Illustration by Google Deepmind

COMP541
DEEP LEARNING

Lecture #13 – Multimodal Pretraining

KOÇ UNIVERSITY

Aykut Erdem // Koç University // Fall 2024

Previously on COMP541

- recap of language modeling
- GPT-3
- understanding in-context learning
- scaling laws
- Llama 3
- other LLMs
- long context models

Illustration by Google Deepmind

Lecture overview

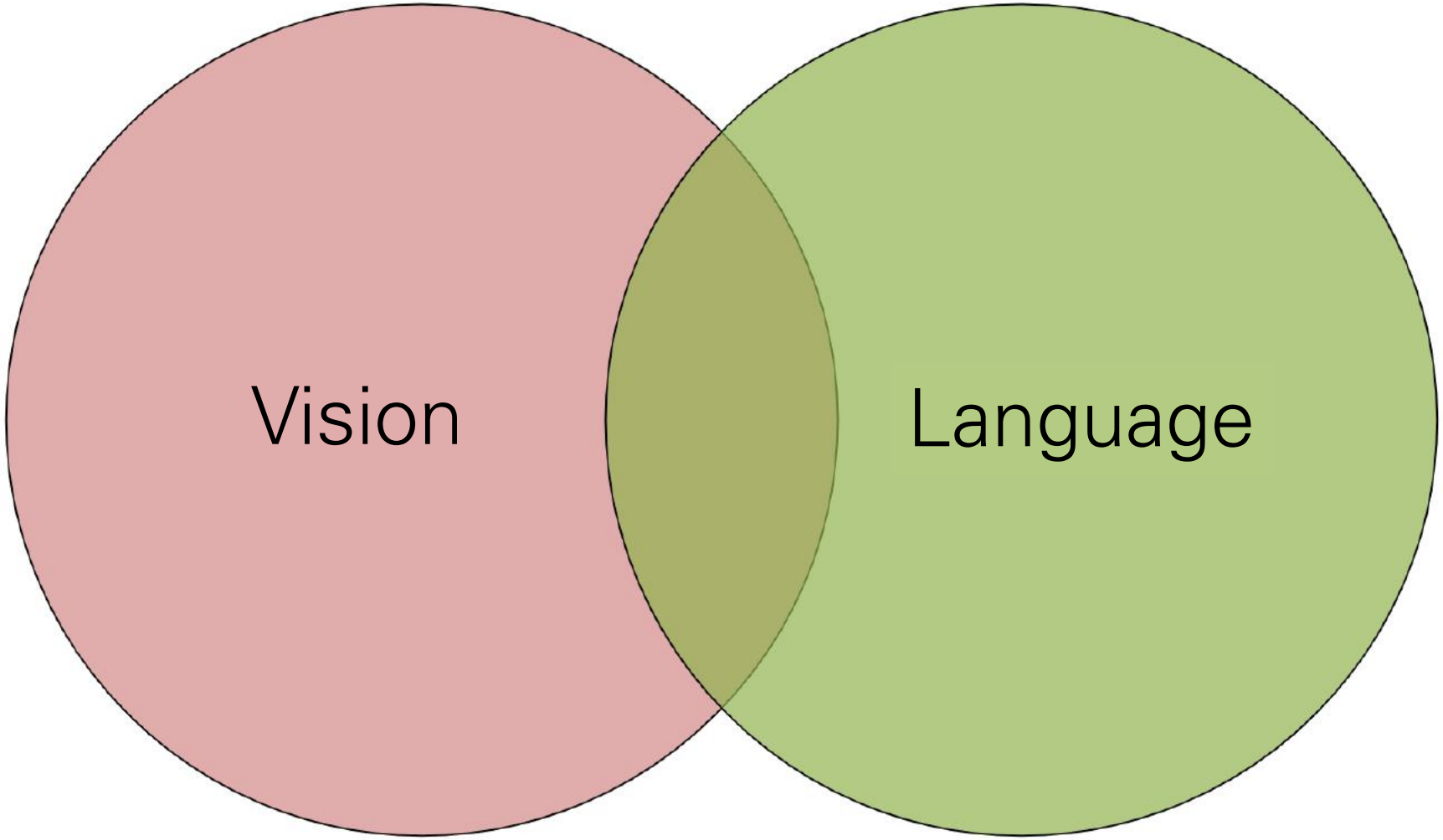
- vision-language landscape before Transformers
- vision-language pretraining
- multimodal large language models

Disclaimer: Much of the material and slides for this lecture were borrowed from

—Aishwarya Agrawal's Umontreal IFT 6765 class

—Wenhu Chen's UWaterloo CS886 class

Vision and Language (VL)



Why vision and language?

- Intuitive:
  - Humans learn in multimodal settings
- Applications:
  - Aid to visually impaired users
  - Online shopping and organizing photos
  - Grounded virtual assistants
- Scientific Curiosity:
  - Visual recognition
  - Language understanding
  - Grounding language into vision
  - Compositional reasoning
  - Commonsense reasoning

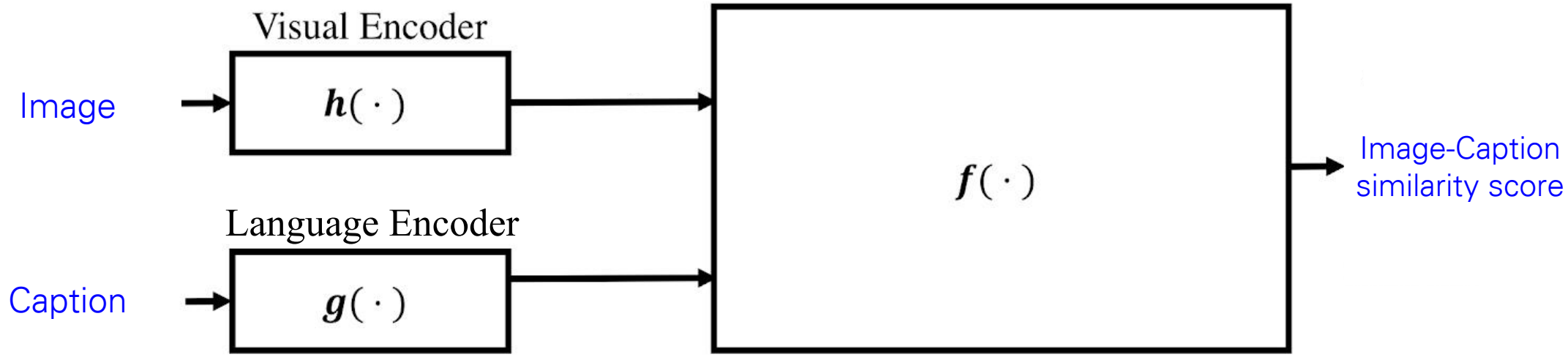
Vision-Language Landscape Before Transformers

Image Retrieval

"Grey haired man in black and yellow tie."

- High level similarity
- Easy evaluation (recall@K)

Basic skeleton of most VL models:
Image Retrieval



Grounding Referring Expressions

"The man who is touching his head."

- Spatial localization
- Finer grained grounding
- Easy evaluation (precision@1)

Basic skeleton of most VL models:
Grounding Referring Expressions

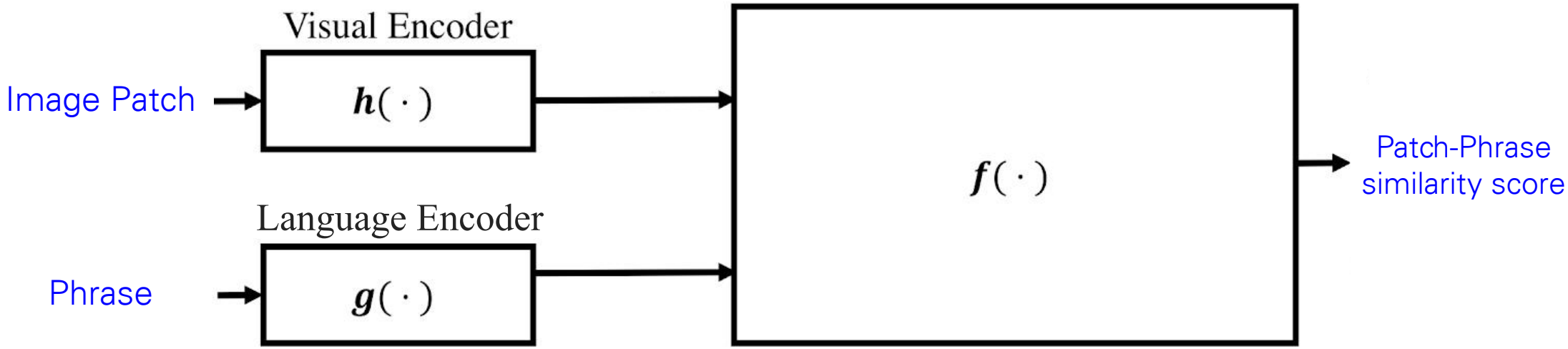


Image Captioning

"A group of young people playing a game of Frisbee."

- Language generation (in addition to visual recognition)
- Difficult automatic evaluation (BLEU, CIDEr)

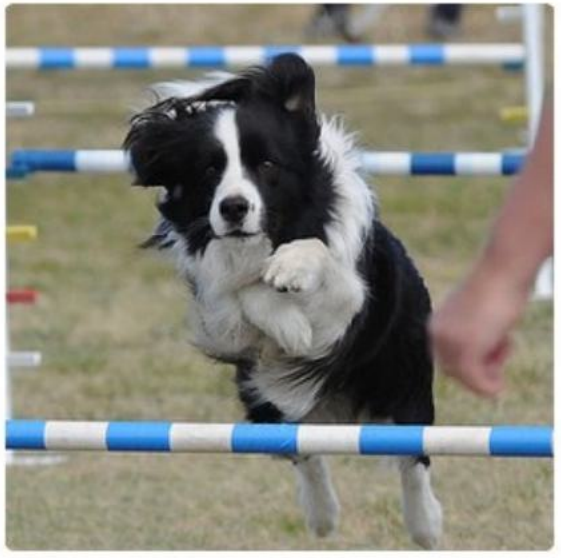
"man in black shirt is playing guitar."

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

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."



"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

Captioning datasets: UIUC Pascal Sentece

[Rashtchian et al., 2010]

- A camouflaged plane sitting on the green grass.
- A plane painted in camouflage in a grassy field
- A small camouflaged airplane parked in the grass.
- Camouflage airplane sitting on grassy field.
- Parked camouflage high wing aircraft.

- 1000 images randomly sampled from PASCAL VOC 2008 training + validation data with 20 object categories.
- 5 generic conceptial descriptions per image.

Captioning datasets: Flickr 8k, Flickr 30k

- A biker in red rides in the countryside.
- A biker on a dirt path.
- A person rides a bike off the top of a hill and is airborne.
- A person riding a bmx bike on a dirt course.
- The person on the bicycle is wearing red.

- 8k images in Flickr8k,[2] >30k images in Flickr30k,[3] with 5 descriptions
- More image sentence pairs to train and test models.
- 21% images (vs 40% images in UIUC Pascal Sentence dataset) have static verbs like *sit*, *stand*, *wear*, *look* or no verbs.

[2][Hodosh et al., 2013], [3][Young et al., 2014]

Captioning datasets: COCO [Lin et al., 2014]

- A baseball winds up to pitch the ball.
- A pitcher throwing the ball in a baseball game.
- A pitcher throwing a baseball on the mound.
- A baseball player pitching a ball on the mound.
- A left-handed pitcher throwing for the San Francisco giants.

- 120k train + validation images [vs 1k (Pascal), 31k (Flickr)].
- Instance level segmentations labels with 91 object classes and 2.5M labelled instances.
- Standard benchmark for image caption generation task.

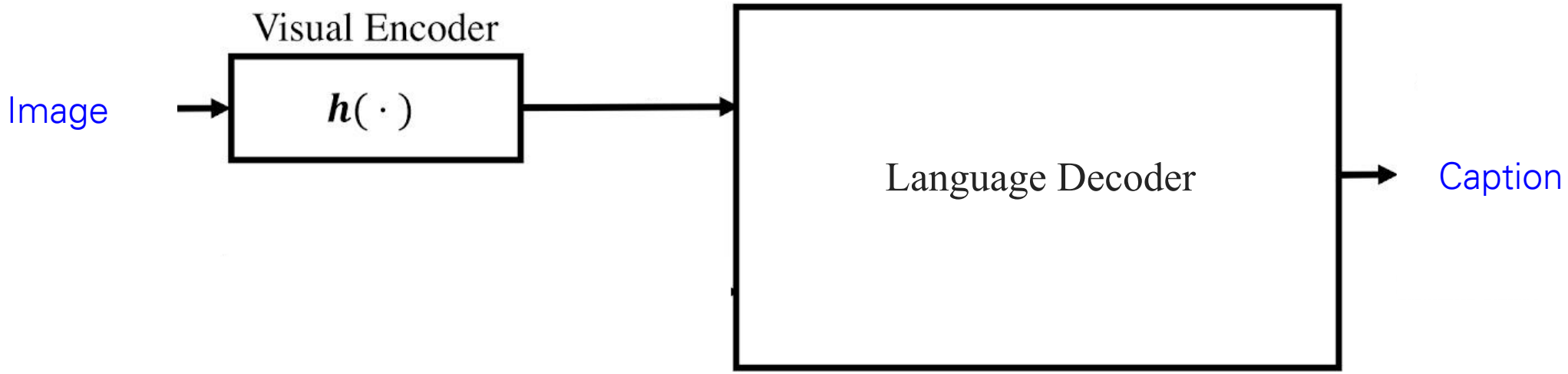
Captioning evaluation metrics

- Automatic Evaluation:
  - N-gram overlap based metrics:
  - BLEU, Rouge, METEOR, CIDEr [Chen et al., 2015]

- Semantic scene-graph based metric: SPICE [Anderson et al., 2016]

- Human Evaluation

Basic skeleton of most VL models:
Image Captioning



Neural Image Caption (NIC) (CVPR 2015)

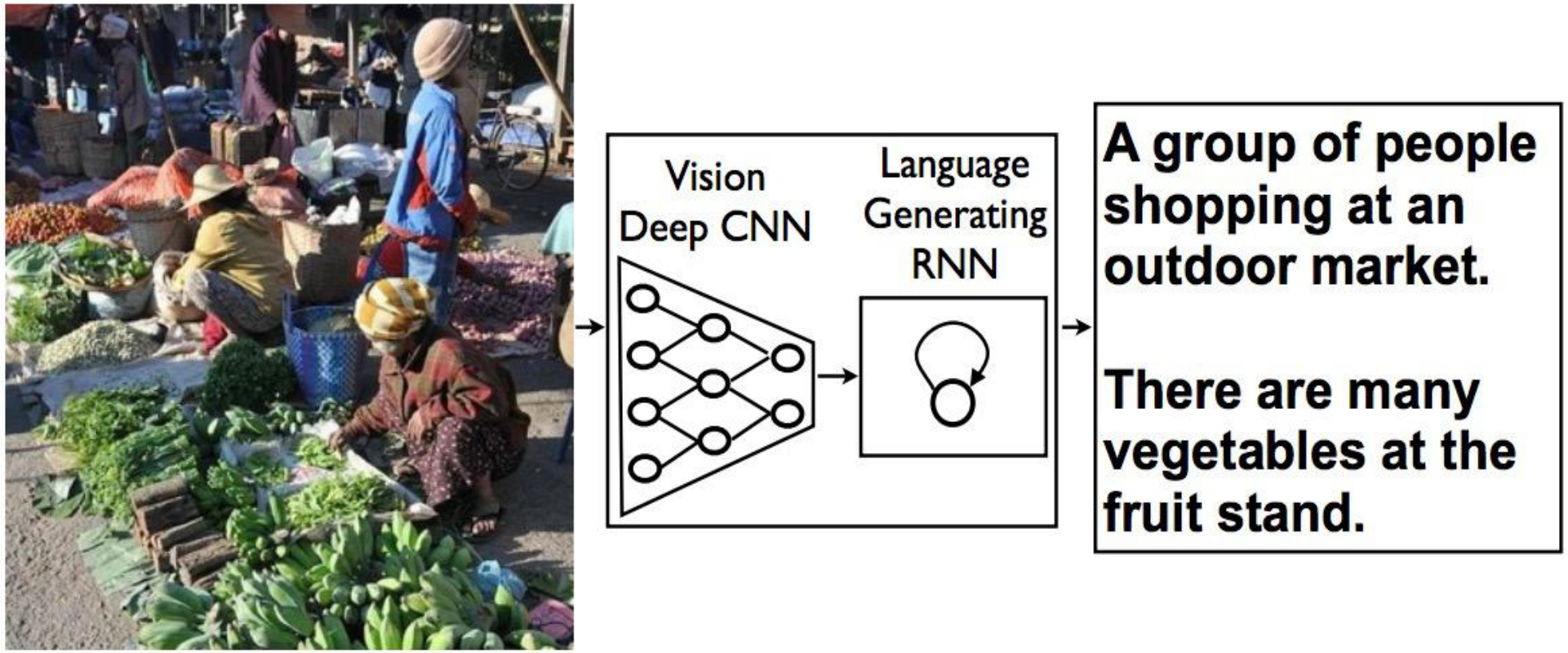


Figure Credits: Show and Tell: A Neural Image Caption Generator

Slides Credits: Jiasen Lu

a man riding a bike on a beach with a dog in the water

a man sitting at a table with a laptop

a vase filled with flowers on top of a table

a man and a woman riding on the back of an elephant

a laptop computer sitting on top of a wooden desk

a little boy standing in a field with a kite

a black and white cat sitting on a bench

Show, Attend and Tell (ICML 2015)

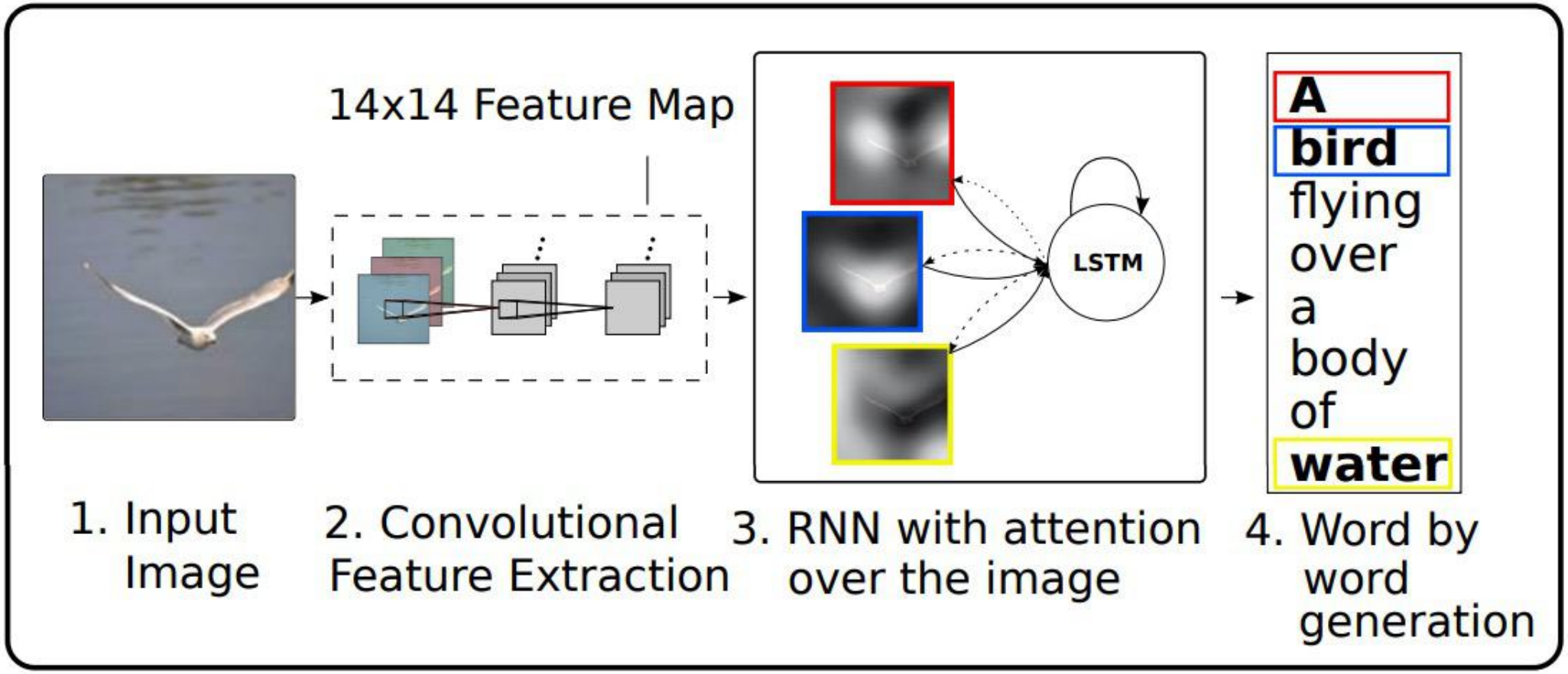


Image Credits: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Slides Credits: Jiasen Lu

Show, Attend and Tell (ICML 2015)

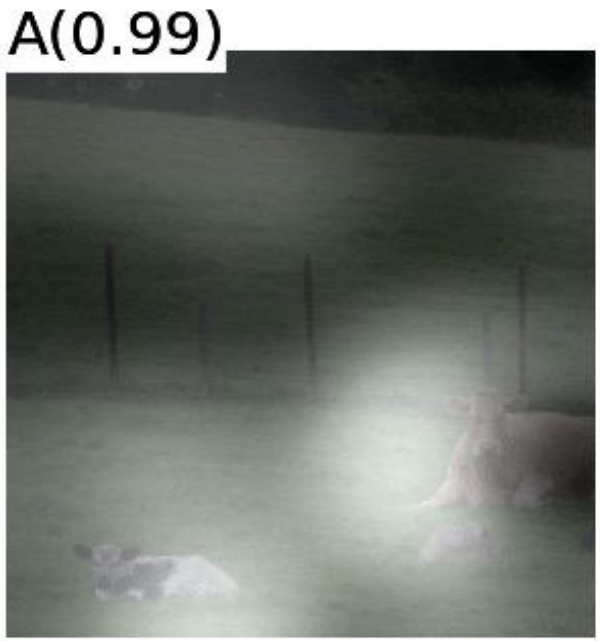
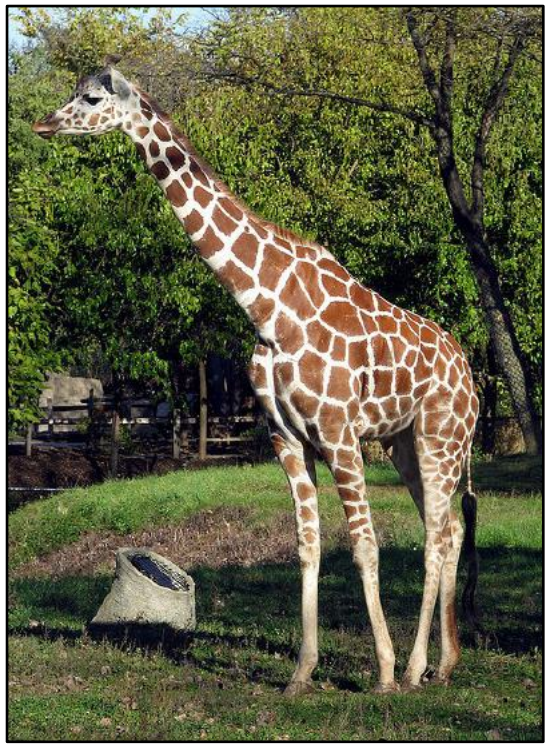
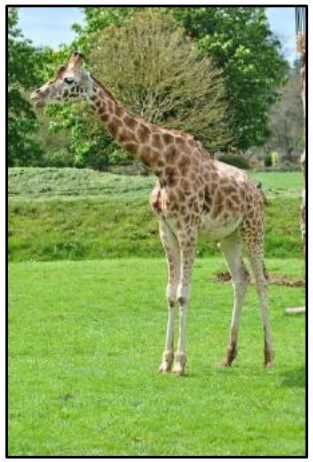


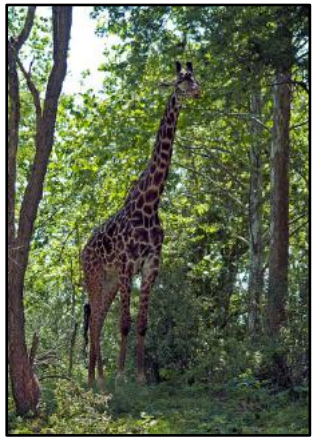
Image Credits: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Slides Credits: Jiasen Lu



A giraffe standing in the grass next to a tree.





Problems with Image Captioning

- Image captions tend to be generic
- Coarse understanding of image + simple language models can suffice

- Answer questions about the scene
  - Q: How many buses are there?
  - Q: What is the name of the street?
  - Q: Is the man on bicycle wearing a helmet?

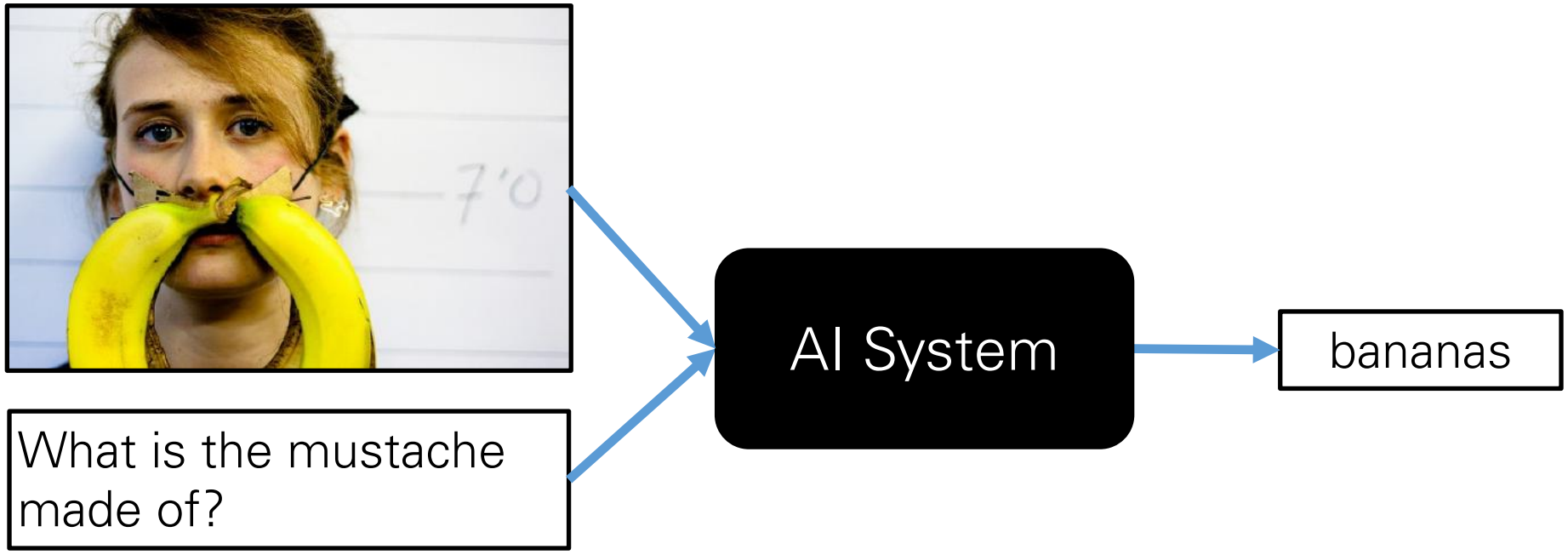
Visual Question Answering

Q: "What is the mustache made of?"

A: *"bananas"*

- Elicit specific information from images
- Relatively easier evaluation (accuracy using string matching)

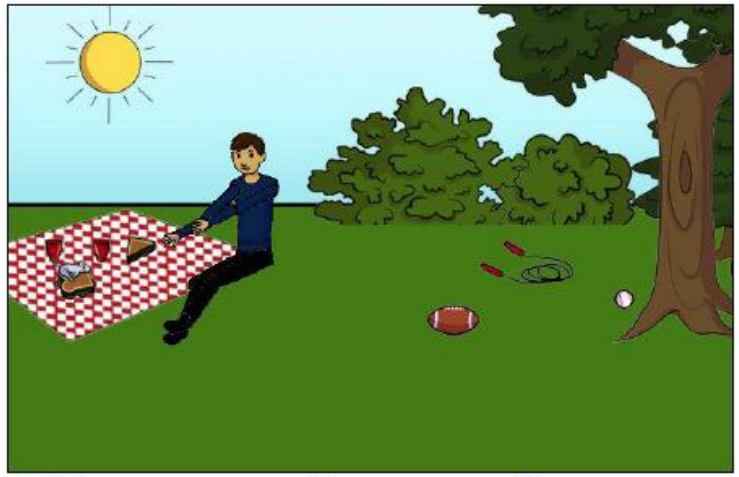
VQA Task



VQA Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

[Antol et al., ICCV15]

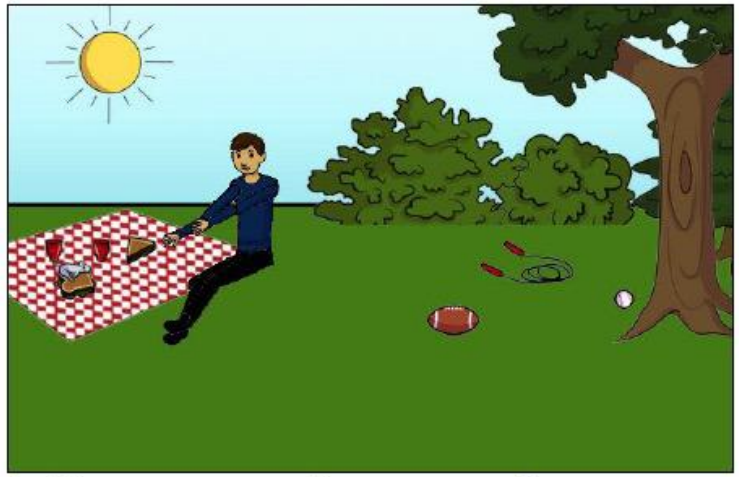
VQA Dataset

About objects

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Counting
Fine-grained recognition



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Commonsense

[Antol et al., ICCV15]

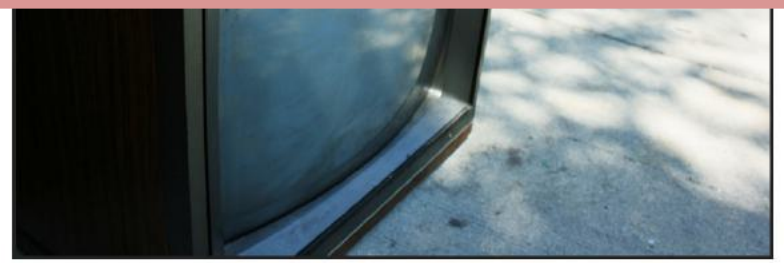
VQA Task

- Multimodal inputs – Image and Question
- Details of the image
- Common sense + knowledge base
- Task-driven
- Holy-grail of automatic image understanding

Accuracy Metric

$$\text{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

Human Accuracy: 83.3%



Ground Truth Answers:

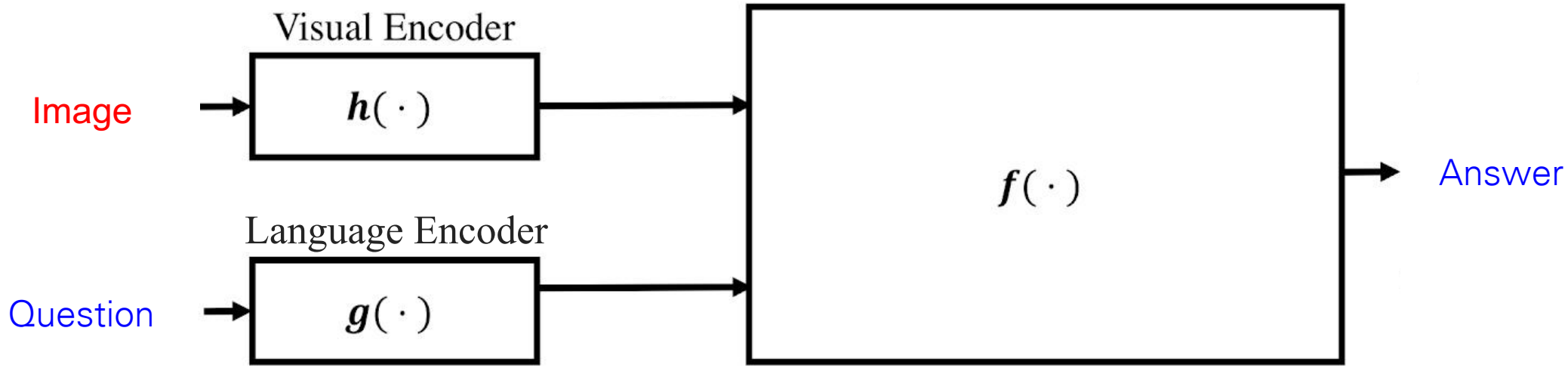
(1) 20 years	(6) old
(2) 35	(7) 80 s
(3) old	(8) 30 years
(4) more than thirty years old	(9) 15 years
(5) old	(10) very old

Q: Is this TV upside-down?

Ground Truth Answers:

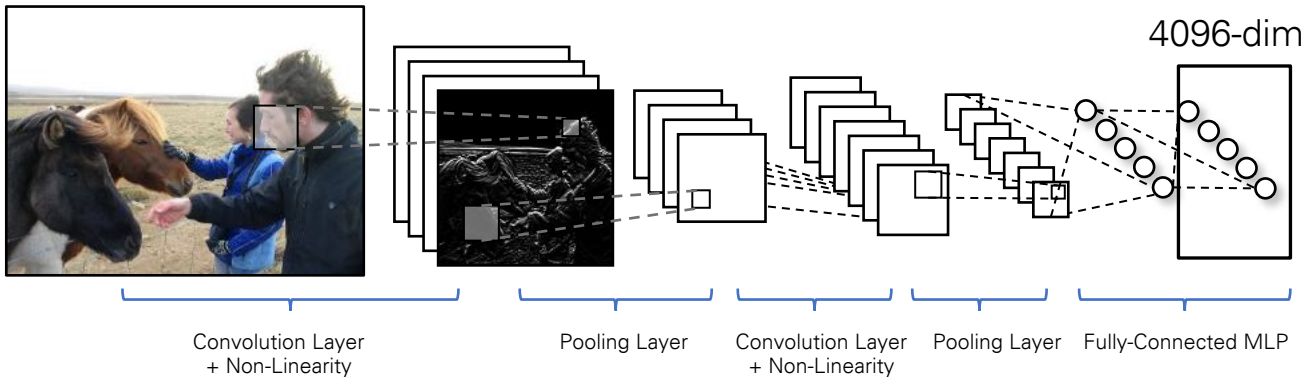
(1) yes	(6) yes
(2) yes	(7) yes
(3) yes	(8) yes
(4) yes	(9) yes
(5) yes	(10) yes

Basic skeleton of most VL models:
VQA

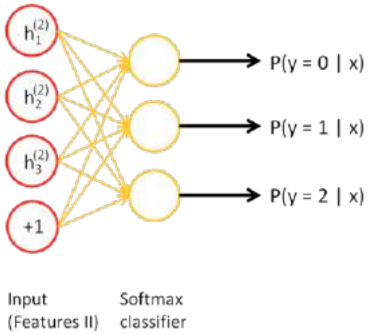


2-Channel VQA Model

ImageEmbedding

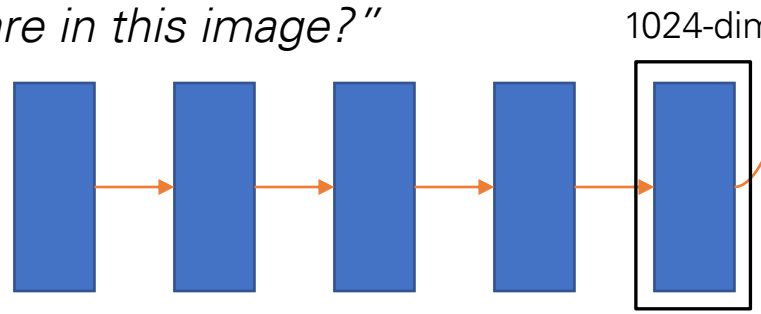


Neural Network
Softmax
over top K answers



QuestionEmbedding

"How many horses are in this image?"



Slides Credits: Jiasen Lu

Human Attention (EMNLP 2016)

What is the name of the cafe? - bagdad

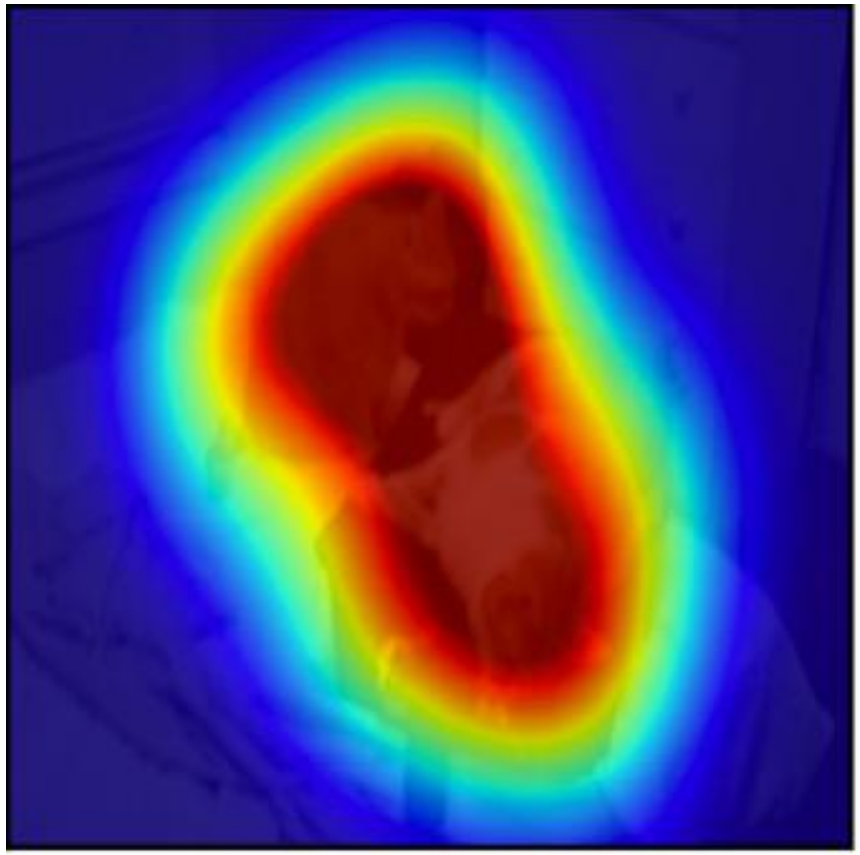
What number of cat is laying on bed? - 2

Image credit: Human Attention in Visual Question Answering:

Slides Credits: Jiasen Lu

Human Attention (EMNLP 2016)

What is the name of the cafe? - bagdad

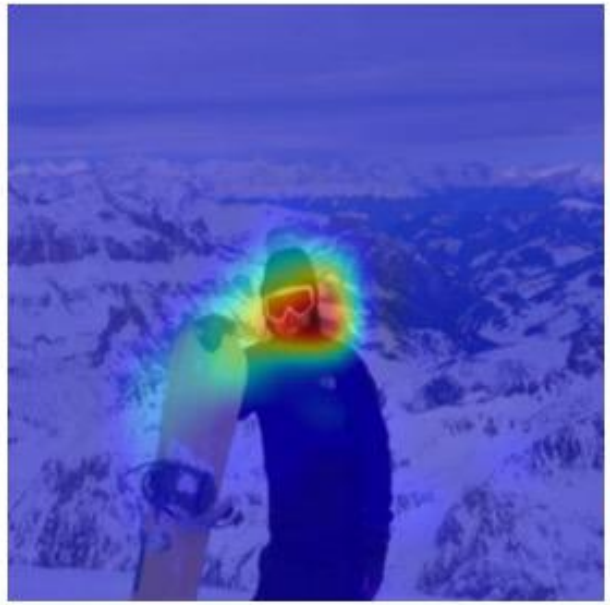


What number of cat is laying on bed? - 2

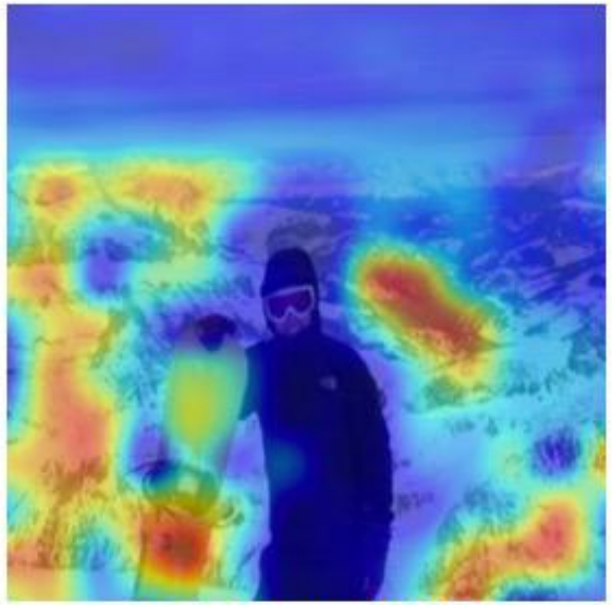
Image credit: Human Attention in Visual Question Answering:

Slides Credits: Jiasen Lu

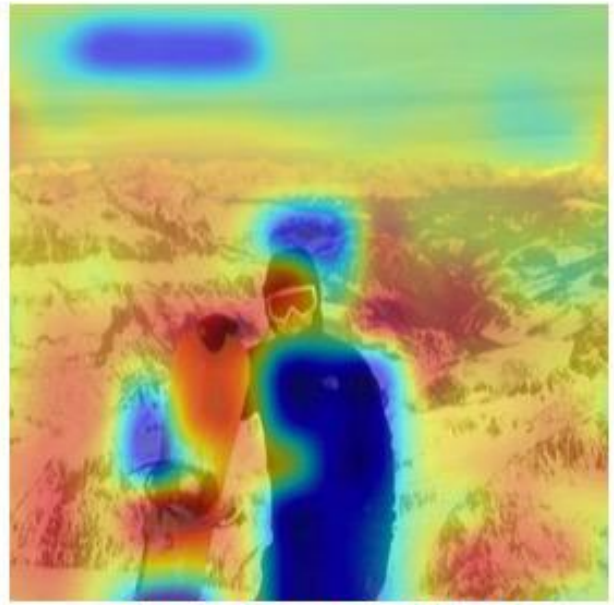
Q: what is the man holding a snowboard on top of a snow covered? **A**: **mountain**



what is the man holding a snowboard on top of a snow covered

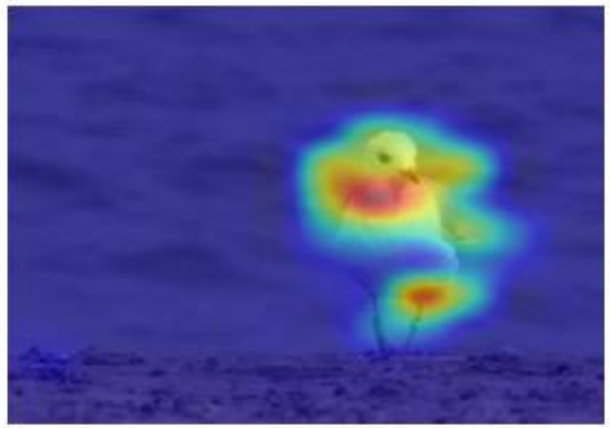


what is the man holding a snowboard on top of a snow covered ?

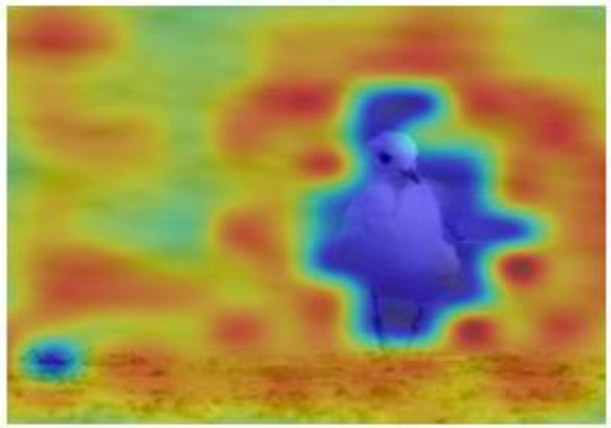


what is the man holding a snowboard on top of a snow covered ?

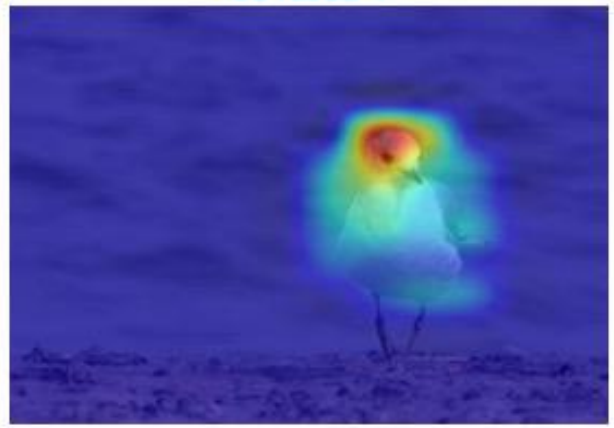
Q: what is the color of the bird? **A**: **white**



what is the color of the bird ?



what is the color of the bird ?



what is the color of the bird ?

Slides Credits: Jiasen Lu

Visual Dialog

Visual Dialog

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Hers is multi-colored

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Hers is multi-colored

How many other people are in the image?

Slides Credits: Abhishek Das

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Hers is multi-colored

How many other people are in the image?

