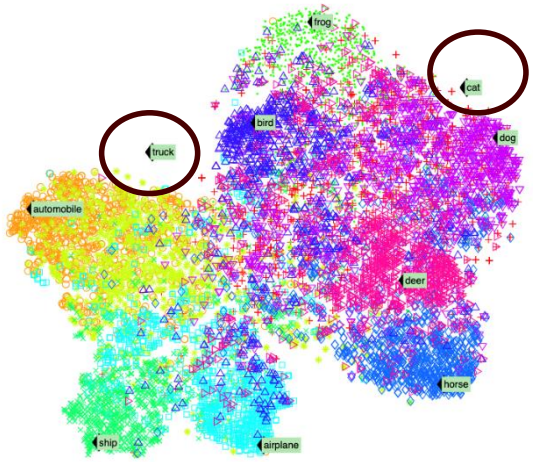
# Vision-language models (VLM) - Evolution

- Zero-shot learning
  - Had the label vocabulary defined, images of all labels not provided
  - The classifier learns the association between labels and can carry over the analogy to new ones
- Image Captioning
  - Vision encoder for objects
  - Text decoder for generation of text
  - Only image at inference
- Visual QA
  - Similar to image captioning
  - Take a question during inference as well
- VLM models
  - Encoders
  - Encoder+Decoders

# Zero shot learning – early attempts to generalize



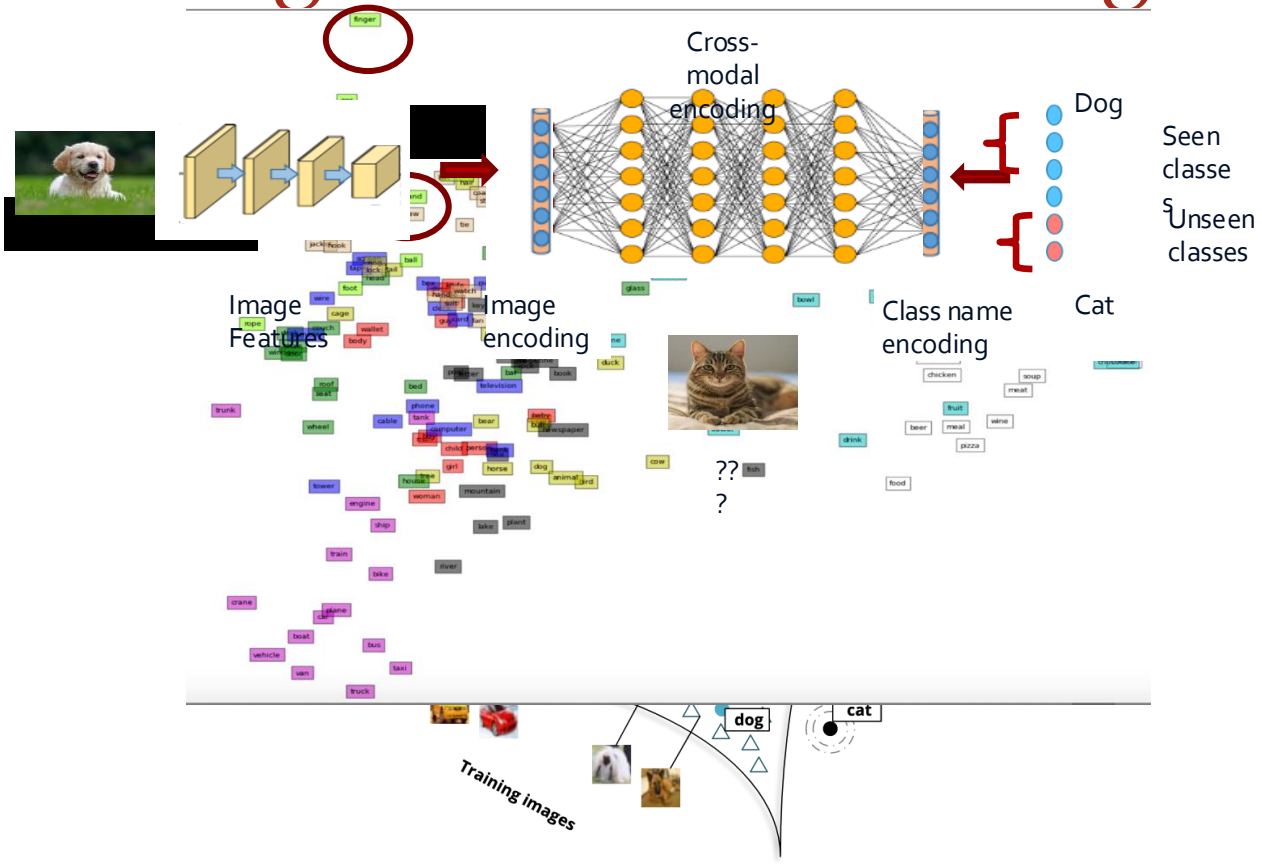

- All class labels known ahead of time
- A simple case of caption as isolated words or class names
- Whole image label prediction


*R. Socher et al., "Zero-shot learning through cross-modal transfer," in Proc. NeurIPS 2013*

# Cross-modal encoding - Zero shot learning

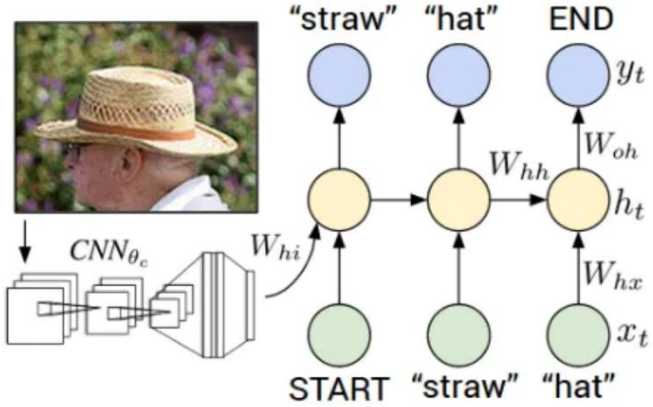
- Leverage the relationship between visual objects and their semantic class names for a joint encoding
  - Allows handling of ad hoc queries
  - unseen classes can still be recognized from the learned encoding
- Class encoding should preserve semantic distance
  - But are the semantic concepts nearby?
  - What if there are distractions nearby?
- How many unseen classes can it tolerate?
- How to scale to the full world of objects?



# Visual captioning – From image to language

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."




https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2

K. Xu et al., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, CVPR 2015

CNN feature extractor + LSTM decoder

- Trained on pairs of image and text
- Associates objects, scenes, color with text
- Caption is a longer phrase
- All vocabulary generated from captions
- Whole image caption prediction
- Visual Encoder + Language Decoder.
- Inference time: Only an image

# Used the concept of soft attention (before transformers)

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.