I think 3. They are occluded

Slides Credits: Abhishek Das

Visual Dialog: Task

Given

- image I
- human dialog history
  $(Q_1, A_1), (Q_2, A_2), \ldots, (Q_{t-1}, A_{t-1})$
- follow-up question Q_t

Predict free-form natural language answer

Q: How many people on wheelchairs?
A: Two
Q: What gender are the people in the wheelchairs?
A: One is female, one is male
Q: Which one is holding the racket?
A: The female

Q: Is the other one holding anything?

A: He is not

Slides Credits: Abhishek Das

Visual Dialog: Evaluation

Given

- image I
- human dialog history
  $(Q_1, A_1), (Q_2, A_2), \ldots, (Q_{t-1}, A_{t-1})$
- follow-up question Q_t
- 100 answer options
  - 50 answers from NN questions
  - 30 popular answers
  - up to 20 random answers

Rank 100 options

Accuracy: mean rank of GT answer, recall@k

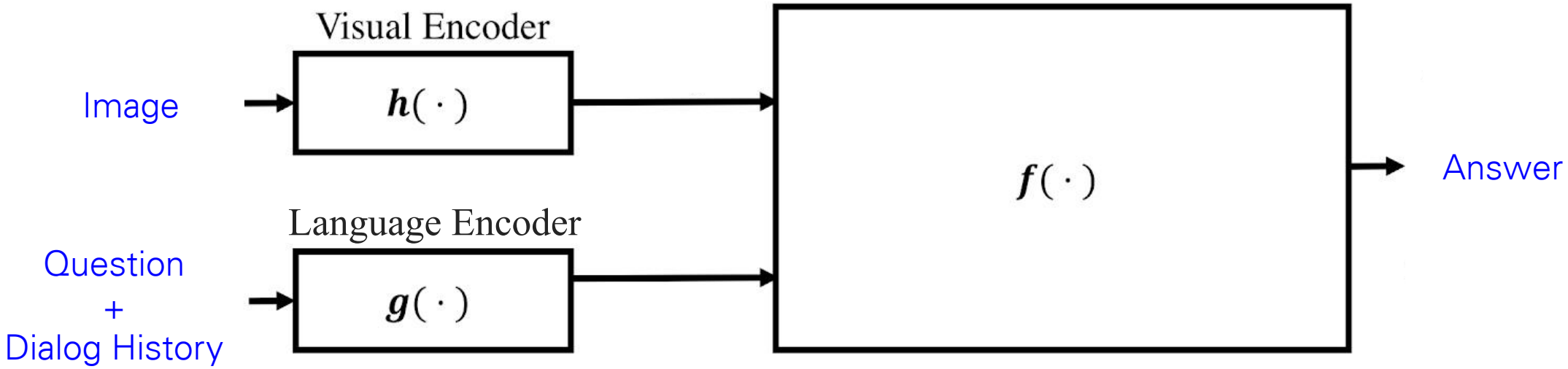
Q: How many people on wheelchairs?
A: Two
Q: What gender are the people in the wheelchairs?
A: One is female, one is male
Q: Which one is holding the racket?
A: The female

Q: Is the other one holding anything?

A: He is not

Slides Credits: Abhishek Das

Basic skeleton of most VL models:
Visual Dialog



Encoder-Decoder models

ENCODERS

Late Fusion Encoder

Hierarchical Recurrent Encoder [Serban et al.]

Memory Network Encoder [Weston et al.]

DECODERS

Generative

Discriminative

Slides Credits: Abhishek Das

Encoder-Decoder models

ENCODERS

Late Fusion Encoder

Hierarchical Recurrent Encoder [Serban et al.]

Memory Network Encoder [Weston et al.]

DECODERS

Generative

Discriminative

Generative Decoding

During training, maximizes likelihood of GT human response

During evaluation, ranks options by LL scores

Slides Credits: Abhishek Das

Encoder-Decoder models

ENCODERS

Late Fusion Encoder

Hierarchical Recurrent Encoder [Serban et al.]

Memory Network Encoder [Weston et al.]

DECODERS

Generative

Discriminative

Discriminative Decoding

Computes dot product between input encoding and LSTM encoding of each of 100 options

Slides Credits: Abhishek Das

Encoder-Decoder models

ENCODERS

Late Fusion Encoder

Hierarchical Recurrent Encoder [Serban et al.]

Memory Network Encoder [Weston et al.]

DECODERS

Generative

Discriminative

Slides Credits: Abhishek Das

Memory Network encoder

Image I

Do you think the woman is with him?

Question Q_t

The man is riding his bicycle on the sidewalk.
Is the man wearing a helmet? No he does not have a helmet on.
How old is the man? He looks around 40 years old.
What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well.
Is anyone else riding a bike? No he's the only one.
Are there any people nearby? Yes there's a woman walking behind him.

t rounds of history

{(Caption), (Q_1,A_1), …, (Q_{t-1}, A_{t-1})}

CNN

4096

512

512

LSTM

LSTM

LSTM

LSTM

LSTM

LSTM

LSTM

Inner product

t x 512

Attention over history

Weighted sum

Fully-connected layer

Decoder

No I don't think they are together

Answer A_t

Slides Credits: Abhishek Das

Vision-Language Pretraining

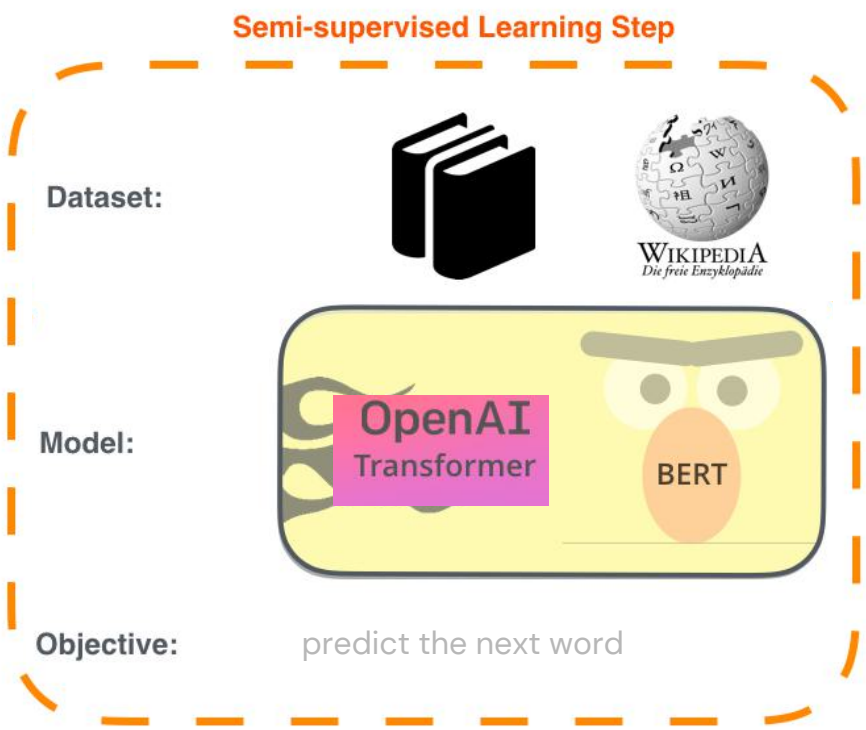
Vision-Language Pretraining

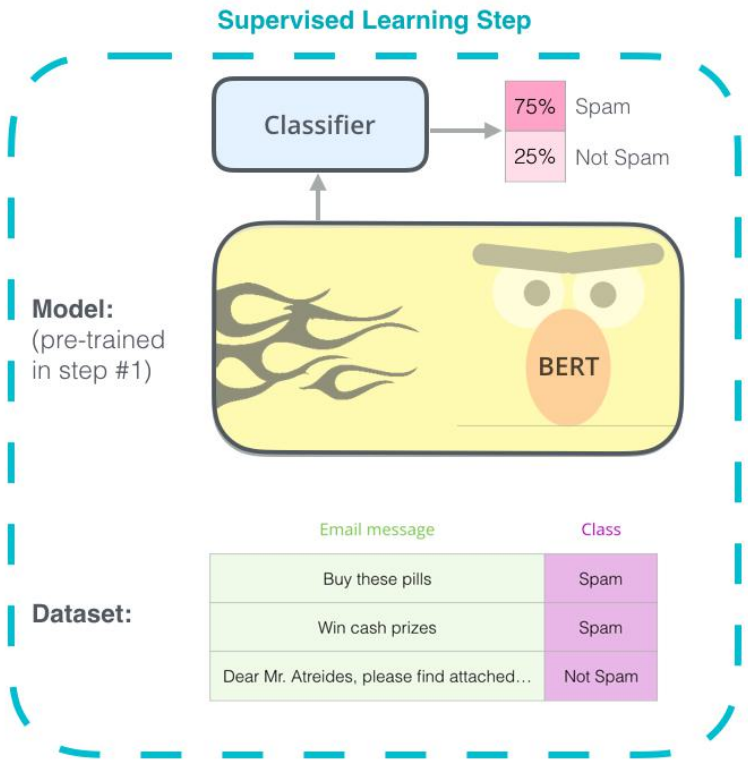
- BERT for Visual Representation Learning
  - ViLBERT, Oscar, VinVL
- Contrastive Language-Image Pre-training
  - CLIP, ALIGN
- Generative Language-Image Pre-training
  - BLIP, CoCa
- Training Scaling Up
  - SigLIP

Vision-Language Pretraining

- BERT for Visual Representation Learning
  - ViLBERT, facebook AI Research August 2019
- Contrastive Language-Image Pre-training
  - CLIP, ALIGN
- Generative Language-Image Pre-training
  - BLIP, CoCa
- Training Scaling Up
  - SigLIP

Success of Pretraining in NLP





- Performance gain is due to architecture innovations & larger data.
  [Peters et al., 2018; Howard & Ruder, 2018; Devlin et al., 2018; Radford et al., 2018; Raffel et al., 2019]

Slide credit: Aida & Lisa

[Images are from http://jalammar.github.io/illustrated-bert/]

Similar Models for Multimodal Pretraining?

Dataset: "The scenic route through mountain ranges includes these unbelievably coloured mountains."

Model:

Objective: predict the next word

Other objectives?

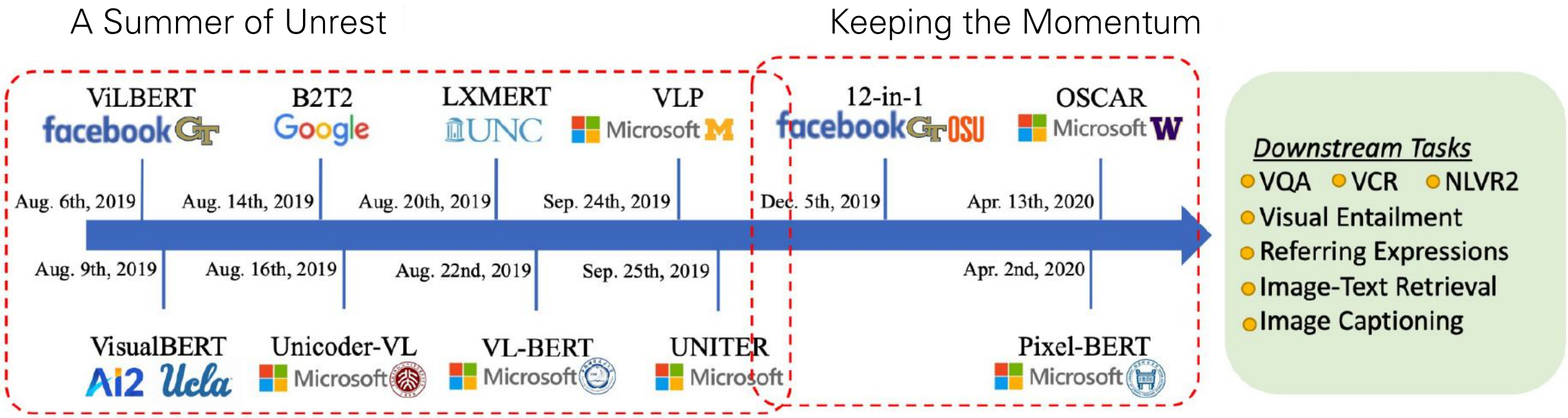
Dataset: image-text pairs where a given text describes its image.

Model: attention mechanisms over both image and text; preprocessing images to "visual tokens".

Objective: loss functions specific to the image modality and image-text pairs.

Slide credit: Aida & Lisa [Images are taken from http://jalammar.github.io/illustrated-bert/ , https://ai.googleblog.com/2018/09/conceptual-captions-new-dataset-and.html, https://visualqa.org/]

Transformer + Pre-training based Methods



- Many more models have been proposed since then…

Borrowed from: https://datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf

Transformer + Pre-training based Methods

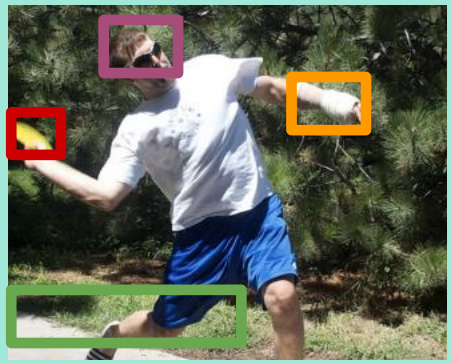
MSCOCO

"The two people are walking down the beach."

MSCOCO/OI Narratives

"In this image we can see a bridge and sea. In the background, we can see trees and the sky. We can see so many people on the bridge. At the bottom of the image, we can see two people. We can see stairs in the right bottom of the image ..."

Visual Genome



small round yellow frisbee, man has cast on his arm, concrete trail path in the park, man wearing black sunglasses

Manually Annotated

Conceptual Captions

"The ***scenic route*** through mountain ranges includes these unbelievably coloured mountains.

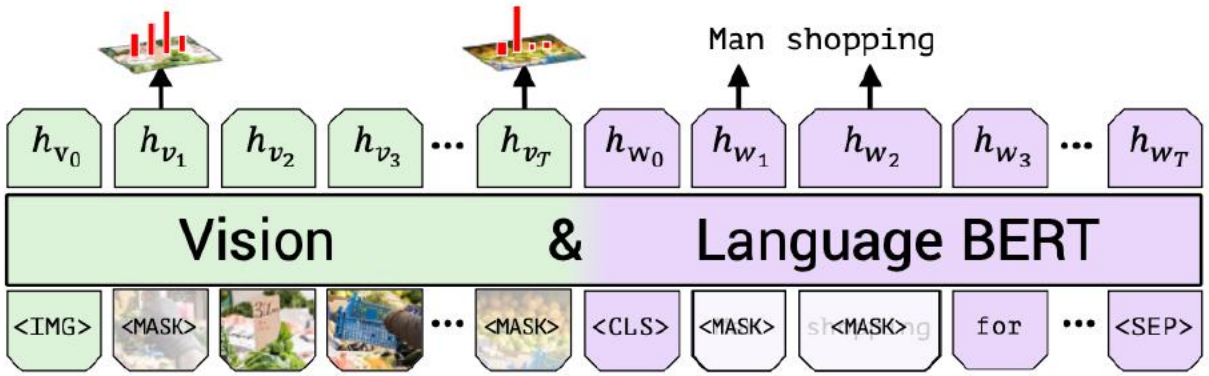
SBU Captions

"***King Arthur's*** beheading rock - right on the sidewalk in the middle of ***town***".

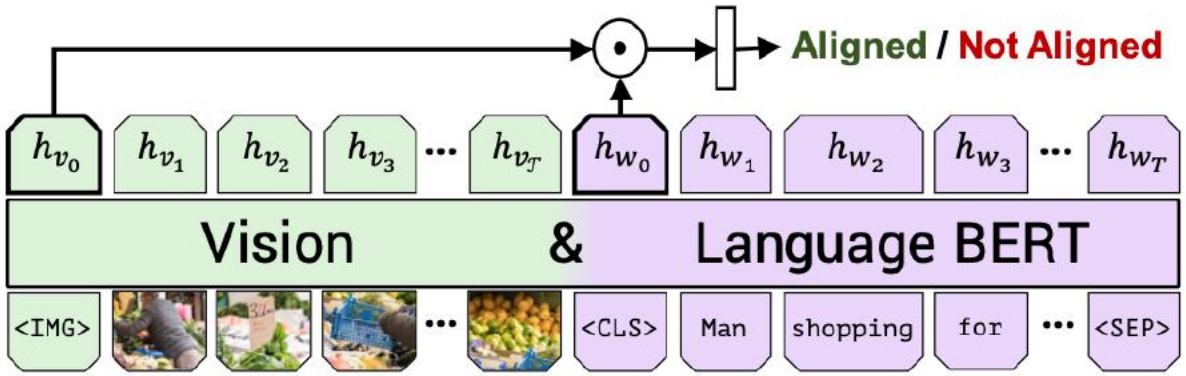
From "the Wild"

Slide credit: Aida & Lisa

Vision-Language BERT (ViLBERT)



(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

https://arxiv.org/pdf/1908.02265.pdf

Vision-Language BERT (ViLBERT)

pop artist performs at the festival in a city.

a worker helps to clear the debris.

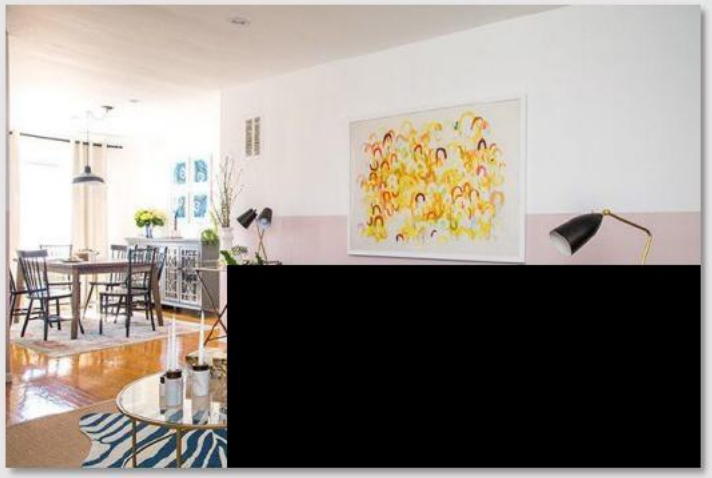
blue sofa in the living room.

https://arxiv.org/pdf/1908.02265.pdf

Vision-Language BERT (ViLBERT)

pop artist performs at the festival in a city.

a worker helps to clear the debris.



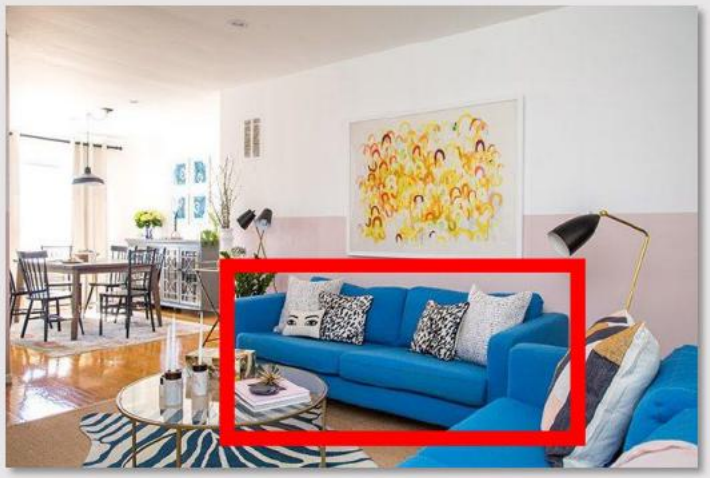
blue sofa in the living room.

https://arxiv.org/pdf/1908.02265.pdf

Vision-Language BERT (ViLBERT)

pop artist performs at the festival in a city.

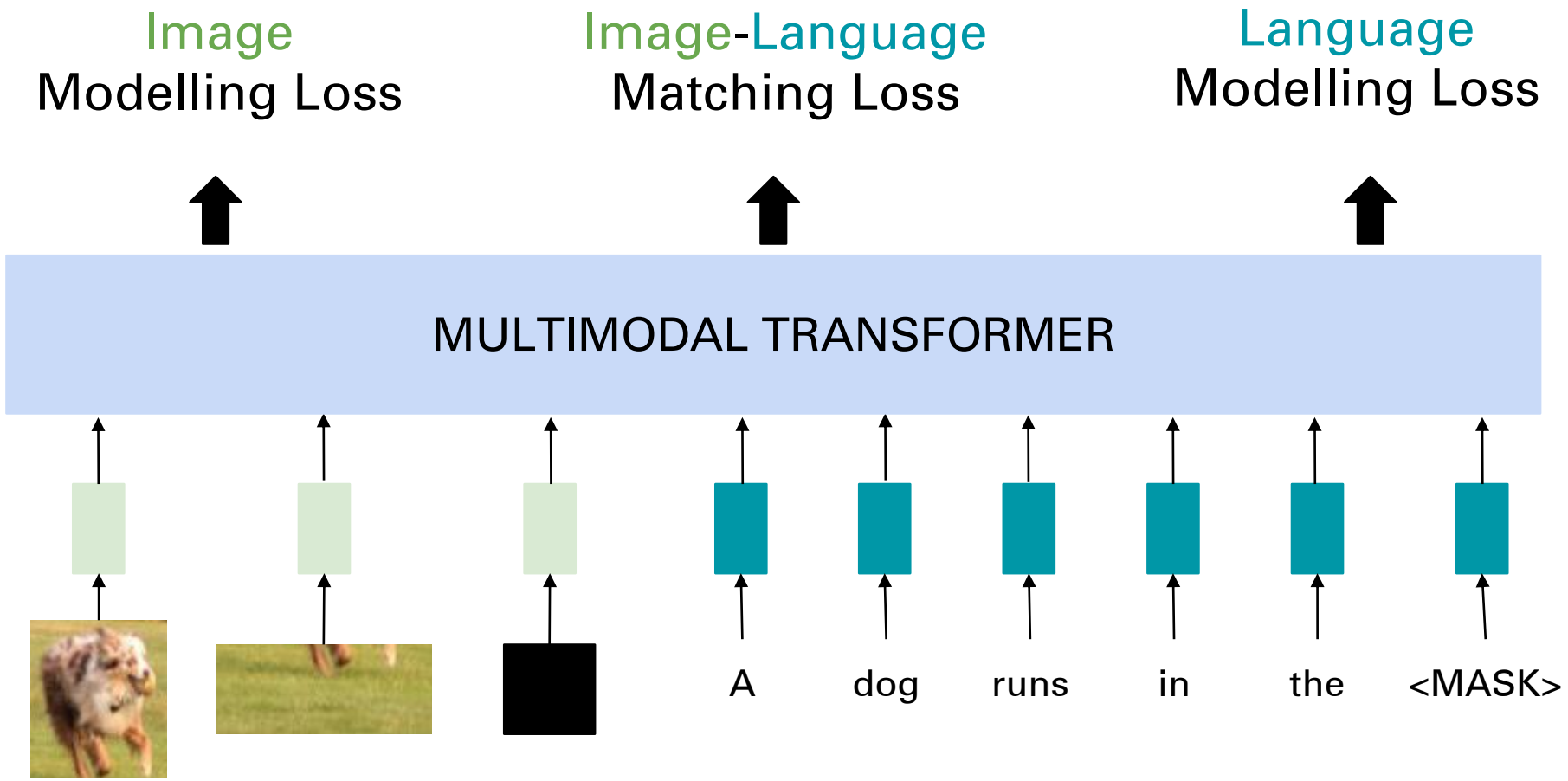
a worker helps to clear the debris.



[redacted] in the living room.

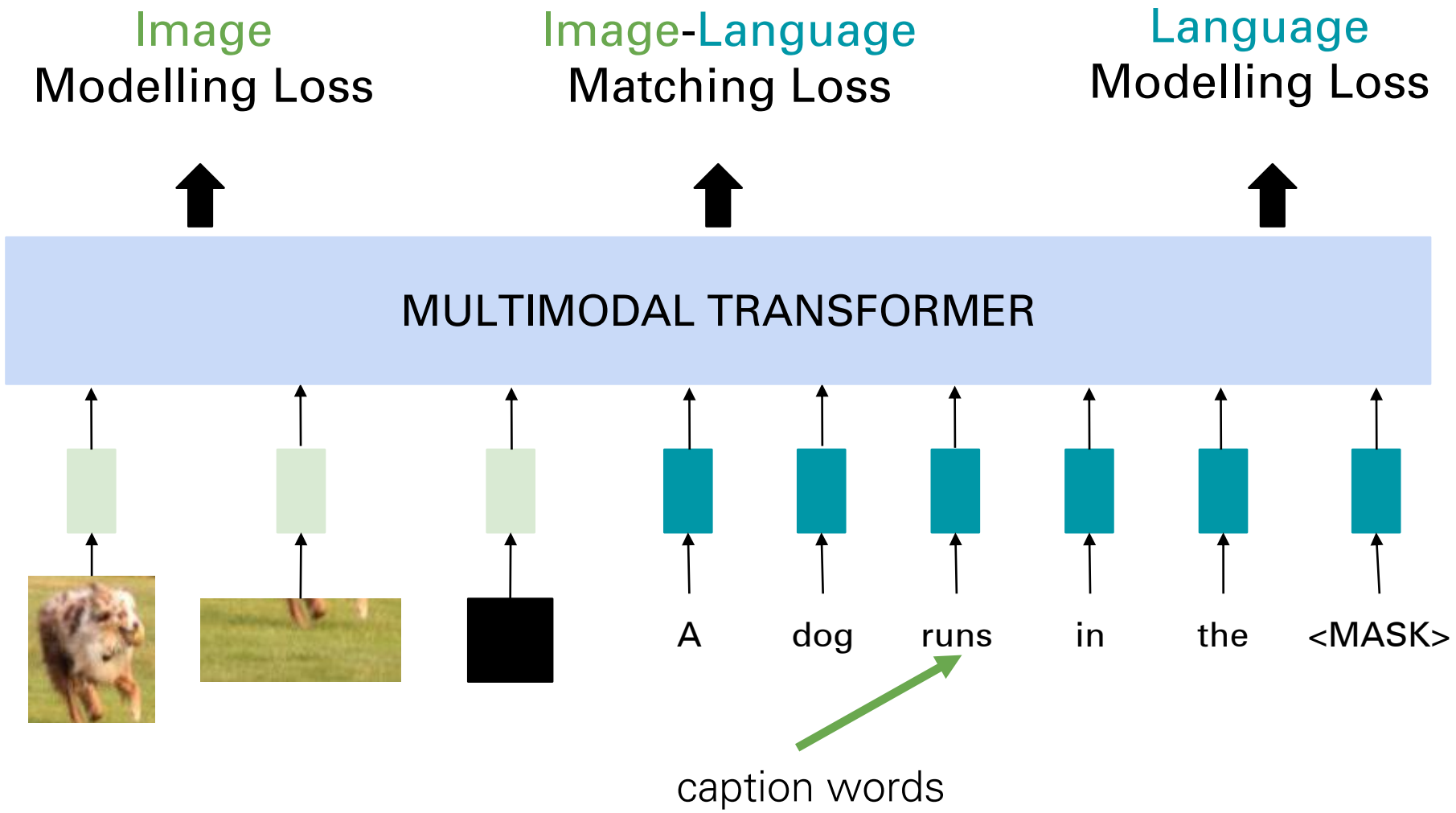
https://arxiv.org/pdf/1908.02265.pdf

ViLBERT Architecture



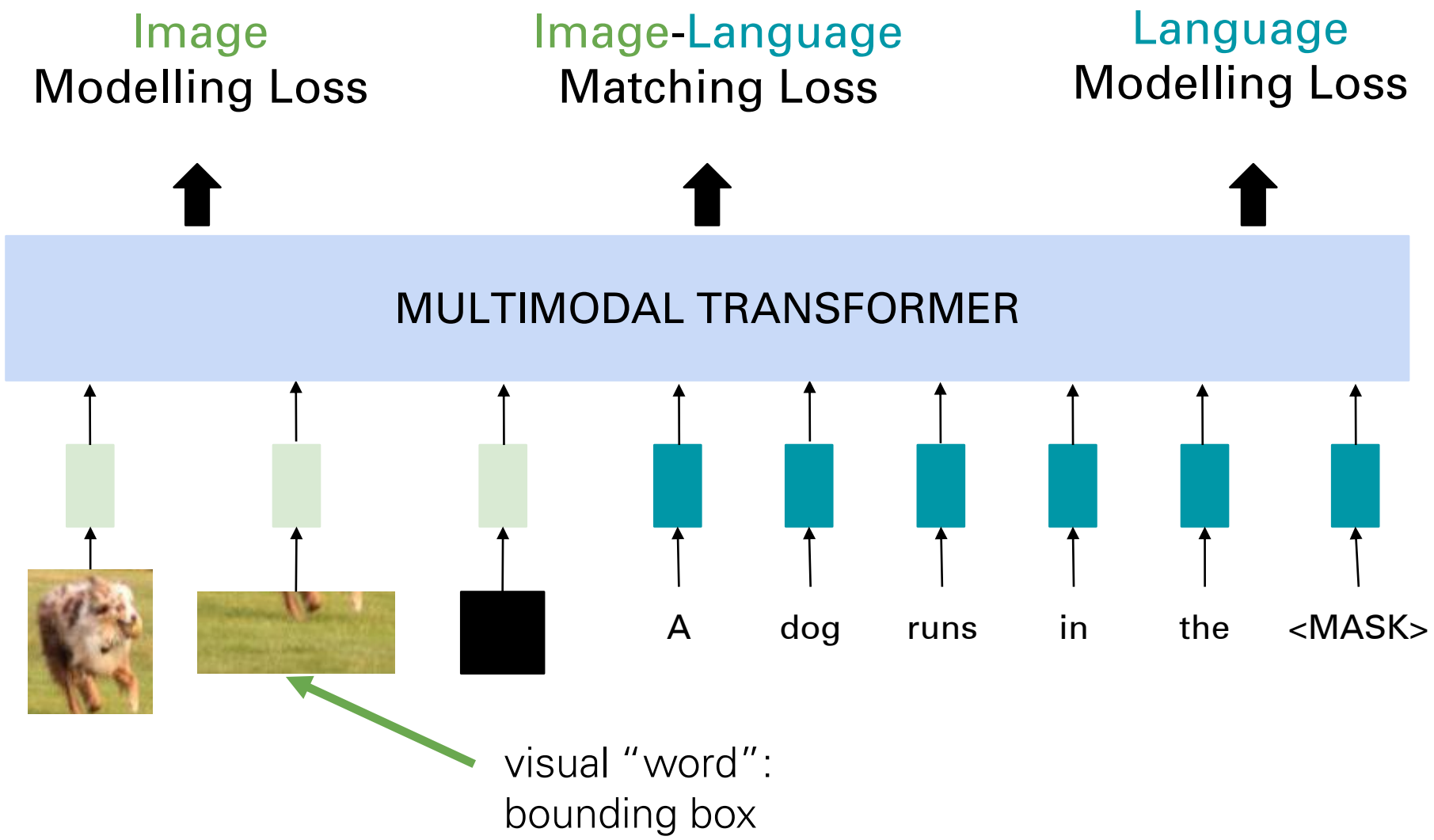
Slide credit: Aida & Lisa

ViLBERT Architecture



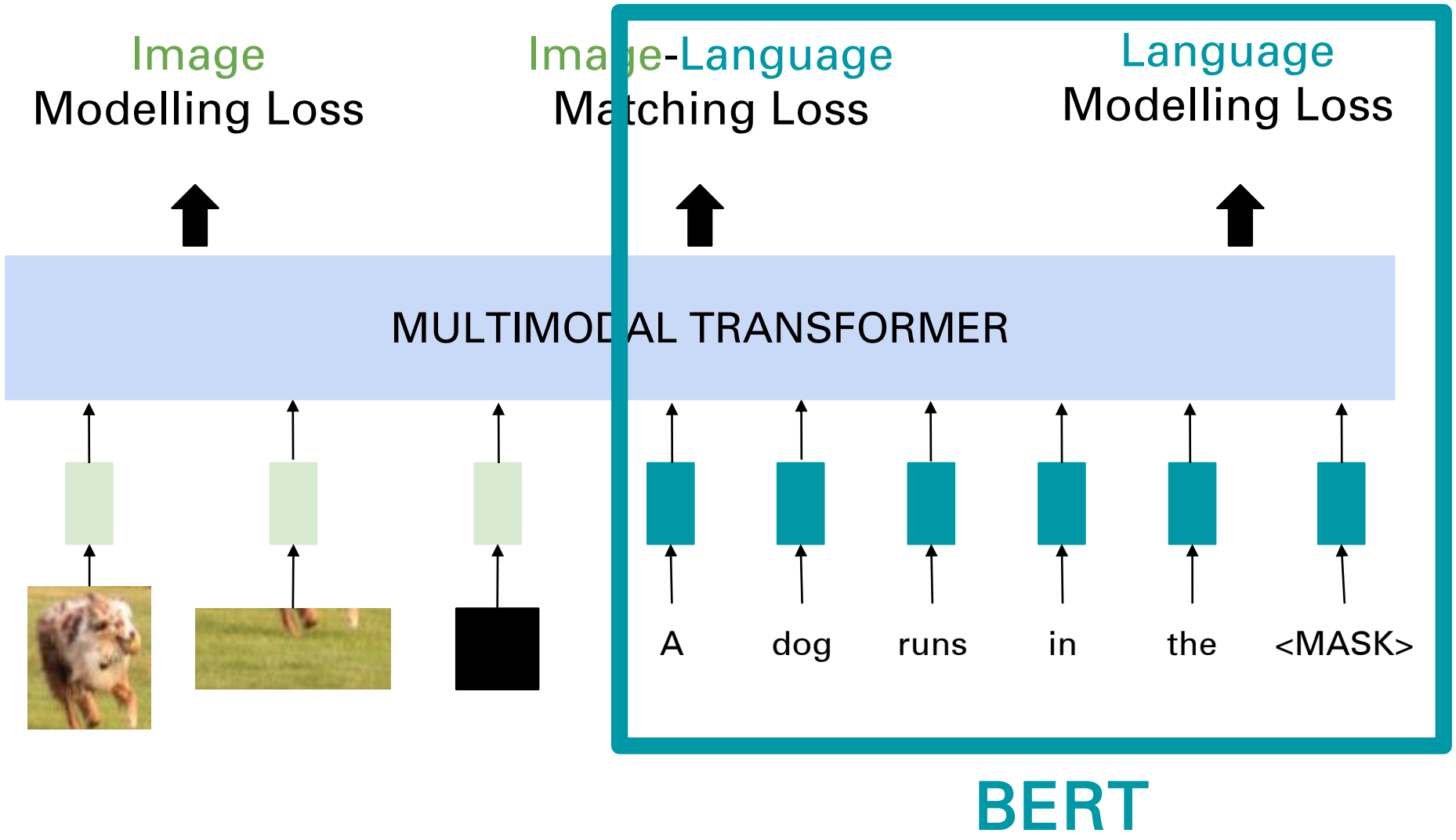
Slide credit: Aida & Lisa

ViLBERT Architecture



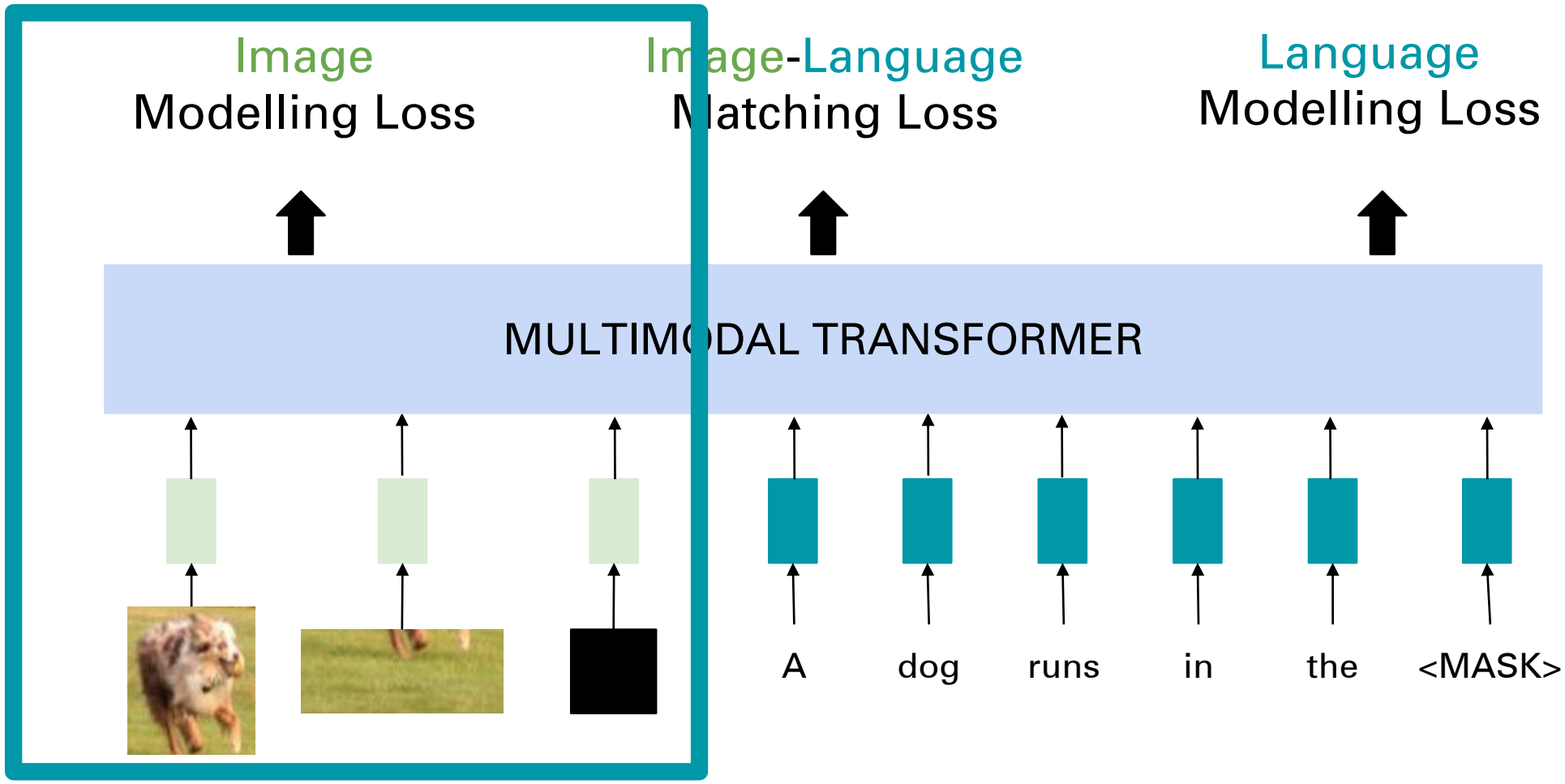
Slide credit: Aida & Lisa

ViLBERT Architecture



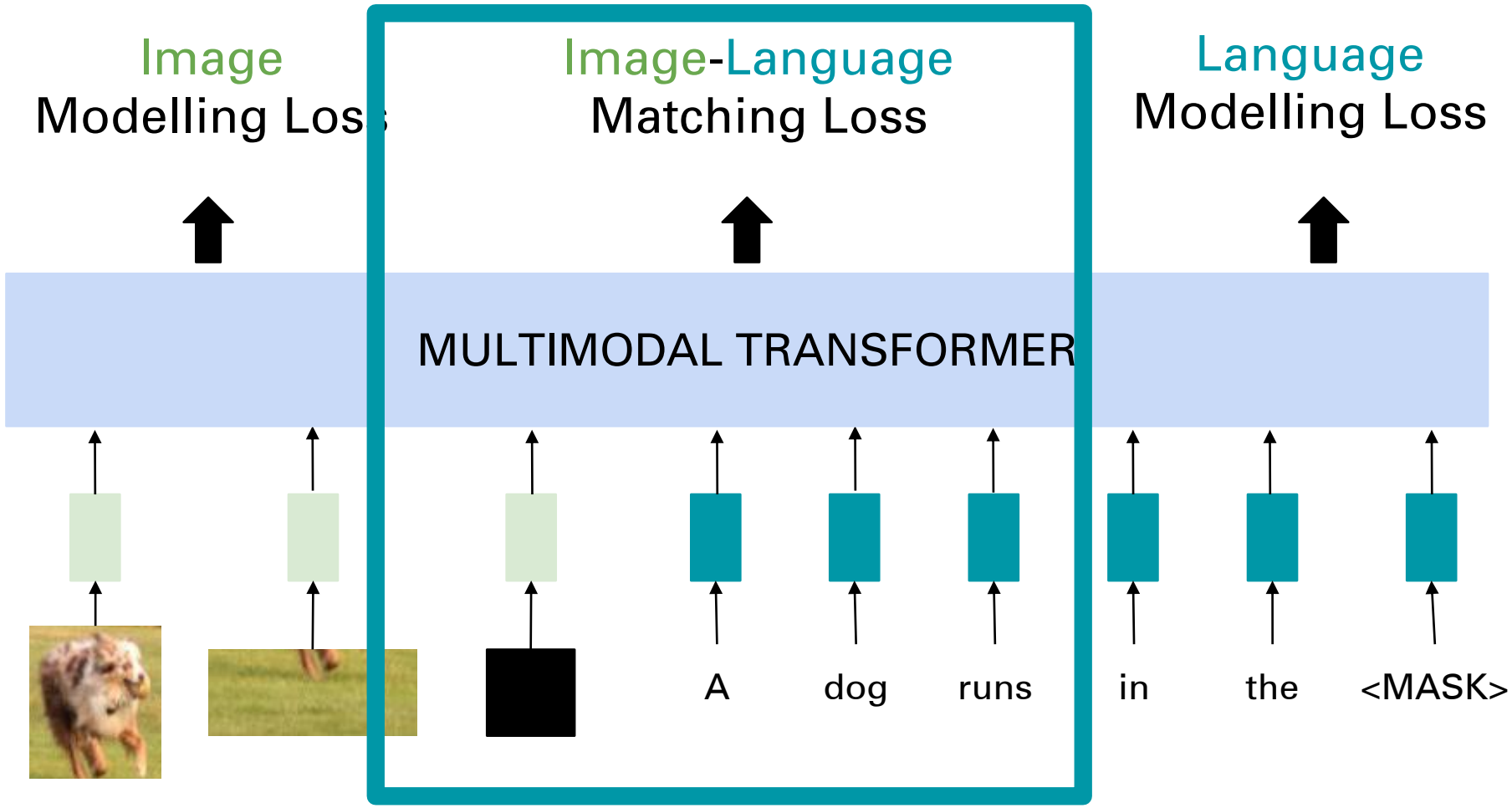
Slide credit: Aida & Lisa

ViLBERT Architecture



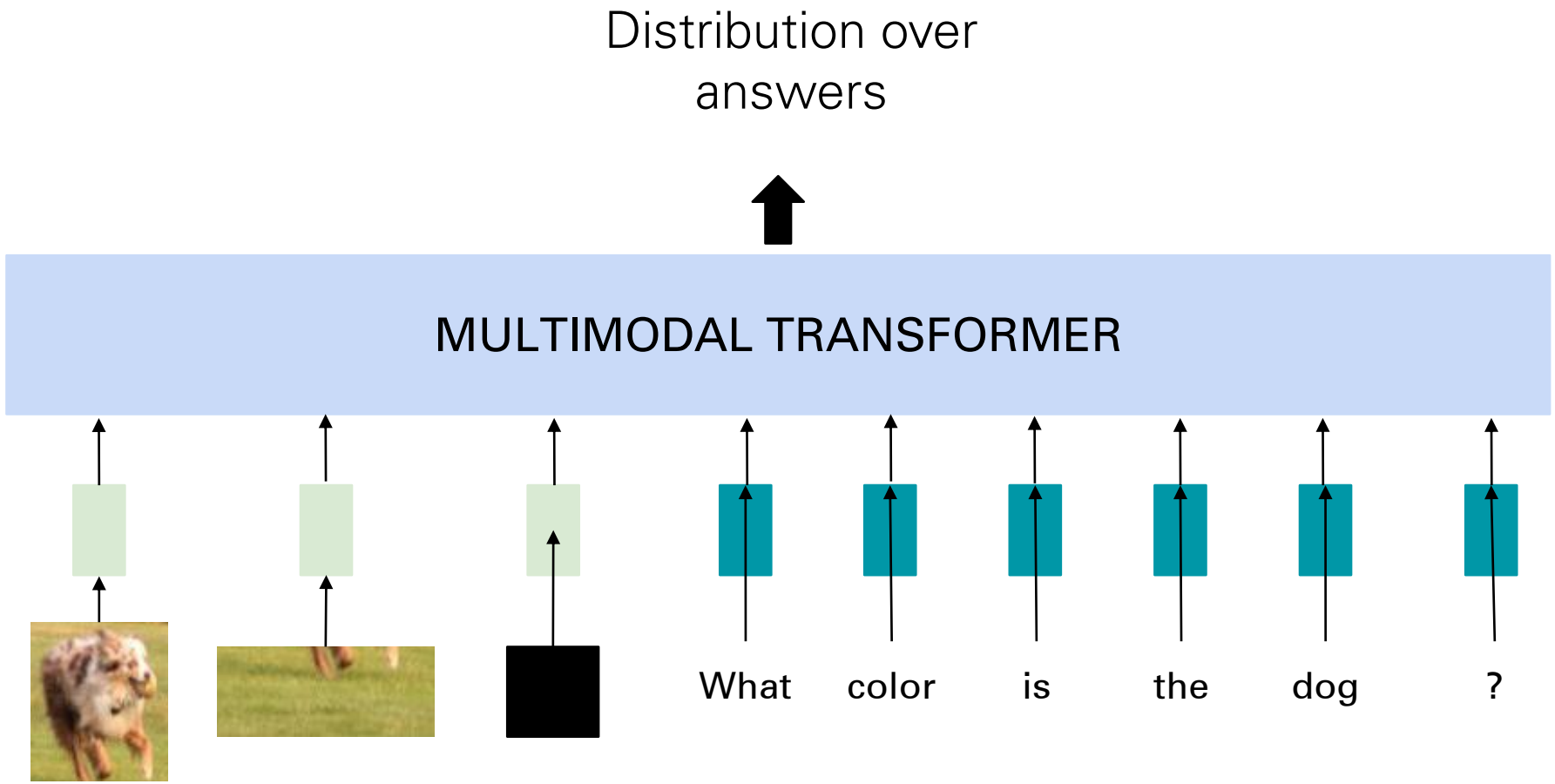
Slide credit: Aida & Lisa

ViLBERT Architecture



Slide credit: Aida & Lisa

ViLBERT



Slide credit: Aida & Lisa

Vision-Language Pretraining

- BERT for Visual Representation Learning
  - VilBERT, **Oscar**,
- Contrastive Langu
  - CLIP, ALIGN
- Generative Language-Image Pre-training
  - BLIP, CoCa
- Training Scaling Up
  - SigLIP