A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

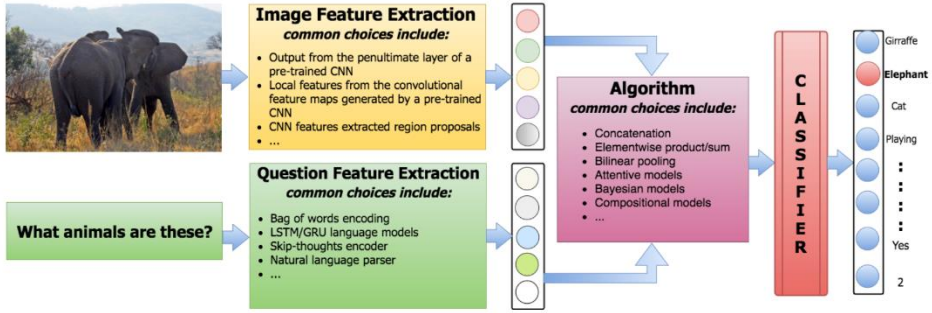
A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

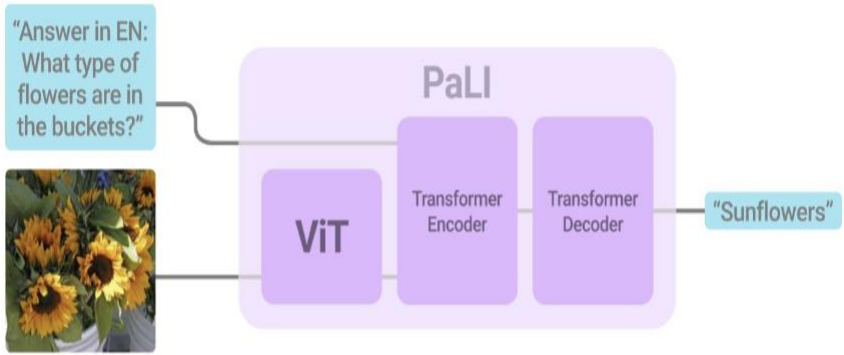
$$\mathbb{E}_{p(s_t|a)}[\hat{\mathbf{z}}_t] = \sum_{i=1}^{L} \alpha_{t,i}\mathbf{a}_i$$

Weighs the context vector to select relevant visual features for textual prediction

# Visual QA




Training : Image + text question input -> Answer
vocabulary is a classification task
Inference: Image + question input -> Answer text



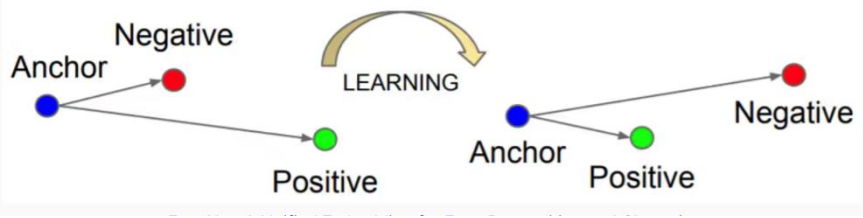

X. Chen, PALI: A jointly-scaled multilingual language-image model, ICLR 2023

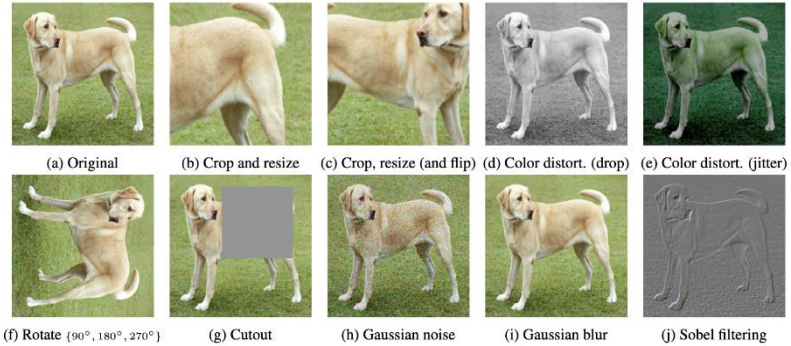
K. Kafle and K. Kanna, Visual Question Answering: Datasets, Algorithms, and Future Challenges, , 2017

# VLM through contrastive learning

Basic idea is to separate positive pairs from negative pairs

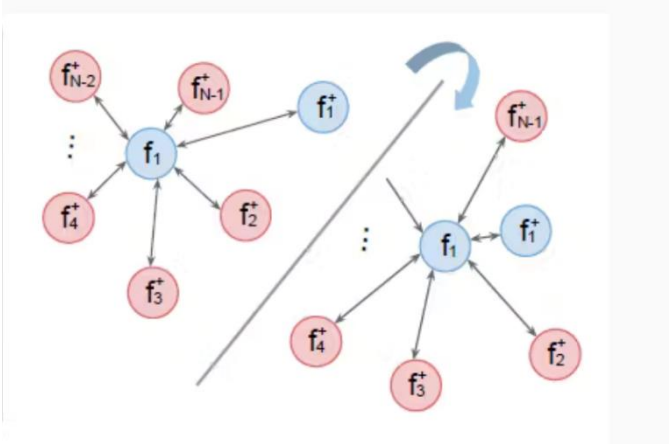
Triplet loss

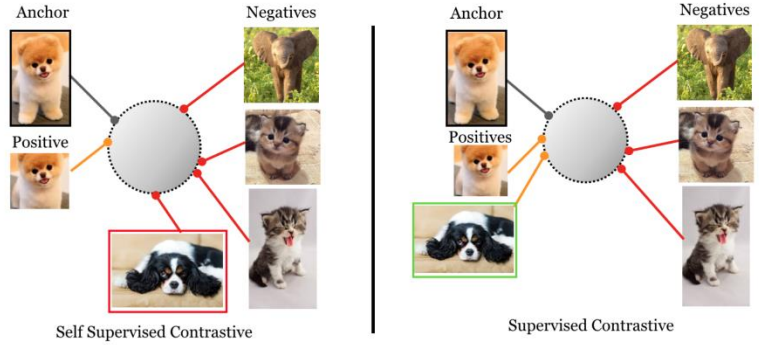



Positive examples through augmentation

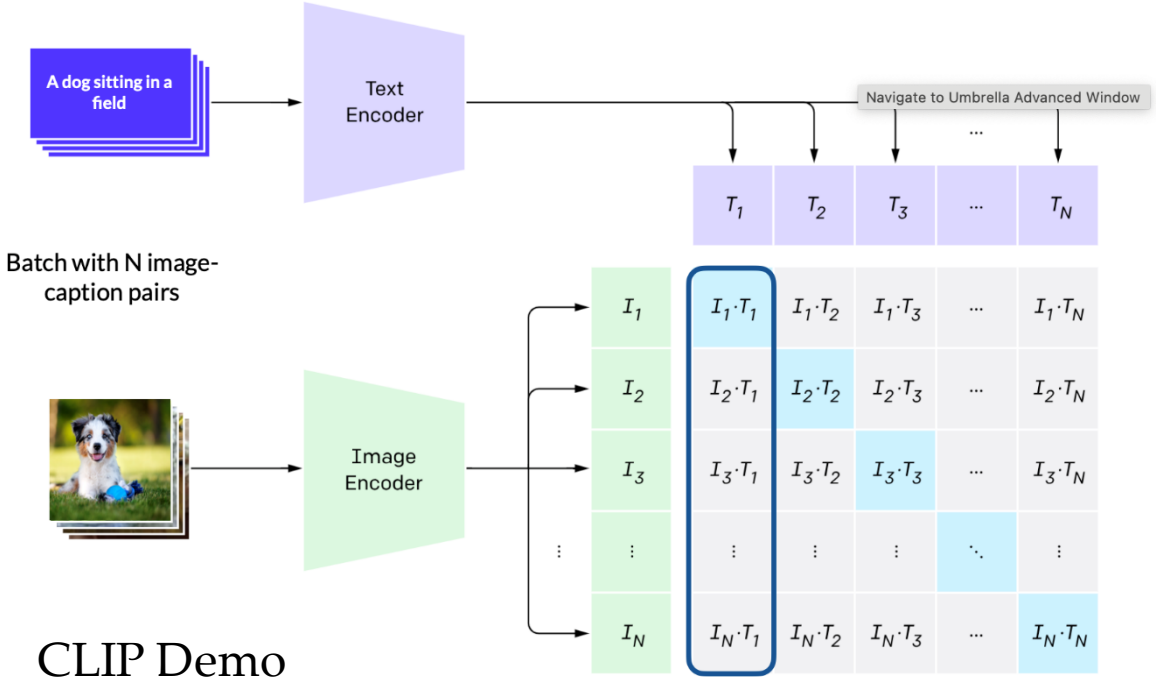
N-pair loss



# Contrastive Language-Image Pretraining (CLIP)

**Key Idea**: Maximize the similarity between true image-text embedding pairs and minimize similarity between mismatched image-text embedding pairs