"Semantic alignments between texts and images using object tags"

March 2020

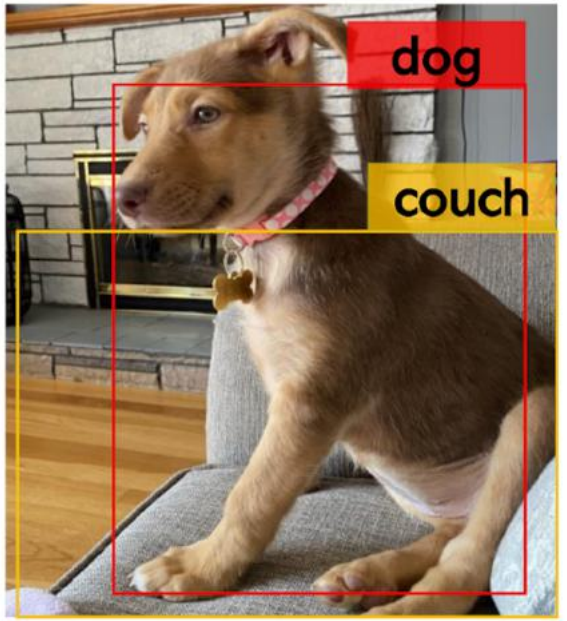
OSCAR - Background & Motivation

Challenges:

- **Ambiguity**: Image region **overlapping** at different positions results in ambiguities for the extracted visual embeddings
- **Lack of explicit alignments**: There is no explicitly labeled alignments between regions or objects in image-text pairs

Motivation:

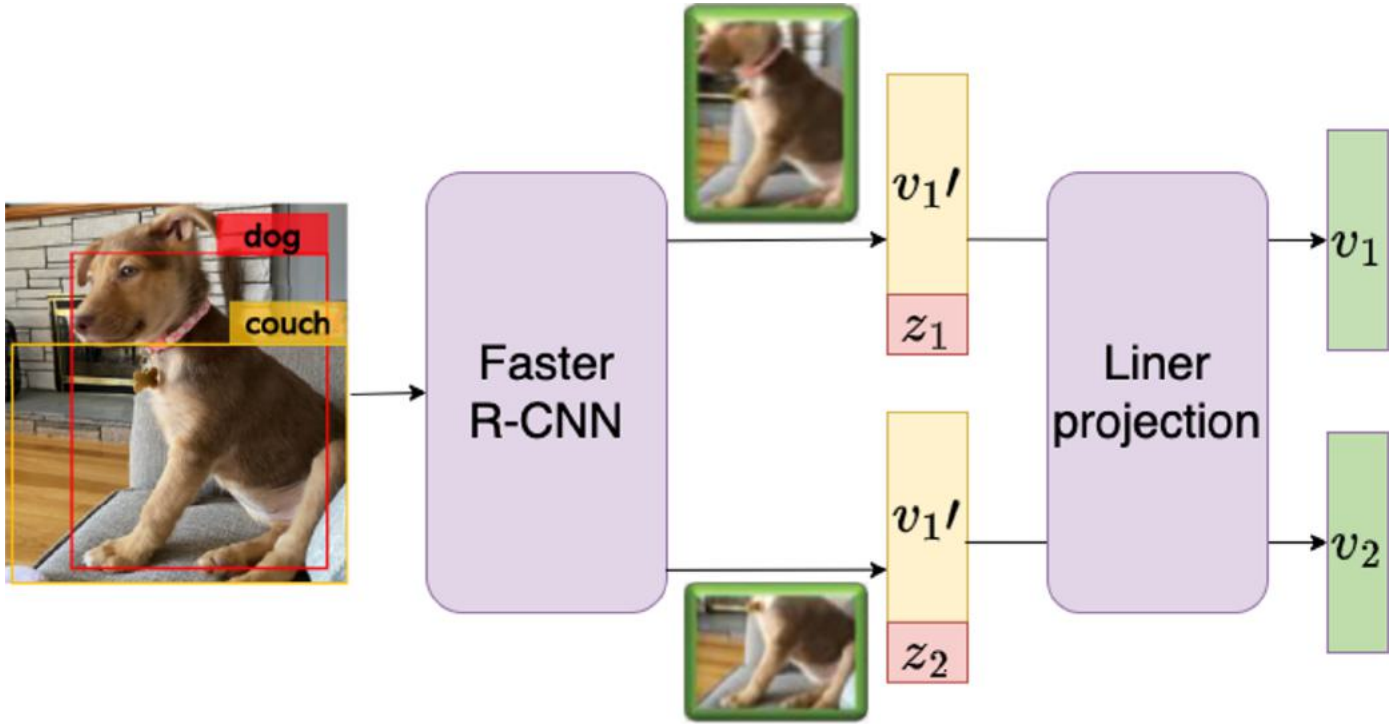
- **Salient objects** can be accurately detected by **object detectors** and are often mentioned in the paired text
- They can be used as **anchor points** for learning semantic alignments between image region features and word embeddings

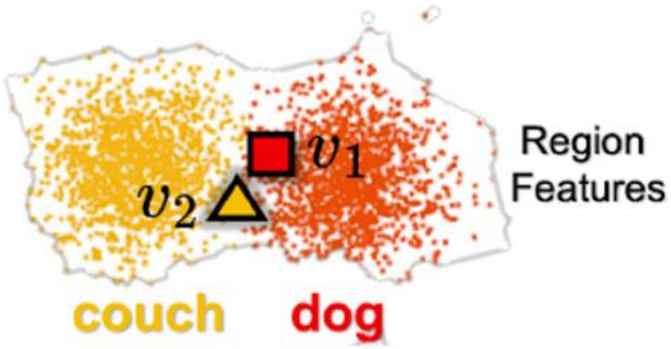


OSCAR - Extracting Anchor Points

To extract visual embeddings v

1. **Faster R-CNN** is used to extract the visual semantics of each region as (v', z)
   a. v': region feature, a vector of dimension P (e.g., 2048)
   b. z: region position, a vector of dimension R (e.g., 4)
2. **Concatenate** v' and z to form a position-sensitive region feature vector
3. Using a trainable **linear projection** to transform [v', z] into v, a vector of dimension H (e.g., 768)

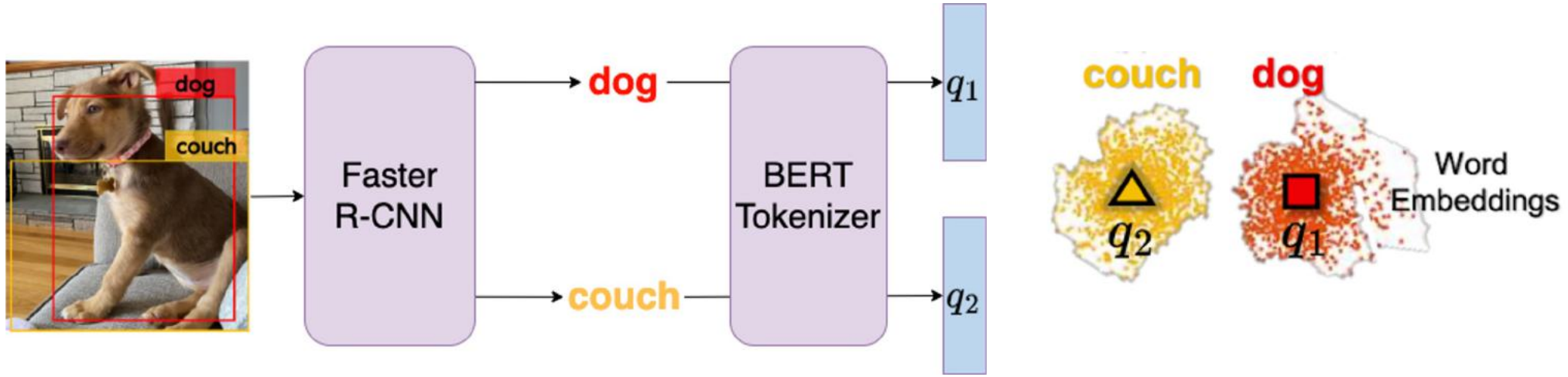




OSCAR - Extracting Word Embeddings

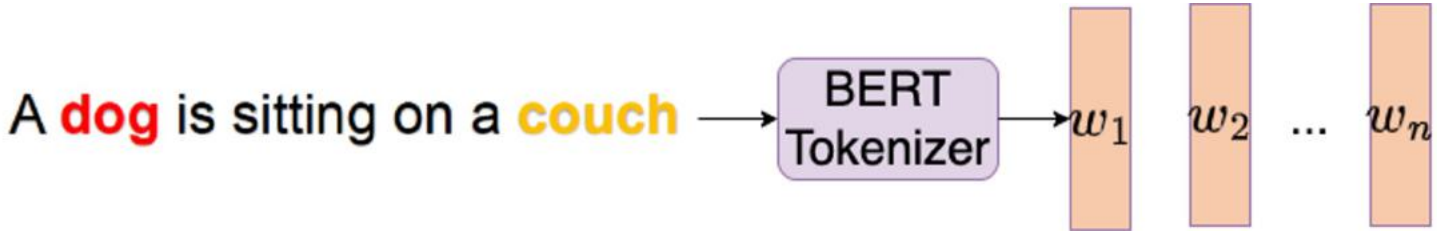
To extract tag embeddings q

1. **Faster R-CNN** is used to extract the tags
2. Embed tags into word tokens q (H-dimensional) using pre-trained **BERT**

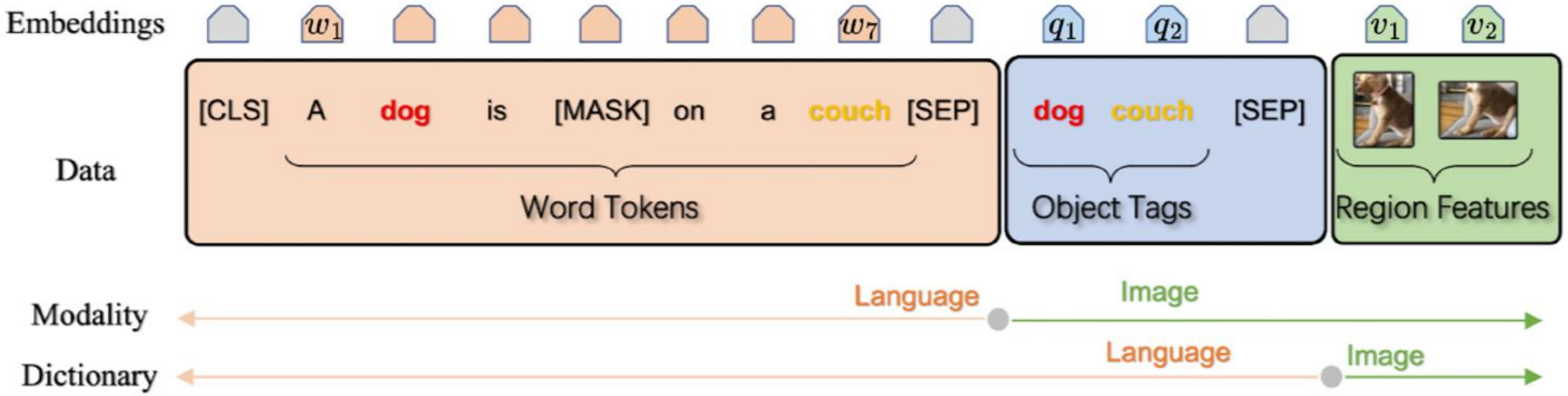


To extract text embeddings w

1. Embed tags into word tokens w (H-dimensional) using pre-trained **BERT**



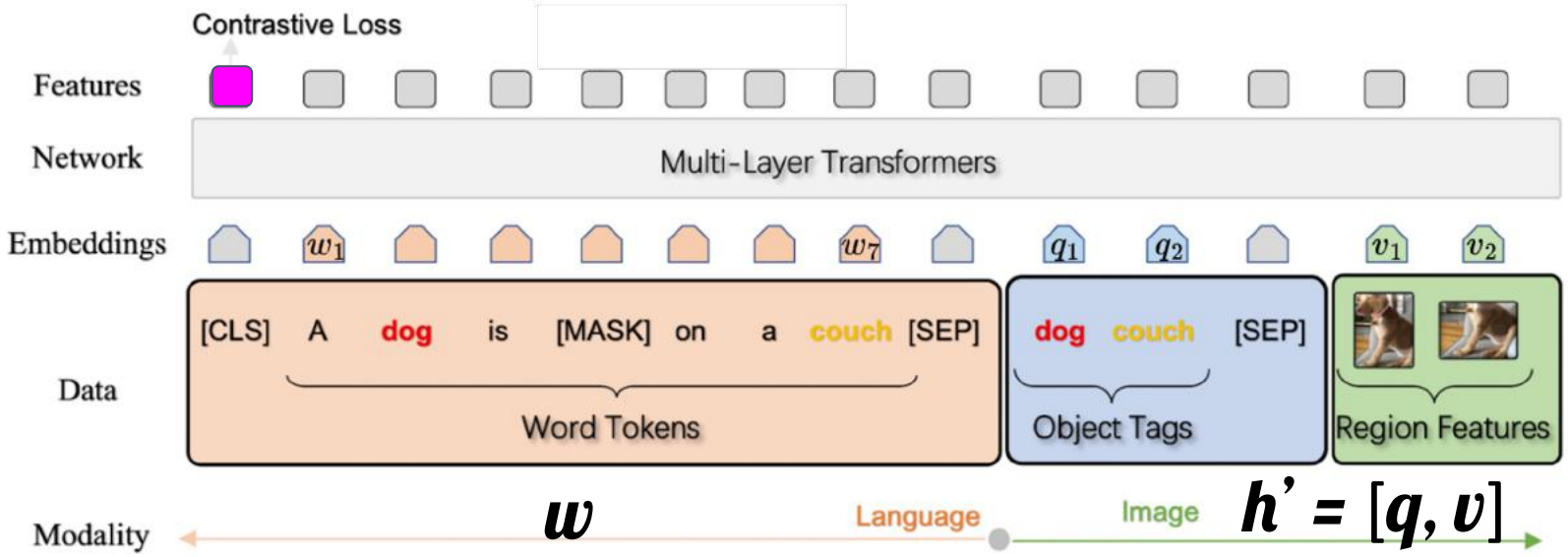
Looking at the same input from 2 perspectives



Now that we have embeddings for texts (w), tags (q) and image regions (v), all in dim H

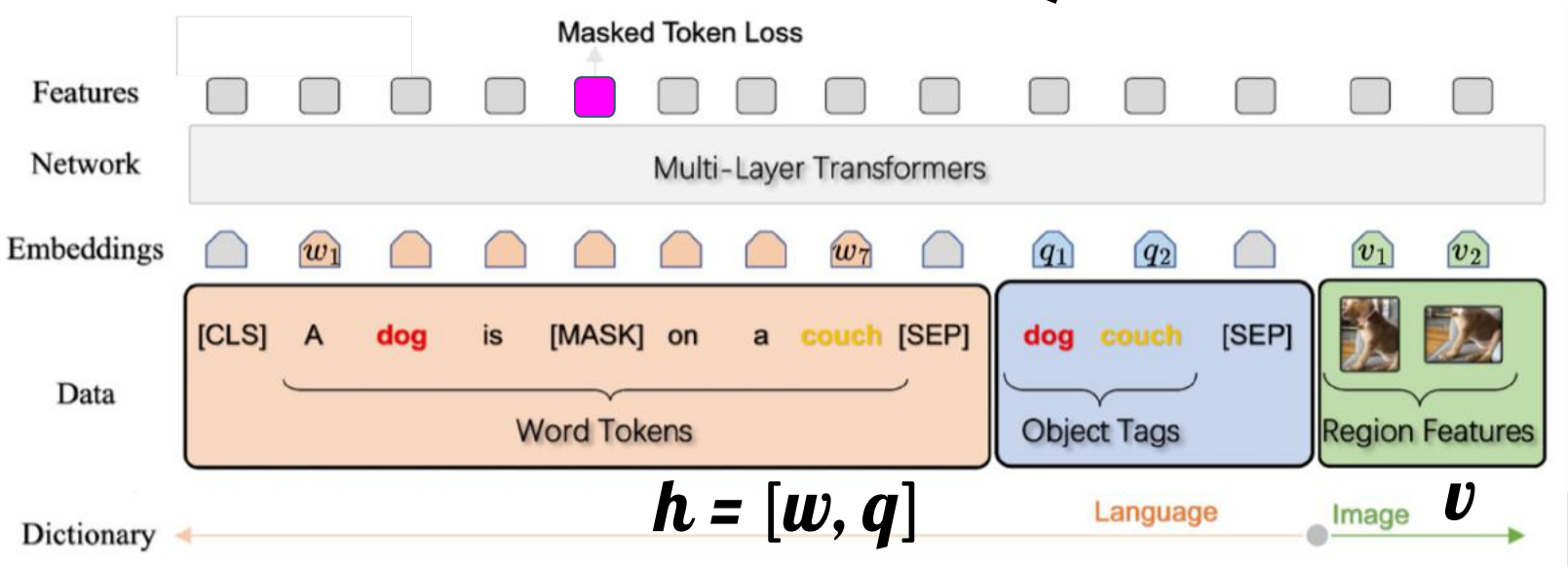
- **Modality view**:
  - Text modality: word tokens (w)
  - Image modality: image region features (v) & associated object tags (q)
  - Goal: to distinguish the **representations** between a text and an image
- **Dictionary view:**
  - Linguistic semantic space: word tokens (w) & object tags (q)
  - Visual semantic space: image region features (v)
  - Goal: to distinguish the **semantic spaces** between text and image

OSCAR - Loss for Modality View



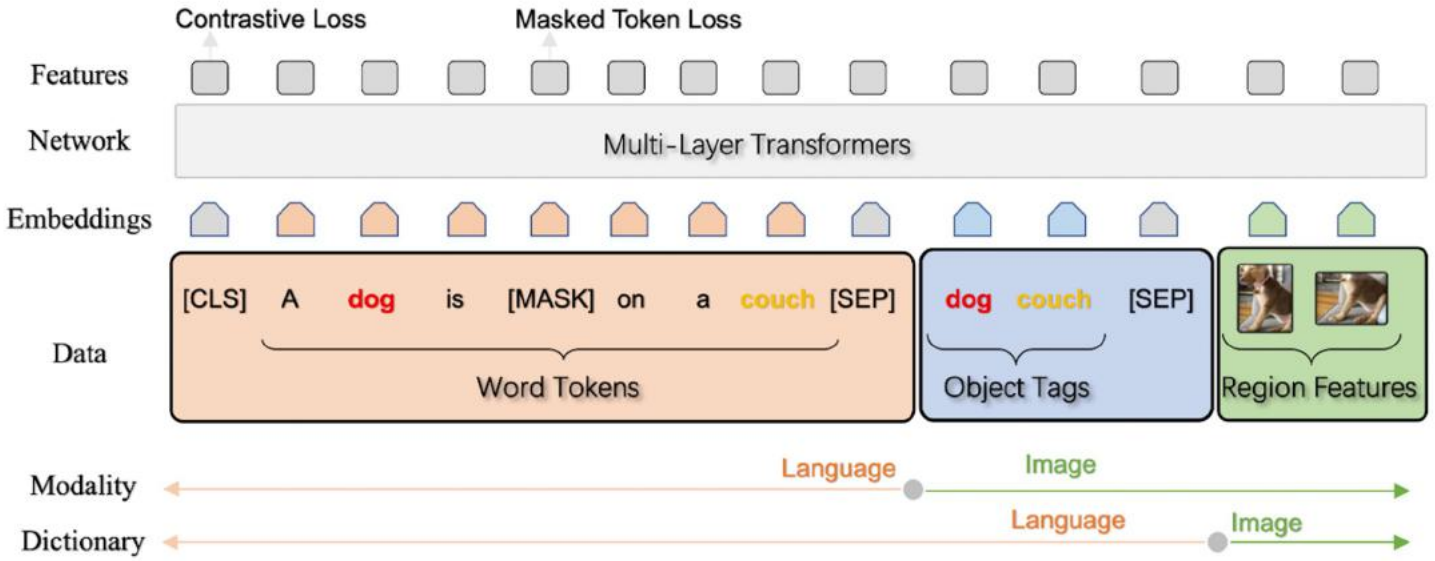
- Text modality representation $\boldsymbol{w}$
- Image modality representation $\boldsymbol{h}' \triangleq [\boldsymbol{q}, \boldsymbol{v}]$
- Pollute $\boldsymbol{h}'$ s.t. it contains a set of images where the 50% tags are replaced with different tags
- Train a binary classifier f to predict whether image-text modality pair $(\boldsymbol{h}', \boldsymbol{w})$ contains the original image or polluted ones
- **Contrastive Loss** $\mathcal{L}_{\mathrm{C}} = -\mathbb{E}_{(\boldsymbol{h}', \boldsymbol{w}) \sim \mathcal{D}} \log p(y | f(\boldsymbol{h}', \boldsymbol{w}))$
- Goal: to adjust word embedding space where a text is similar to its paired image and dissimilar to the polluted images

OSCAR - Loss for Dictionary View



- Linguistic semantic space representation $\boldsymbol{h} \triangleq [\boldsymbol{w}, \boldsymbol{q}]$
- Visual semantic space representation $\boldsymbol{v}$
- 15% tokens in $\boldsymbol{h}$ is replaced with [MASK] token
- Similar to masked language models, we want to predict masked text tokens ($\boldsymbol{h_i}$) based on surrounding text tokens ($h_{\setminus i}$) and all image features ($\boldsymbol{v}$)
- **Masked Token Loss** $\mathcal{L}_{\mathrm{MTL}} = -\mathbb{E}_{(\boldsymbol{v},\boldsymbol{h})\sim\mathcal{D}} \log p(h_i|\boldsymbol{h}_{\setminus i}, \boldsymbol{v})$
- Goal: to ground the learned word embeddings in the vision context

OSCAR - Pre-training



- The total pre-training loss $\mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\text{MTL}} + \mathcal{L}_{\text{C}}$
- Trainable parameters
  - Linear projection matrix
  - BERT
- Datasets:
  - 6.5M image-text pairs consisting of 4M unique images
  - COCO, Conceptual Captions (CC), SBU captions, Flickr30k, GQA

OSCAR - Quantitative Results

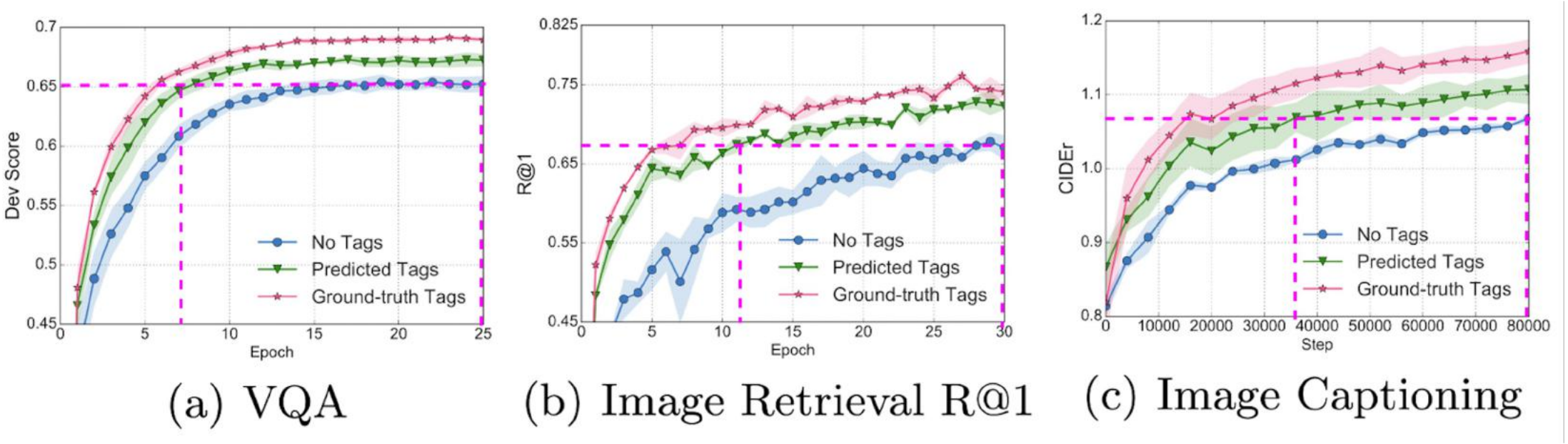
Task	Image Retrieval			Text Retrieval			Image Captioning				NoCaps		VQA	NLVR2
	R@1	R@5	R@10	R@1	R@5	R@10	B@4	M	C	S	C	S	test-std	test-P
SoTA$_S$	39.2	68.0	81.3	56.6	84.5	92.0	38.9	29.2	129.8	22.4	61.5	9.2	70.90	53.50
SoTA$_B$	48.4	76.7	85.9	63.3	87.0	93.1	39.5	29.3	129.3	23.2	73.1	11.2	72.54	78.87
SoTA$_L$	51.7	78.4	86.9	66.6	89.4	94.3	–	–	–	–	–	–	73.40	79.50
OSCAR$_B$	**54.0**	**80.8**	**88.5**	**70.0**	**91.1**	**95.5**	**40.5**	**29.7**	**137.6**	22.8	**78.8**	**11.7**	**73.44**	78.36
OSCAR$_L$	**57.5**	**82.8**	**89.8**	**73.5**	**92.2**	**96.0**	**41.7**	**30.6**	**140.0**	**24.5**	**80.9**	**11.3**	**73.82**	**80.37**
Δ	**5.8** ↑	**4.4** ↑	**2.9** ↑	**6.9** ↑	**2.8** ↑	**1.7** ↑	**2.2** ↑	**1.3** ↑	**10.7** ↑	**1.3** ↑	**7.8** ↑	**0.5** ↑	**0.42** ↑	**0.87** ↑

Note that the dataset size of

- OSCAR: 6.5M image-text pairs
- Counterpart: over 9M image-text pairs

With fewer image-text pairs than SoTA$_L$, OSCAR$_B$ achieves higher score than its counterpart in 5 out of 6 tasks, highlighting OSCAR's parameter efficiency, partially because **the use of object tags as anchor points eases the learning of semantic alignments** between images and texts

OSCAR - The Effect of Object Tags



- Training using predicted tags takes **less than half of the training time** to achieve the final performance of the baseline (no tags), showing the efficiency of utilizing object tags for VLP
- Training using ground truth tags further reduces the training time by over 50% to achieve the final performance of the predicted tags

OSCAR - Qualitative Results

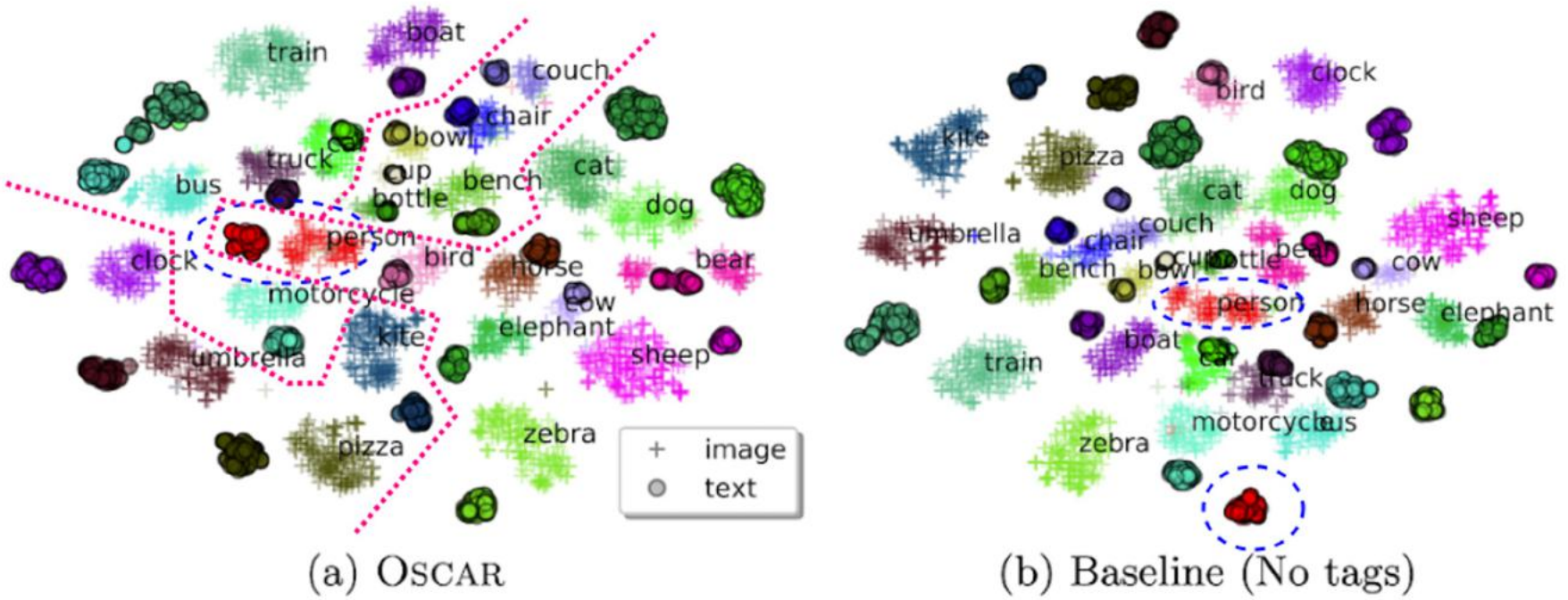


Fig. 4: 2D visualization using t-SNE. The points from the same object class share the same color. Please refer Appendix for full visualization.

Intra-class: same object between two modalities is **closer** (e.g., person)
Inter-class: classes of related semantics are closer but still **distinguishable**, such as animal (zebra, elephant, sheep), transportation (train, car, truck), furniture (couch, chair, bowl).

OSCAR - Limitations

- Requires a powerful object detector to handle complex scenes
- Does not work well when salient objects are missing in the text

A few good reasons to start with country line dance

Vision-Language Pretraining

- **BERT for Visual Representation Learning**
  - ViLBERT, Oscar, **VinVL**
- Contrastive Language-Im
  - CLIP, ALIGN
- Generative Language-Image Pre-training
  - BLIP, CoCa
- Training Scaling Up
  - SigLIP

"Extract better visual representation rather than just fuse multi-modal information"

Microsoft

March 2021

VinVL: Background & Motivation

- Success of visual language pre-training (VLP) in visual language (VL) tasks
  - *VilBERT* and *OSCAR*
  - object detection (OD) model + cross-modal fusion model
- Vision-language fusion model
  - OD model improvement untouched
  - significance of visual features
- OD
  - **large-scale object-attribute detection model - *ResNeXt-152 C4 (X152-C4)***
  - *Visual Genome (VG)* dataset

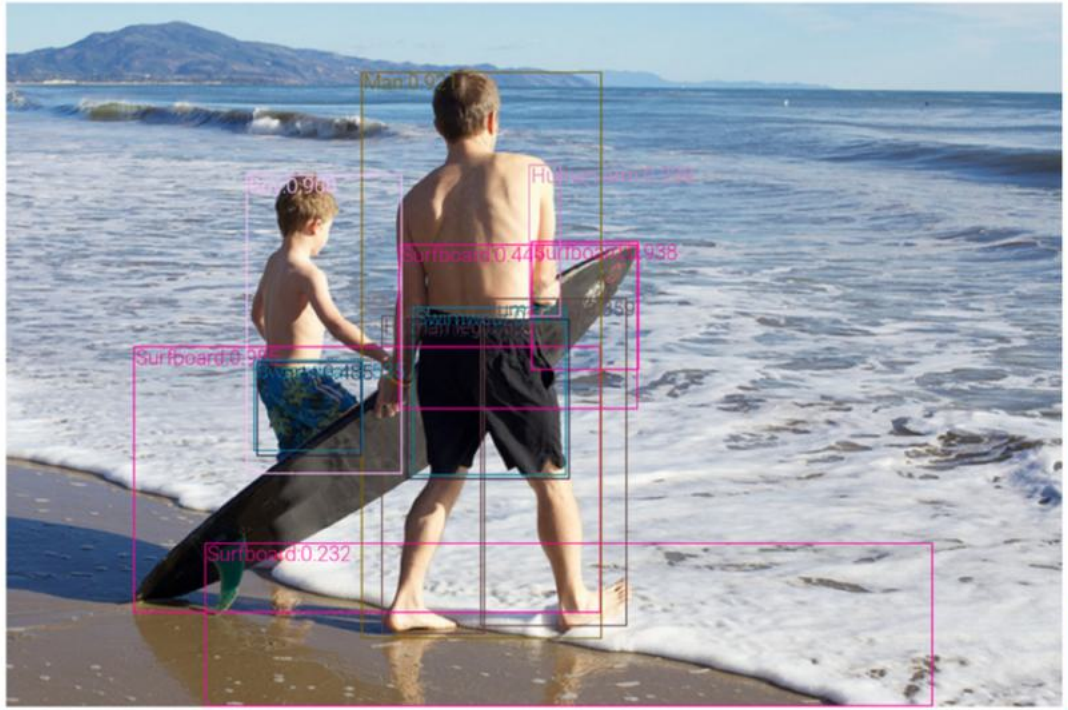
VinVL: Improve Vision in Vision Language

- Mainstream
  - **Vision** as a black box
  - larger training dataset: *OpenImages* and *Objects 365*
  - new insights in OD algorithms: *feature pyramid network*
  - Powerful GPUs for bigger models

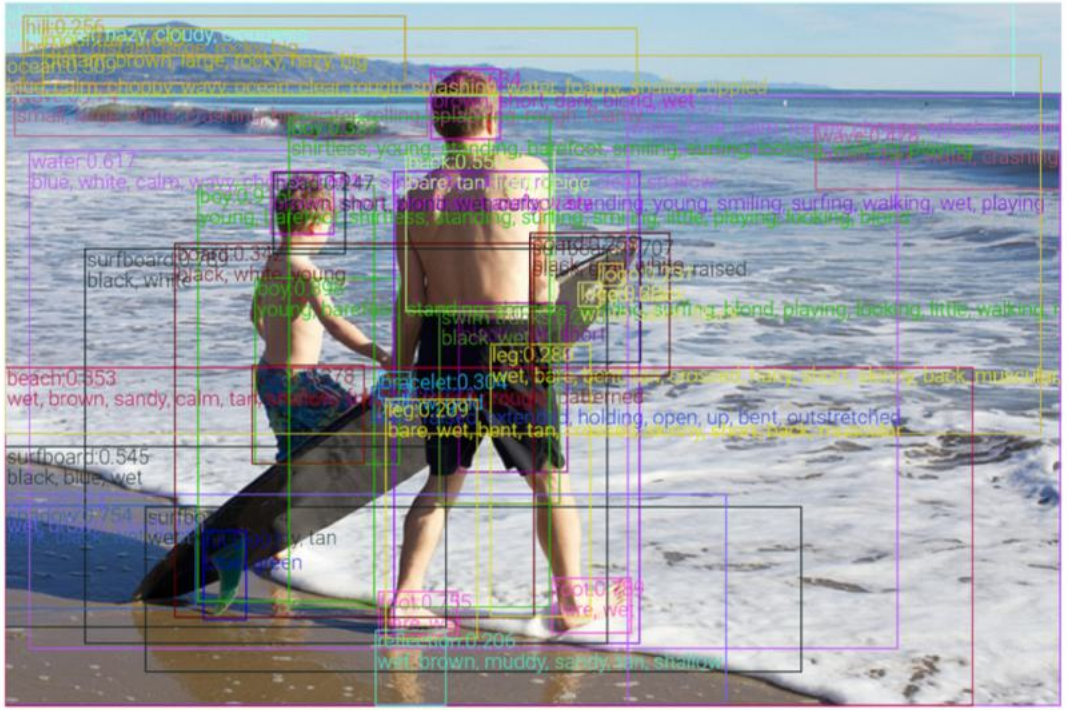
- Idea:
  - improve **Vision** for better visual representations
  - enrich the visual object and attribute categories
  - enlarge the model size
  - train on much larger OD dasetset

VinVL: Improve Vision in Vision Language

- A new object detection model
- more accurate **object-attribute** detection results and better visual features for VL applications



- *X152-FPN*: "boy"



- *X152-C4*: "young barefoot shirtless standing surfing smiling little playing looking blond boy"
- More than 20 additional object concepts

Zhang, Pengchuan, et al. "Vinvl: Revisiting visual representations in vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

VinVL: Revisit VL Models

1. Pre-training
   - data
     - 4 public complementary dataset - *COCO, OpenImagesV5 (OI), Objects365V1, and Visual Genome (VG)*
     - build a unified corpus with VG vocabulary - sampling, balancing and merging
   - model architecture: ***X152-C4***
   - model pre-training
     - freezing: first conv layer, first res block and all batch-norm layers
     - data augmentation: horizontal flipping and multi-scale training
     - initialization from an ImageNet-5K checkpoint

2. Fine-tuning
   - fine-tune the new OD model on ***VG*** to inject attribute information

VinVL: OSCAR+ Pre-training

- Deep learning-based VL models: $(\boldsymbol{q}, \boldsymbol{v}) = \mathbf{Vision}(Img), \quad y = \mathbf{VL}(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{v})$
  - **Vision**: image understanding module
  - **VL**: cross-modal understanding module
  - *Img*: vision
  - ***q***: semantic representation of the image - object tag
  - ***v***: distributional representation of the image - visual representation
  - **w**: language - text (question in VQA)
  - y: output - text (answer to be predicted in VQA)
- Convention
  1. unify vision and language modeling **VL** with *Transformer*
  2. pre-train the unified **VL** with large-scale text-image corpora

VinVL: OSCAR+ Pre-training

Pre-train an *OSCAR+* to learn the joint image-text representations using image tags as anchors for image-text alignment.

- Pre-training corpus
  - three types of existing vision and VL dataset
    - 8.85 million (***w-q-****img*) triples
    - image captioning dataset
    - visual QA dataset
    - image tagging dataset

OSCAR+ pre-training loss: $\mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\text{MTL}} + \mathcal{L}_{\text{CL3}}$

VinVL: OSCAR+ Pre-training

OSCAR+ pre-training loss: $\mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\text{MTL}} + \mathcal{L}_{\text{CL3}}$

Masked Token Loss: $\mathcal{L}_{\text{MTL}} = -\mathbb{E}_{(\boldsymbol{v},\boldsymbol{h})\sim\mathcal{D}} \log p(h_i|\boldsymbol{h}_{\setminus i}, \boldsymbol{v})$

- defined on the text modality (***w*** and ***q***)
- define the *discrete token sequence* as $\boldsymbol{h} \triangleq [\boldsymbol{w}, \boldsymbol{q}]$
- apply the Masked Token Loss (MTL)
- randomly mask each input token with probability 15% and replace the masked one with a special token [MASK].
- predict the masked tokens based on their surrounding tokens and image features

VinVL: OSCAR+ Pre-training

OSCAR+ pre-training loss: $\mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\text{MTL}} + \mathcal{L}_{\text{CL3}}$

Three-way Contrastive Loss: $\mathcal{L}_{\text{CL3}} = -\mathbb{E}_{(\boldsymbol{w},\boldsymbol{q},\boldsymbol{v};c)\sim\tilde{\mathcal{D}}} \log p(c|f(\boldsymbol{w},\boldsymbol{q},\boldsymbol{v}))$

- optimize the objectives for VQA and text-image matching
- training samples: $\boldsymbol{x} \triangleq (\underbrace{\boldsymbol{w}}_{\text{caption}}, \underbrace{\boldsymbol{q},\boldsymbol{v}}_{\text{tags\&image}})$ or $(\underbrace{\boldsymbol{w},\boldsymbol{q}}_{\text{Q\&A}}, \underbrace{\boldsymbol{v}}_{\text{image}})$

Negative examples for contrastive learning:

- polluted "captions": (***w'***, ***q***, ***v***) for text-image matching task
- polluted "answers": (***w***, ***q'***, **v**) for VQA
- apply a FC layer on top as a 3-way classifier $f(\cdot)$ given encoding of [CLS]
  - triplet is matched ($c = 0$)
  - triplet contains a polluted w ($c = 1$)
  - triplet contains a polluted q ($c = 2$)

VinVL: OSCAR+ Pre-training

- Pre-trained models
  - language tokens = [$\boldsymbol{w}$, $\boldsymbol{q}$]
  - region features = $\boldsymbol{v}$
  - *BERT base* and *BERT large*
  - ensure that the features have the same input embedding size using a linear projection via matrix $\boldsymbol{W}$
  - trainable parameters are $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}_{BERT}, \boldsymbol{W}\}$

VinVL: Adapt to VL Tasks

- Generation tasks - Image Captioning
  - fine-tune
    - training sample converted to a triplet: a set of captions, a set of image region features and a set of object tags
    - seq2seq objective + uni-directional prediction with mask of 15% of the caption
  - inference
    - encode the image regions, object tags, and [CLS] as input
    - generate a caption by feeding in a [MASK]

VinVL: Adapt to VL Tasks

- Understanding tasks - VQA & GQA
  - construct the input by concatenating a given question, object tags and object region features
  - feed the [CLS] output from *OSCAR+* to a task-specific linear classifier with a softmax layer


```
{'question': 'Where is he looking?',
 'question_type': 'none of the above',
 'question_id': 262148000,
 'image_id': 'COCO_val2014_000000262148.jpg',
 'answer_type': 'other',
 'label': {
     'ids': ['at table', 'down', 'skateboard', 'table'],
  'weights': [0.30000001192092896,
   1.0,
   0.30000001192092896,
   0.30000001192092896]
   }
}
```

Vision-Language Pretraining

- BERT for Visual Representation Learning
  - ViLBERT, Oscar, VinVL
- **Contrastive Language-Image Pre-training**
  - **CLIP**, "Introduce self-supervised signals widely used in NLP into Vision"
- Generative Language-Image Pre-training
  - BLIP, CoCa
- Training Scaling Up
  - SigLIP

OpenAI

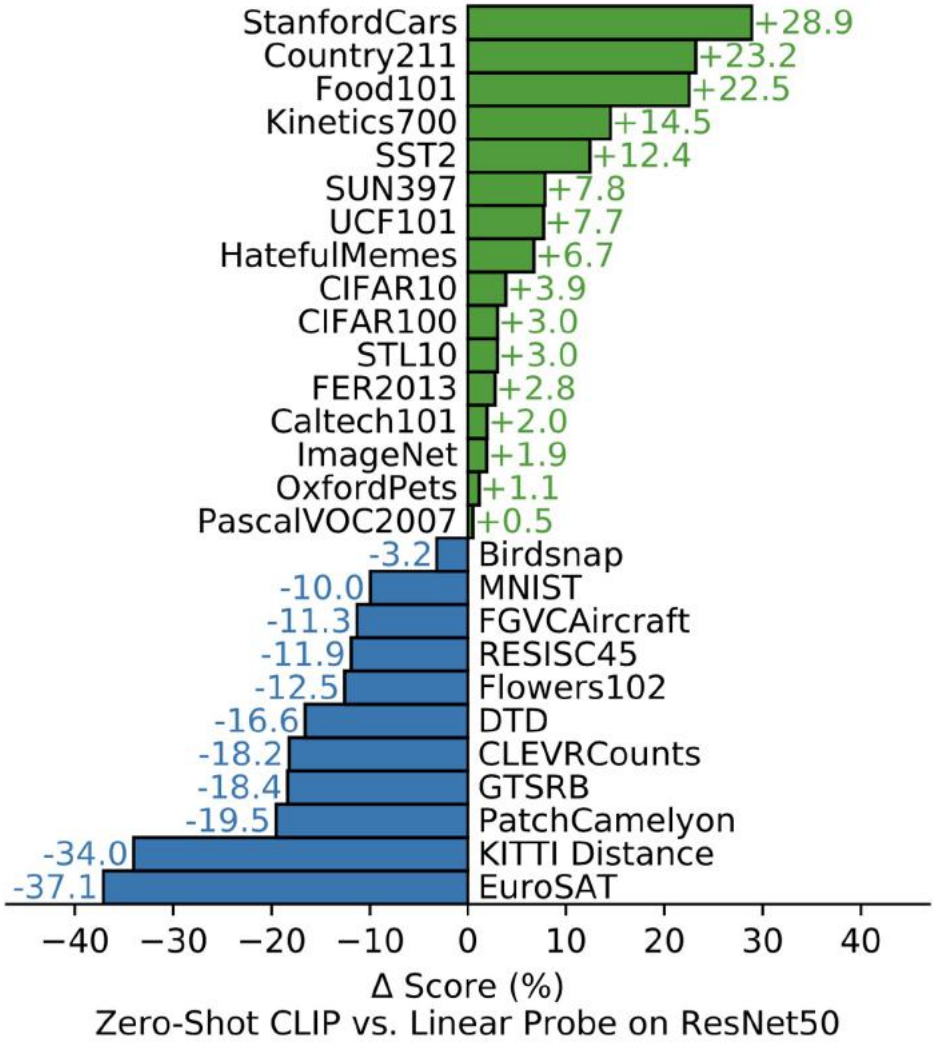
January 2021

CLIP: Background & Motivation

- Success of pre-trained models in NLP
  - *GPT* family
- Zero-shot CV tasks
  - 11.5% accuracy on ImageNet in 2017
  - Improved performance in narrower and more targeted weak supervision
- SOTA CV systems
  - Fixed set of predetermined object categories
  - Low generality and usability
- CLIP-like methods
  - *VirTex*, *ICMLM*, and *ConVIRT*: small scale training (< 1 million images)
- Close the gap
  - Big data set: 400 million image-text pairs
  - Large model size: *ViT-large*

CLIP: Contribution

- Contrastive language-image pre-training
- Zero-shot beats task-specific supervised models
- Linear-probe with good performance
- Better generalization performance
  - combine representation natural language and image



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

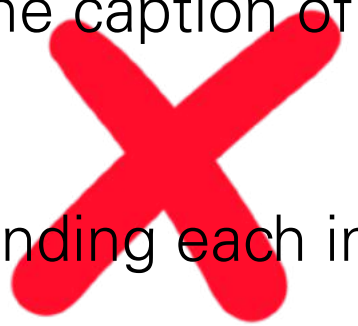
CLIP: Method

- Idea
  - use natural language supervision signals to train a better visual model
- *Why?*
  - no need to label data anymore
  - images and text bound together to form a multi-modal feature

- Method
  1. create a sufficiently large dataset: *Wikipedia-based Image Text (WIT)* dataset - over 400 million image-text pairs
  2. select an efficient pre-training method
  3. choose and scale a model
  4. train

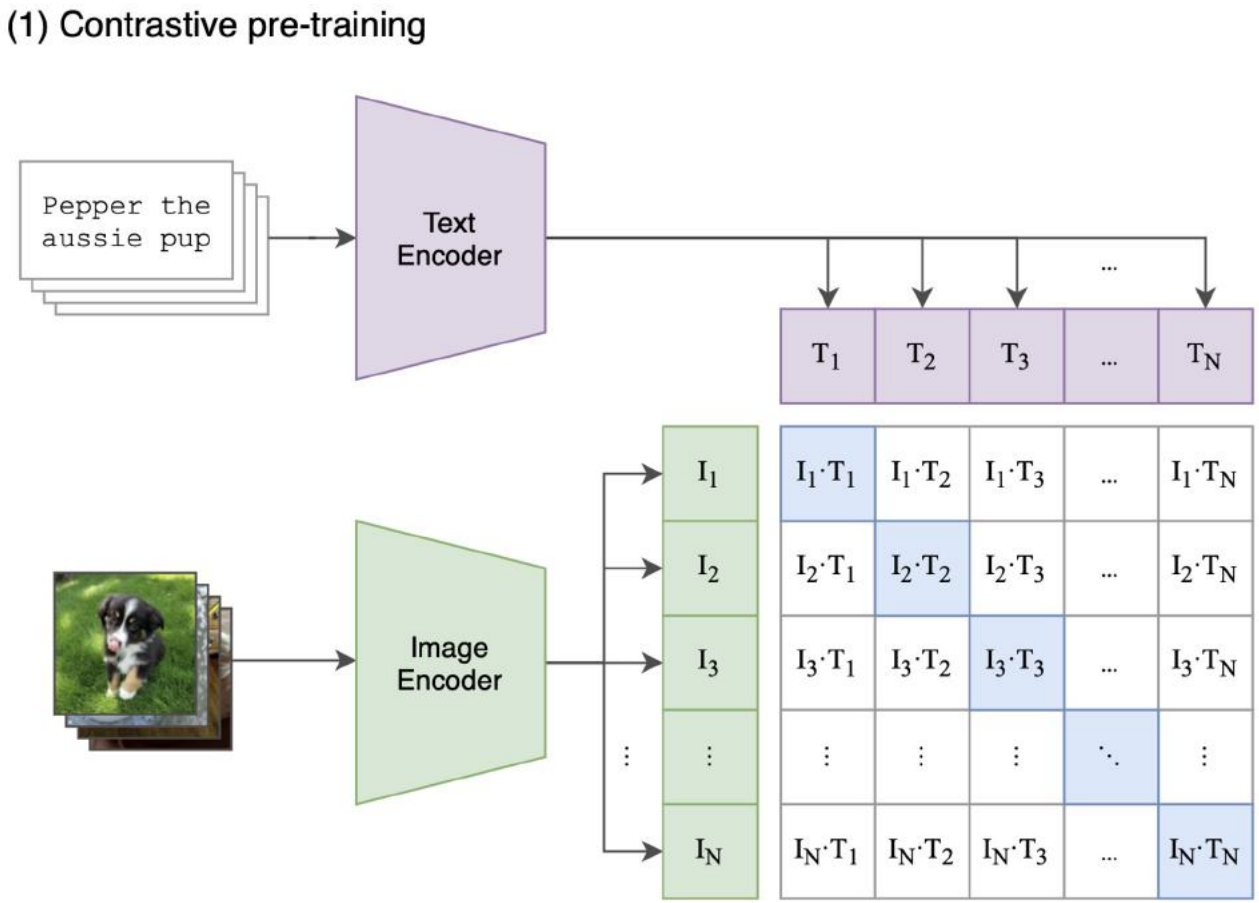
CLIP: Efficient Pre-Training Method

- Trial
  - predictive task: training from scratch and predicting the caption of the image with CNN for image & Transformer for text
- Problem
  - difficult and slow to predict the exact words corresponding each image



- Solution
  - **contrastive learning**
- *Why?*
  - easy to only predict which text as a whole is paired with which image, instead of the exact words of the text

CLIP: Pre-training

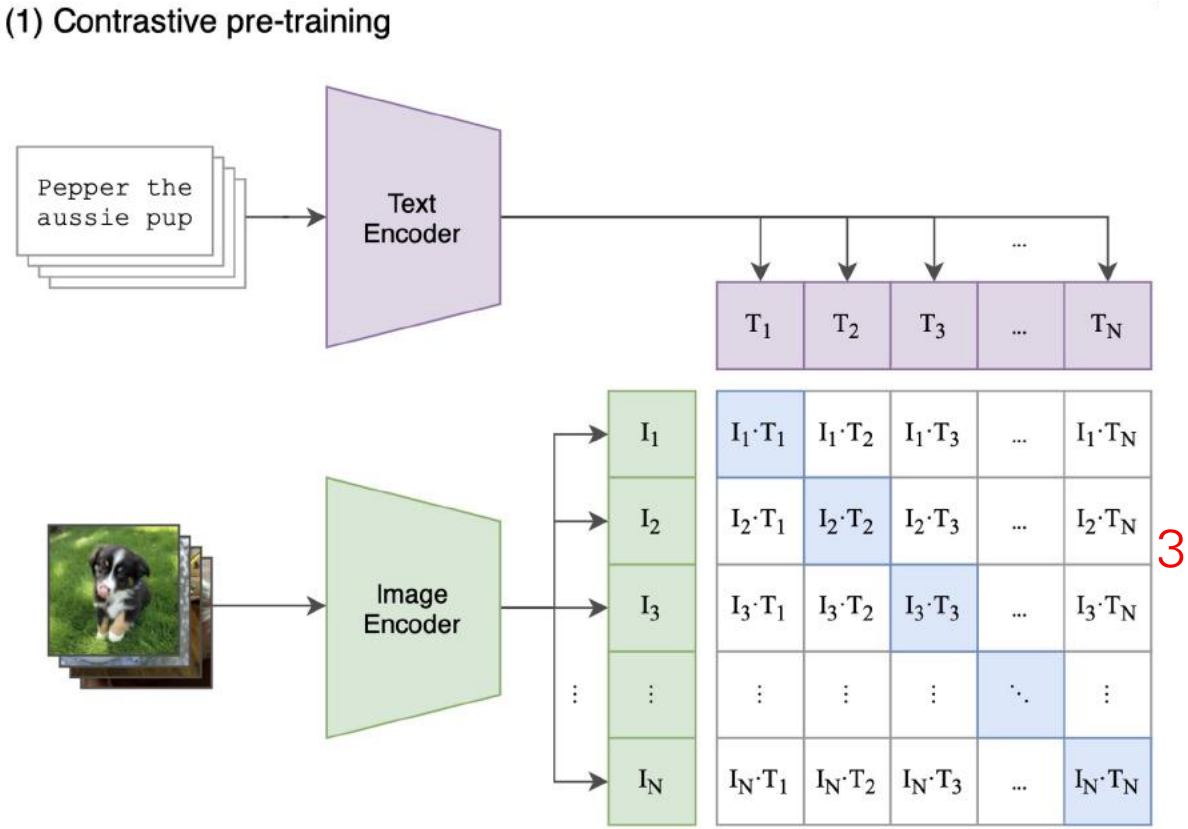


- Image encoder
  - *ResNet / Vision Transformer*
- Text encoder
  - *Transformer*

- Contrastive pre-training
  - contrastive learning on $n \times n$ features
  - positive samples: image-text pairs on the diagonal
  - negative samples: image-text pairs not on the diagonal

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

CLIP: Contrastive Training

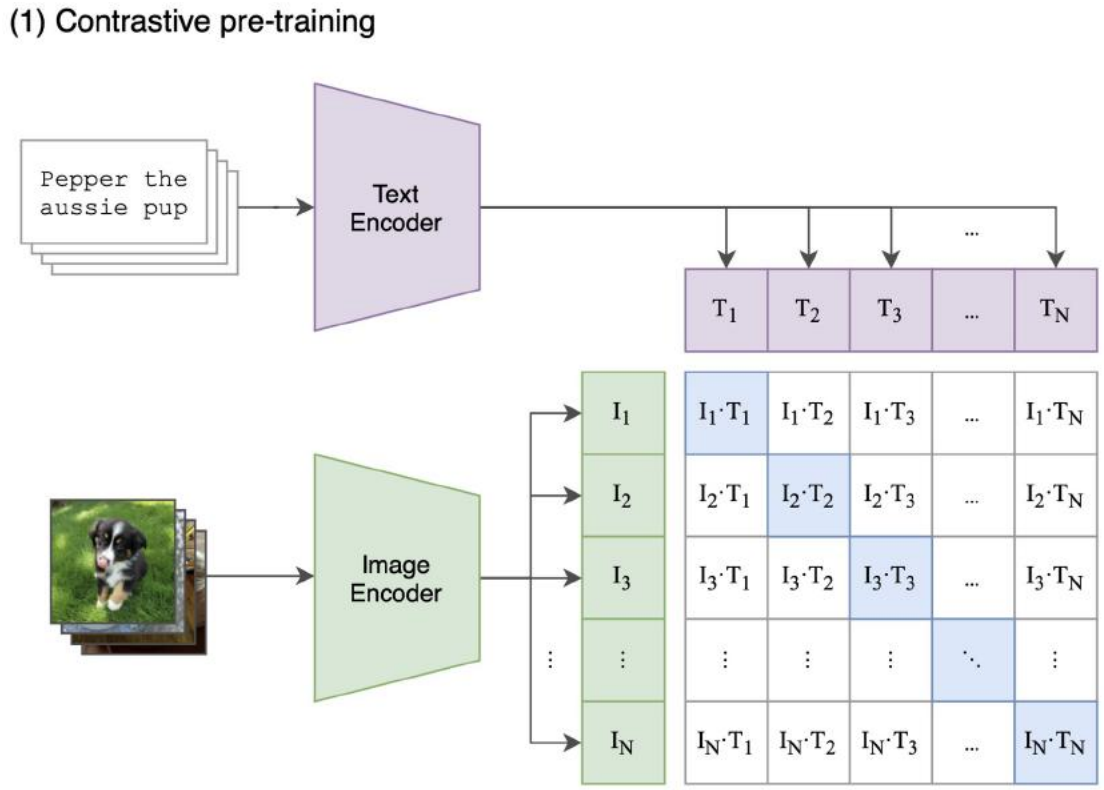



```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

1 2 3 4 5

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

CLIP: Contrastive Training



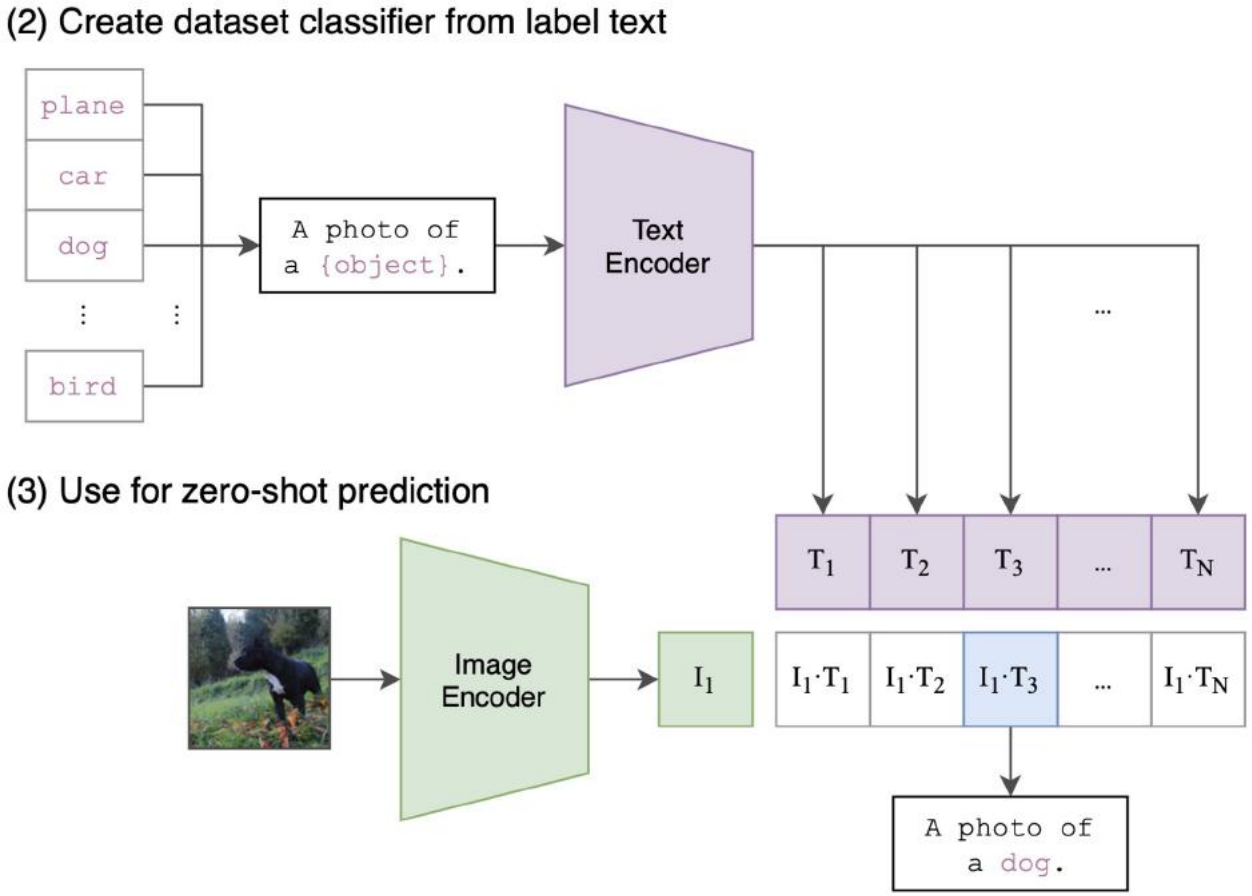

```
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

5

```
def contrastive_loss(logits: torch.Tensor) -> torch.Tensor:
    return nn.functional.cross_entropy(logits, torch.arange(logits.shape[0]))

def clip_loss(logits: torch.Tensor) -> torch.Tensor:
    loss_i = contrastive_loss(logits)
    loss_t = contrastive_loss(logits.t())
    return (loss_i + loss_t) / 2.0
```

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

CLIP: Inference

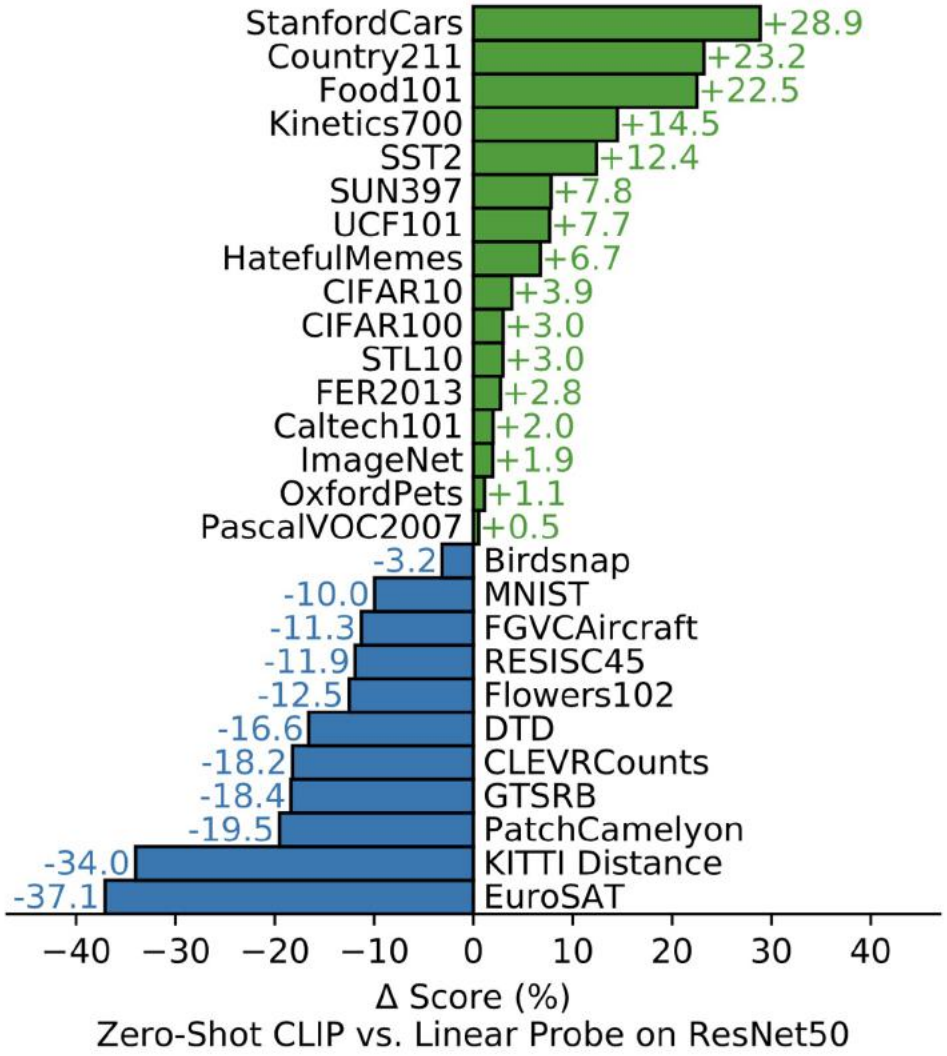


- Inference without classification header
  - cosine similarity

- prompt template
  - "A photo of a {label}, a type of pet"

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

CLIP: Zero-Shot Classification Results



Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

CLIP: Limitations

- Computational efficiency
  - SOTA performance on general dataset requires 1000x computation
- Weak zero-shot performance
  - fine-grained classification
  - abstract concepts: counting tasks
  - new tasks un-existed in pre-training dataset
- Static classification
  - non-generative model: image description
- Data efficiency
  - 12.8 billion imagers in total requires 405 years with training one image/second
- Unrealistic zero-shot
  - *ImageNet*

Vision-Language Pretraining

- BERT for Visual Representation Learning
  - VilBERT, Oscar, VinVL
- **Contrastive Language-Image Pre-training**
  - CLIP, **ALIGN**
- Generative Lang
  - BLIP, CoCa
- Training Scaling Up
  - SigLIP

"Scale of the corpus makes up for noise and leads to SoTA representations"

Google

May 2021

ALIGN: Background & Motivation

- Non-trivial data collection / cleaning in VL field
  - *CLIP*
- Scaling of the corpus makes up for noise
  - noisy dataset of over one billion image alt-text pairs: *Conceptual Captions* dataset
- An objective aligning the visual and language representations
  - dual-encoder
  - Image and text encoders learnt with contrastive loss
  - a shared latent embedding space
- Aligned representations for cross-modality matching/retrieval tasks
  - Zero-shot image classification

ALIGN: Noisy Image-Text Dataset

"motorcycle front wheel"

"*thumbnail for version as of 21 57 29 june 2010*"

"file frankfurt airport skyline 2017 05 jpg"

"file london barge race 2 jpg"

"moustache seamless wallpaper design"

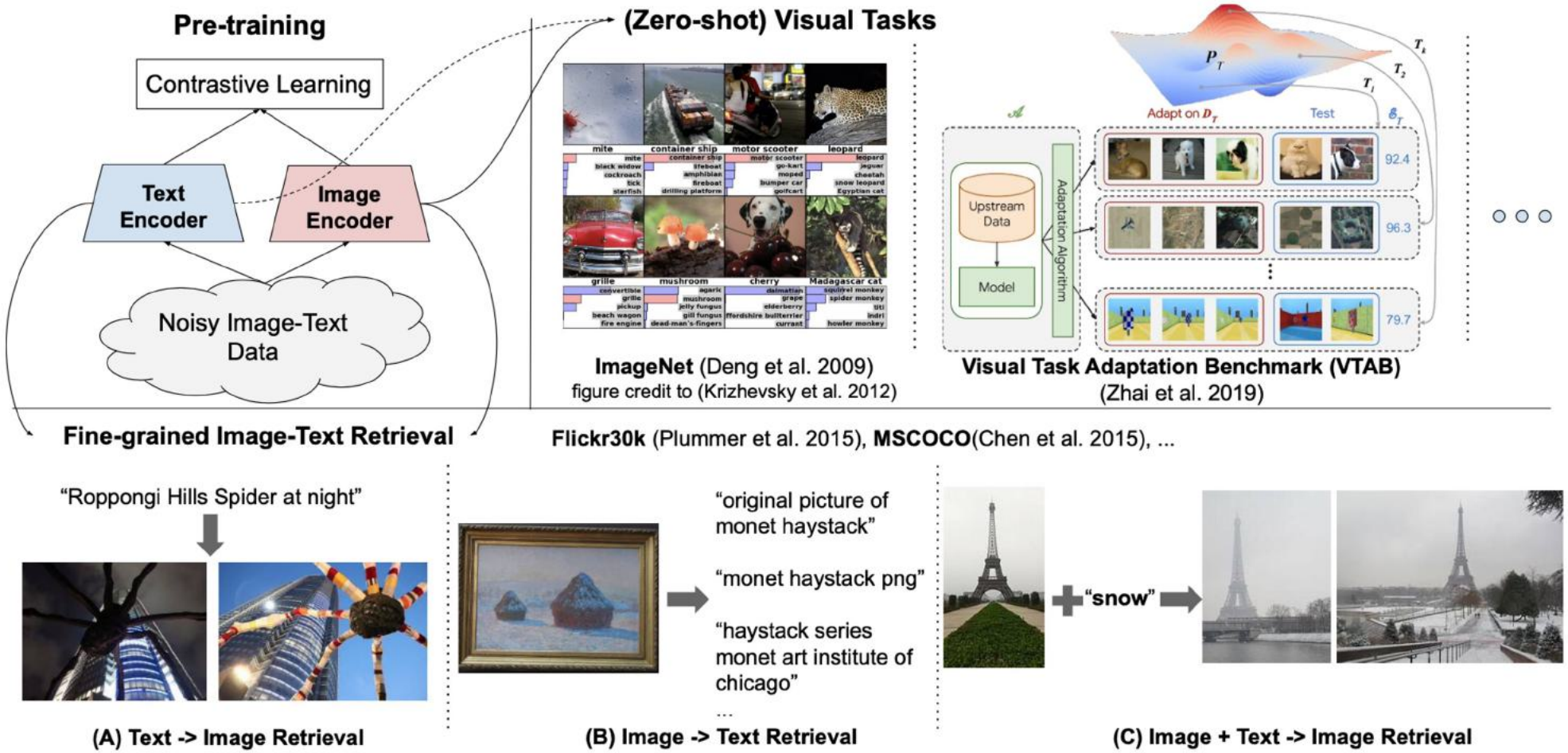
"st oswalds way and shops"

Scale up visual and vision-language representation learning.

- Trade quality for scale by relaxing most of the cleaning steps in the original work of *Conceptual Captions* dataset.
- Only apply minimal frequency-based filtering: aspect ratio, short dimension, content relevancy, text length, …

Jia, Chao, et al. "Scaling up visual and vision-language representation learning with noisy text supervision." International conference on machine learning. PMLR, 2021.

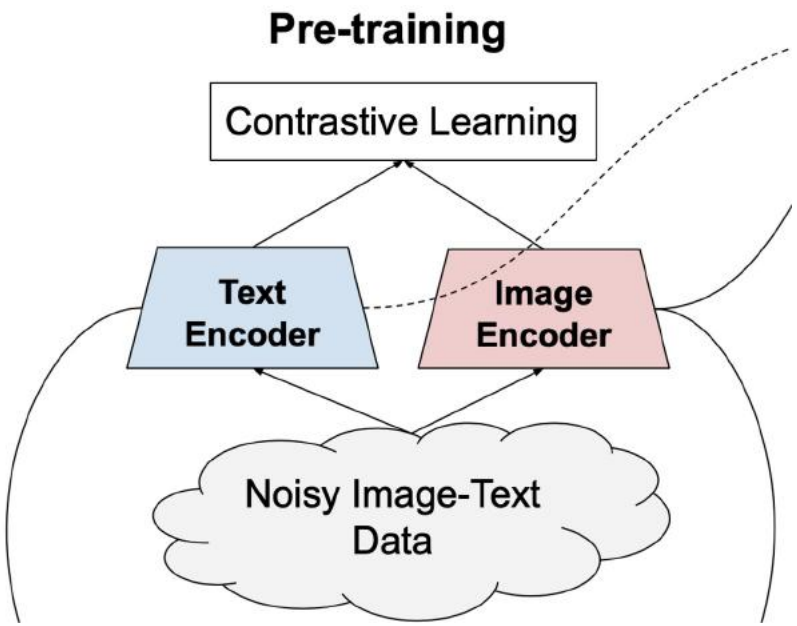
ALIGN: Method



Visual and language representations are jointly learned from noisy image alt-text data. The representations can be used for vision-only or vision-language task transfer. Without any fine-tuning, ALIGN powers zero-shot visual classification and cross-modal search including image-to-text search, text-to-image search and even search with joint image+text queries.

Jia, Chao, et al. "Scaling up visual and vision-language representation learning with noisy text supervision." International conference on machine learning. PMLR, 2021.

ALIGN: Pre-training on Noisy Data



- Image encoder
  - *EfficientNet*
  - global pooling
  - without training the 1x1 conv layer in the classification head
- Text encoder
  - *BERT* with [CLS] token embedding
  - A fully-connected layer on top
- Cosine-similarity combination function on top
- optimized via normalized softmax loss

- Training
  - matched image-text pairs as positive and other as negative

Jia, Chao, et al. "Scaling up visual and vision-language representation learning with noisy text supervision." International conference on machine learning. PMLR, 2021.

ALIGN: Pre-training on Noisy Data

Minimize the sum of two losses:

Image-to-text classification:

$$L_{i2t} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)}$$

Text-to-image classification:

$$L_{t2i} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)}$$

- Parameters
  - x_i and y_j: normalized embedding of image in the *i*-th pair and that of text in the *j*-th pair respectively
  - N: batch size
  - σ: learnable temperature to scale the logits

ALIGN: Transferring

- Image-text matching & retrieval
  - w/wo fine-tuning
  - dataset: *Flickr30K*, *MSCOCO* and *CxC*
  - four intra- and inter-modal retrieval tasks
  - three semantic similarity tasks
- Visual classification
  - *ALIGN* zero-shot transfer
    - dataset: same set (or a subset) of *ImageNet* classes
  - Image encoder transfer
    - dataset: *ImageNet*
    - fine-grained classification dataset: *Flowers-102*, *Oxford-IIIT Pet*, *Stanford Cars* and *Food101*
  - *ImageNet*
    - training the top classification layer only with frozen ALIGN image encoder
    - fully fine-tuned

ALIGN: Results

		Flickr30K (1K test set)						MSCOCO (5K test set)					
		image → text			text → image			image → text			text → image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot	ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
	UNITER	83.6	95.7	97.7	68.7	89.2	93.9	-	-	-	-	-	-
	CLIP	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
	ALIGN	**88.6**	**98.7**	**99.7**	**75.7**	**93.8**	**96.8**	**58.6**	**83.0**	**89.7**	**45.6**	**69.8**	**78.6**
Fine-tuned	GPO	88.7	98.9	99.8	76.1	94.5	97.1	68.1	90.2	-	52.7	80.2	-
	UNITER	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
	ERNIE-ViL	88.1	98.0	99.2	76.7	93.6	96.4	-	-	-	-	-	-
	VILLA	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
	Oscar	-	-	-	-	-	-	73.5	92.2	96.0	57.5	82.8	**89.8**
	ALIGN	**95.3**	**99.8**	**100.0**	**84.9**	**97.4**	**98.6**	**77.0**	**93.5**	**96.9**	**59.9**	**83.3**	**89.8**

Image-text retrieval

Model	ImageNet	ImageNet-R	ImageNet-A	ImageNet-V2
CLIP	76.2	88.9	**77.2**	**70.1**
ALIGN	**76.4**	**92.2**	75.8	**70.1**

Zero-shot Visual Classification

Jia, Chao, et al. "Scaling up visual and vision-language representation learning with noisy text supervision." International conference on machine learning. PMLR, 2021.

ALIGN: Ablation Study

Model + Data	MSCOCO I2T R@1	MSCOCO T2I R@1	ImangeNet KNN R@1
B7 + BERT-base			
+ ALIGN full data	55.4	41.7	69.3
+ ALIGN 10% data	52.0	39.2	68.8
+ CC-3M data	18.9	15.5	48.7
B3 + BERT-mini			
+ ALIGN full data	37.4	24.5	56.5
+ ALIGN 10% data	36.7	24.4	55.8
+ CC-3M data	22.1	17.3	48.9

A large scale training set is essential to allow scaling up of the models and to achieve better performance. A larger model is required to fully utilize the larger dataset.

Model quality improves nicely with larger backbones. As expected, scaling up image encoder capacity is more important for vision tasks. In image-text retrieval tasks the image and text encoder capacities are equally important.

Model	MSCOCO I2T R@1	MSCOCO T2I R@1	ImangeNet KNN R@1
B5 + BERT-base	51.7	**37.5**	64.6
w/ embedding dim=320	50.3	34.1	64.0
w/ embedding dim=160	47.0	34.4	63.7
w/ embedding dim=80	42.0	29.3	61.9
w/ 50% in-batch negs	50.2	37.0	63.8
w/ 25% in-batch negs	48.7	35.8	63.3
w/ softmax temp=1/128	**52.2**	36.5	**64.8**
w/ softmax temp=1/64	**52.2**	37.3	**64.8**
w/ softmax temp=1/32	39.6	26.9	61.2

Jia, Chao, et al. "Scaling up visual and vision-language representation learning with noisy text supervision." International conference on machine learning. PMLR, 2021.

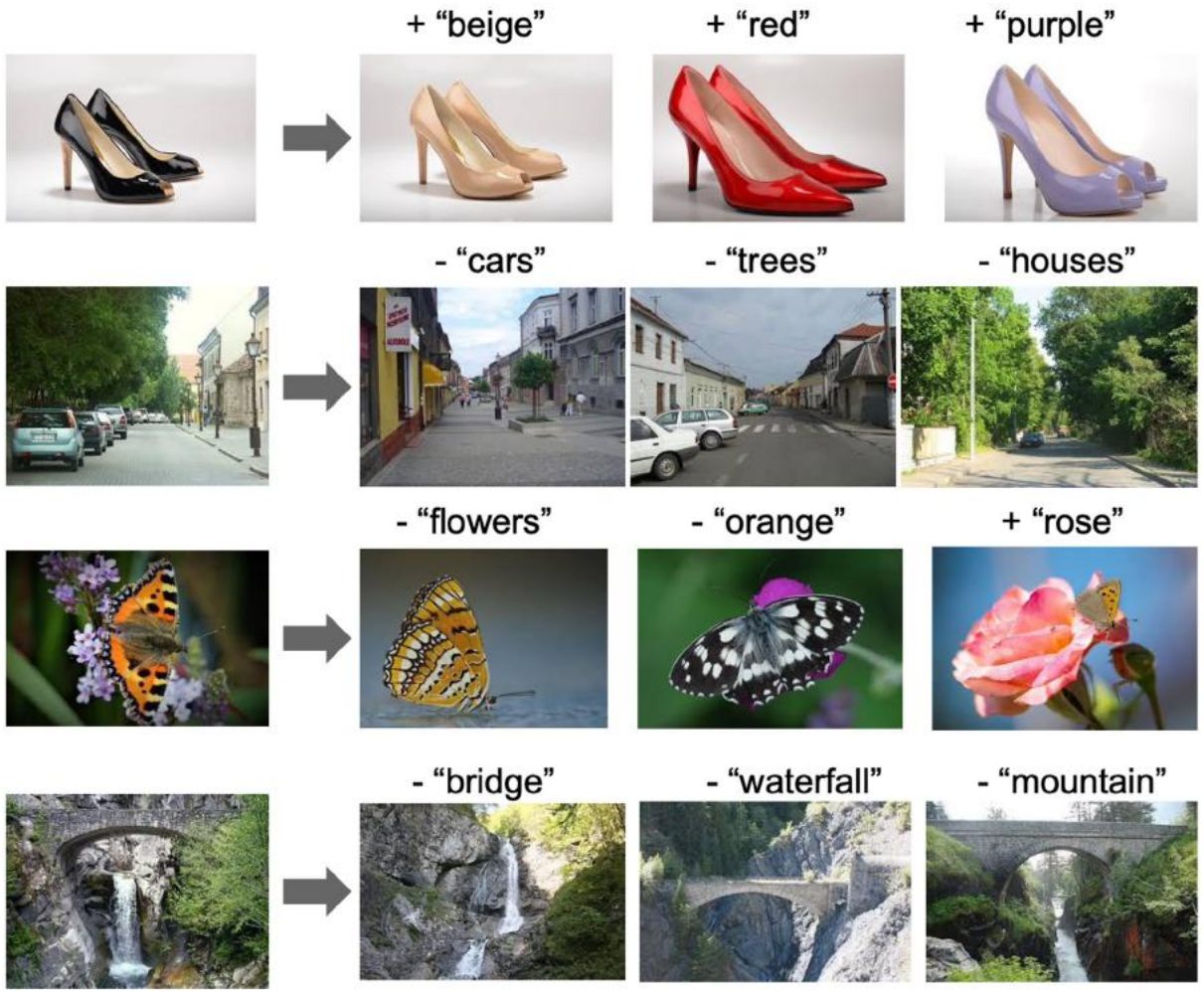
ALIGN: Analysis of Learned Embeddings

A simple image retrieval system to study the behaviors of embeddings trained by *ALIGN*.

ALIGN can align images and texts with similar semantics and generalize to novel complex concepts.

Jia, Chao, et al. "Scaling up visual and vision-language representation learning with noisy text supervision." International conference on machine learning. PMLR, 2021.

ALIGN: Analysis of Learned Embeddings



ALIGN shows that *word2vec*-like linear relationships between word vectors emerge as a result of training them to predict adjacent words in sentences and paragraphs.

Given a query image and a text string, add their *ALIGN* embeddings together and use it to retrieve relevant images.

Jia, Chao, et al. "Scaling up visual and vision-language representation learning with noisy text supervision." International conference on machine learning. PMLR, 2021.

Vision-Language Pretraining

- BERT for Visual Representation Learning
  - ViLBERT, Oscar, VinVL
- Contrastive Language-Image Pre-training
  - CLIP, ALIGN
- **Generative Language-Image Pre-training**
  - **BLIP**, "Improving text quality by bootstrapping contrastive training"
- Training Scaling Up
  - SigLIP

January 2022

BLIP - Background & Motivation

To improve CLIP and ALIGN from 2 perspectives:

1. From model perspective: CLIP & ALIGN adopt encoder-based models
    a. Encoder-based models are **not easily** transferred directly to **text generation tasks**, such as image captioning
    b. Encoder-decoder models have not been successfully adopted for image-text retrieval tasks

BLIP - Background & Motivation

2. From data perspective:
    a. The number of high-quality human-annotated image-text pairs (e.g., COCO) is not enough for large multimodal model training
    b. CLIP & ALIGN are pre-trained on **noisy web text**, which can only yield **suboptimal results**

"Congratulations. You're now the branch manager."

BLIP - Improving Caption Quality

To solve the text quality issue, a natural approach is to build

- A **discriminator** to distinguish between good and bad image-text pairs
- A **generator** to synthesize better quality captions to replace noisy captions
- A **unimodal encoder** to align vision and language representations (similar to ALIGN)

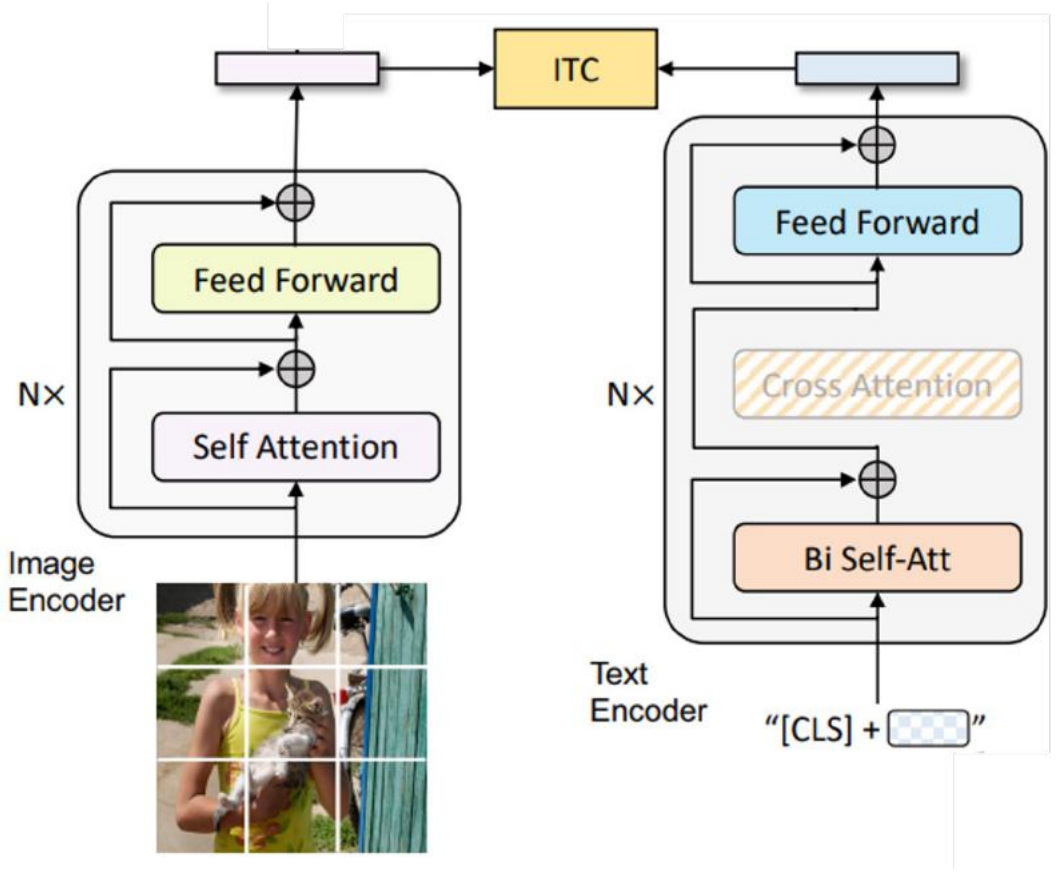
BLIP - Unimodal Encoder

A **multimodal alignment task** to encourage matched image-text pairs to have similar representations in contrast to the negative pairs

Image-Text Contrastive (ITC) Loss

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\Big(\underbrace{\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)}}_{\text{image-to-text}} + \underbrace{\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)}}_{\text{text-to-image}}\Big)$$

- x_i and y_j are normalized low-dimensional representations of [CLS] embeddings of text in the i-th pair and image in the j-th pair mapped by linear transformations
- Sum of 2 InfoNCE (Noise Contrastive Estimation) losses for I2T and T2I



BLIP - Discriminator (Filter)

A **binary classification task** to predict whether an image-text pair is matched or nor, given multimodal features

Hard negative sampling strategy: negative pairs with higher contrastive similarity from ITC are more likely to be selected so that training is meaningful

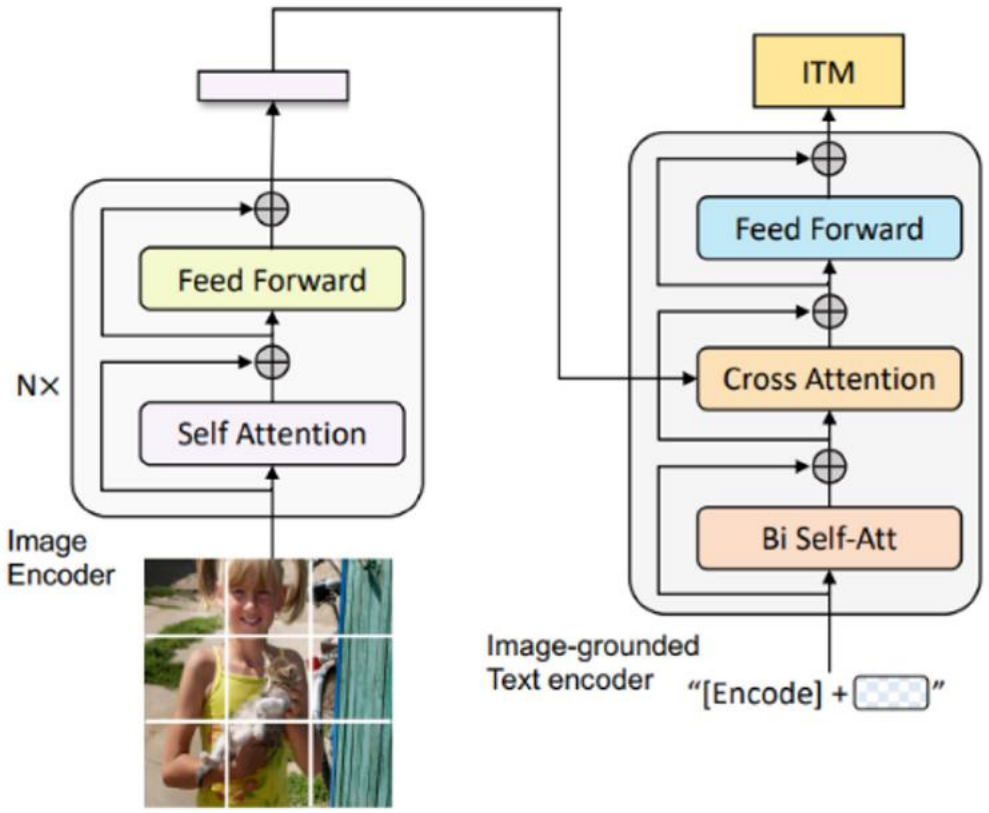


Image-Text Matching (ITM) Loss

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(I,T)\sim D}\text{H}(\boldsymbol{y}^{\text{itm}}, \boldsymbol{p}^{\text{itm}}(I,T))$$

- p^{itm} is the predicted two-class probability
- y^{itm} is a 2-D one-hot vector representing the ground-truth
- H is cross-entropy loss

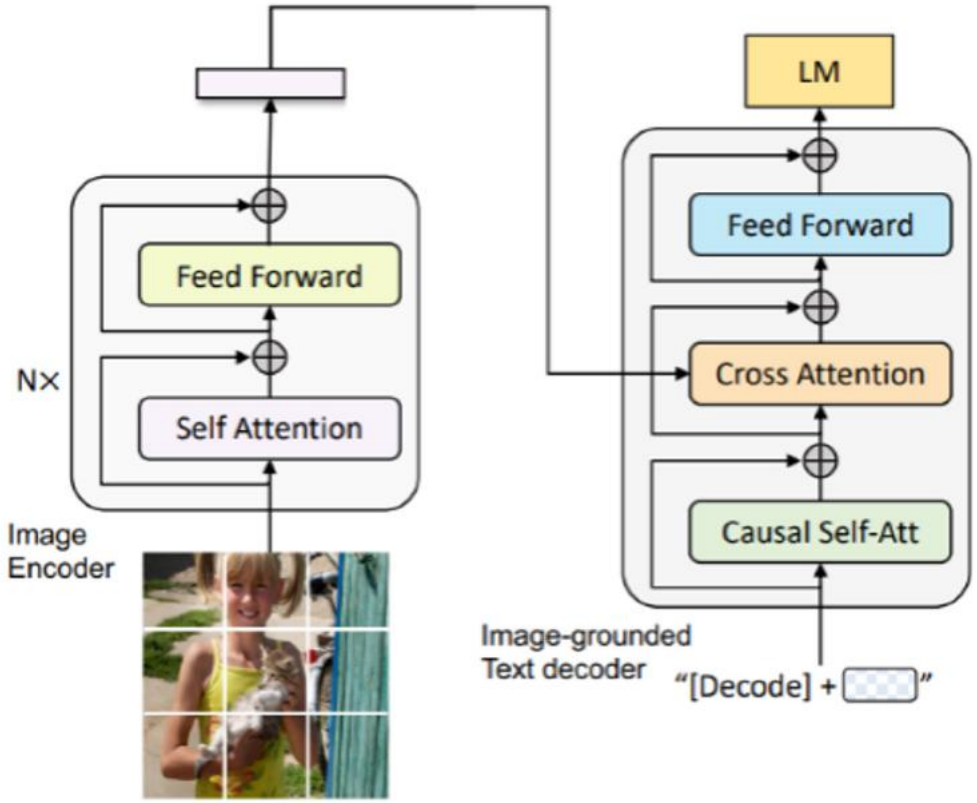
BLIP - Generator (Captioner)

A **generative task** to produce textual descriptions in an **autoregressive** manner given an image