CLIP Demo

**Objective:** InfoNCE Loss Function

$$L_{I \to T} = \sum_{k=1}^{N} -\log \frac{exp(I_k \cdot T_k / \tau)}{\sum_{j=1}^{N} exp(I_k \cdot T_j / \tau)}$$

$$L_{T \to I} = \sum_{k=1}^{N} -\log \frac{exp(I_k \cdot T_k / \tau)}{\sum_{j=1}^{N} exp(I_j \cdot T_k / \tau)}$$

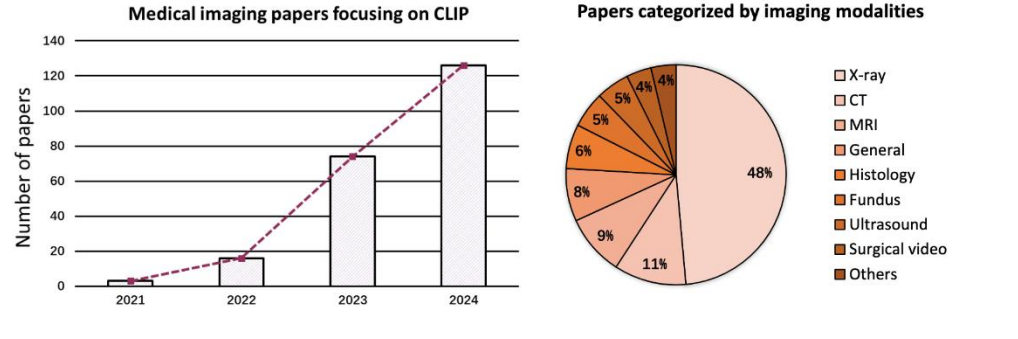
$$L = L_{T \to I} + L_{I \to T}$$

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

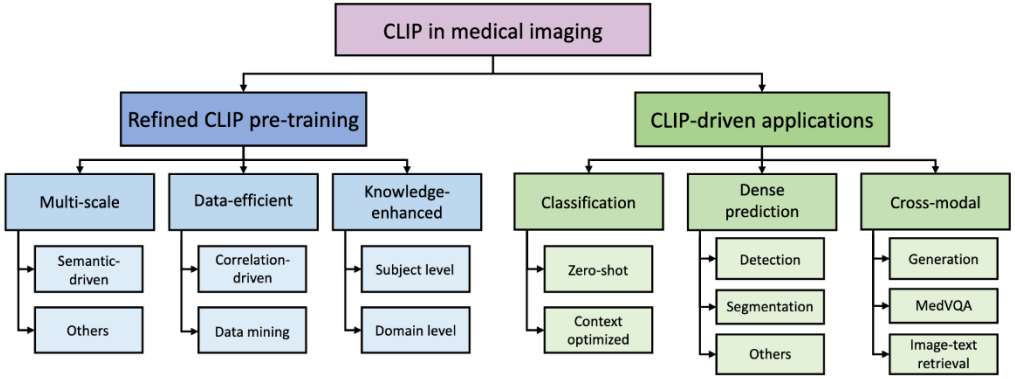
# What can CLIP-style VLMs be used for?

- Feature encoders
- Object detection, recognition
- Auto-tagging images with captions
  - Zero shot image classification
- Image-text matching
- Cross-modal retrieval
- Multimodal VLM decoder models
- Image generation (DALL-E, Stable Diffusion)

# CLIP widely adopted in medical imaging






Zhao et al. [CLIP in medical imaging: A survey]



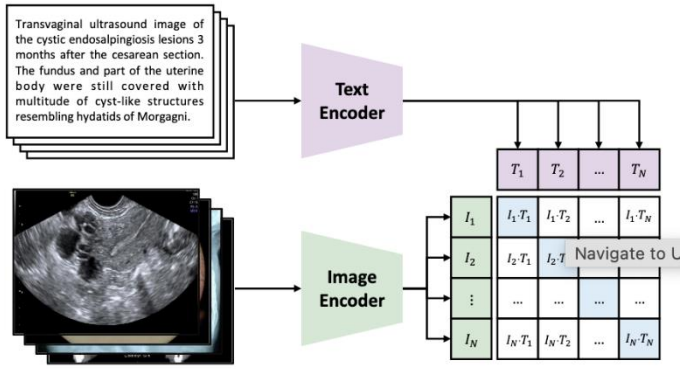



**Fig. 3. Illustration of CLIP in medical imaging, with an example from the PMC-OA dataset.**

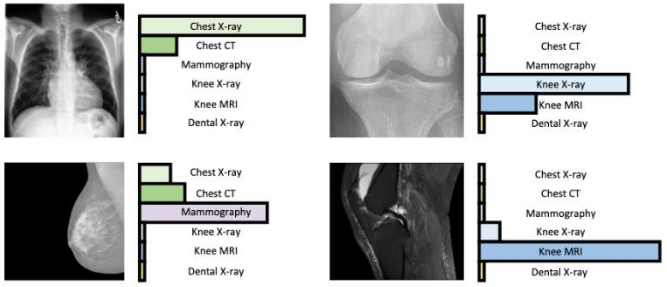



**Fig. 4. Illustration of CLIP's generalizability via domain identification.**

# CLIP training issues

Everything on the diagonal is for matched pair of image and text (positive example), and non-diagonal are mismatched pair of image and text (negative examples)

- But are other text really negative examples?
  - What if they were alternate ways of describing the same image?
  - Given very large pairs of images and text from web collections, this is likely.
- Prompts do matter – why?
- Pairs used for training were not provided.
  - So prediction needs fresh vocabulary.
- Which way does the image gravitate?
  - Since a picture contains many objects