Language Modeling (LM) Loss

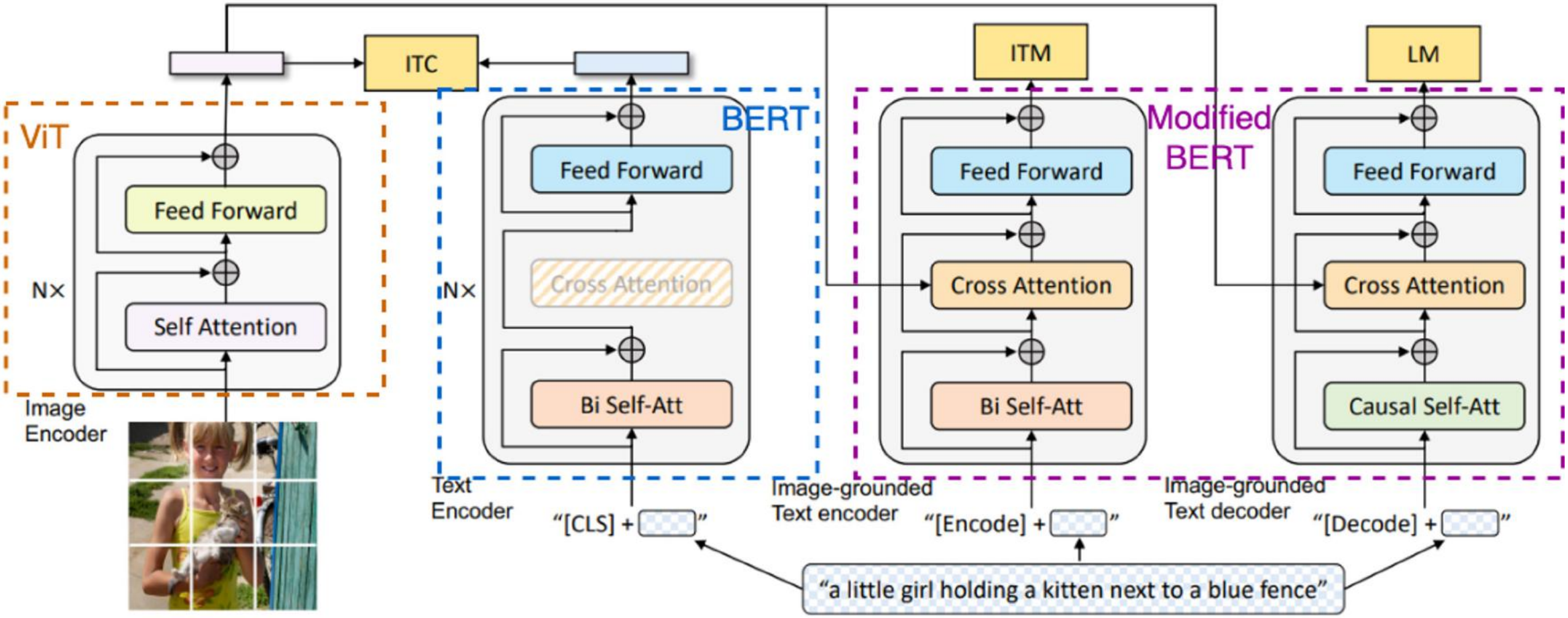
$$\mathcal{L}_{\text{lm}} = -\sum_{t=1}^{T} \log P_\theta(y_t | y_{<t}, x)$$

- y is the language tokens
- x is the image embedding



BLIP - Architecture

$$\mathcal{L} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{lm}} + \mathcal{L}_{\text{itm}}$$



The same color of blocks indicates shared parameters

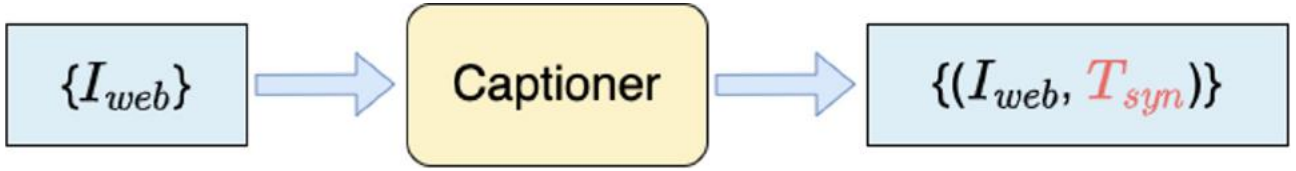
BLIP - Bootstrapping Dataset with CapFilt

1. Pre-train encoder & decoder with noisy web-scale dataset

2. Fine-tune filter and captioner using human annotated dataset (e.g., COCO)

Noisy text
Clean text

3. Generate synthetic caption for web dataset.



BLIP - Bootstrapping Dataset with CapFilt

4. Filter synthetic and web captions to get high quality image-text pairs

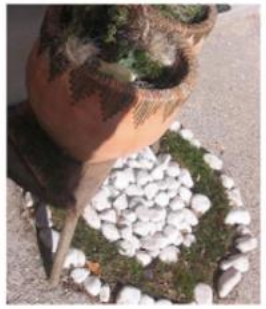
5. Use high quality image-text pairs (129M, larger and cleaner) to pre-train a new model

Continue training does not help. This observation agrees with the common practice in knowledge distillation, where the student model cannot be initialized from the teacher

BLIP - Bootstrapping Dataset with CapFilt

T_w: "from bridge near my house"

T_s: "a flock of birds flying over a lake at sunset"



T_w: "in front of a house door in Reichenfels, Austria"

T_s: "a potted plant sitting on top of a pile of rocks"

T_w: "the current castle was built in 1180, replacing a 9th century wooden castle"

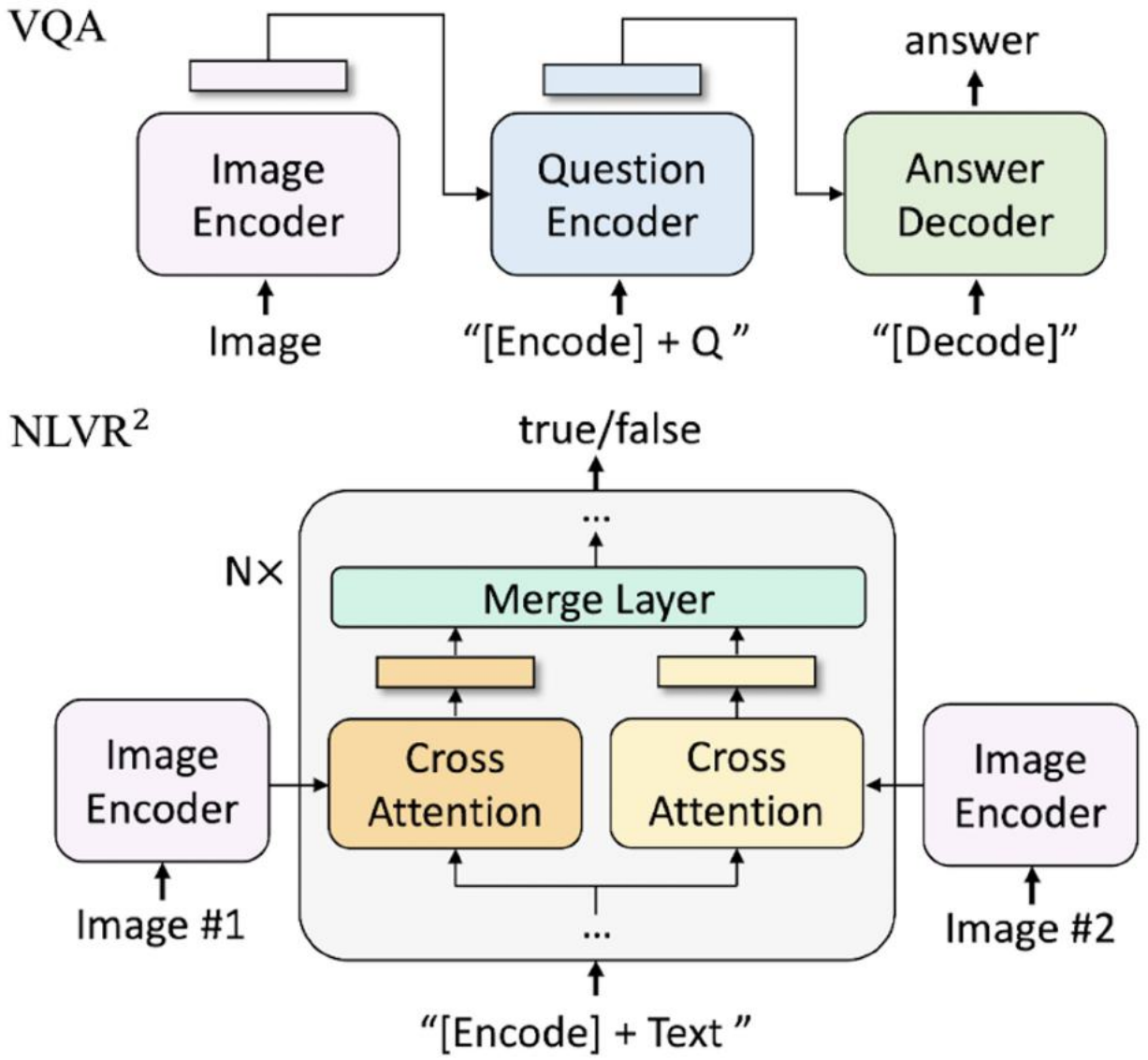
T_s: "a large building with a lot of windows on it"

Figure 4. Examples of the web text T_w and the synthetic text T_s. Green texts are accepted by the filter, whereas red texts are rejected.

These examples show the effectiveness of both captioner and filter

- Captioner is able to generate reasonable descriptions given an image
- Filter is able to accurately identify the more matched text

BLIP - Downstream Tasks



BLIP - Quantitative Results

Pre-train dataset	Bootstrap C	Bootstrap F	Vision backbone	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
COCO+VG +CC+SBU (14M imgs)	✗	✗	ViT-B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	✗	$\checkmark_B$		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	$\checkmark_B$	✗		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
	$\checkmark_B$	$\checkmark_B$		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
COCO+VG +CC+SBU +LAION (129M imgs)	✗	✗	ViT-B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	$\checkmark_B$	$\checkmark_B$		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	$\checkmark_L$	$\checkmark_L$		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
	✗	✗	ViT-L/16	80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
	$\checkmark_L$	$\checkmark_L$		82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8

Table 1. Evaluation of the effect of the captioner (C) and filter (F) for dataset bootstrapping. Downstream tasks include image-text retrieval and image captioning with finetuning (FT) and zero-shot (ZS) settings. TR / IR@1: recall@1 for text retrieval / image retrieval. $\checkmark_{B/L}$: captioner or filter uses ViT-B / ViT-L as vision backbone.

Comparison between using captioner only and using filter only

- Captioner generates more diverse captions, which contain more new information that the model could benefit from

BLIP - Quantitative Results

Pre-train dataset	Bootstrap C	Bootstrap F	Vision backbone	Retrieval-FT (COCO) TR@1	Retrieval-FT (COCO) IR@1	Retrieval-ZS (Flickr) TR@1	Retrieval-ZS (Flickr) IR@1	Caption-FT (COCO) B@4	Caption-FT (COCO) CIDEr	Caption-ZS (NoCaps) CIDEr	Caption-ZS (NoCaps) SPICE
COCO+VG +CC+SBU (14M imgs)	✗	✗	ViT-B/16	78.4	60.7	93.9	82.1	38.0	127.8	102.2	13.9
	✗	$\checkmark_B$		79.1	61.5	94.1	82.8	38.1	128.2	102.7	14.0
	$\checkmark_B$	✗		79.7	62.0	94.4	83.6	38.4	128.9	103.4	14.2
	$\checkmark_B$	$\checkmark_B$		80.6	63.1	94.8	84.9	38.6	129.7	105.1	14.4
COCO+VG +CC+SBU +LAION (129M imgs)	✗	✗	ViT-B/16	79.6	62.0	94.3	83.6	38.8	130.1	105.4	14.2
	$\checkmark_B$	$\checkmark_B$		81.9	64.3	96.0	85.0	39.4	131.4	106.3	14.3
	$\checkmark_L$	$\checkmark_L$		81.2	64.1	96.0	85.5	39.7	133.3	109.6	14.7
	✗	✗	ViT-L/16	80.6	64.1	95.1	85.5	40.3	135.5	112.5	14.7
	$\checkmark_L$	$\checkmark_L$		82.4	65.1	96.7	86.7	40.4	136.7	113.2	14.8

Table 1. Evaluation of the effect of the captioner (C) and filter (F) for dataset bootstrapping. Downstream tasks include image-text retrieval and image captioning with finetuning (FT) and zero-shot (ZS) settings. TR / IR@1: recall@1 for text retrieval / image retrieval. $\checkmark_{B/L}$: captioner or filter uses ViT-B / ViT-L as vision backbone.

Comparison between using CapFilt base and using CapFilt large

- Scaling up CapFilt from base to large only improves generative task performance
- Improvements of retrieval tasks is achieved by scaling up the vision backbone

BLIP - Quantitative Results

Method	Pre-train # Images	Flickr30K (1K test set) TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.8B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP	14M	94.8	99.7	**100.0**	84.9	96.7	98.3
BLIP	129M	**96.0**	**99.9**	**100.0**	85.0	**96.8**	98.6
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**96.0**	**99.9**	**100.0**	**85.5**	**96.8**	**98.7**
$\text{BLIP}_{\text{ViT-L}}$	129M	96.7	100.0	100.0	86.7	97.3	98.7

Table 6. Zero-shot image-text retrieval results on Flickr30K.

- The smallest BLIP outperforms ALIGN, despite using less than 1% of the data
- The smallest BLIP also outperforms ALBEF, which adopts encoder-based design and uses the same 14M images as BLIP without bootstrapping text

Vision-Language Pretraining

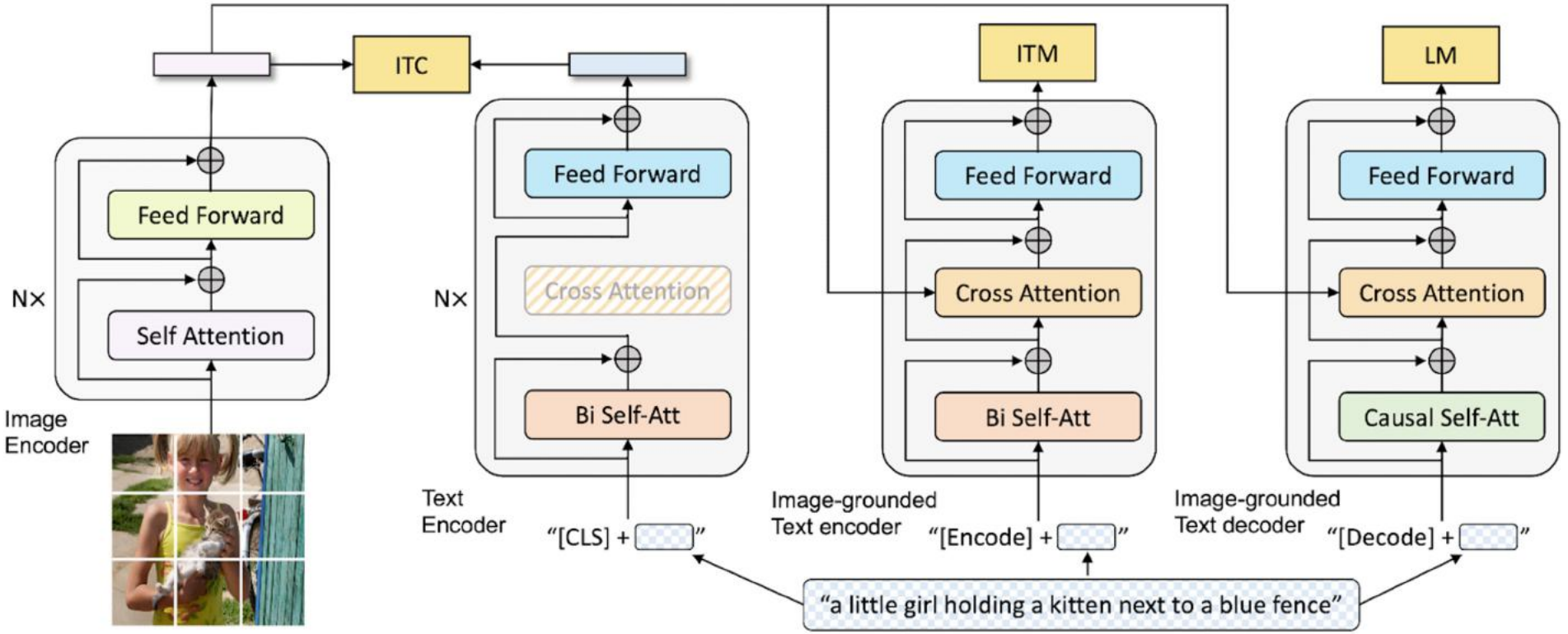
- BERT for Visual Representation Learning
  - ViLBERT, Oscar, VinVL
- Contrastive Language-Image Pre-training
  - CLIP, ALIGN
- **Generative Language-Image Pre-training**
  - BLIP, **CoCa** "Combining contrastive training + generative training"
- Training Scaling Up
  - SigLIP

Google

May 2022

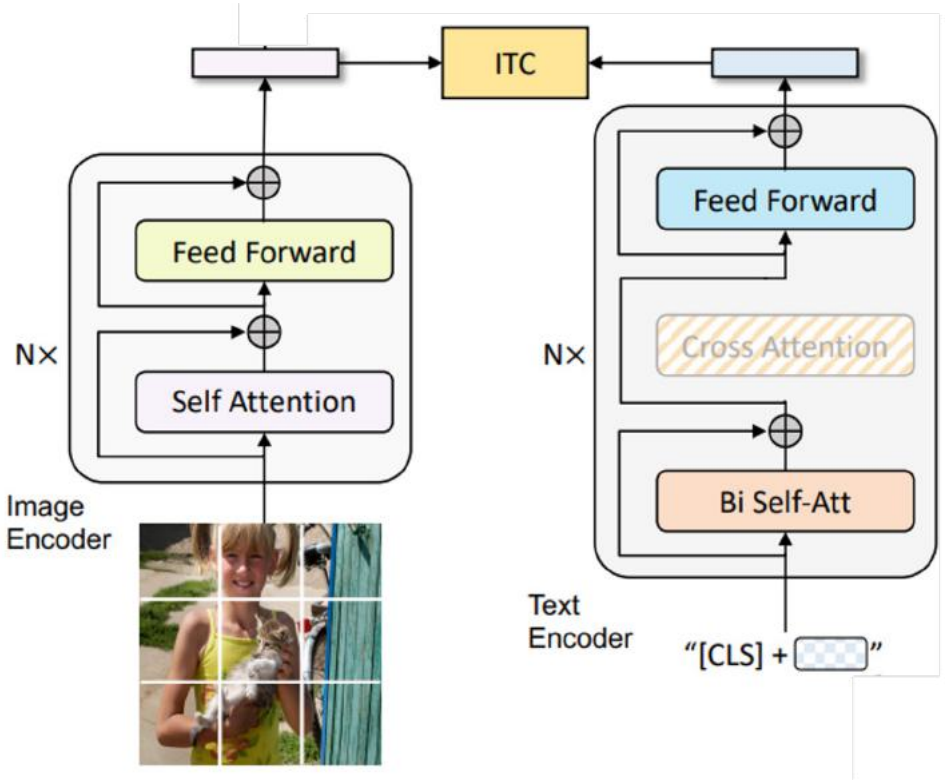
CoCa - Background & Motivation

- Recall: for each image-text pair, BLIP pre-training requires 1 forward pass through visual transformer and 3 forward passes through text transformers
- Need **a minimalist design** of BLIP to improve training efficiency

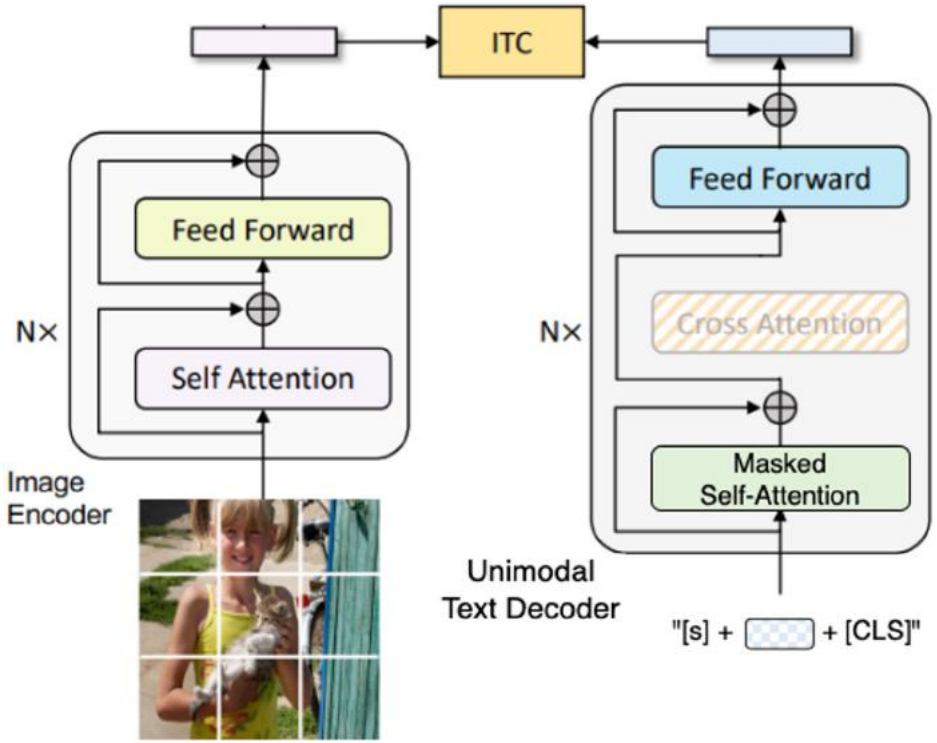


CoCa - Replacing Text Encoder with Decoder

Append a [CLS] token at the end of input sentence and use its corresponding output of decoder as the text embedding

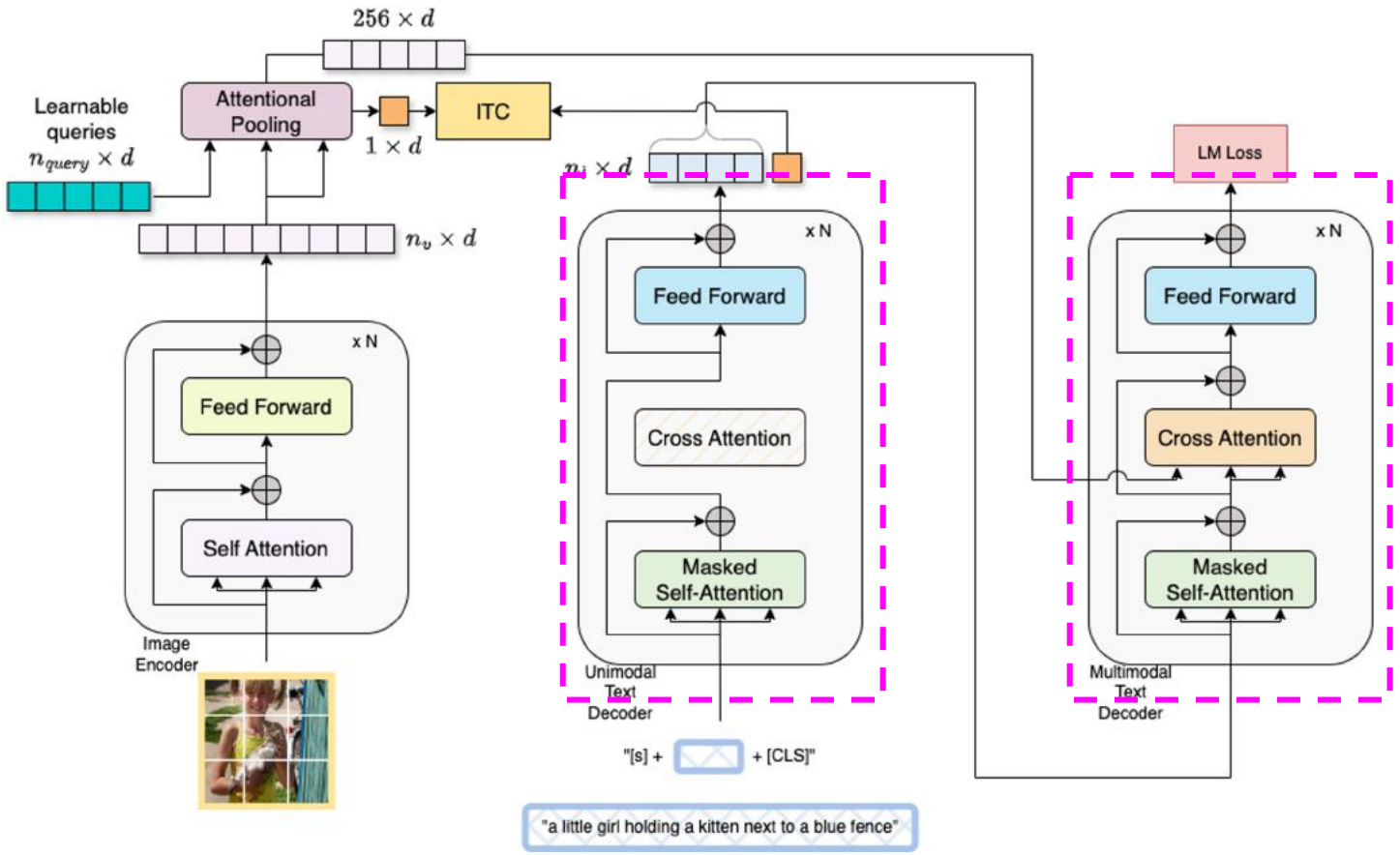


BLIP



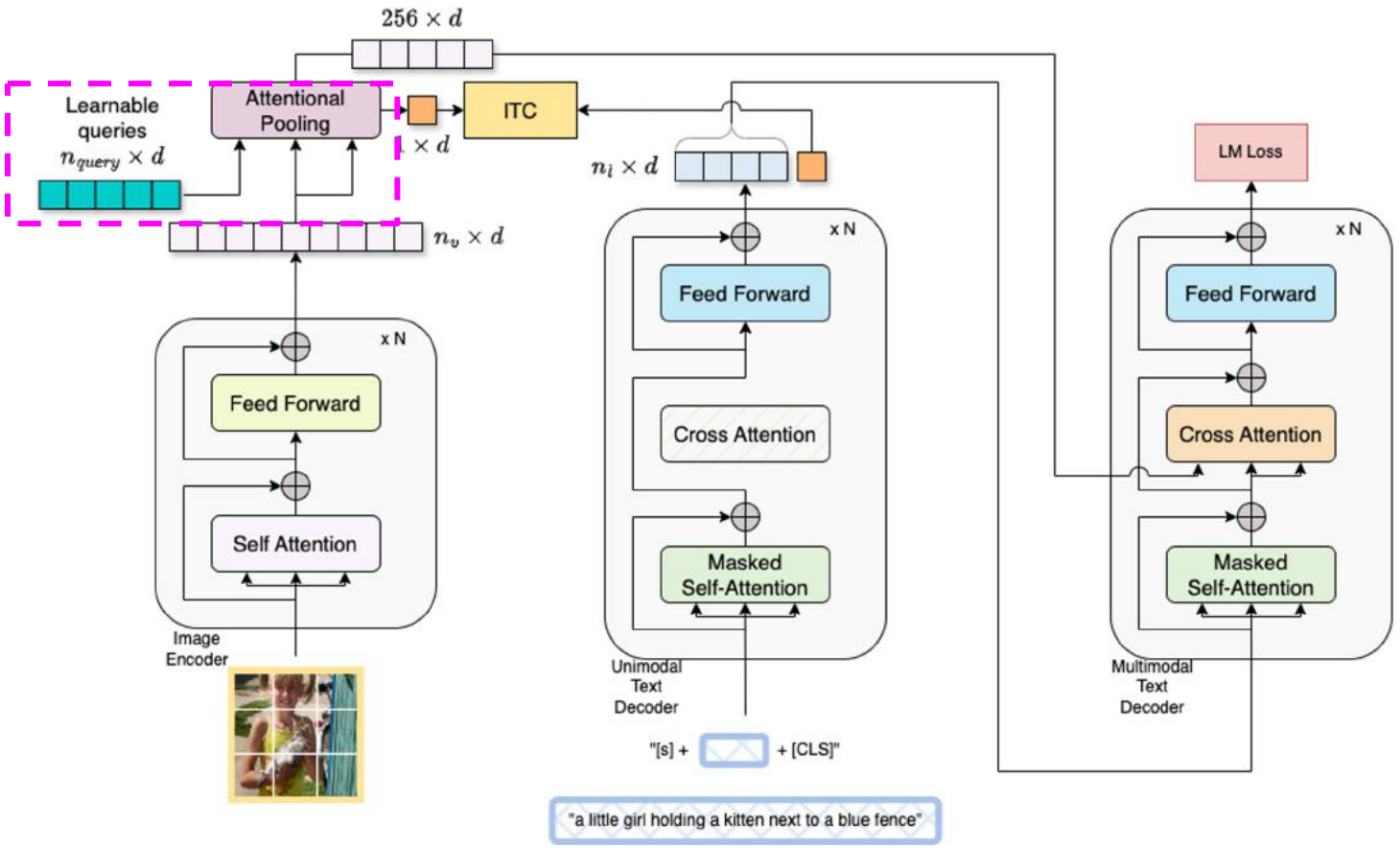
CoCa

CoCa - Decoupled Decoder



- Split the decoder into **unimodal** and **multimodal** components, by skipping the cross-attention mechanism in the unimodal decoder layers
- Benefit: the text decoder can efficiently generate outputs for both contrasive and generative losses **with a single forward pass**, compared to two passes for in BLIP

CoCa - Attentional Poolers



Task-specific attentional pooling

- A multi-head attention layer with n_{query} learnable queries, with image encoder output as both keys and values
- n_{query} = 1 for ITC loss. Pooled image embedding as a global representation
- n_{query} = 256 for LM loss. More visual tokens are beneficial for region-level features

CoCa - Benefits of Attentional Poolers

Adaptor for downstream tasks

- E.g., for video classification, a single query-token is learned to weight outputs of all tokens of spatial patches x temporal frames

Enhanced frozen-feature evaluation

- Linear probing struggles to accurately measure learned representations
- Learning a new pooler to aggregate features enables the model to obtain strong performance as a frozen encoder
- It can also benefit to multi-task problems that share the same frozen image encoder but different task-specific heads

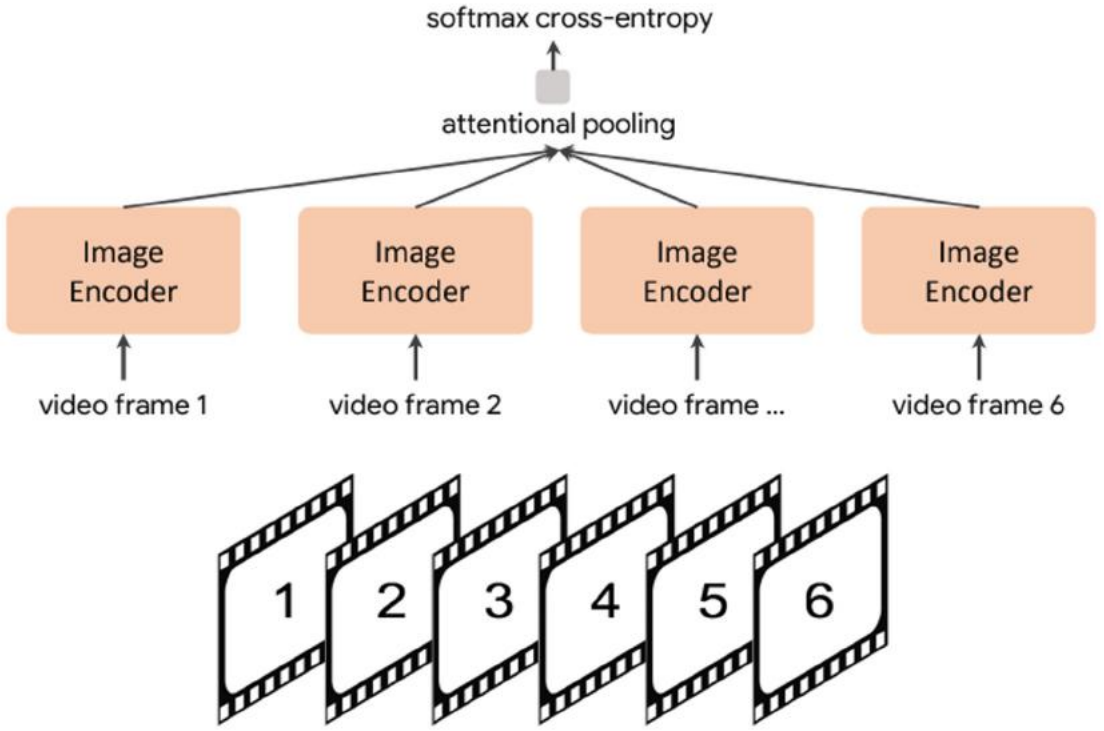


Figure 3: CoCa for video recognition.

CoCa - Pre-Training Details

Loss Function

$$\mathcal{L}_{CoCa} = \lambda_{ITC} \cdot \mathcal{L}_{ITC} + \lambda_{LM} \cdot \mathcal{L}_{LM}$$

- λ are loss weighting hyper-parameters
- Empirically, a larger LM loss weight is better (**λ_{LM}: λ_{ITC} = 2:1**)
- Explanation: the ITC loss can be interpreted as a special case of the generative approach applied on image, when the vocabulary is the set of all captions

Number of unimodal and multimodal decoder layers

- **$N_{unimodal\_decoder} = N_{multimodal\_decoder}$**
- Intuitively, fewer unimodal text layers leads to worse zero-shot classification due to lack of capacity for good unimodal text understanding
- Fewer multimodal layers reduces the model's power to reason over multimodal inputs such as VQA

Dataset (4.8B): ALIGN (1.8B) + JFT-3B (internal Google dataset)

CoCa - Evaluations

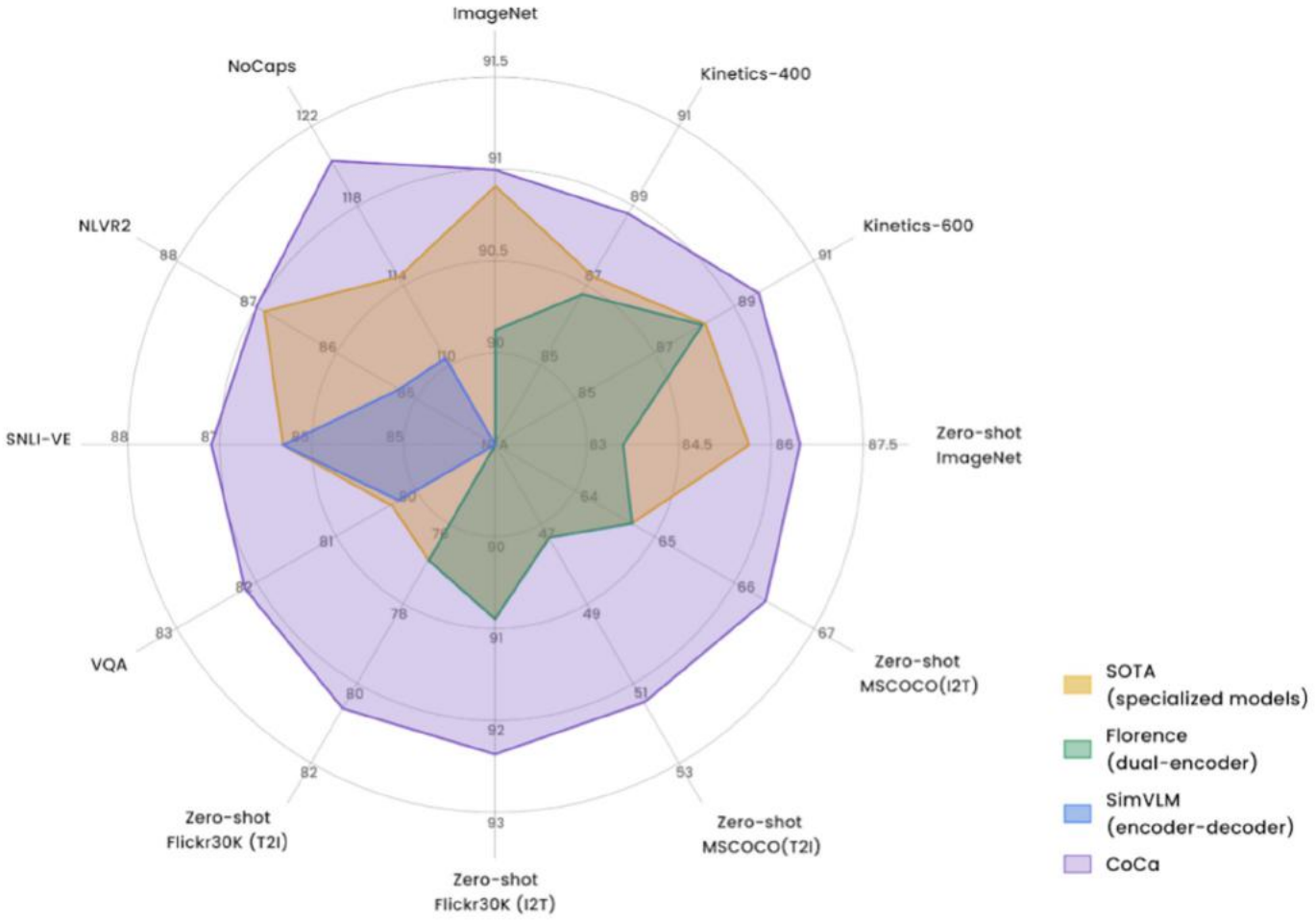


Figure 4: Comparison of CoCa with other image-text foundation models (without task-specific customization) and multiple state-of-the-art task-specialized models.

CoCa outperforms foundation models and task-specialized models on 12 benchmarks including significant improvements in image-text retrieval, image captioning and VQA

Vision-Language Pretraining

- BERT for Visual Representation Learning
  - ViLBERT, Oscar, VinVL
- Contrastive Language-Image Pre-training
  - CLIP, ALIGN
- Generative Language-Image Pre-training
  - BLIP, CoCa
- **Training Scaling Up**
  - **SigLIP** "Scaling up training with sigmoid loss"

Google

May 2022

SigLIP: Background & Motivation

- Contrastive pre-training
  - weak supervision
  - aligned representation space for images and texts
  - *CLIP* and *ALIGN*
  - contrastive objective
- Batch-level softmax-based contrastive loss
  - pairwise similarity scores across all images, then all texts
  - numerically unstable
  - stabilization requiring additional pass over the full batch
- Sigmoid loss
  - simplifying the distributed loss implementation
  - symmetric sigmoid loss requiring just a single pass
  - boosting efficiency
  - decoupling batch size from definition of task

SigLIP: Softmax-based Contrastive Loss

Given a mini-batch $\mathcal{B} = \{(I_1, T_1), (I_2, T_2), \dots\}$ of image-text pairs.

When using the softmax loss to formalize this objective, an image model $f(\cdot)$ and a text model $g(\cdot)$ are trained to minimize the following objective:

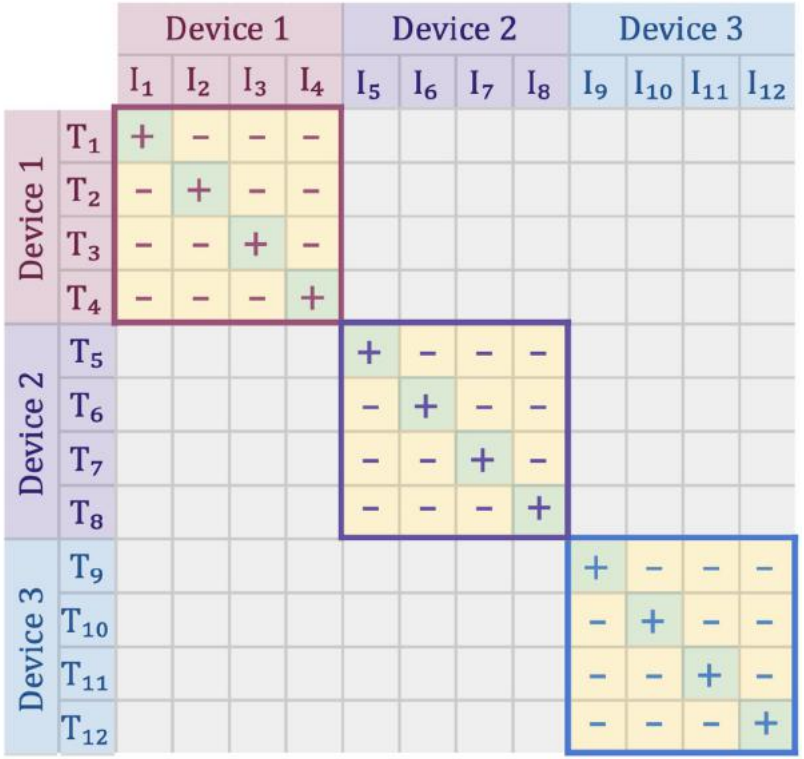
$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

$\mathbf{x}_i = \frac{f(I_i)}{\|f(I_i)\|_2}$, $\mathbf{y}_i = \frac{g(T_i)}{\|g(T_i)\|_2}$, scalar t is parametrized as $\exp(t')$ and t' is a global freely learnable parameter.

Due to the asymmetry of the softmax loss, the normalization is independently performed two times: across images and across texts.

SigLIP: Softmax-based Contrastive Loss

Contrastive training typically utilizes data parallelism. Computing the loss when data is split across D devices necessitates gathering all embeddings with expensive all-gathers and the materialization of a memory-intensive $|\mathcal{B}| \times |\mathcal{B}|$ matrix of pairwise similarities.



$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

Zhai, Xiaohua, et al. "Sigmoid loss for language image pre-training." arXiv preprint arXiv:2303.15343 (2023).

SigLIP: Sigmoid Loss

Sigmoid loss does not require computing global normalization factors. It processes every image-text pair independently, effectively turning the learning problem into the standard binary classification on the dataset of all pair combinations, with a positive labels for the matching pairs (I_i, T_i) and negative labels for all other pairs $(I_i, T_{j \neq i})$. The loss is defined as:

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

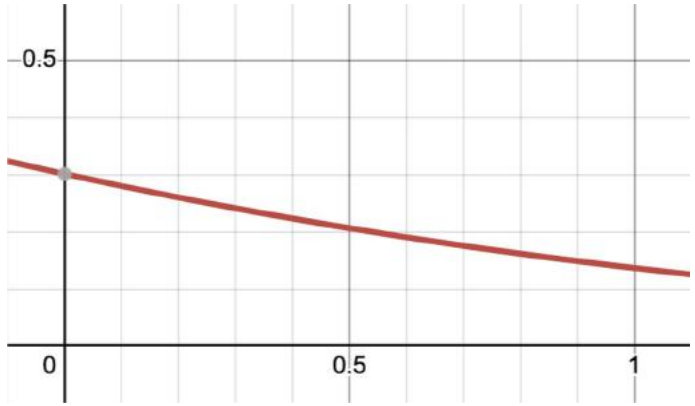
z_{ij} is the label for a given image and text input, which equals 1 if they are paired and −1 otherwise.

An additional learnable bias term b similar to the temperature t is introduced to overcome heavy imbalance coming from the many negatives dominating the loss.