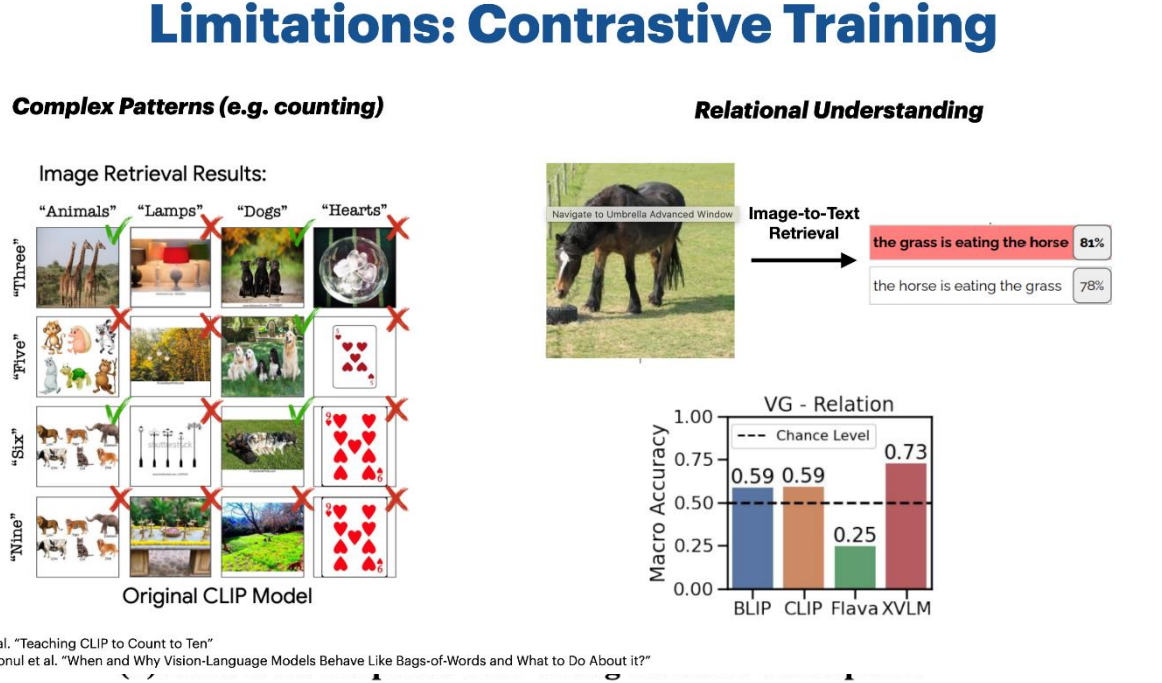
**Textual Prompts**

*Example text prompts used by CLIP for zero-shot classification on CIFAR-10*

```
templates = [
    'a photo of a {}.',
    'a blurry photo of a {}.',
    'a black and white photo of a {}.',
    'a low contrast photo of a {}.',
    'a high contrast photo of a {}.',
    'a bad photo of a {}.',
    'a good photo of a {}.',
    'a photo of a small {}.',
    'a photo of a big {}.',
    'a photo of the {}.',
    'a blurry photo of the {}.',
    'a black and white photo of the {}.',
    'a low contrast photo of the {}.',
    'a high contrast photo of the {}.',
    'a bad photo of the {}.',
    'a good photo of the {}.',
    'a photo of the small {}.',
    'a photo of the big {}.',
]
```

# CLIP Improvements

- OpenClip
  - Released a large image-text pair dataset (LAION)
- NegCLIP
- KnowledgeCLIP
- SigLIP
- SemCLIP

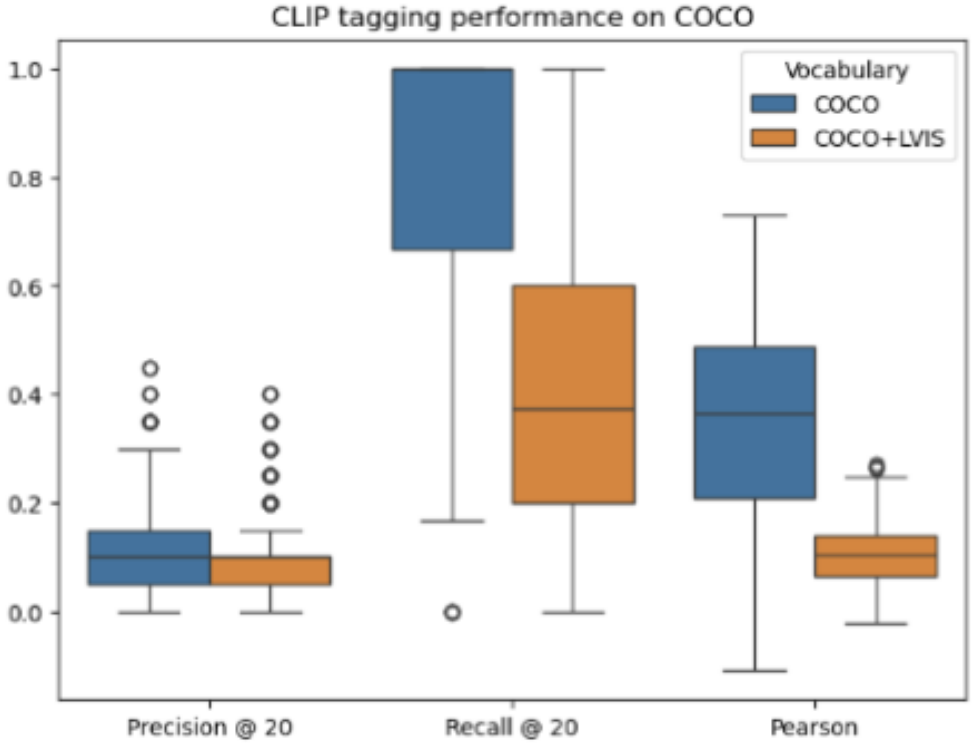



Paiss et al. "Teaching CLIP to Count to Ten"
Yuksekgonul et al. "When and Why Vision-Language Models Behave Like Bags-of-Words and What to Do About it?"

When and Why Vision–Language Models Behave Like Bags-Of-Words, and What to Do About It?" (ICLR 2023).

Contrastive Language-Image Pre-Training with Knowledge Graphs (NeurIPS'22)

# CLIP worsens with large vocabularies




COCO vocabulary = 80
LVIS vocabulary = over 1200 object categories

# Problems with large vocabularies

Visual Genome vocabulary >80,000 tags

Ground truth tags

'shadow', 'benches', 'palm tree', 'blue sky', 'cage', 'flag', 'dirt', 'light', 'lighting', 'leaves', 'trees', 'tree', 'ground', 'stone wall', 'cloud', 'pole', 'stadium', 'grass', 'sign', 'building'

VLM predicted top 50 tags

'rodeo stadium', 'stadium stands', 'amphitheater', 'grandstands', 'stadium', 'stadium area', 'spectator stands', 'stadium complex', 'stadiums', 'fans stands', 'stadium stand', 'spectator seating', 'upper stands', 'spectator area', 'grass stadium', 'track area', 'baseball arena', 'stadium region', 'soccer stadium', 'spectator seats', 'skate venue', 'arena', 'several bleachers', 'stadium box', 'dome stadium', 'spectator stand', 'cement seats', 'part of the ground', 'athletic field', 'part of track', 'baseball stadium', 'bowl area', 'track section', 'tennis arena', 'stadium seat', 'audience seats', 'grand stand', 'seating section', 'stadium wall', 'grandstand', 'ventura', 'audience area', ....

Too many redundant matches

Matching requires understanding semantics!

# CLIP with semantic clustering

This type of grouping also needed for radiology reports

1. 'stadium area', 'stadium seat', 'soccer stadium', 'stadium', 'grass stadium', 'stadium region', 'stadium wall', 'skate venue', 'dome stadium', 'rodeo stadium', 'stadiums', 'baseball stadium', 'stadium complex', 'stadium box', 'baseball arena', 'tall stadium'
2. 'stadium stands', 'fans stands', 'grandstand', 'grand stand', 'stadium stand', 'grandstands'
3. 'tennis arena', 'arena', 'copa davis', 'amphitheater'
4. 'spectator area', 'upper stands', 'spectator stands', 'spectator stand'
5. 'audience area', 'bleachers entrance', 'cement seats', 'seating section', 'spectator seats', 'spectator seating', 'audience seats', 'several bleachers'
6. 'part of track', 'track side', 'track area', 'track section'
7. 'part of the ground'
8. 'sports area', 'players field', 'football field', 'athletic field'
9. 'bowl area', 'ventura'
10. 'emptybleachers'

Textual embeddings can be used to semantic cluster top K matches

Verifying matches needs semantics

| Representative label per cluster | Ground truth label |
|---|---|
| Stadium | Stadium |
| Emptybleachers | Benches |
| Part of the ground | Ground |
| Stadium wall | Stone wall |
| Track area | Grass |
| Amphitheater | Building |

# CLIP retrieval improved with semantic clustering and matching

(a) Image for tagging

[sweni lodges, cocos properties, above ground pool, cabana area, jacuzzi, surf pool, cabanas, bathing tub, spa tub, exterior of the tub, pool area, tub exterior, water pool, swimming pool, cabana roof, cabana, an outdoor image, hot_spring, small pool, hot tub]

(b) Top 20 tags from CLIP

[pool, palm tree, deck, it, leg object, side, hair, glass, chair, bikini top, hand, lounge chair, water, table, tree]

( c) Ground truth tags

[terrace, outdoor photo, water pool, brasil, large basin, bath tub, lagoon, person swimming, top of chair, apia, family home, lounges, parage, outdoor area, water point, view, wood deck, circular area, holiday, lifeguard stand]

(d) Top 20 semantic cluster representative tags

Pool -> water pool
Lounge chair -> lounges
Table ->lifeguard stand
Woman -> person swimming
Deck -> wood deck
Building -> family home
Water -> water point

(e) Ground tags matched with semantic matching

| Dataset | Images | Labels | CLIP | | | | SigLIP | | | | CLIP-ViT-BigG | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | A | B | C | D | A | B | C | D | A | B | C | D |
| MS-COCO | 40670 | 568456 | 0.01 | 0.09 | 0.11 | **0.16** | .03 | 0.12 | 0.15 | **0.19** | .07 | 0.14 | 0.16 | **0.21** |
| VGenome [8] | 7554 | 83404 | 0.01 | 0.11 | 0.11 | **0.18** | .03 | 0.15 | 0.17 | **0.22** | .06 | 0.17 | 0.19 | **0.25** |
| SUN [16] | 16637 | 567 | 0.02 | 0.11 | 0.10 | **0.15** | 0.02 | 0.16 | 0.15 | **0.19** | 0.04 | 0.19 | 0.19 | **0.21** |
| CUB [12] | 11788 | 200 | 0.04 | 0.13 | 0.12 | **0.22** | 0.08 | 0.16 | 0.15 | **0.31** | 0.09 | 0.19 | 0.17 | **0.38** |
| AWA2 [15] | 6985 | 50 | 0.16 | 0.18 | 0.32 | **0.35** | 0.28 | 0.31 | 0.35 | **0.41** | 0.29 | 0.34 | 0.36 | **0.45** |

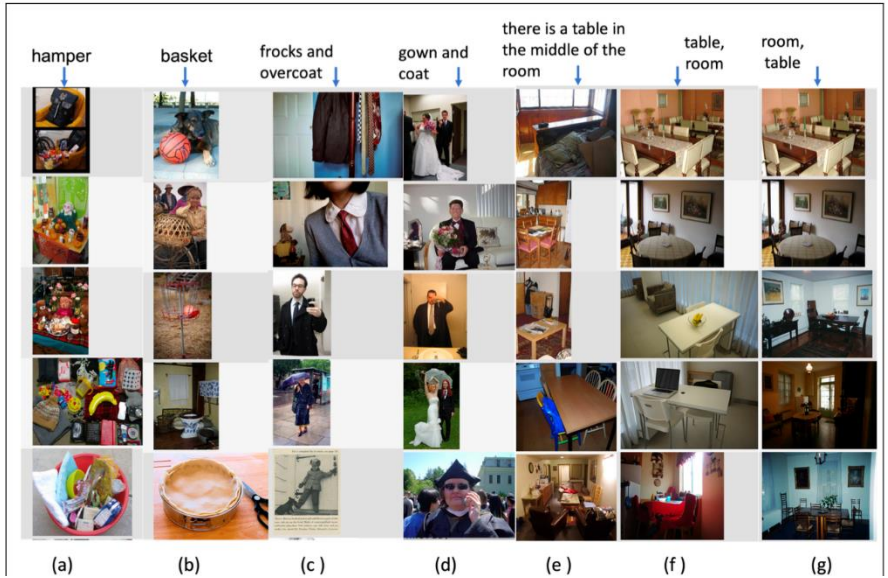
# Does CLIP understand text semantics?



https://huggingface.co/spaces/vivien/clip

CLIP demo

# Synonyms not well recognized by CLIP

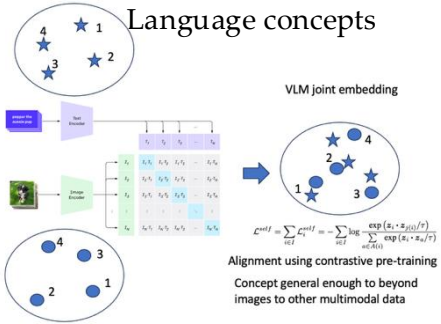


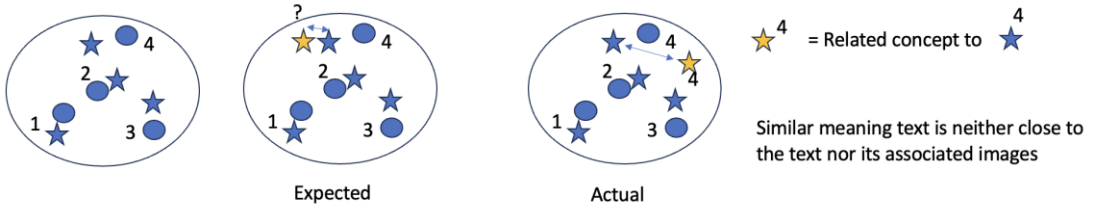

Problem is with text embeddings!

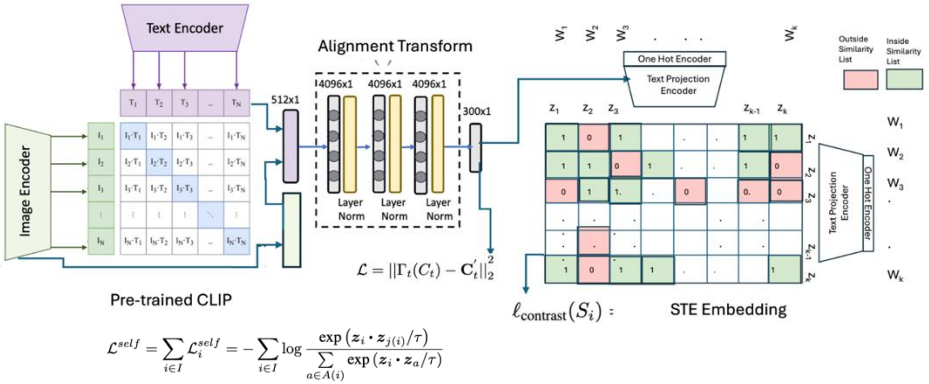
Table 2: Illustration of synonym recognition across text embeddings.