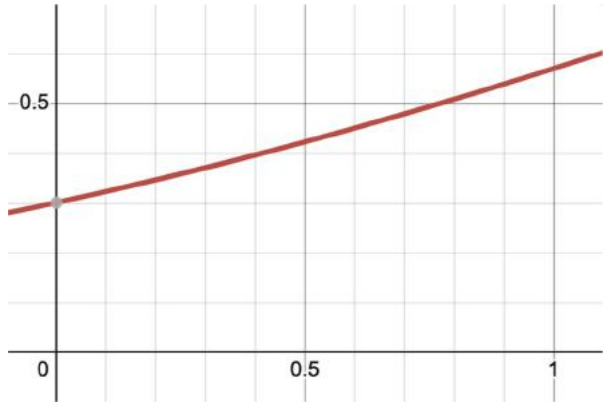
SigLIP: Sigmoid Loss

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

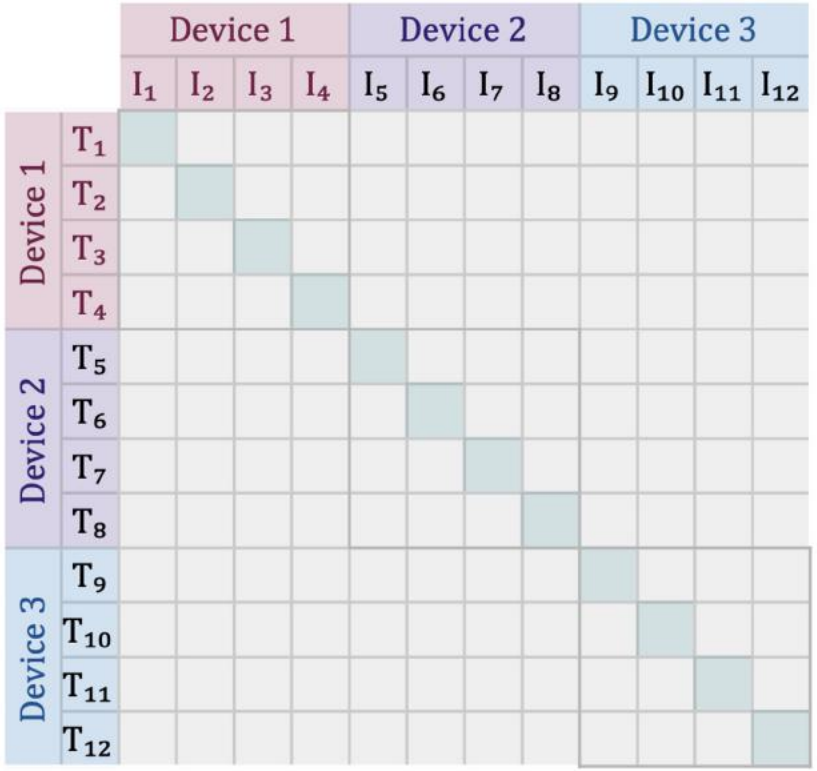
pair: $-\mathcal{L}_{ij} = \log\left(\frac{1}{1+e^{1\cdot(-x)}}\right)$

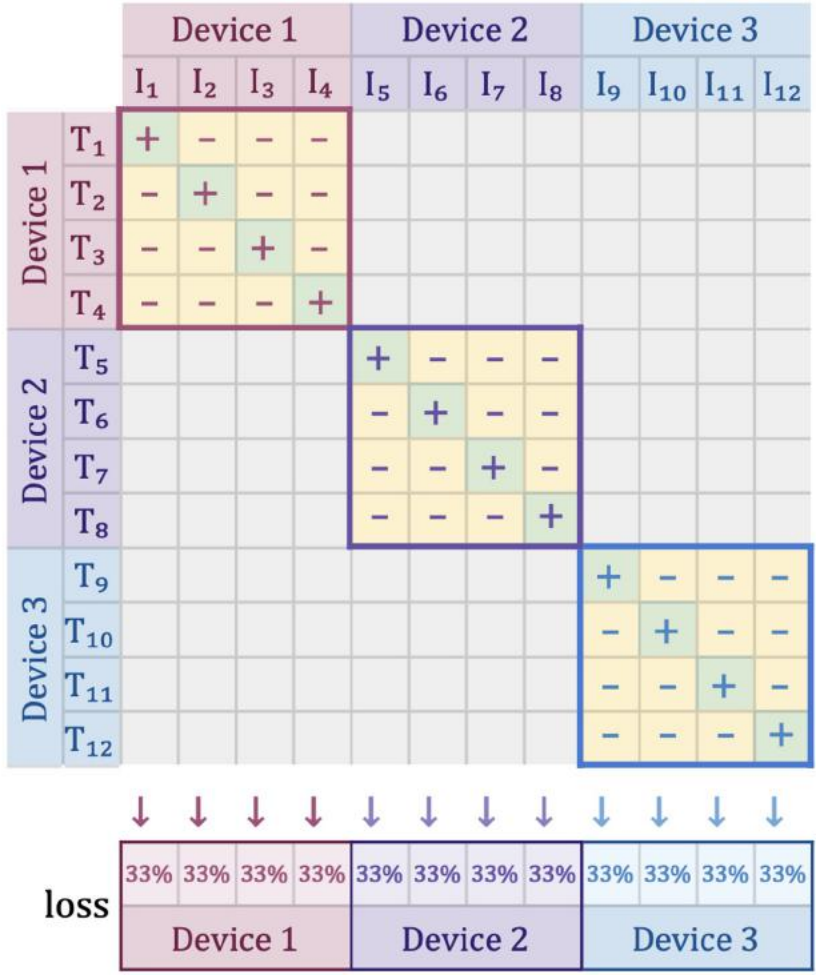


unpair: $-\mathcal{L}_{ij} = \log\left(\frac{1}{1+e^{-1\cdot(-x)}}\right)$



SigLIP: Efficient Loss Implementation



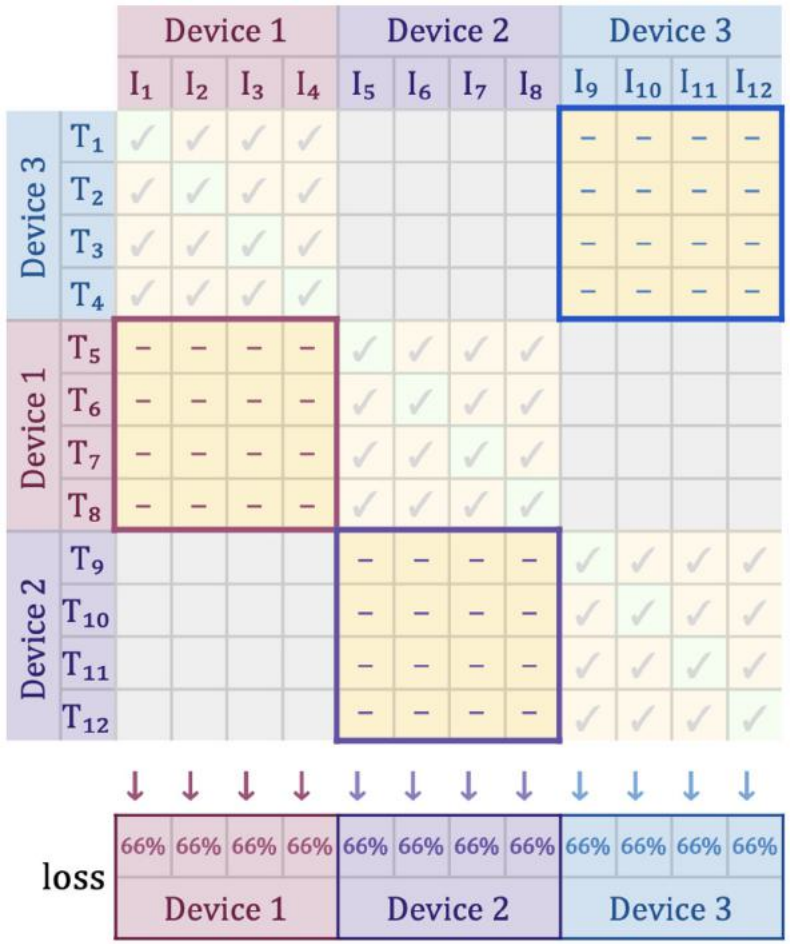


(a) Initially each device holds 4 image and 4 text representations. Each device needs to see the representations from other devices to calculate the full loss.

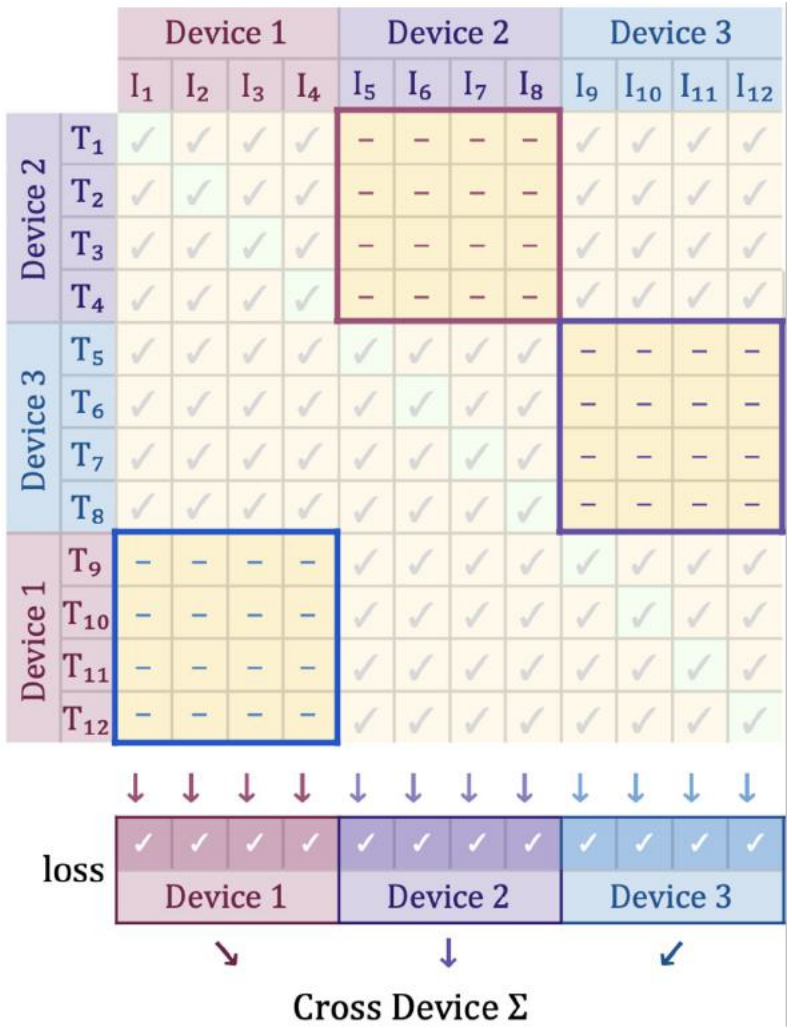
(b) They each compute the component of the loss (highlighted) for their representations, which includes the positives.

Zhai, Xiaohua, et al. "Sigmoid loss for language image pre-training." arXiv preprint arXiv:2303.15343 (2023).

SigLIP: Efficient Loss Implementation



(c) Texts are swapped across the devices, so device 1 now has $I_{1:4}$ and $T_{5:8}$ etc. The new loss is computed and accumulated with the previous.

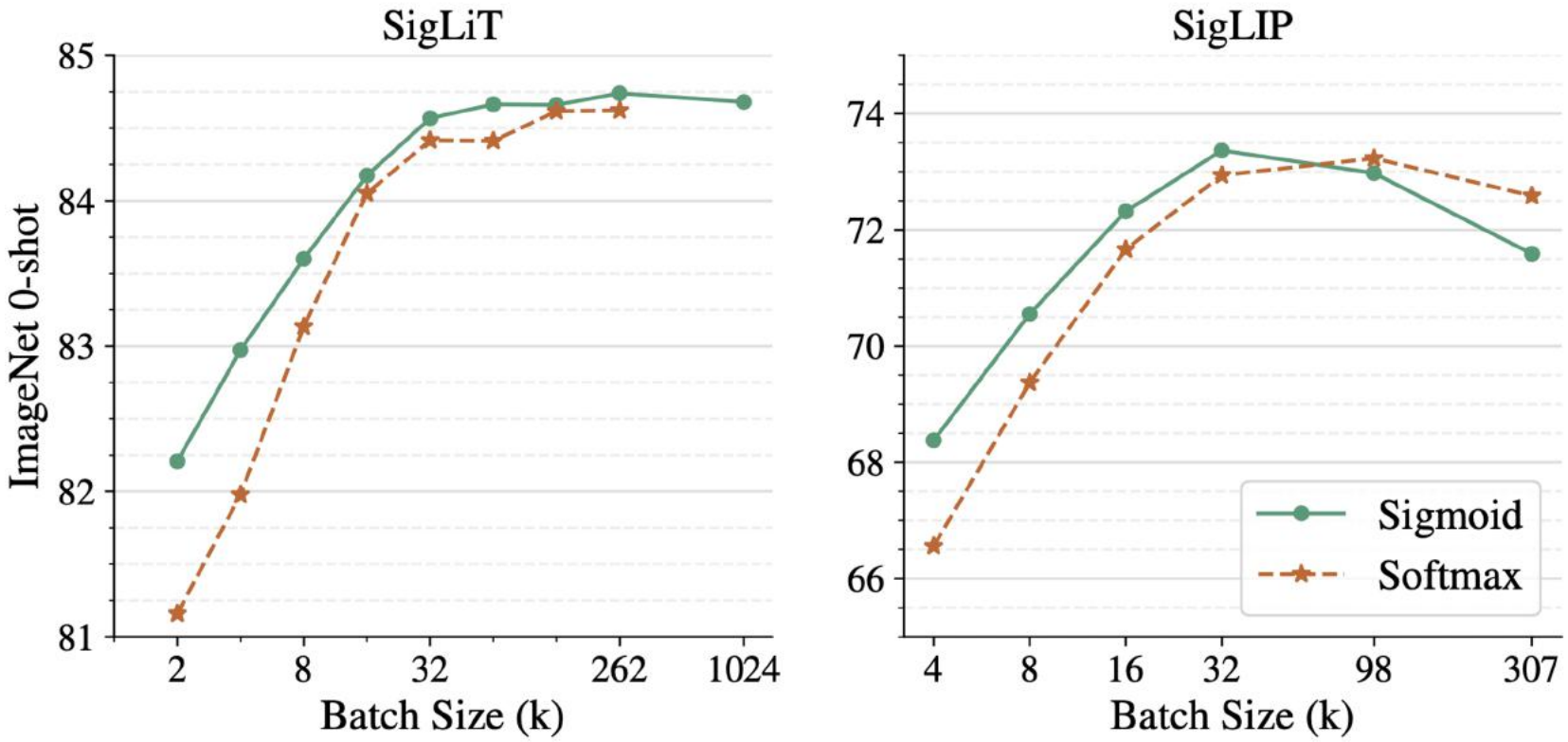


(d) This repeats till every image & text pair have interacted, e.g. device 1 has the loss of $I_{1:4}$ and $T_{1:12}$. A final cross-device sum brings everything together.

Zhai, Xiaohua, et al. "Sigmoid loss for language image pre-training." arXiv preprint arXiv:2303.15343 (2023).

SigLIP: Batch Size

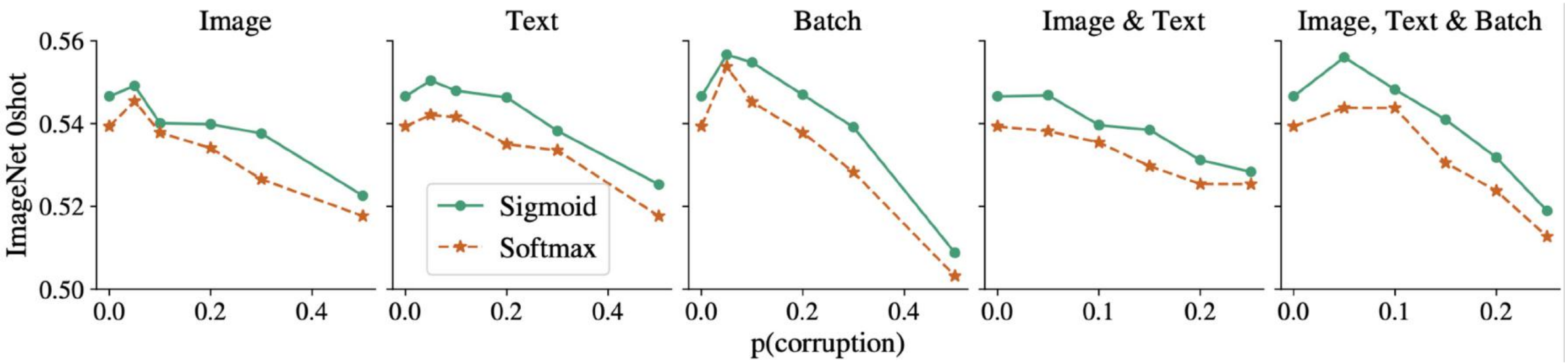
Apply sigmoid-based loss with *CLIP* and *LiT*:



SigLiT results: Sigmoid loss outperforms the softmax loss significantly with small batch sizes, and performs similarly at larger batch sizes.

SigLIP results: Both sigmoid loss and softmax loss saturate at a reasonable batch size, while the peak of the sigmoid loss comes earlier and slightly outperforms the peak of the softmax loss.

SigLIP: Label Noise Robustness



Sigmoid-training increases robustness to data noise.

Titles show the type of corruption applied, and x-axes show the probability with which they are applied. With increasing corruption severity, M-scale models trained with sigmoid loss for 3.6 billion examples retain superiority over corresponding softmax baseline.

Models trained with sigmoid loss are increasingly robust to all kinds of added noise.

Zhai, Xiaohua, et al. "Sigmoid loss for language image pre-training." arXiv preprint arXiv:2303.15343 (2023).

Summary

Contrastive loss

- ITC: sum of I2T and T2I InfoNCE loss to contrast paired text against others in the sampled batch (e.g., CLIP, BLIP)
- Sigmoid loss: binary classification of all pair combinations (SigLip)
- Binary classification to predict whether text-tag-image triplet contains the original tag or polluted tag (e.g., OSCAR)

Image-Text Matching (ITM) loss

- Binary classification to predict whether an image-text pair is matched or unmatched (e.g., BLIP)

Language Modeling (LM) loss

- A generative task to produce textual descriptions in an autoregressive manner given an image (e.g., BLIP)

Masked Language Modeling (MLM) loss

- Predict masked text tokens based on surrounding text tokens and image features (e.g., OSCAR, VinVL)

Uni-Encoder Family

OSCAR

- Feeding the sequence of texts, tags and image regions embeddings to BERT
- Semantic alignments between texts and images using object tags
- Image-Text Contrastive (ITC) loss, Masked Language Modeling (MLM) loss

VinVL

- Improving OSCAR with a more powerful object detection model
- 3-way contrastive loss
- Same MLM loss as OSCAR

Architecture	First published	Model Name	Image-text Pairs (M)	VQA (test-dev)	GQA (test-dev)	NLVR2 (dev)	I2T retrieval (COCO R@1)	T2I retrieval (COCO R@1)	Image Captioning (BLEU@4)	NoCaps (Valid CIDEr)	NoCaps (Valid SPICE)
Uni-encoder	2020/03	OSCAR	7	73.82	61.58	80.37	73.5 (FT)	57.5 (FT)	**41.7**	80.9	11.3
	2021/01	VinVL	9	**76.6**	**65.05**	**82.7**	**75.4 (FT)**	**58.8 (FT)**	41	**105.1**	**14.4**

Dual-Encoder Family

CLIP

- Introducing a learnable text encoder to encode free-form texts
- Image-Text Contrastive (ITC) loss

ALIGN

- Sacrificing quality to gain quantity – scaling up the corpus to 1.8B
- Extends dataset to multilingual to train *ALIGNmling*

SigLIP

- Changing softmax-based contrastive loss to sigmoid loss
- Advantages: memory efficient, fast, and numerically stable implementation

Architecture	First published	Model Name	Image-text Pairs (M)	I2T retrieval (COCO R@1)	T2I retrieval (COCO R@1)	I2T retrieval (Flickr R@1)	T2I retrieval (Flickr R@1)
Dual-encoder	2021/02	CLIP	400	58.4 (ZS)	37.8 (ZS)	88.0 (ZS)	68.7 (ZS)
	2021/02	ALIGN	1800	58.6 (ZS)	45.6 (ZS)	**88.6 (ZS)**	**75.7 (ZS)**
	2023/03	SigLIP	40000	**70.6 (ZS)**	**52.7 (ZS)**	--	--

Encoder-Decoder Family

BLIP

- Adding natural language generation capabilities
- ITC, LM, ITM loss
- Quality also matters – improving text quality by bootstrapping text

CoCa

- Minimalist design of BLIP, reducing the number of forward passes through transformer blocks
- ITC, LM loss
- Pre-trained with 4.8B images

Architecture	First published	Model Name	Image-text Pairs (M)	VQA (test-dev)	NLVR2 (dev)	I2T retrieval (Flickr R@1)	T2I retrieval (Flickr R@1)	Image Captioning (BLEU@4)	NoCaps (CIDEr)
Encoder-decoder	2022/01	BLIP	129	78.3	82.2	**96.7 (ZS)**	**86.7 (ZS)**	40.4	113.2 (ZS)
	2022/05	CoCa	4800	**82.3**	**86.1**	92.5 (ZS)	80.4 (ZS)	**40.9**	**122.4 (ZS)**

Multimodal LLMs

Motivation

The few-shot dream

Aspect of intelligence: ability to quickly learn tasks given short instructions

- Model Learning environment to make better use of data

We like the multimodal systems (vision and language) that achieve this property

Dominant computer vision paradigm:

Large-scale pretraining + Task specific fine-tuning

But current fine-tuning approaches require:

- Thousands of training samples
- Task specific hyperparameter tuning
- Significant computational resources

Can we train a multimodal model that has good performance in "few-shot" regime ?

Open-ended task abilities

Multimodal models like CLIP and ALIGN show good zero shot performance

But they are not flexible, they lack the ability to generate language

Inspiration from NLP: large language models like GPT-3 are flexible few-shot learners

Given a few examples of a task as a prompt + query input the language model generates a continuation to produce the predicted output

A key factor of their success is large-scale pretraining.

In principle: image/video understanding tasks (e.g. classification, captioning, question answering) are text prediction problems with visual input conditioning.

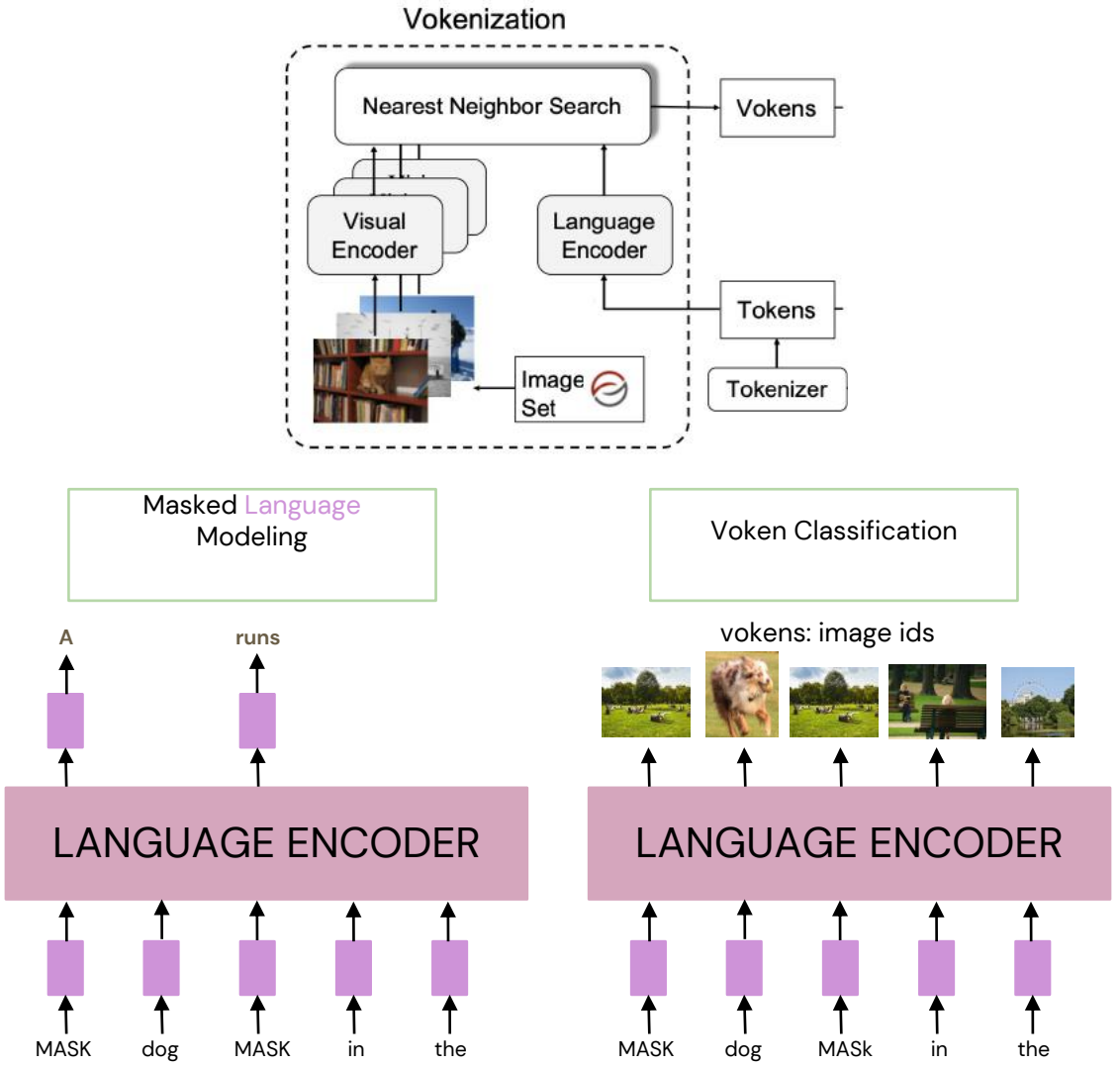
Can we learn a models capable of open-ended multimodal task via pretraining?

Language Encoders

A language modeling setup:

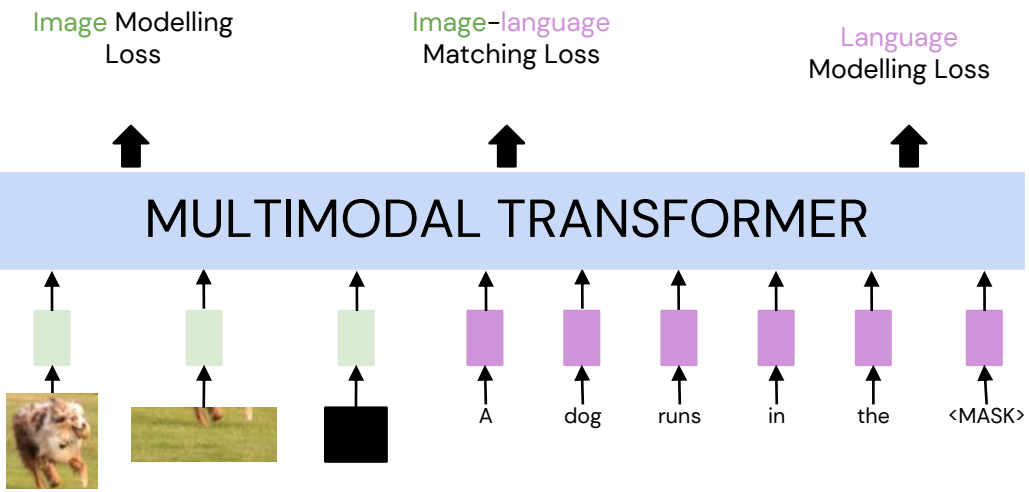
- Vokenization: map each language token to a visual token (voken)
  [Tan & Bansal, 2020]

Uses vision as supervision for language pretraining.

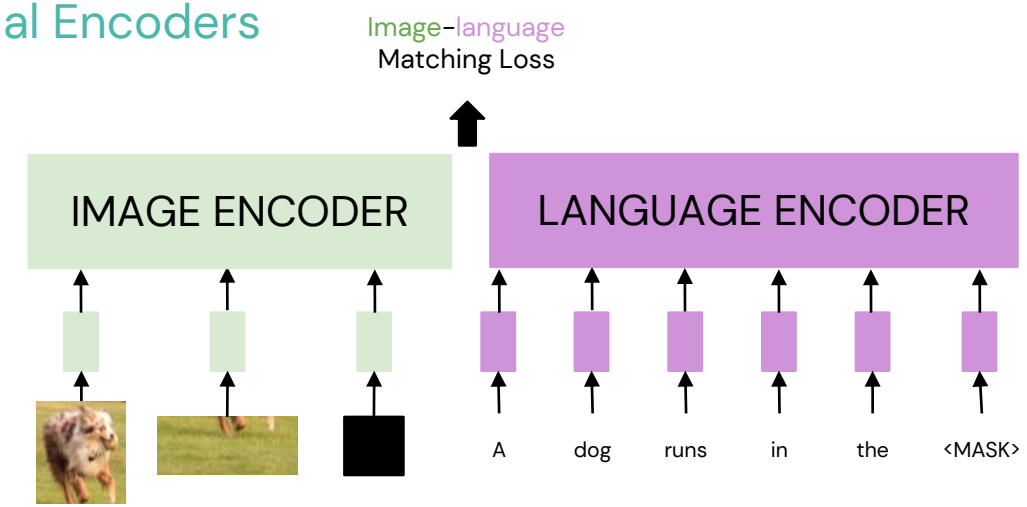


Slide credit: Aida & Lisa

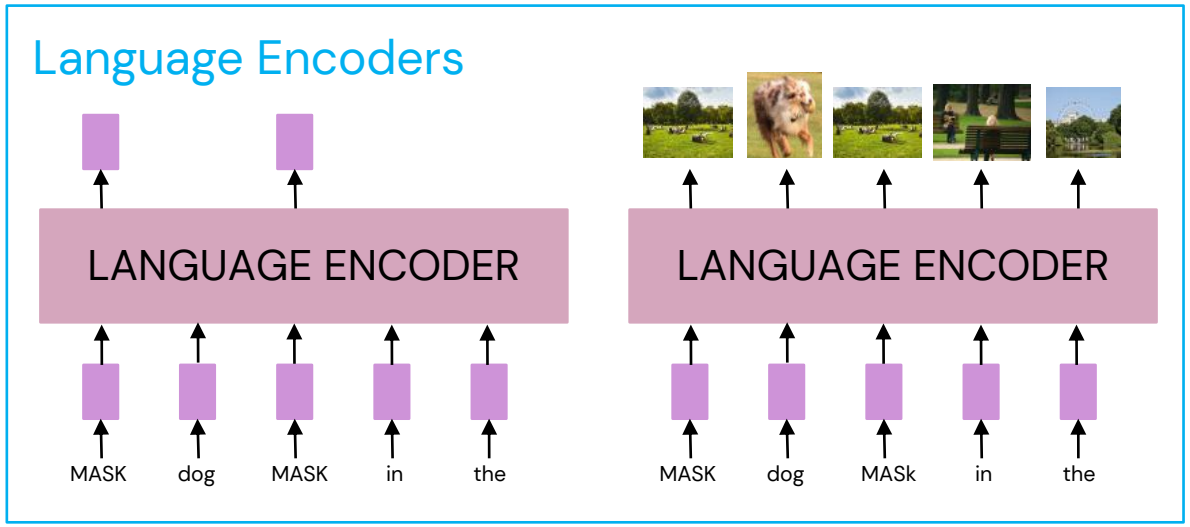
Joint Encoders



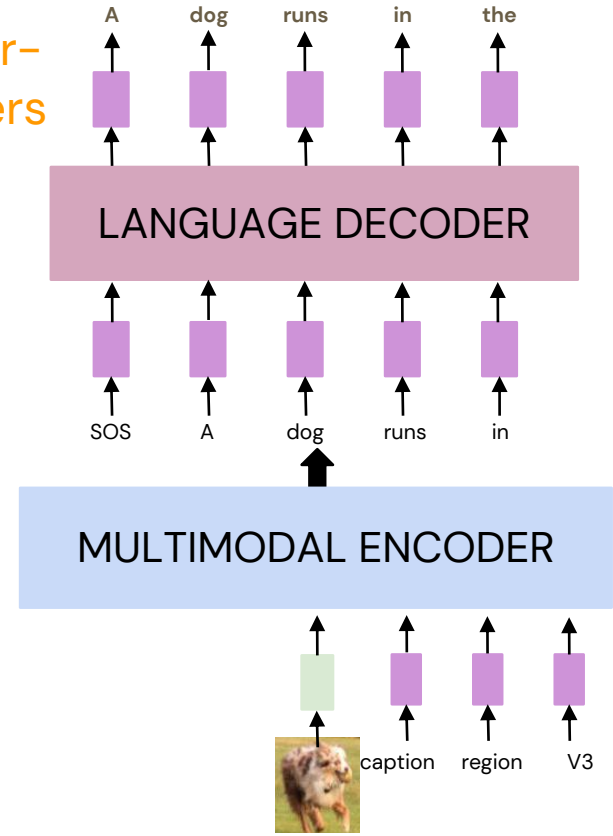
Dual Encoders



Language Encoders

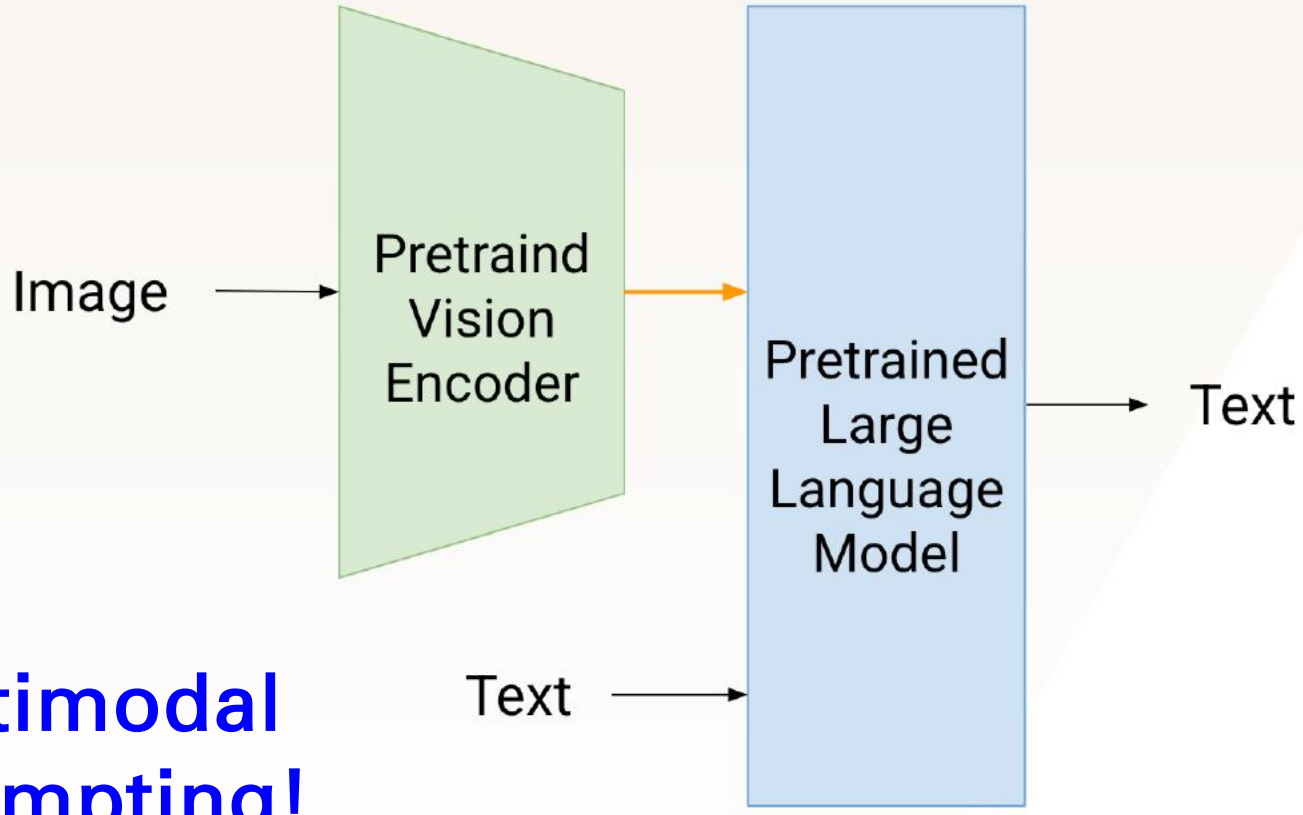


Encoder-Decoders



Slide credit: Aida & Lisa

Large Language Models Based Methods



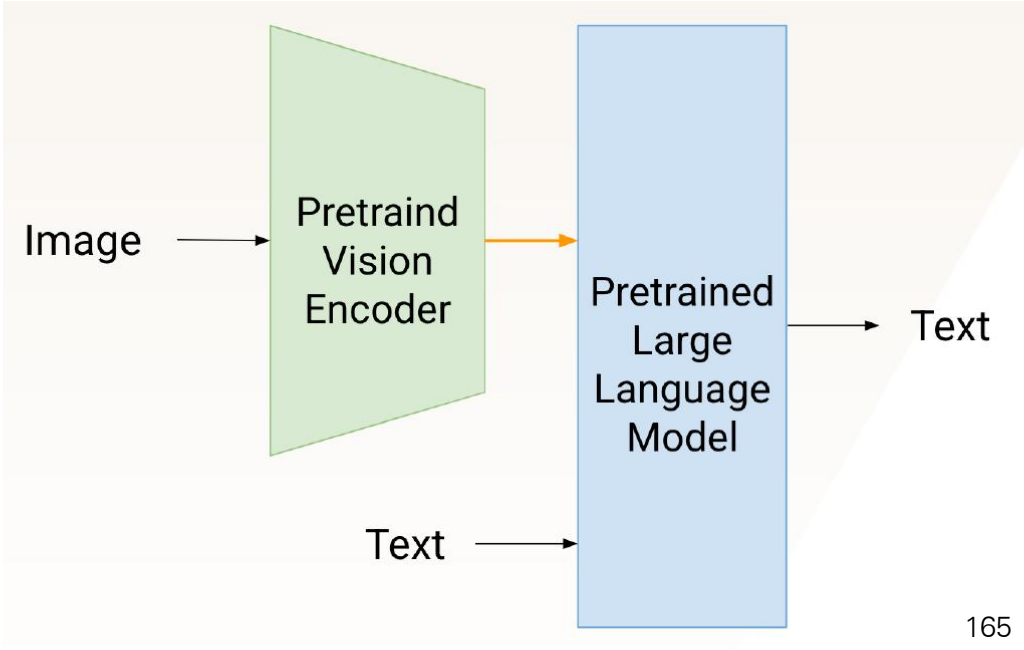
Enables multimodal few-shot prompting!

Different Types of Methods

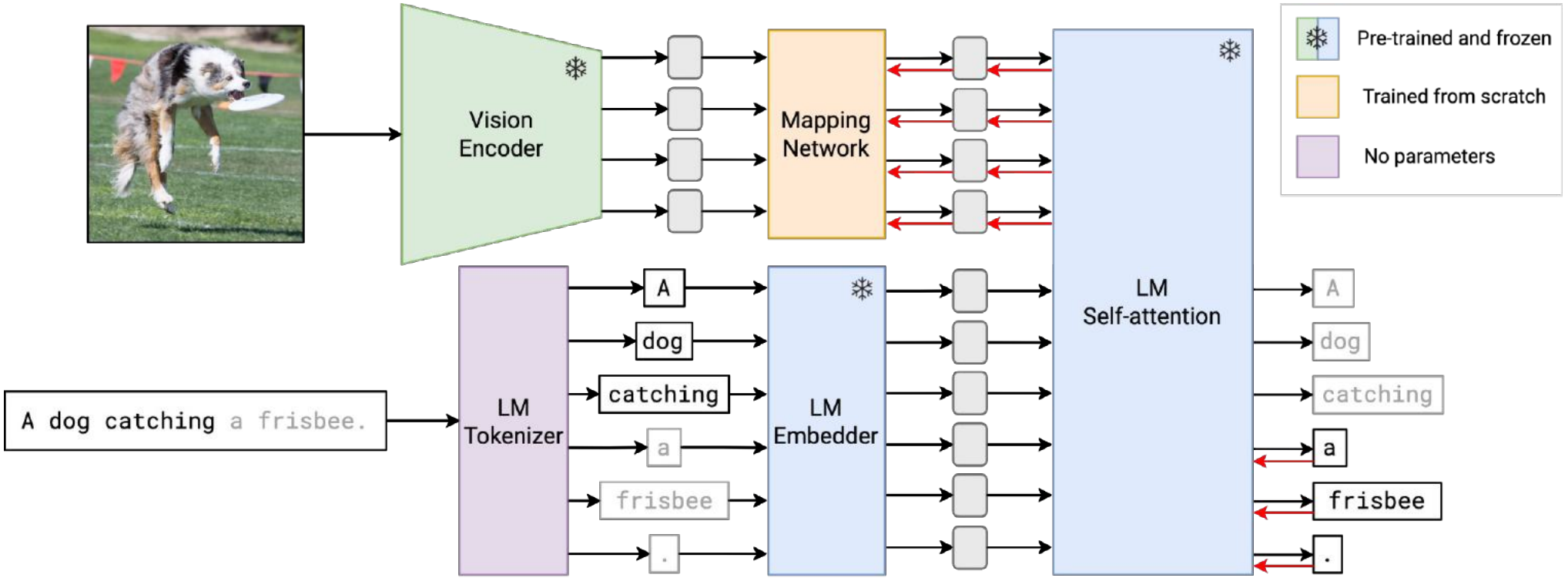
- Finetune the entire language model [Dai et al. 2022, Hao et al. 2022]
- Insert and train adapter layers in the language model [MAGMA, Flamingo]
- Learn vision encoder from scratch [Frozen]
- Only learn the mapping network [MAPL, BLIP-2]

Tradeoff:

- Performance vs. parameter count

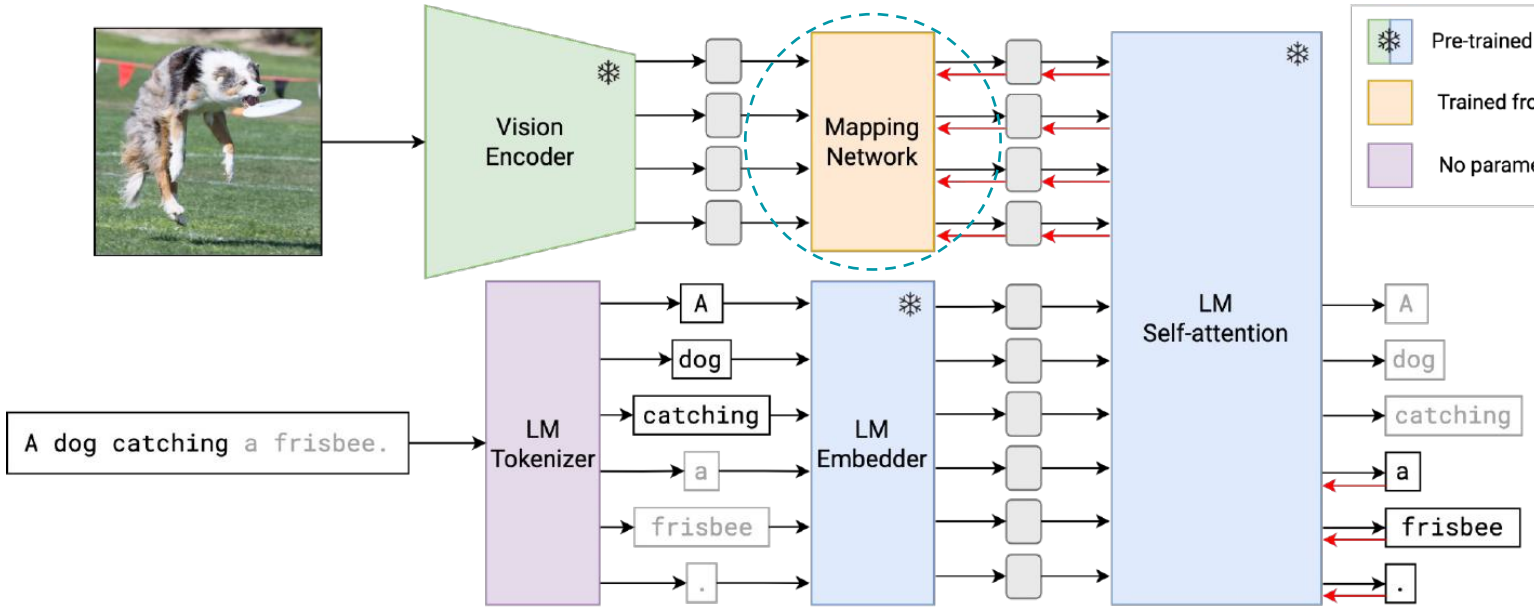


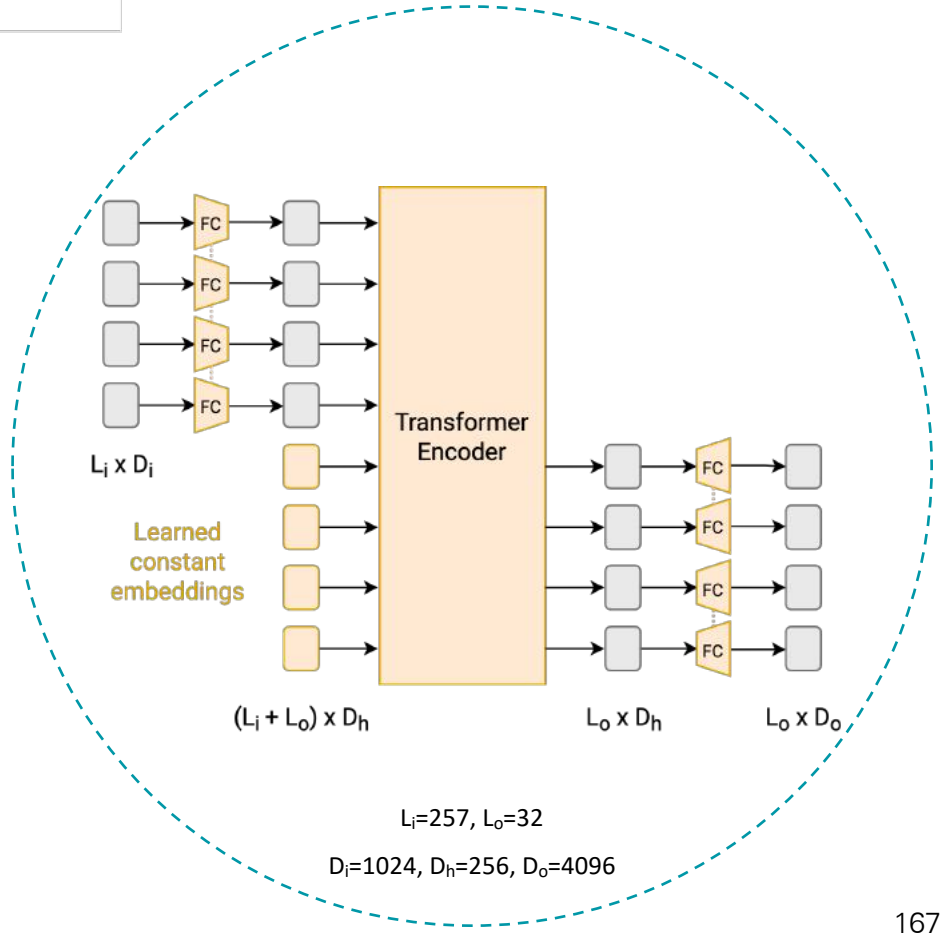
MAPL 🍁: Method



Slide credits: Oscar Mañas

MAPL 🍁: Method

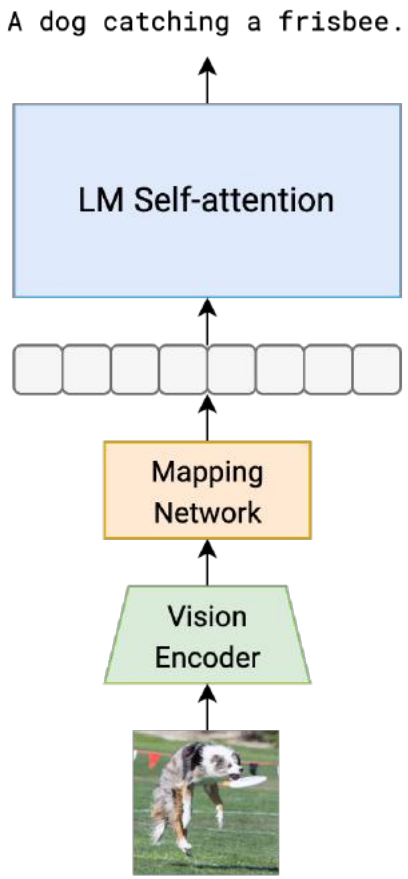




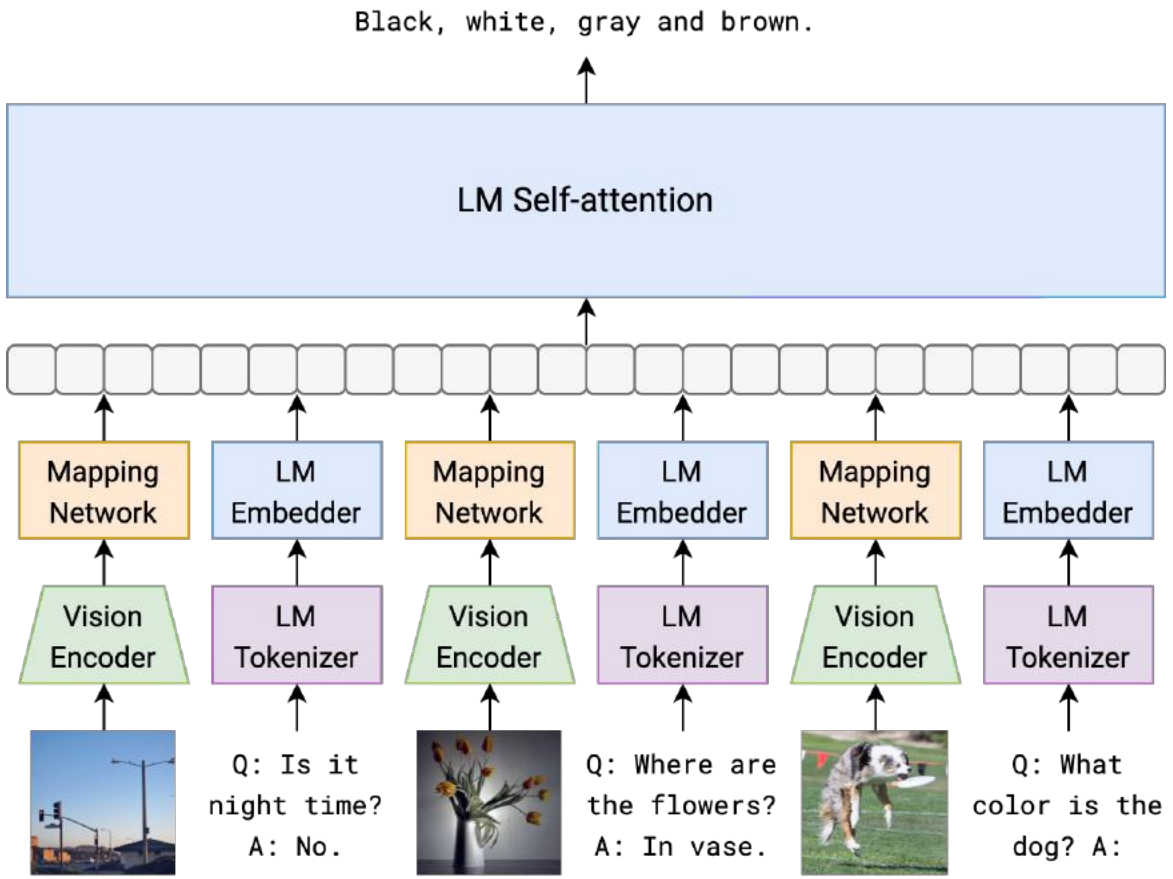
- Dimensionality bottleneck.
- Shared projection layers.
- Learned constant embeddings.