| Embedding | # Queries | Synonyms in Top10 | %age synonyms covered |
|---|---|---|---|
| CLIP (Radford et al., 2021) | 71895 | 28070 | 49.27% |
| SBERT (Reimers & Gurevych, 2019b) | 71895 | 37888 | 52.7% |
| Ours | 71895 | 67309 | 87.7% |

# SemCLIP – Semantic Vision Language Model




$$\mathcal{L}^{self} = \sum_{i \in I} \mathcal{L}_i^{self} = -\sum_{i \in I} \log \frac{\exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}$$

$$\mathcal{L} = ||\Gamma_t(C_t) - C_t'||_2^2$$

$$\ell_{\text{contrast}}(S_i) =$$

Aligns to semantic memory

$$L_{contrast}(S_i) = \sum_{W_j \in S_i} \log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}$$

$$L_{contrast} = \sum_{j}^{|V|} L_{contrast}(S_j)$$

# Many choices of loss functions

- Contrastive variants
  - Contrastive loss (ITC), Matching loss (ITM)
  - Self-supervision loss (SimCLR)
  - Text-to-pixel contrastive (TPC)
  - Region-word alignment loss (RWA)
  - Region-word contrastive loss (RWC)
  - Unified contrastive learning (Uni
  - Supervised contrastive loss
  - Sigmoidal Loss (SigLIP)
- Generative variants
  - MLM (bi-directional)
  - LM (auto-regressive)
  - Cap Loss (predict next token given tokens and whole image)
  - Flamingo Loss (predict next token given previous image and text tokens)

$$\mathcal{L}^{self} = \sum_{i \in I} \mathcal{L}_i^{self} = -\sum_{i \in I} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_{j(i)}/\tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a/\tau\right)}$$

$$\mathcal{L}_{out}^{sup} = \sum_{i \in I} \mathcal{L}_{out,i}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p/\tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a/\tau\right)}$$

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{x^t \sim D}\left[\log p(x^t | \hat{x}^t)\right].$$

$$\mathcal{L}_{\text{LM}} = -\mathbb{E}_{x^t \sim D}\left[\sum_{l=1}^{L} \log p(x_l^t | x_{<l}^t)\right]$$

$$\mathcal{L}_{\text{Cap}} = -\mathbb{E}_{x \sim D} \sum_{l=0}^{L} \log p(x_l^t | x_{<l}^t, x^v),$$

$$-\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \underbrace{\log \frac{1}{1 + e^{z_{ij}(-t\mathbf{x}_i \cdot \mathbf{y}_j + b)}}}_{\mathcal{L}_{ij}}$$

# Language generation from images

- A pure LLM is trained with causal language modeling for next token prediction:
  - Maximize $P(w_t \mid w_1, w_2, \ldots, w_{t-1})$ $\mathcal{L}_{batch} = \frac{1}{B} \sum_{i=1}^{B} \sum_{t=1}^{T_i} -\log P_\theta(x_t^{(i)} \mid x_1^{(i)}, \ldots, x_{t-1}^{(i)})$
- VLM for text:
  - Align visual and textual representations
  - Generate text conditioned on images
    - Prefixing tokens in transformers
    - Instructional tuning through VQA
    - Adapter layer -> prefixing, deep mixing

# Aligning visual textual representation

- Aligning visual-textual representations
  - Zero-shot learning
  - Contrastive loss (e.g. CLIP, SimCLR)
  - Image-text matching (e.g. in **BLIP**, **UNITER**, **ViLBERT**, **FLAVA)**
    - (image, text)->Match/NoMatch
- Multimodal language modeling:

$$P(text \mid image)$$

input → A dog in a park

Target

- Masked language modeling with vision context: (**ViLBERT**, **LXMERT**, **UNITER**, **FLAVA)**

A - is siting on a table → cat

# How are pure LLM for text trained?

- Maximize $P(w_t \mid w_1, w_2, \ldots, w_{t-1})$

Next token prediction through softmax and linear layer $z_t = W_{out} \cdot h_t + b$

- The model produces logits $z_t \in \mathbb{R}^V$ for a vocabulary of size $V$.
- Probabilities via softmax:

$$P_\theta(x_t = v \mid x_{<t}) = \frac{\exp(z_{t,v})}{\sum_{v'=1}^{V} \exp(z_{t,v'})}$$

- Loss uses the log of the probability for the **ground truth token** $x_t$.

$$\mathcal{L}_{batch} = \frac{1}{B} \sum_{i=1}^{B} \sum_{t=1}^{T_i} -\log P_\theta(x_t^{(i)} \mid x_1^{(i)}, \ldots, x_{t-1}^{(i)})$$

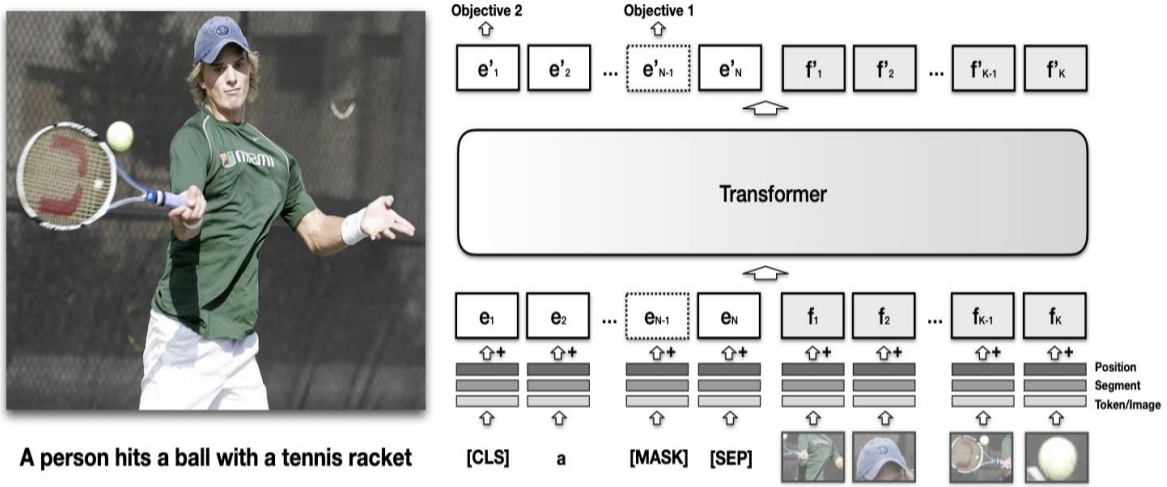
Minimizing this loss is equivalent to **maximizing the likelihood of the observed data**:

$$\hat{\theta} = \arg\max_\theta \prod_{i=1}^{B} \prod_{t=1}^{T_i} P_\theta(x_t^{(i)} \mid x_{<t}^{(i)})$$

MLE

# VLM- Combining image and text in a transformer

Translating images into embedding features that can be jointly trained with token embeddings.