Slide credits: Oscar Mañas

MAPL 🍁: Interface at Inference Time



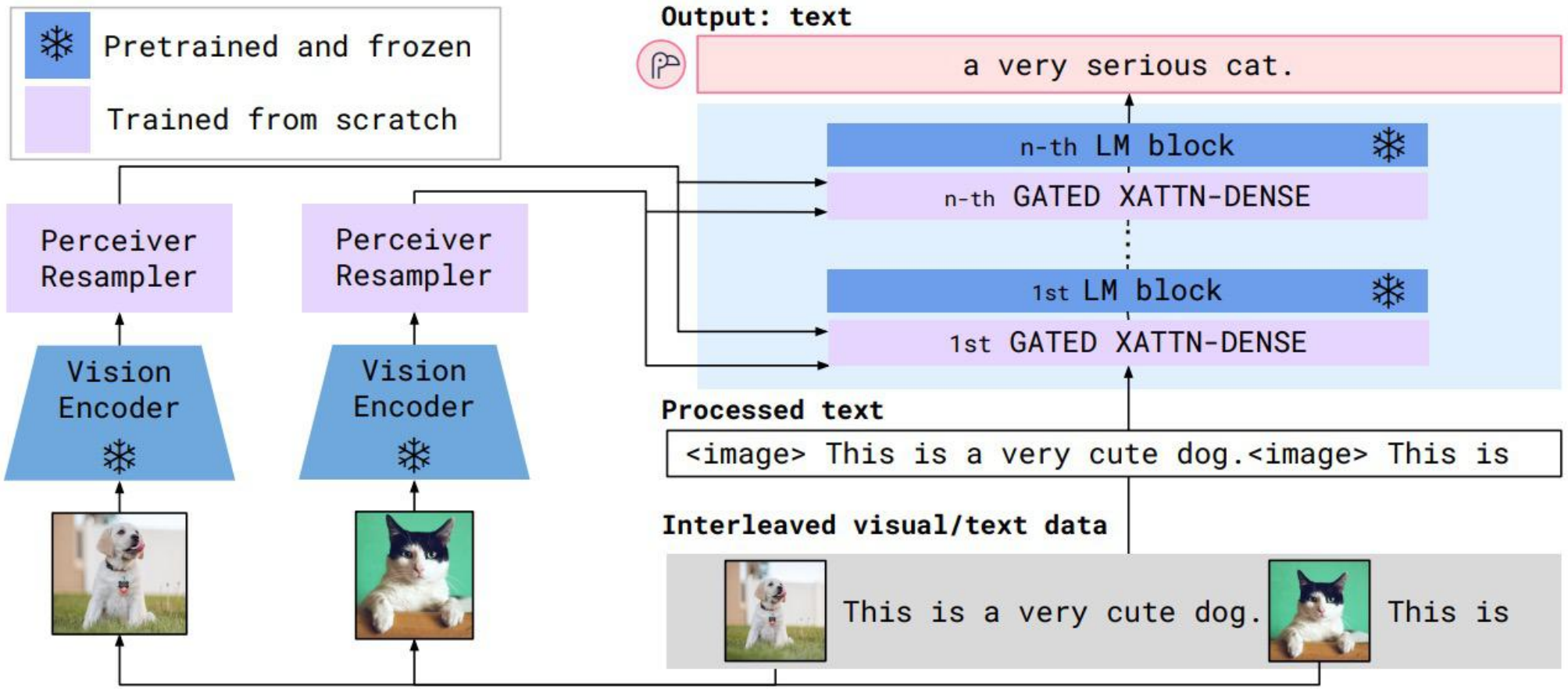
0-shot image captioning.



2-shot VQA.

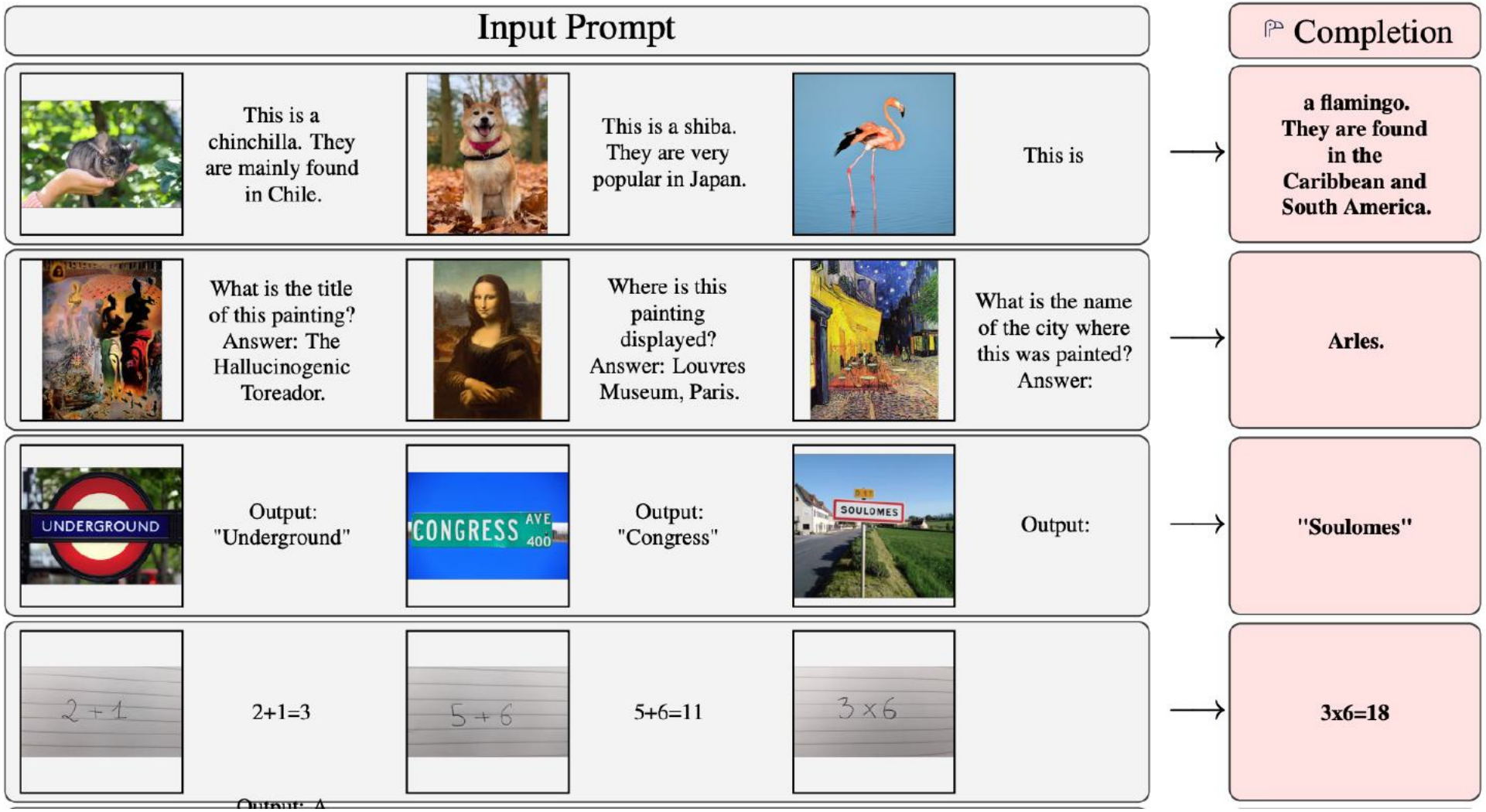
Slide credits: Oscar Mañas

Flamingo



https://arxiv.org/pdf/2204.14198.pdf

Flamingo



https://arxiv.org/pdf/2204.14198.pdf

Flamingo

https://arxiv.org/pdf/2204.14198.pdf

Flamingo: VQA

https://arxiv.org/pdf/2204.14198.pdf

Flamingo: Visual Dialogue

https://arxiv.org/pdf/2204.14198.pdf

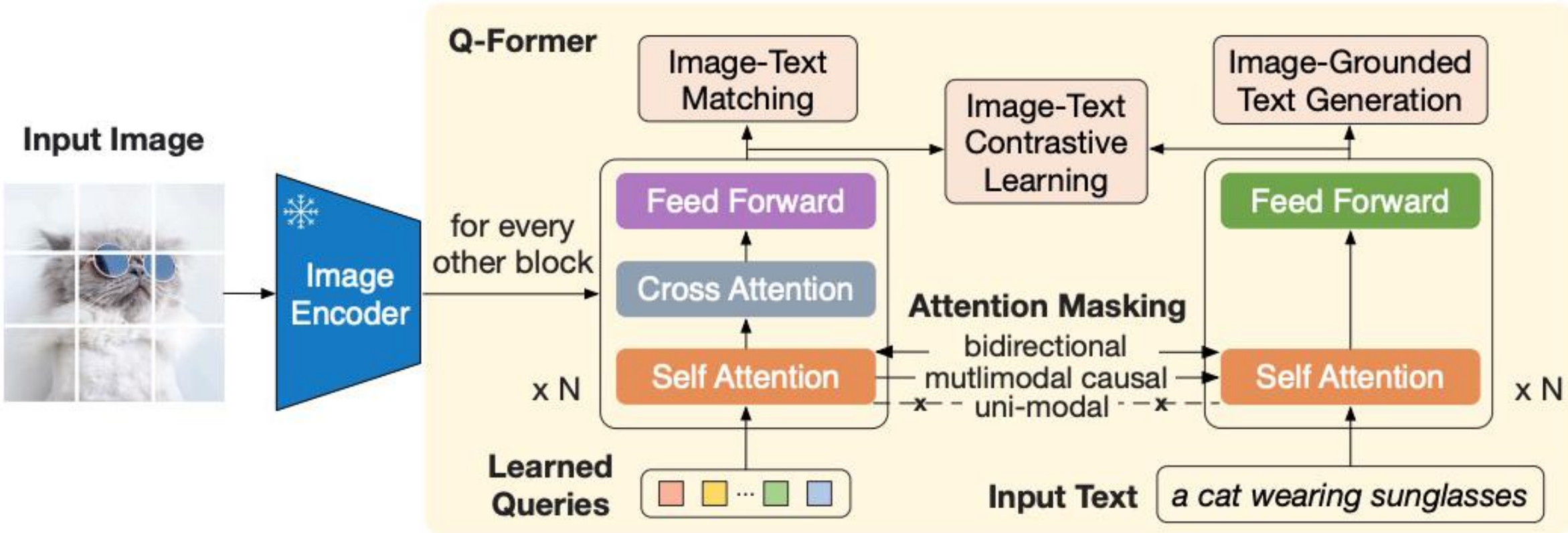
Flamingo: Video Prompt

https://arxiv.org/pdf/2204.14198.pdf

BLIP-2: Two Stage Pre-training

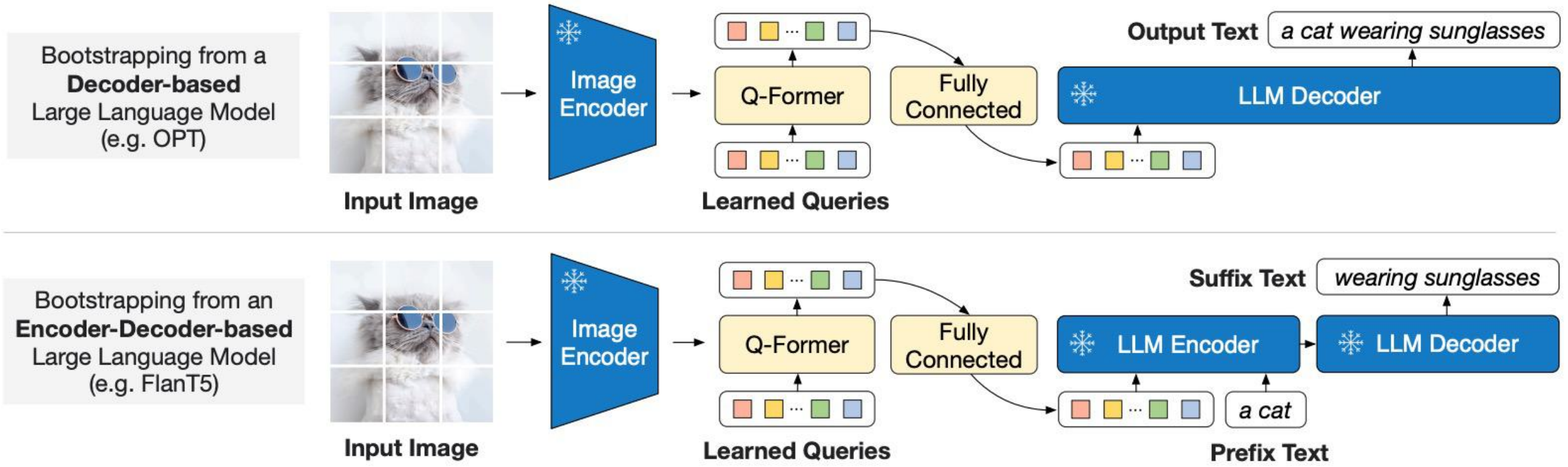
https://arxiv.org/pdf/2301.12597.pdf

BLIP-2: Stage 1



https://arxiv.org/pdf/2204.14198.pdf

BLIP-2: Stage 2



https://arxiv.org/pdf/2204.14198.pdf

GPT-4(V)

GPT-4 accepts prompts consisting of both images and text, which – parallel to the text-only setting – lets the user specify any vision or language task. Specifically, the model generates text outputs given inputs consisting of arbitrarily interlaced text and images. Over a range of domains – including documents with text and photographs, diagrams, or screenshots – GPT-4 exhibits similar capabilities as it does on text-only inputs. An example of GPT-4's visual input can be found in Table 3. The standard test-time techniques developed for language models (e.g. few-shot prompting, chain-of-thought, etc) are similarly effective when using both images and text - see Appendix G for examples.

https://arxiv.org/pdf/2303.08774.pdf

GPT-4(V)

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

Source: https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/

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.

More qualitative explorations:

https://arxiv.org/pdf/2309.17421.pdf

Gemini

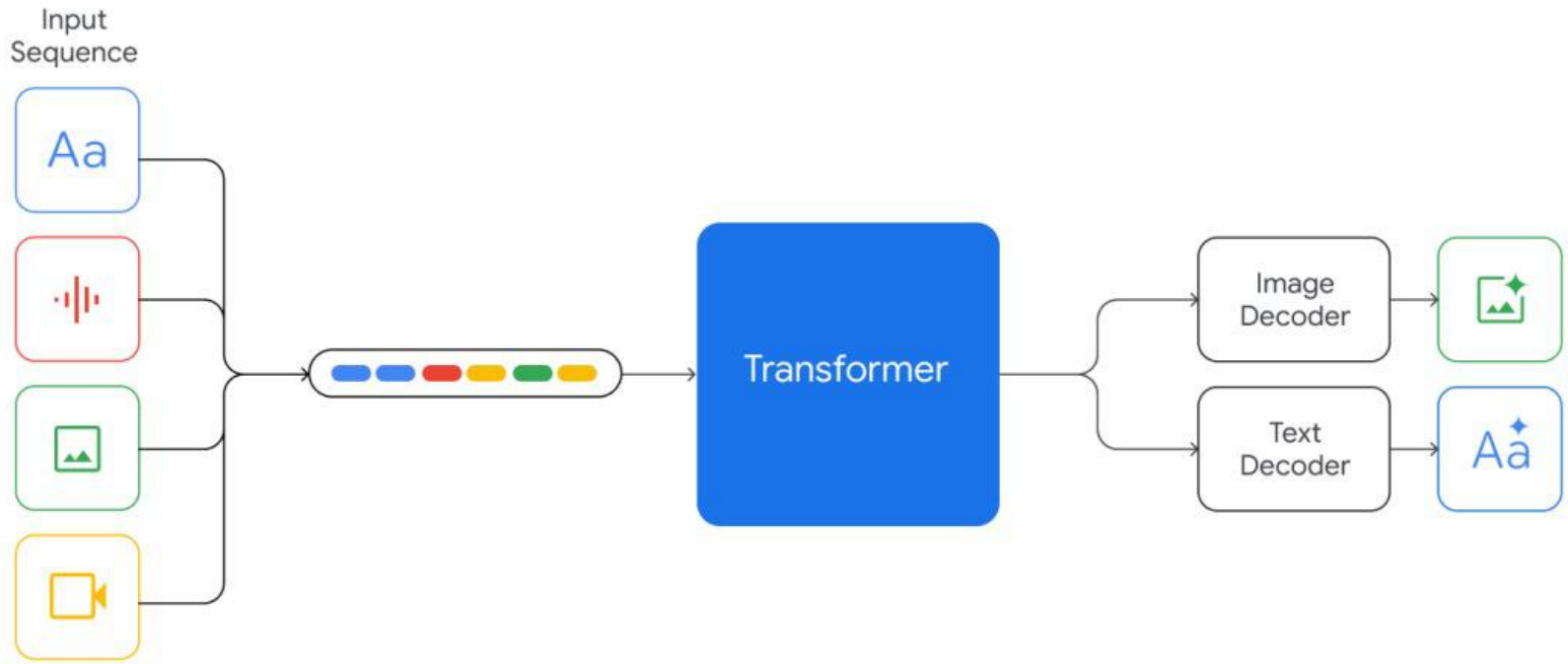


Figure 2 | Gemini supports interleaved sequences of text, image, audio, and video as inputs (illustrated by tokens of different colors in the input sequence). It can output responses with interleaved image and text.

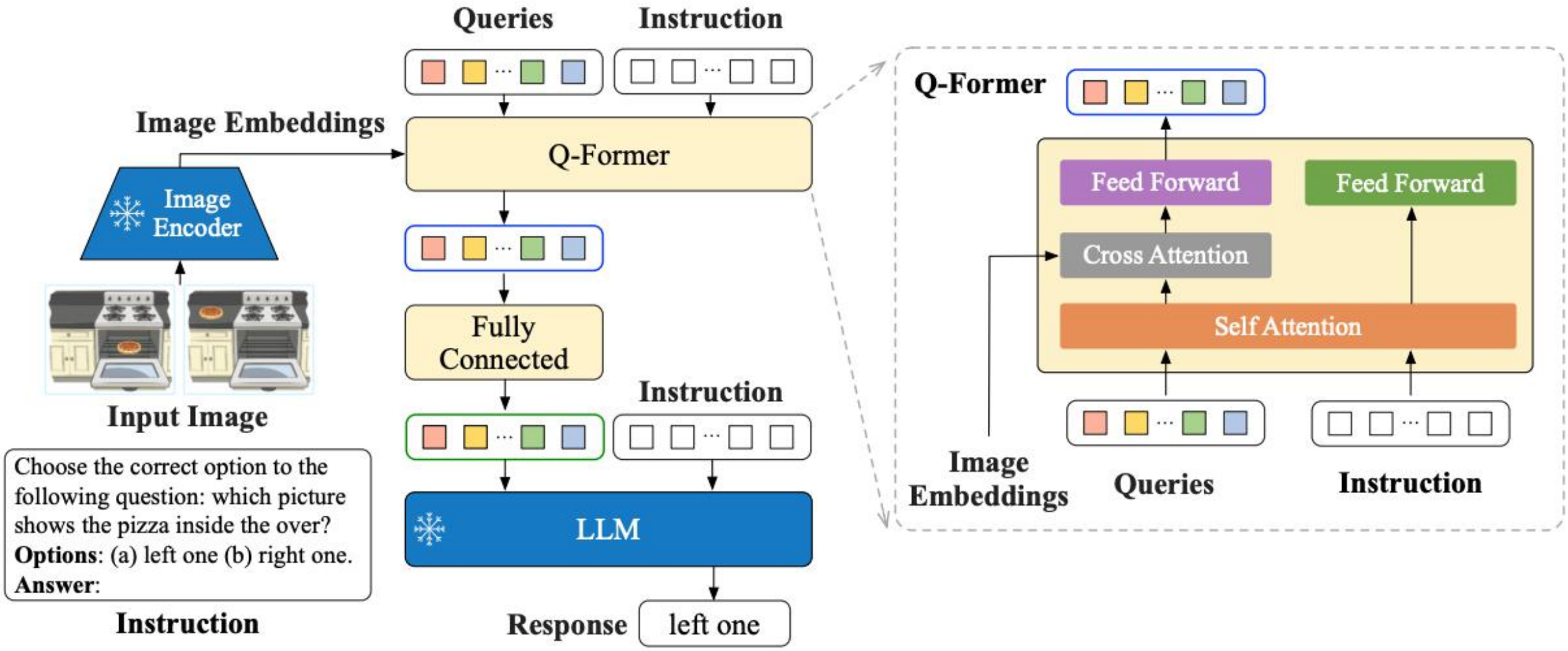
https://arxiv.org/pdf/2312.11805.pdf

InstructBLIP: Instruction Tuning

Task	Instruction Template
Image Captioning	<Image>A short image caption: <Image>A short image description: <Image>A photo of <Image>An image that shows <Image>Write a short description for the image. <Image>Write a description for the photo. <Image>Provide a description of what is presented in the photo. <Image>Briefly describe the content of the image. <Image>Can you briefly explain what you see in the image? <Image>Could you use a few words to describe what you perceive in the photo? <Image>Please provide a short depiction of the picture. <Image>Using language, provide a short account of the image. <Image>Use a few words to illustrate what is happening in the picture.
VQA	<Image>{Question} <Image>Question: {Question} <Image>{Question} A short answer to the question is <Image>Q: {Question} A: <Image>Question: {Question} Short answer: <Image>Given the image, answer the following question with no more than three words. {Question} <Image>Based on the image, respond to this question with a short answer: {Question}. Answer: <Image>Use the provided image to answer the question: {Question} Provide your answer as short as possible: <Image>What is the answer to the following question? "{Question}" <Image>The question "{Question}" can be answered using the image. A short answer is
VQG	<Image>Given the image, generate a question whose answer is: {Answer}. Question: <Image>Based on the image, provide a question with the answer: {Answer}. Question: <Image>Given the visual representation, create a question for which the answer is "{Answer}". <Image>From the image provided, craft a question that leads to the reply: {Answer}. Question: <Image>Considering the picture, come up with a question where the answer is: {Answer}. <Image>Taking the image into account, generate an question that has the answer: {Answer}. Question:

https://arxiv.org/pdf/2305.06500.pdf

InstructBLIP: Instruction Tuning



https://arxiv.org/pdf/2305.06500.pdf

LLaVA: Training only the projection layer

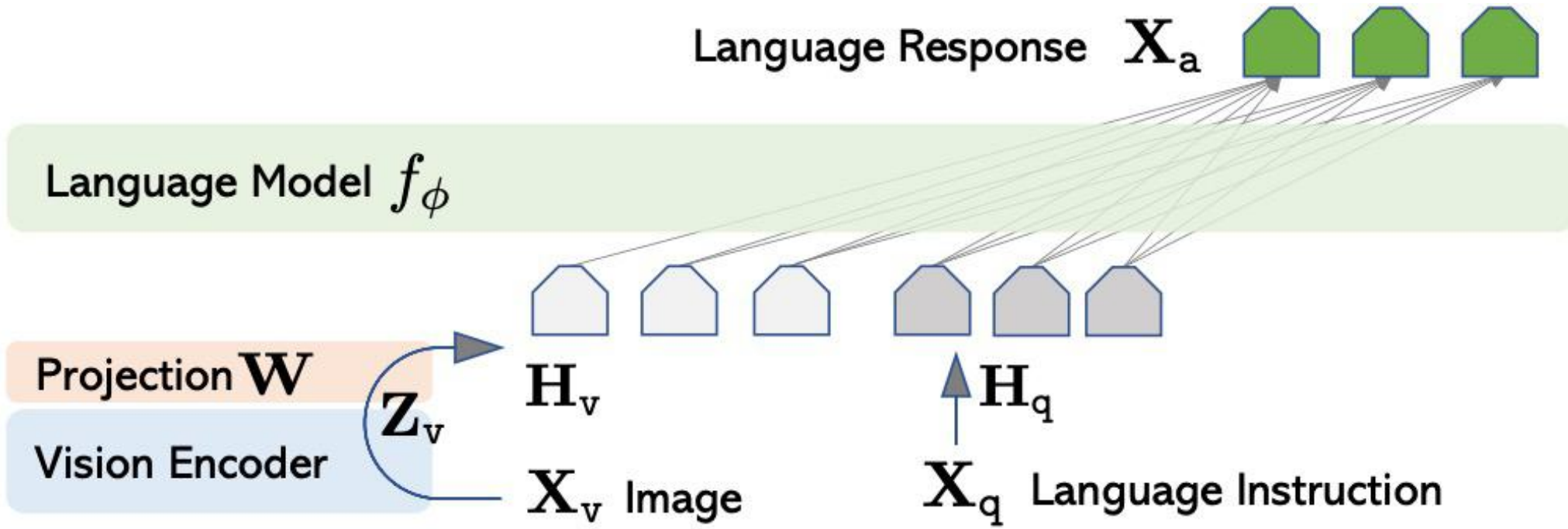


Figure 1: LLaVA network architecture.

https://arxiv.org/pdf/2304.08485.pdf

RLHF: Reinforcement Learning from Human Feedback

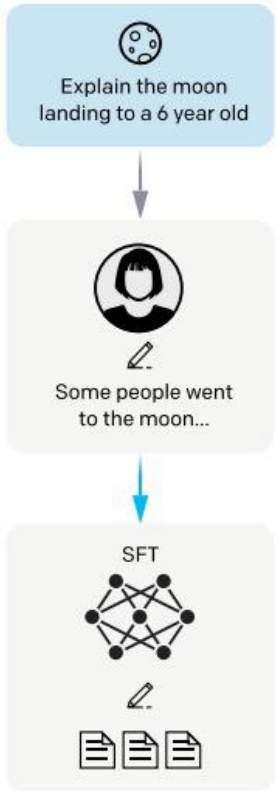
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



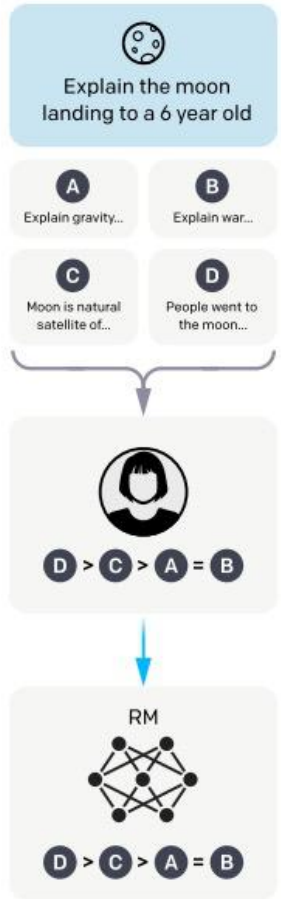
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

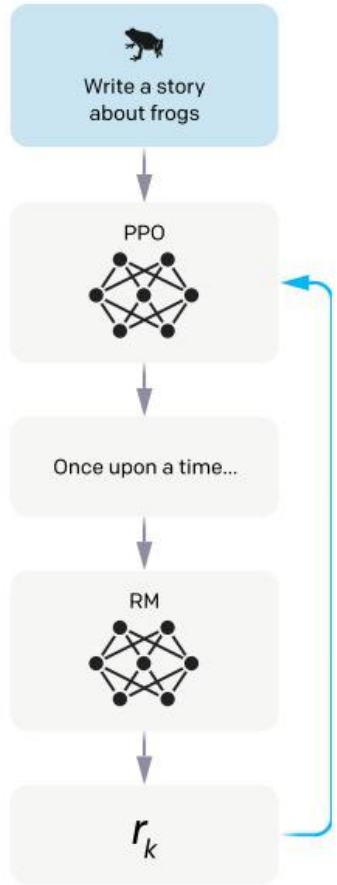
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



LLaVA-RLHF: RLHF applied to VL models

LLaVA-RLHF

Aligning Large Multimodal Models with Factually Augmented RLHF

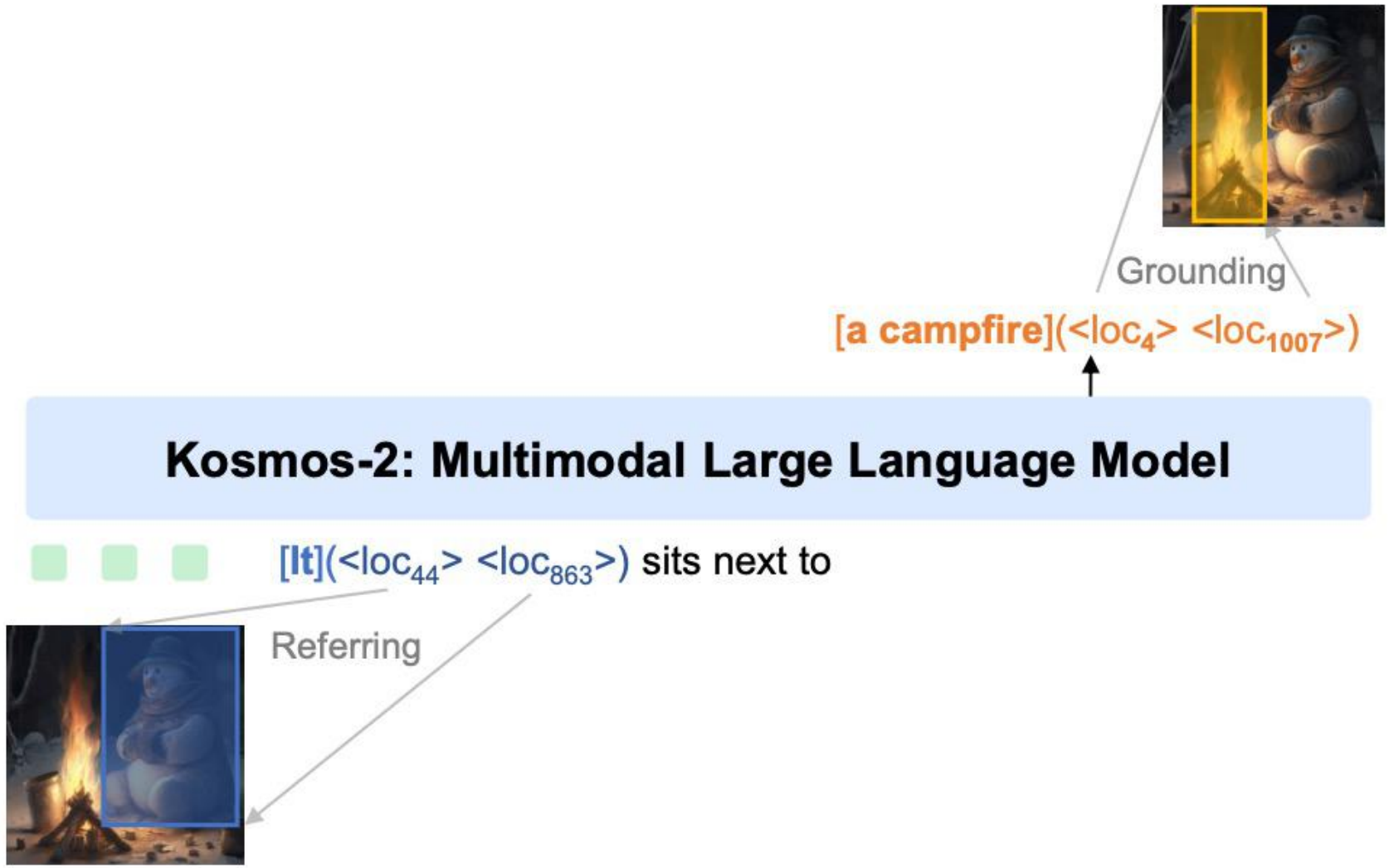
Zhiqing Sun*, Sheng Shen*, Shengcao Cao*,
Haotian Liu, Chunyuan Li, Yikang Shen,
Chuang Gan†, Liang-Yan Gui†, Yu-Xiong Wang†, Yiming Yang†, Kurt Keutzer†, Trevor Darrell†,

UC Berkeley CMU UIUC UW–Madison Microsoft Research MIT-IBM Watson AI Lab
*Equal Contribution, †Equal Advising

arXiv | Code | Demo | Dataset (RM) | Dataset (SFT) | MMHal-Bench | Model (13b) | Model (7b)

https://llava-rlhf.github.io/

Models that can do grounding of language into images



https://arxiv.org/pdf/2306.14824.pdf

PaLI: Motivation

Observation

Increasing the network capacity has been a successful trend for:

- Language models (T5, GPT-3)
- Vision models (ViT)
- Language-Vision models (Flamingo)

But

- But the scale distribution is not equitable in large-capacity language-vision models
- Language component is larger
- English-only

PaLI Design

- Reuse of large unimodal backbones
- Benefits from jointly scaling vision and language and a More balanced parameter share between language and vision components
- Enable knowledge-sharing b/w tasks by casting them to generalized VQA-like task
- Train on a new high-volume dataset of tens of billions image-text pairs across 100 languages

PaLI: What can it do?

Input: Generate the alt_text in EN
Output: A cellar filled with barrels of wine

Input: Generate the alt_text in EN
Output: a clock on a building that says 'lyvania' on it

Input: Generate the alt_text in EN
Output: Two helicopters are flying in the sky and one has a yellow stripe on the tail

Input: Generate the alt_text in FR
Output: Un arbre debout dans un champ avec un ciel violet

(A tree standing in a field with a purple sky)

Input: Generate the alt_text in TH
Output: ลา สี เทา เดิน ไป ตาม ถนน

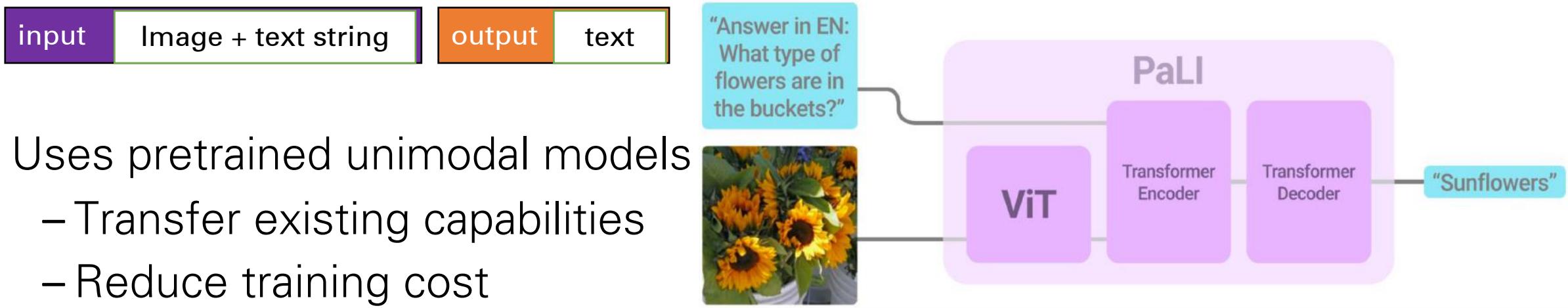
(A gray donkey walks down the street)

Input: Generate the alt_text in ZH
Output: 一辆 电动 汽车 停 在 充电 桩 上 。

(An electric car parking on a charging station)

PaLI: Architecture

- PaLI aims to do both unimodal and multimodal tasks
- Enable knowledge-sharing by casting all tasks to a generalized VQA-like task



- Uses pretrained unimodal models
  - Transfer existing capabilities
  - Reduce training cost
- Visual token are passed to encoder-decoder via cross-attention

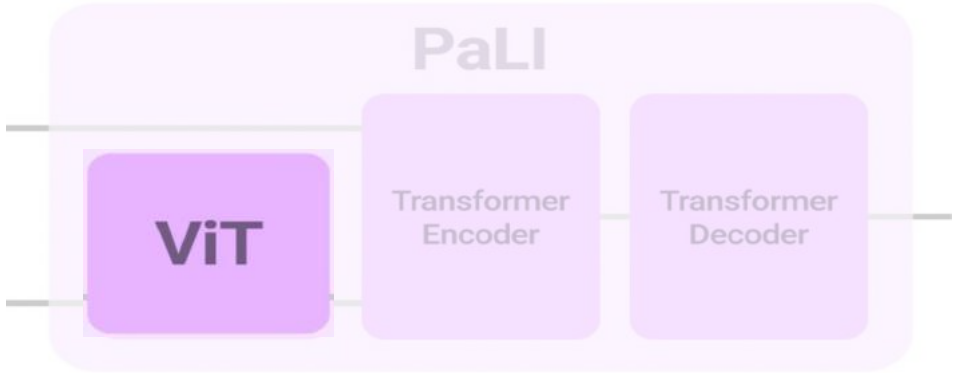
PaLI: Architecture

The visual component:
- Largest vanilla ViT called ViT-e
- 4B parameters
- Scaling up ViT on multimodal data not only does not saturate but has higher return (accuracy improvement per parameter/FLOP)

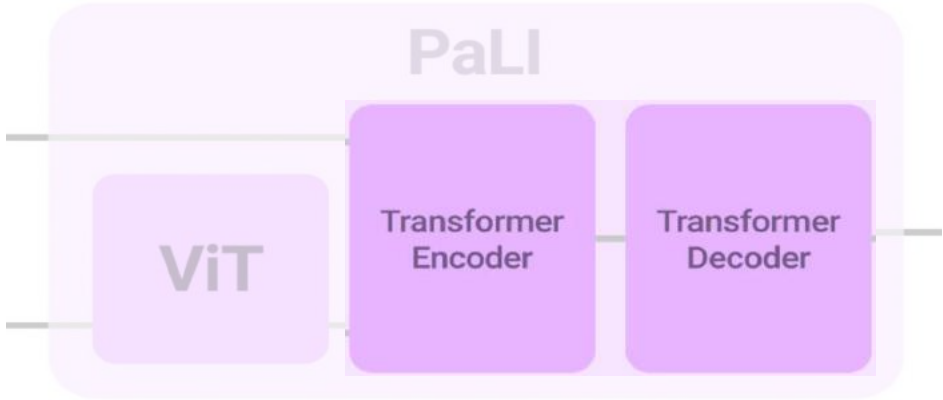
The language component:
- mT5 backbone
- Train on a mix of task to avoid catastrophic forgetting

The overall model:
(ViT-e or ViT-G) and (mT5-Large or mT5-XXL)

Visual Component



Language Component



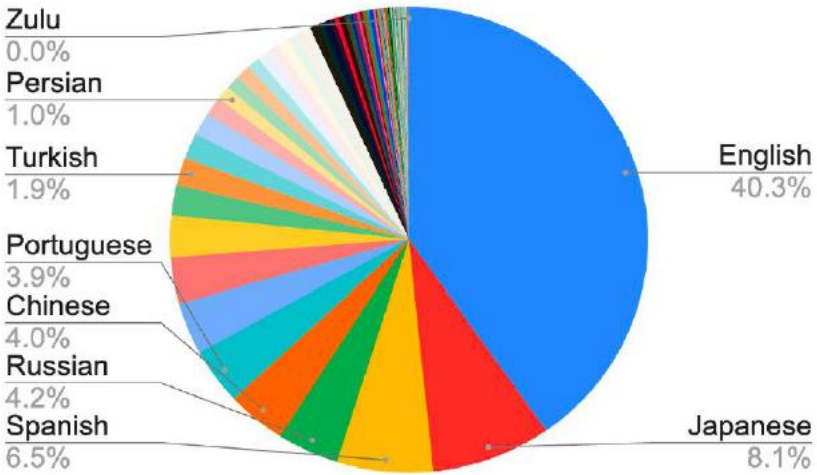
PaLI: Data

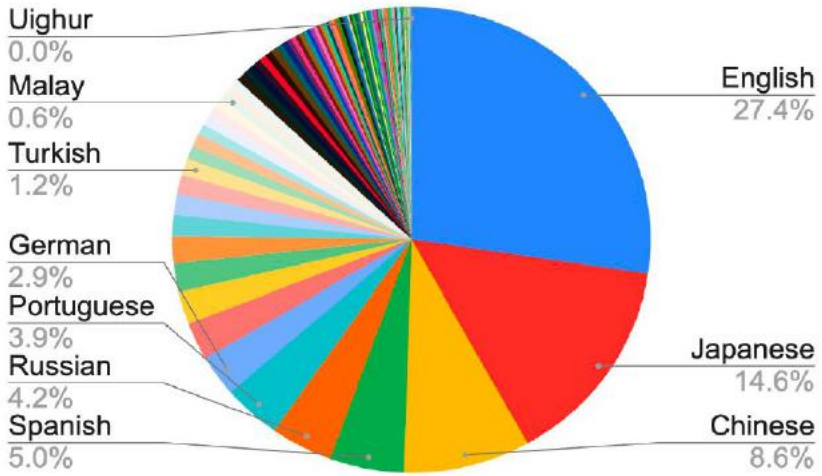
	English	French	Thai	Chinese
				
Alt-text	"free stock photo of matrix and sidekick"	"carte joyeux noël anges et étoiles"	"ทานตะวันเป็นดอกไม้ที่หันหน้าเข้าหาดวงอาทิตย์"	"太行山 脉 长治 太行山 大峡谷 林州 河北 平原 长城"
OCR	"card", "telecom", "5624"	"joyeux noël"	n/a	n/a

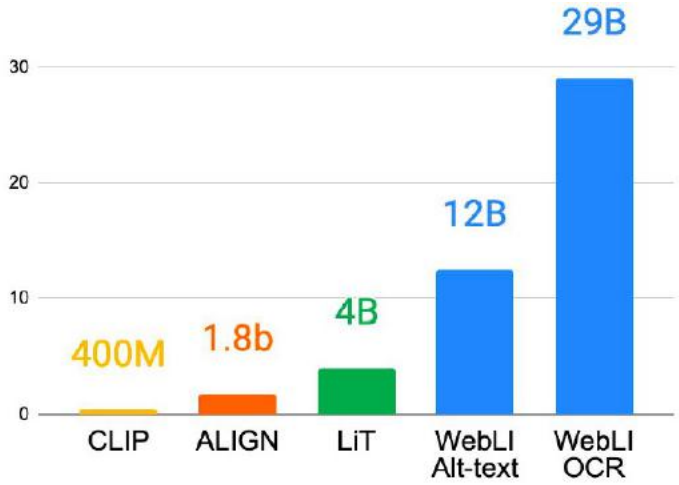
WebLI dataset:

- Build from image-text on public web Covering 109 languages
- 10B images, 12B alt-text, and 29B image-OCR pairs
- Only top 10% scoring, 1B, used for training

PaLI: Data







WebLI dataset:

- Build from image-text on public web Covering 109 languages
- 10B images, 12B alt-text, and 29B image-OCR pairs
- Only top 10% scoring, 1B, used for training

PaLI: Quantitative Results

	COCO	NoCaps		TextCaps		VizWiz-Cap	
Model	Karpathy-test	val	test	val	test	test-dev	test-std
LEMON (0.7B)	139.1	117.3	114.3	-	-	-	-
SimVLM	143.3	112.2	110.3	-	-	-	-
CoCa (2.1B)	143.6	122.4	120.6	-	-	-	-
GIT (0.7B)	144.8	125.5	123.4	143.7	138.2	113.1	114.4
GIT2 (5.1B)	145.0	**126.9**	**124.8**	148.6	145.0	119.4	120.8
OFA (0.9B)	145.3	-	-	-	-	-	-
Flamingo (80B)	138.1	-	-	-	-	-	-
BEiT-3 (1.9B)	147.6	-	-	-	-	-	-
PaLI-3B	145.4	121.1	-	143.6	-	117.2	-
PaLI-15B	146.2	121.2	-	150.1	-	121.7	-
PaLI-17B	**149.1**	**127.0**	**124.4**	**160.0**	**160.4**	**123.0**	**124.7**

PALI-3: Smaller, Faster, Stronger

- Motivation:
  - scaling of vision-language models (VLM) to tens and even hundreds of billions of parameters has shown ever-increasing performance
  - models at a smaller scale remain critical
  - present PaLI-3 with only 5B parameters

- 3 key components to achieve competitive performance:
  - contrastive pretraining of image encoder on web-scale image-text data
  - an improved dataset mixture inherited from PaLI
  - training at higher resolutions

- 2 dominant ways to pretrain image encoders are compared using the PaLI framework
  - classification pretraining using large weakly labeled datasets (JFT)
  - contrastive pretraining on web-scale noisy data

PALI-3: Architecture

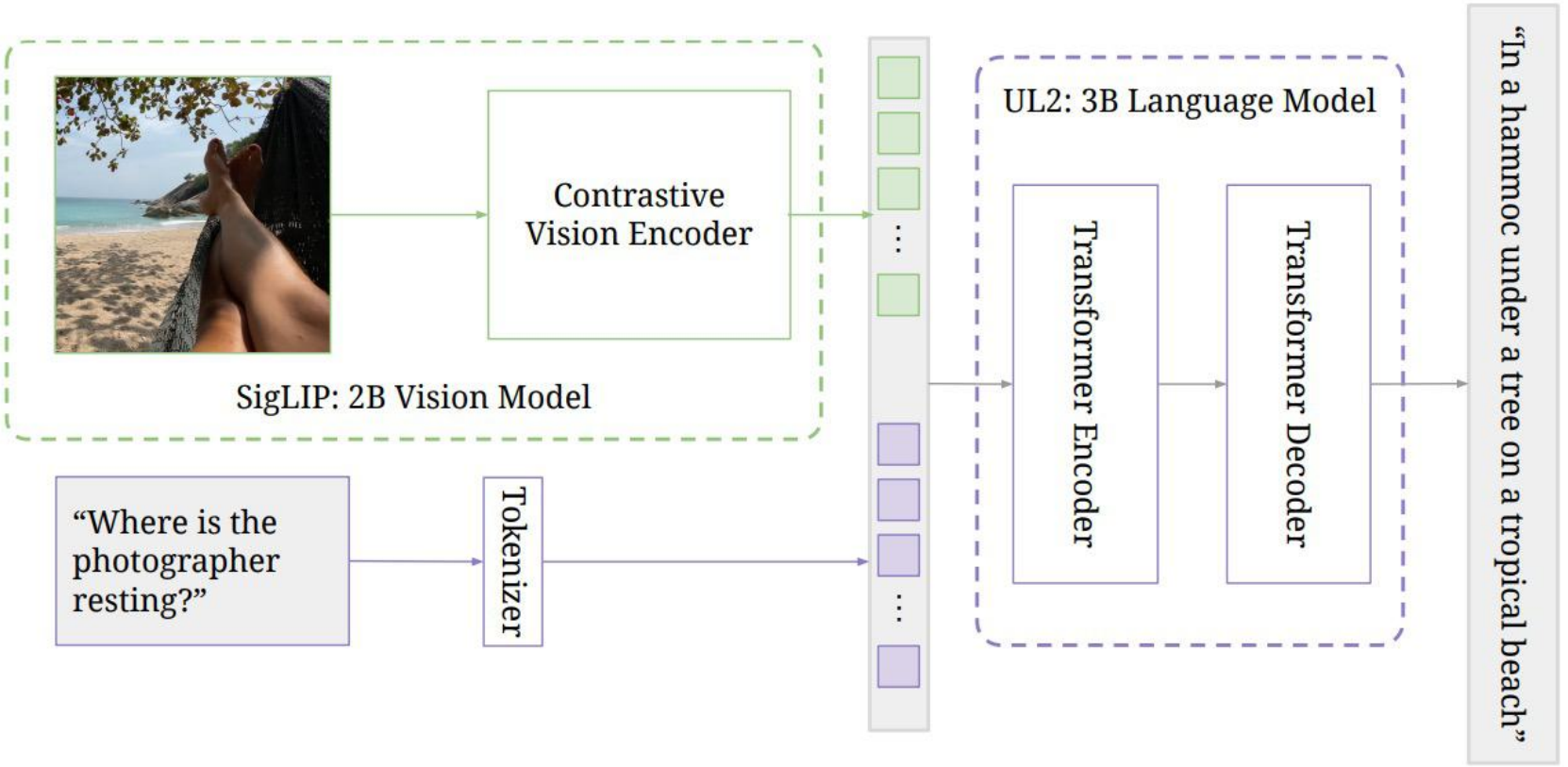


Figure 1: Overview of the PaLI-3 (5B) model: images are encoded into visual tokens individually by the contrastively pretrained 2B SigLIP vision model. Along with a query, these visual tokens are passed to an 3B encoder-decoder UL2 Transformer which produces the desired answer. In such a setup, a contrastively pretrained model provides significantly more useful tokens than one classification pretrained model as in previous PaLI models.