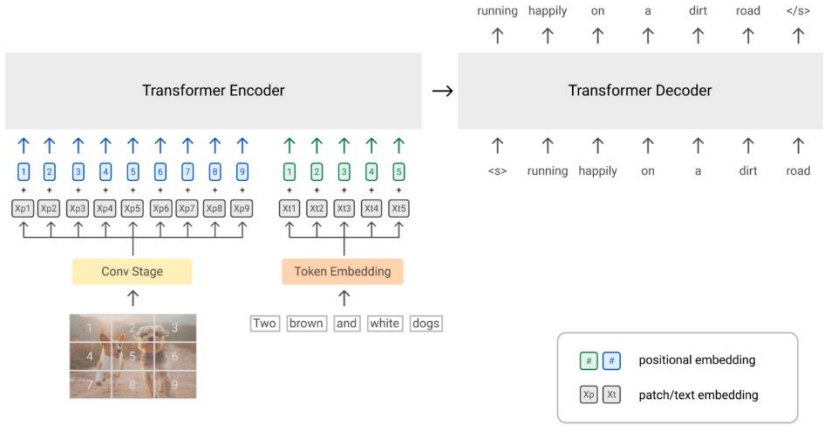


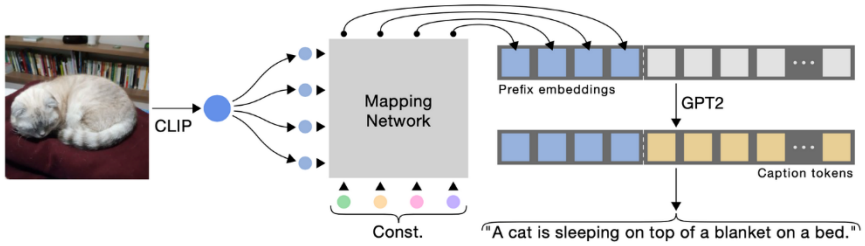

L. H. Li et al. VisualBERT: A Simple and Performant Baseline for Vision and Language, ACL 2020

2019 -2021

# VLM- Vision prefix for an LLM

Learning good image embeddings that can work as a prefix for a frozen, pre-trained language model.

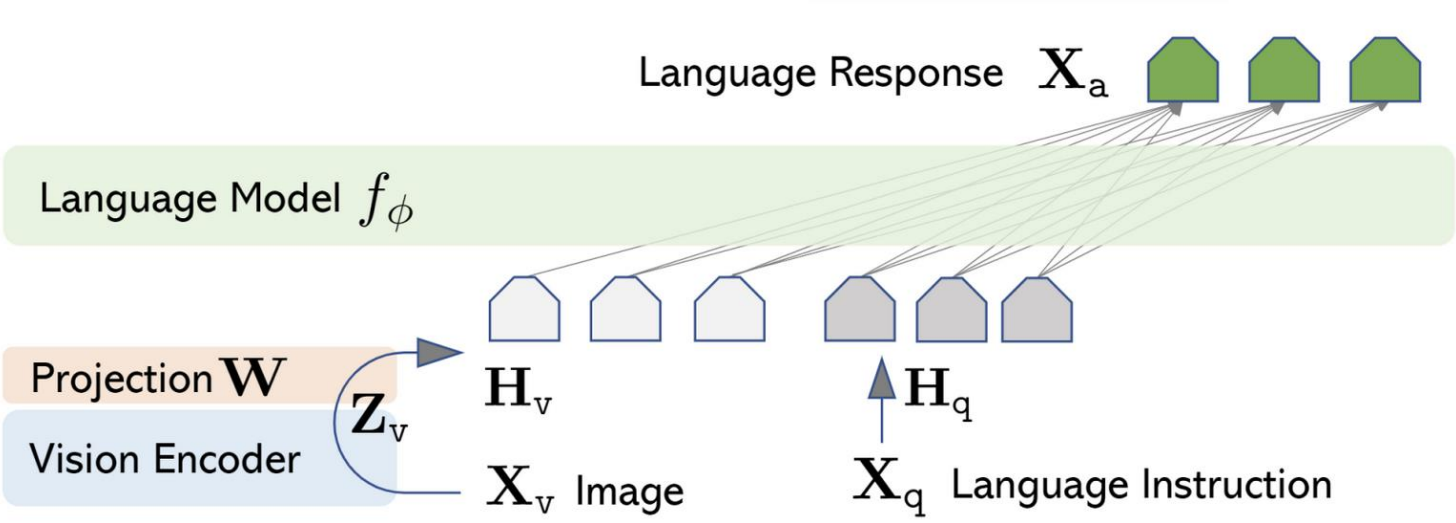



Frozen, CLIPCAP, language models with vision encoders

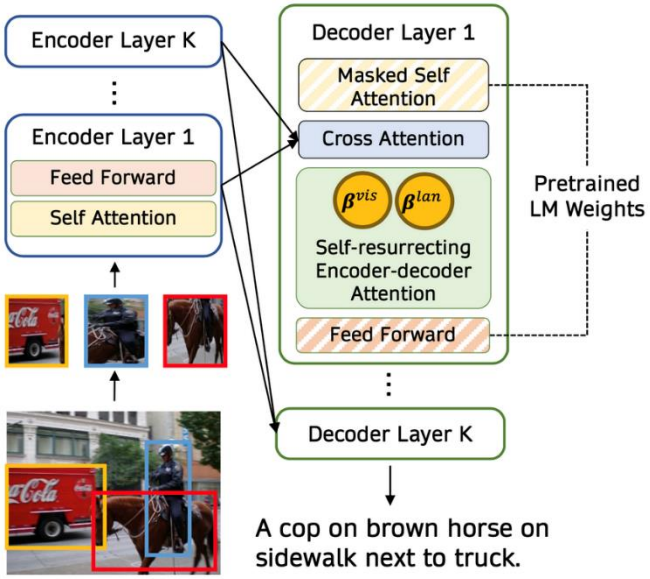
2019 -2021

# Llava Model

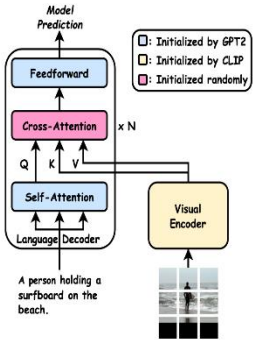



Visual Instruction Tuning, NeurIPS2023

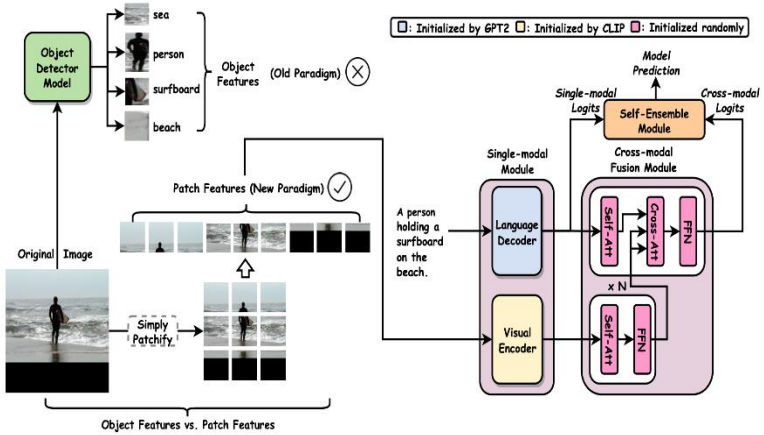
# VLM - Text-Image Cross-Attention Fuse Mechanisms




VisualGPT




(a) The Vanilla Framework.

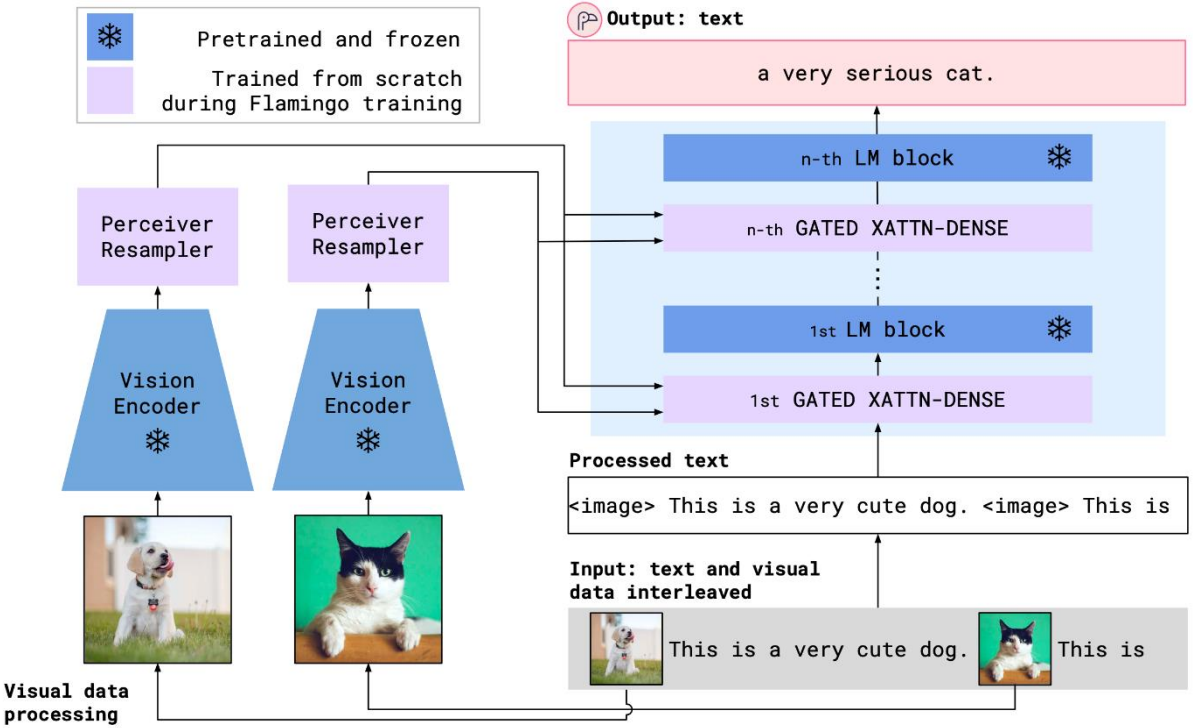



(b) Our End-to-End VC-GPT Framework.

VC-GPT

# VLM - Multimodal Generative AI

Flamingo




Interleaved placement of multimodal input

# What does the adapter do?

- **Dimensional mismatch** between visual embeddings and LLM token embeddings.
- **Information selection**, by letting the LLM attend only to the most relevant visual features.
- **Low-data learning**, allowing instruction-following behaviors to emerge without massive retraining.

# How to generate text from images?

- Older methods focused on label generation beyond classification
  - Zero-shot learning
  - Aligning visual and textual representation
- Basic approach:
  - Image encoder
  - Language decoder
  - Transition from image to language (during training)
    - By prefixing image tokens to language tokens
    - Cross-attention where the language model attends to visual tokens looking back at relevant parts of the image
    - By mapping from image to language through an adapter layer
- Training datasets:
  - Image-label, Image-caption datasets
  - Instruction following multimodal datasets
  - VQA

# Generating text conditioned on images

- Prefixing tokens in transformers
- Instructional tuning through VQA
- Adapter layer

```
(Image + Instruction text) → Response text
```

```
User: What is the person holding?
<Image>
Assistant: The person is holding a tennis racket.
```

# Typical modern VLM training

- Gemini, Llama, Granite, etc. follow Llava paradigm of projecting image to text through an MLP layer
  - Freeze LLM, vision encoder but train MLP
- Do vision language alignment using image-text captions
  - E.g. LIAON-5B captions for alignment
- Continue with multimodal instruction tuning

```
User: What is the person holding?
<Image>
Assistant: The person is holding a tennis racket.
```

[<img_tokens>] [<instruction_text_tokens>]

Prefixing (Llava)

Cross-attention (Flamingo)

# Instructional styles

| Model Style | Token Ordering | Notes |
| --- | --- | --- |
| LLaVA / MiniGPT-4 | `[image tokens]` + `[instruction tokens]` | Simple prefix; uses standard LLM attention |
| Flamingo / OpenFlamingo | `[instruction tokens]` + cross-attend `[image tokens]` | Image tokens in memory; not part of main LLM sequence |
| BLIP-2 | `[image tokens]` → Q-Former → LLM prefix | Adapter condenses visual tokens before prefix |

# Human-Annotated Instruction Datasets

| Dataset | Size | Focus | Notes |
|---|---|---|---|
| **VQAv2 Instruction Pairs** | ~200K Q-A | Visual Question Answering | Converted to instruction–response format (e.g., "What is the person holding?" → "The person is holding a tennis racket") |
| **GQA Instruction Pairs** | ~100K | Compositional reasoning | Helps model learn reasoning over objects and relations |
| **TextVQA / TextCaps** | 100K+ | OCR + image reasoning | Instruction: "Read the text in this image" |
| **COCO / Flickr30k Instruction Conversion** | ~100K | Captioning framed as instructions | Prompt engineering converts captions to instruction-response pairs |

# LLM-generated synthetic datasets

| Dataset | Size | Focus | Notes |
| --- | --- | --- | --- |
| **LLaVA Synthetic Instruction Dataset** | ~595K | Instruction-response pairs for images | GPT-4 takes an image caption or VQA pair and generates diverse instructions and answers |
| **MiniGPT-4 Synthetic Conversations** | 1M+ | Multi-turn dialogue over images | GPT-4 converts COCO/Flickr captions into conversations |
| **OIG (Open Instruction Generalist)** | 100K+ text-instruction templates | Adapted to multimodal by pairing images | Reuses textual instruction templates for visual context |

# Multi-turn instructional datasets

| Dataset | Focus | Example |
| --- | --- | --- |
| **VQA + dialogue pairs** | Multi-turn reasoning | User: “How many people are wearing hats?” Assistant: “Three people are wearing hats.” |
| **MiniGPT-4 conversations** | Multi-turn multimodal dialogue | User: “Describe this image in detail.” → Assistant: “The image shows…” |
| **LLaVA-Instruction-Enhanced** | Chain-of-thought reasoning | Image + instruction → stepwise answer |

# Where is the field of VLM going?

- Towards deeper visual reasoning
  - **Vision Encoder → LLM**:
    - Most models use pretrained vision encoders and feed embeddings into LLMs.
  - **Instruction or Chain-of-Thought Training**:
    - Reasoning emerges when LLMs are fine-tuned with step-by-step visual instructions.
  - **Cross-Attention / Multi-layer Fusion**:
    - Deep integration (Flamingo, Kosmos-2) enables iterative reasoning rather than one-shot captioning.
  - **Scene or Object Awareness**:
    - Some models (BLIP-2, Kosmos-2) benefit from object-level features for logical reasoning.
  - Segmentation helps reasoning

- Multimodal extensions to video, audio, time series, etc.

# Summary

- LLM were the earliest foundational models.
- Vision-language models are one large class of foundational models
- Many different architectures and loss functions for capturing vision-language associations.
- Active field of research to generate text for visual imagery
  - Enterprise applications are difficult