Unifying Language Learning Paradigms (UL2)

Motivation: why should the choice of the pre-trained LM depend on the downstream task?

Recap: Pre-training Objectives for Large Language Models

- Causal LM: use all previous time-steps as **inputs** to the model to predict the next token, which is the **target**

- prefixLM: use past tokens as **inputs**, but consume the inputs bidirectionally

- Span corruption: leverages all uncorrupted tokens from the past and futures as **inputs** for predicting the corrupted span (**targets**)

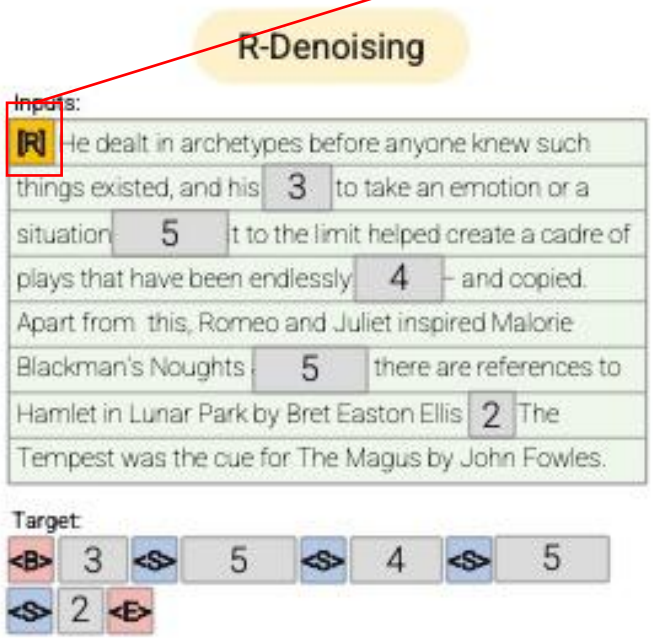
Can reduce one pre-training objective to another

UL2 pre-training objective

Extra paradigm token that helps for mode switching

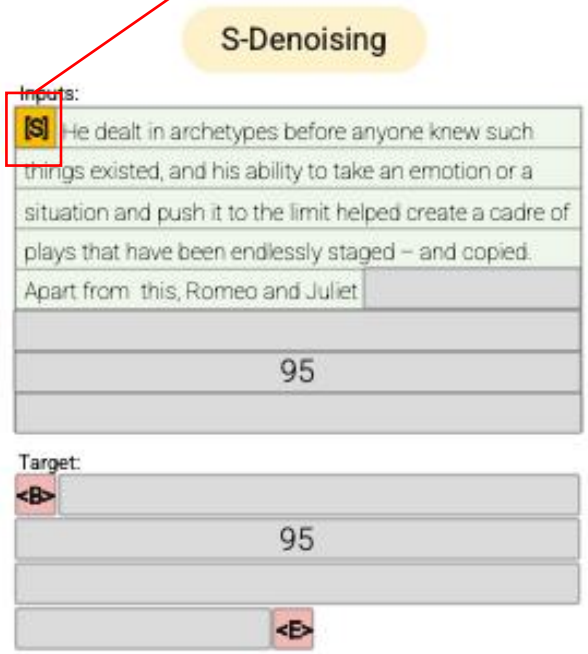
- **Mixture of Denoisers (MoD)**

R-Denoiser: span corruption



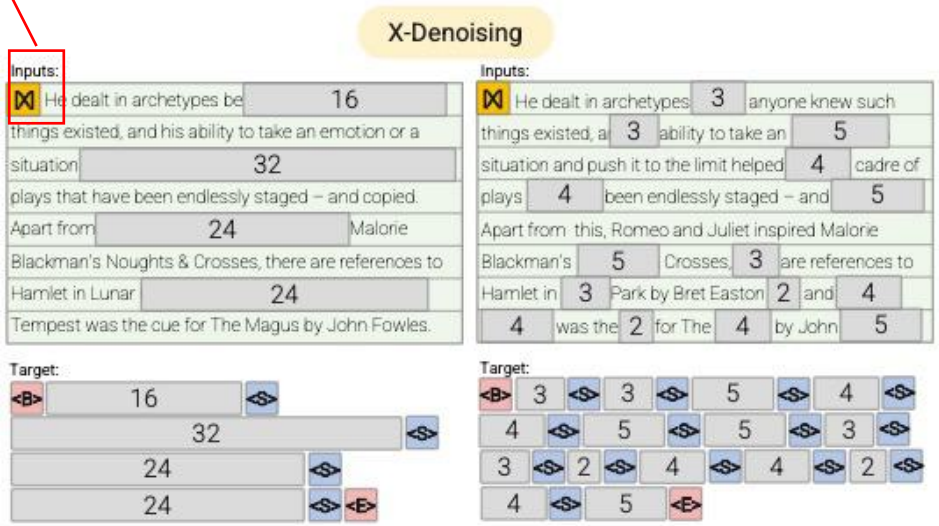
Spans are short and potentially useful to acquire knowledge instead of learning to generate fluent text

S-Denoiser: Prefix-LM



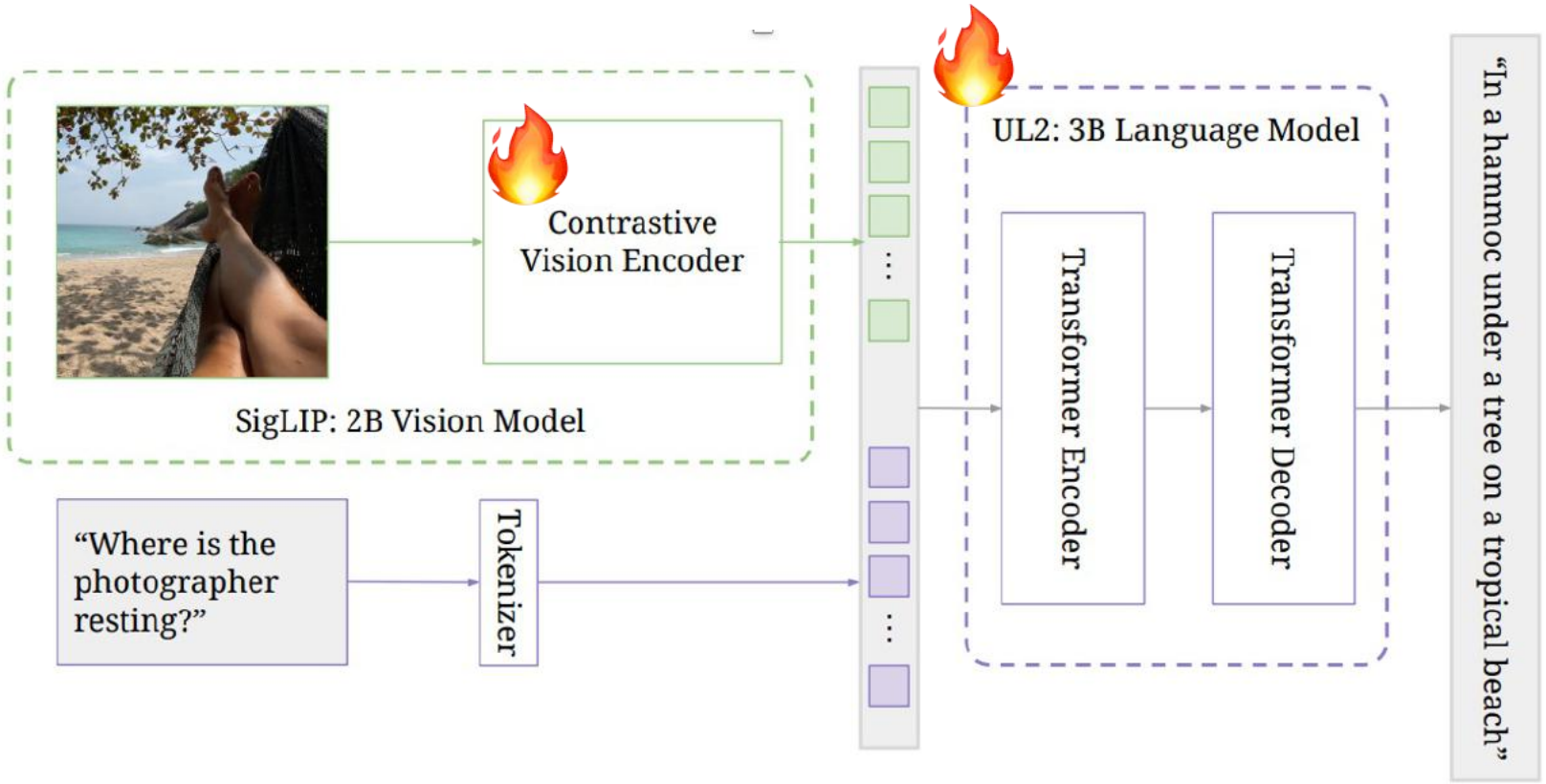
The context(prefix) retains a bidirectional receptive field

X-Denoiser: recover a larger part of the input, given a small part of it



Interpolation between regular span corruption and language model like objectives

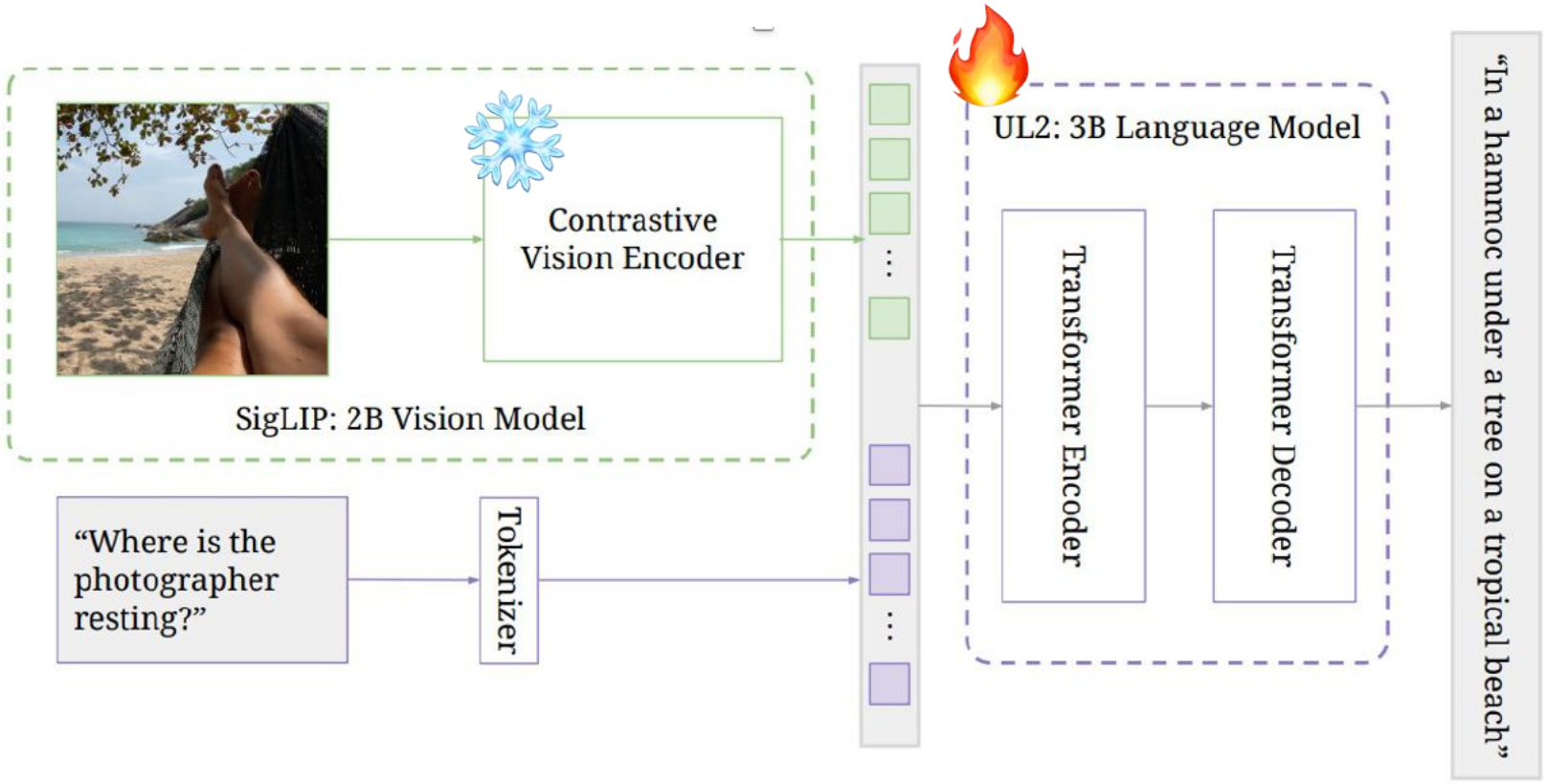
PALI-3: Stages of Training



Stage 0: Unimodal pretraining

- image encoder: pretrained contrastively on image-text pairs from the web, following the SigLIP training protocol
- text encoder-decoder: 3B UL2 model trained following the mixture of denoisers

PALI-3: Stages of Training

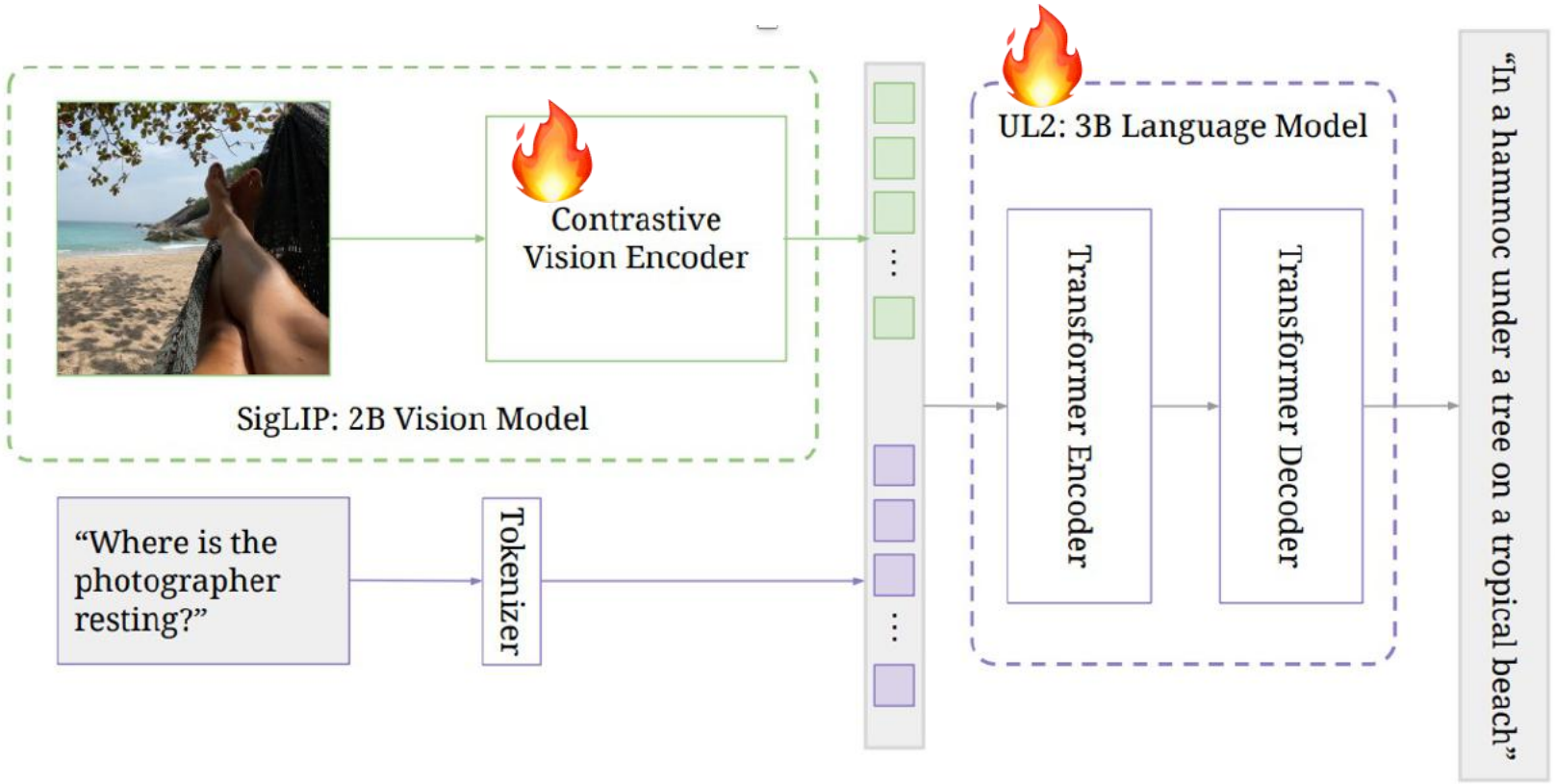


Stage 1: Multimodal training

- trained on a multimodal task and data mixture(retained from PALI) while keeping the image encoder frozen (224 x224 resolution)

Note: PALI-3 is not trained with task or data derived from video

PALI-3: Stages of Training



Stage 2: Resolution increase

- fine-tune the whole model (unfreeze the image encoder) to 812 x 812 and 1064 x 1064 resolution

PALI-3: Quantitative Results

- Comparison of different ViT models within the PaLI framework

Table 1: Performance comparison between contrastively pre-trained ("SigLIP") models and classification pre-trained ("Classif") ViT image encoders using the same PaLI setup, across a wide range of tasks. While linear classification few-shot probing (first column) suggests SigLIP encoders are worse across many tasks, when plugged into PaLI and transferred, they show clear improvements. On the most complicated and detailed image understanding tasks, SigLIP models outperform Classif models by a large margin. Captioning numbers are CIDEr scores, where XM3600 shows the English performance in the first column, and the average across other languages in the second column. RefCOCO numbers are mIoU scores (details in Section 4.3).

		Probe	Captioning			VQA			RefCOCO		
		8 tasks	COCO	XM3600		v2	OK	Text	val	+	g
G/14	Classif	88.1	139.9	94.5	44.7	76.7	57.2	31.9	51.6	43.5	43.4
	SigLIP	-2.5	+0.4	+1.6	+0.7	+0.8	+1.4	+18.7	+15.1	+19.1	+17.7
L/16	Classif	86.2	132.6	93.0	42.3	73.7	55.6	24.9	46.9	38.8	38.8
	SigLIP	-2.8	+3.2	+1.4	+1.4	+1.9	+1.9	+16.2	+17.4	+20.9	+20.1
B/16	Classif	83.7	127.7	91.7	40.7	72.3	54.7	22.5	46.3	38.1	38.4
	SigLIP	-2.6	+3.6	-2.0	-0.2	+1.4	+0.9	+13.3	+16.8	+19.6	+19.3

SigLIP models provide large gains for more "complicated" scene-text and spatial understanding tasks

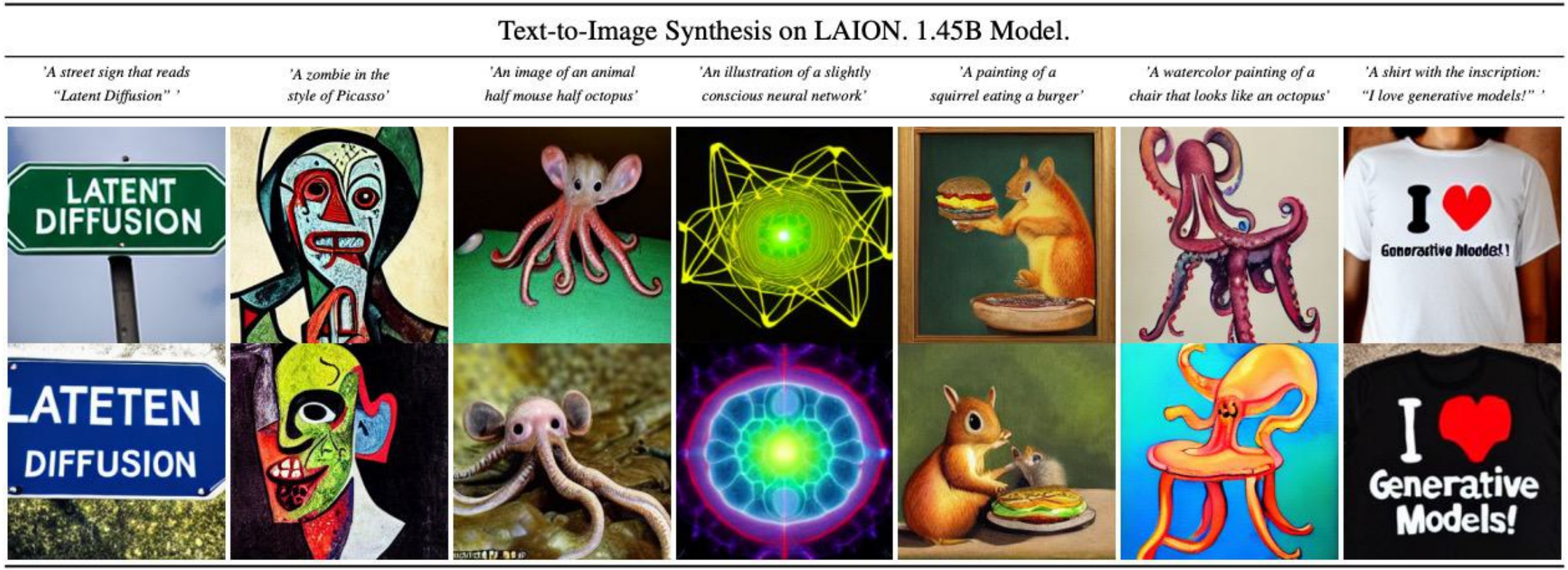
PALI-3: Quantitative Results

Table 5: Results for Video Captioning and Video-QA using up to 16 frames. †GIT2 directly optimizes the CIDEr metric. mPLUG-2 is Xu et al. (2023), PaLI-X is Chen et al. (2023a), GIT2 is Wang et al. (2022a), and Flamingo-32 is the 32-shot variant of Alayrac et al. (2022).

	MSR-VTT		Activity-Net		VATEX	SMIT	NExT-QA
Method	Caption	QA	Caption	QA	Caption	Caption	QA
Prior SOTA	**80.3** mPLUG-2	48.0 mPLUG-2	**54.9** PaLI-X	49.4 PaLI-X	94.0† GIT2	**43.5** PaLI-X	**38.3** Flamingo-32
PaLI-3	78.3	**49.3**	50.8	**51.2**	66.9	39.6	37.7

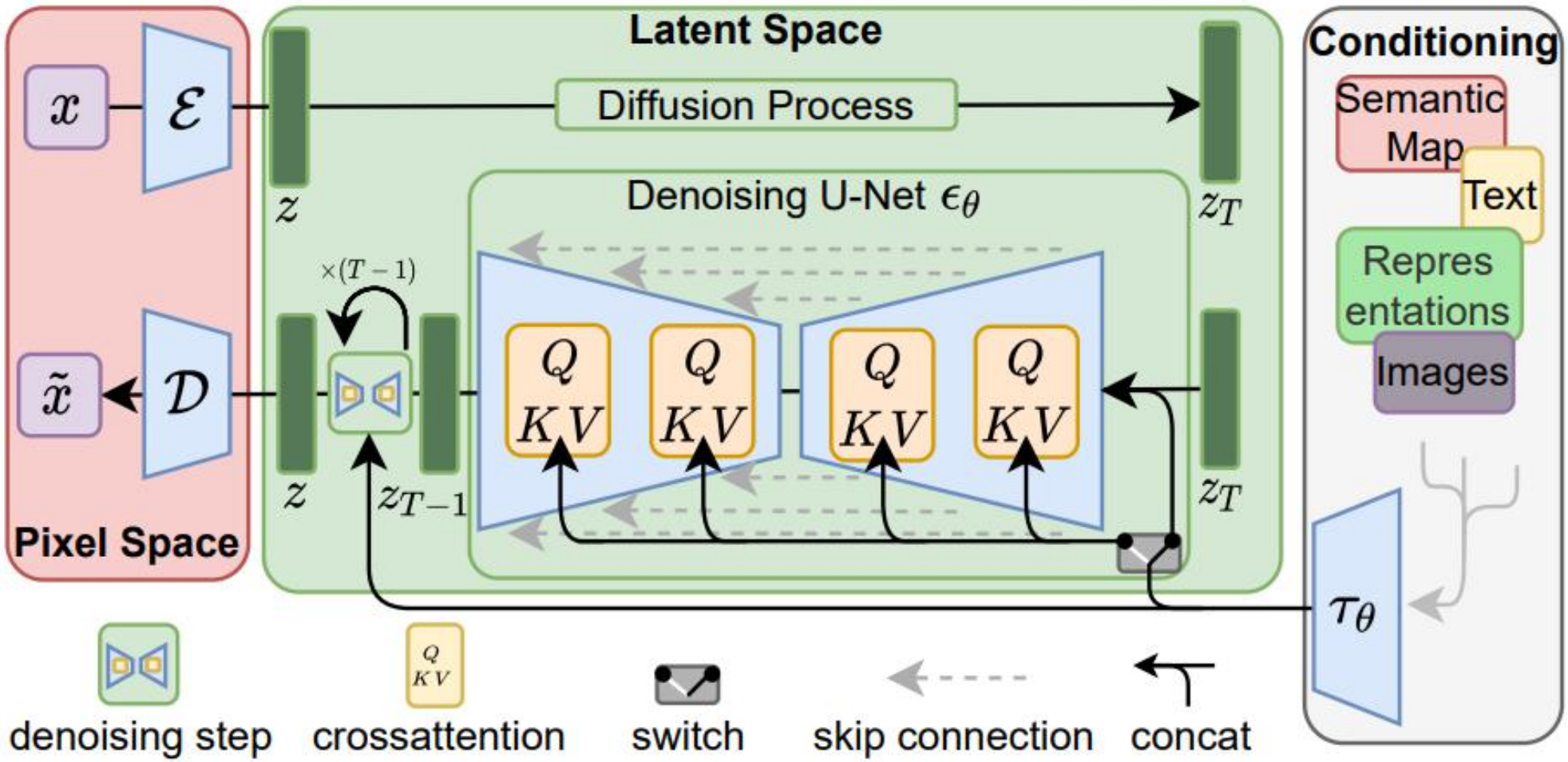
Video captioning and question answering

Stable Diffusion: Text to image generation models



https://arxiv.org/pdf/2112.10752.pdf

Stable Diffusion: Text to image generation models



https://arxiv.org/pdf/2112.10752.pdf

Imagen

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

Teddy bears swimming at the Olympics 400m Butterfly event.

A cute corgi lives in a house made out of sushi.

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

https://arxiv.org/pdf/2205.11487.pdf

Parti

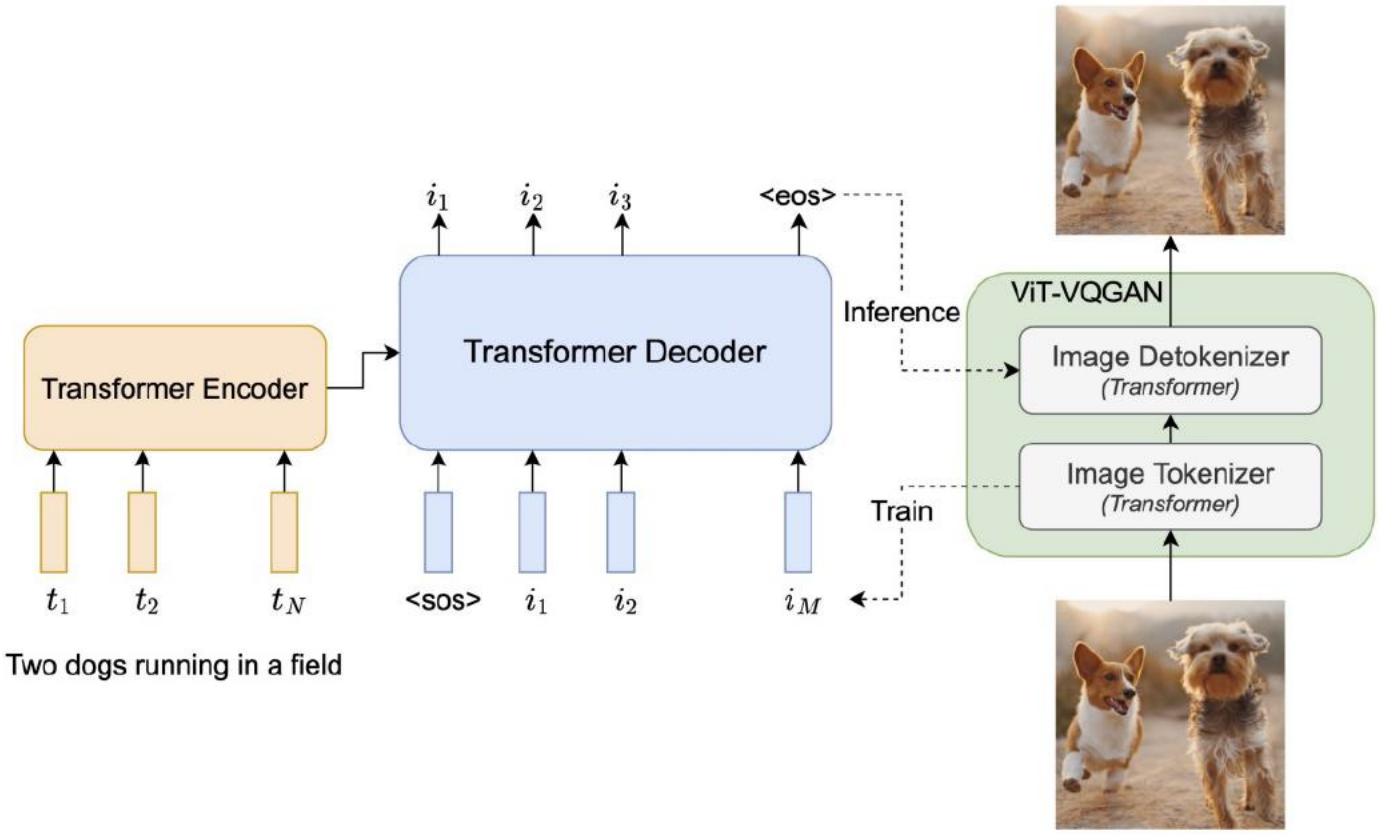
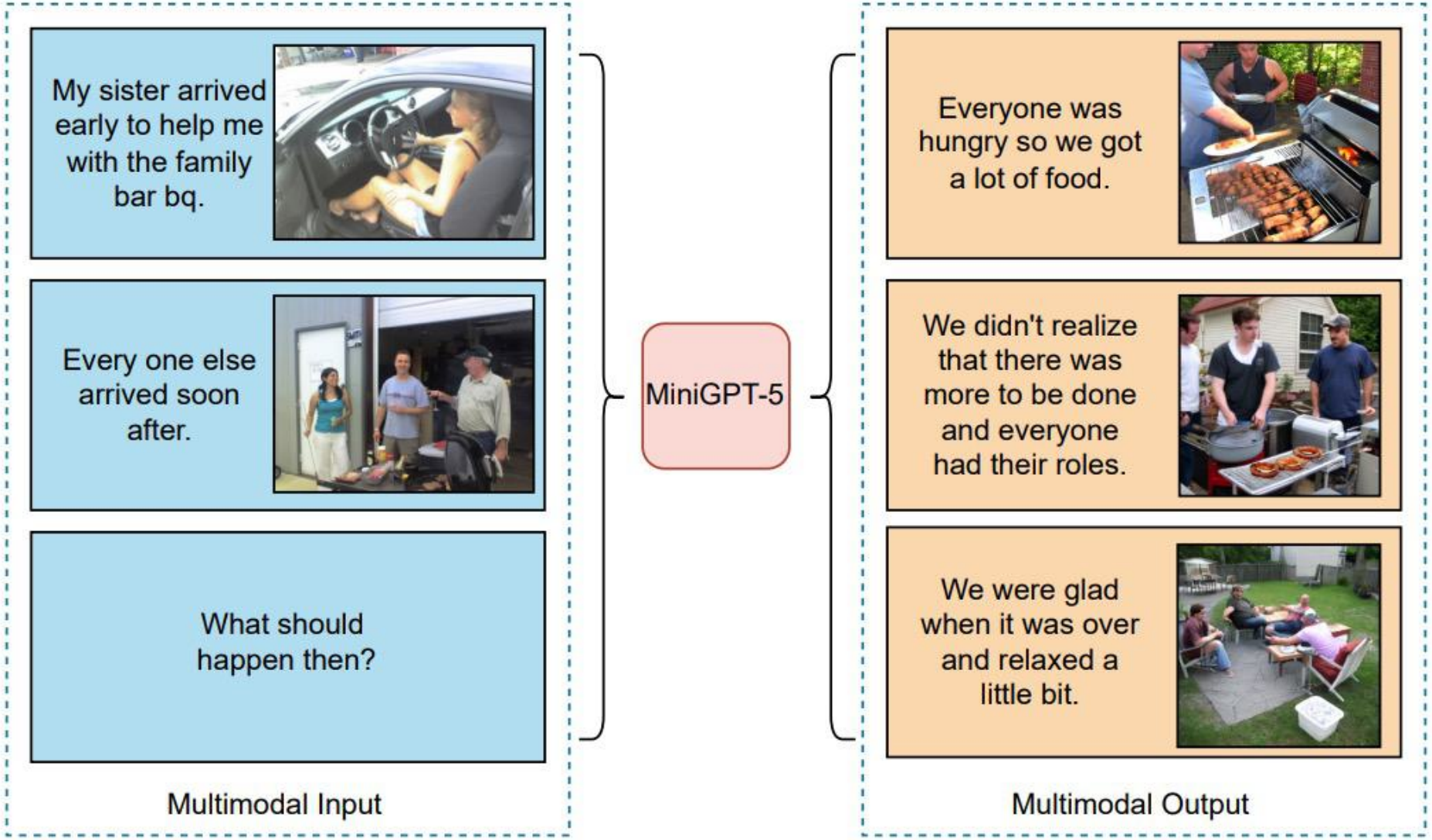


Figure 3: Overview of Parti sequence-to-sequence autoregressive model (left) for text-to-image generation with ViT-VQGAN as the image tokenizer [21] (right).

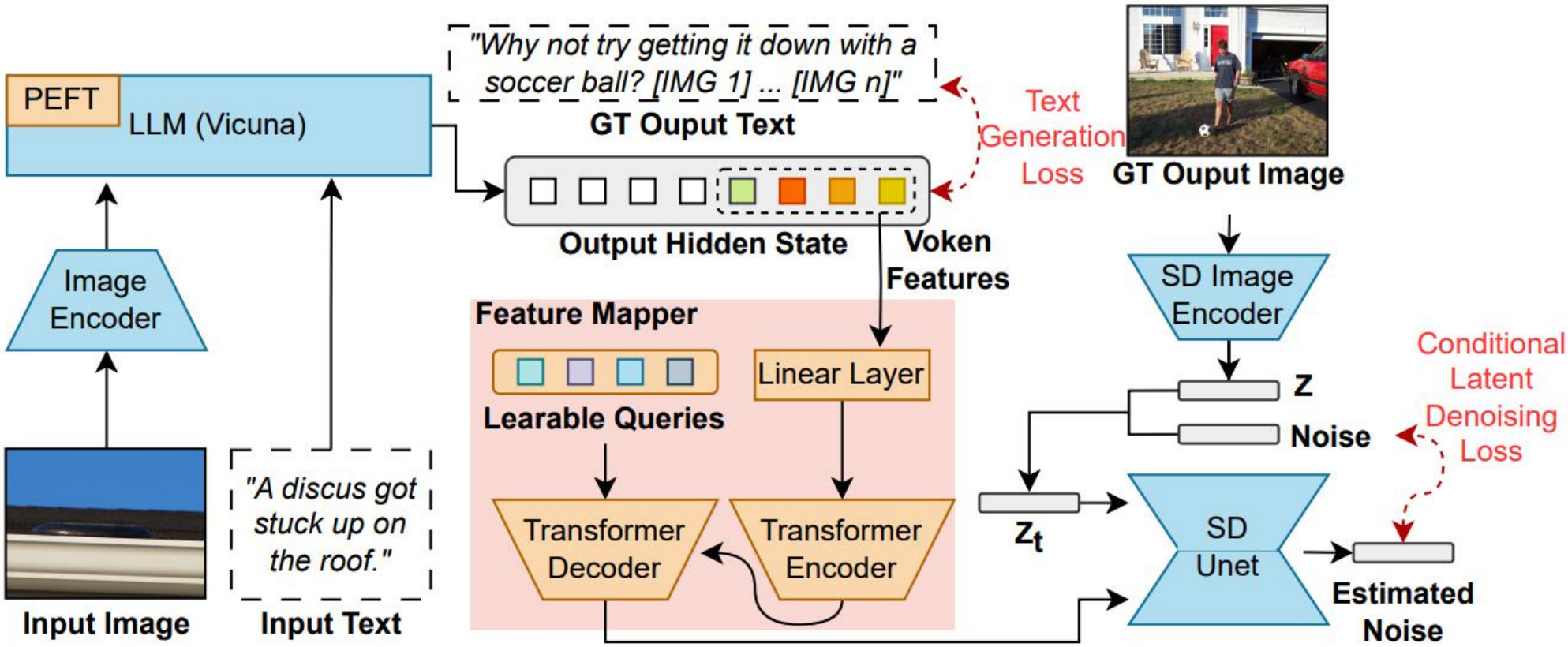
https://arxiv.org/pdf/2206.10789.pdf

Models that can generate images along with text



https://arxiv.org/pdf/2310.02239.pdf

Models that can generate images along with text

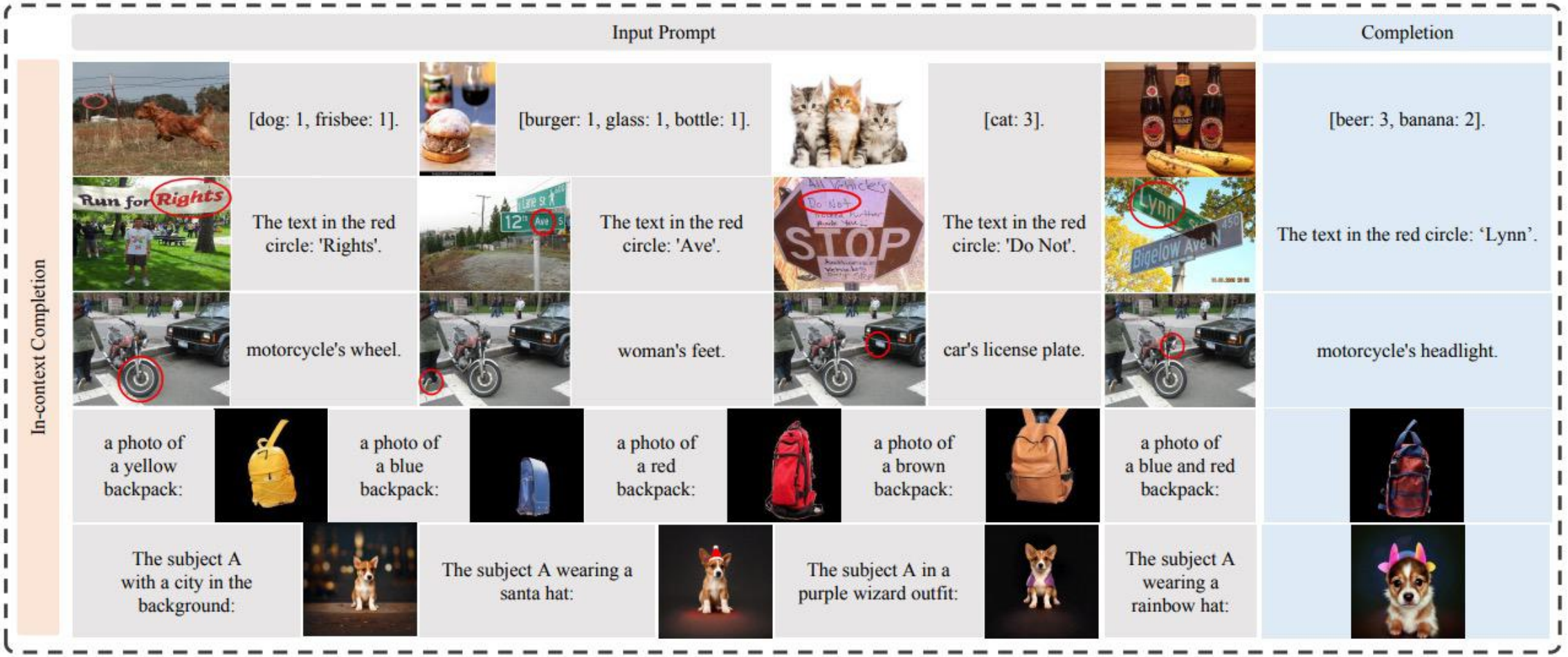


https://arxiv.org/pdf/2310.02239.pdf

Emu2: Generative Multimodal Models are In-Context Learners

- What is In-Context Learning?

  Ability to solve multimodal tasks in context (i.e., with only a few demonstrations or simple instructions)



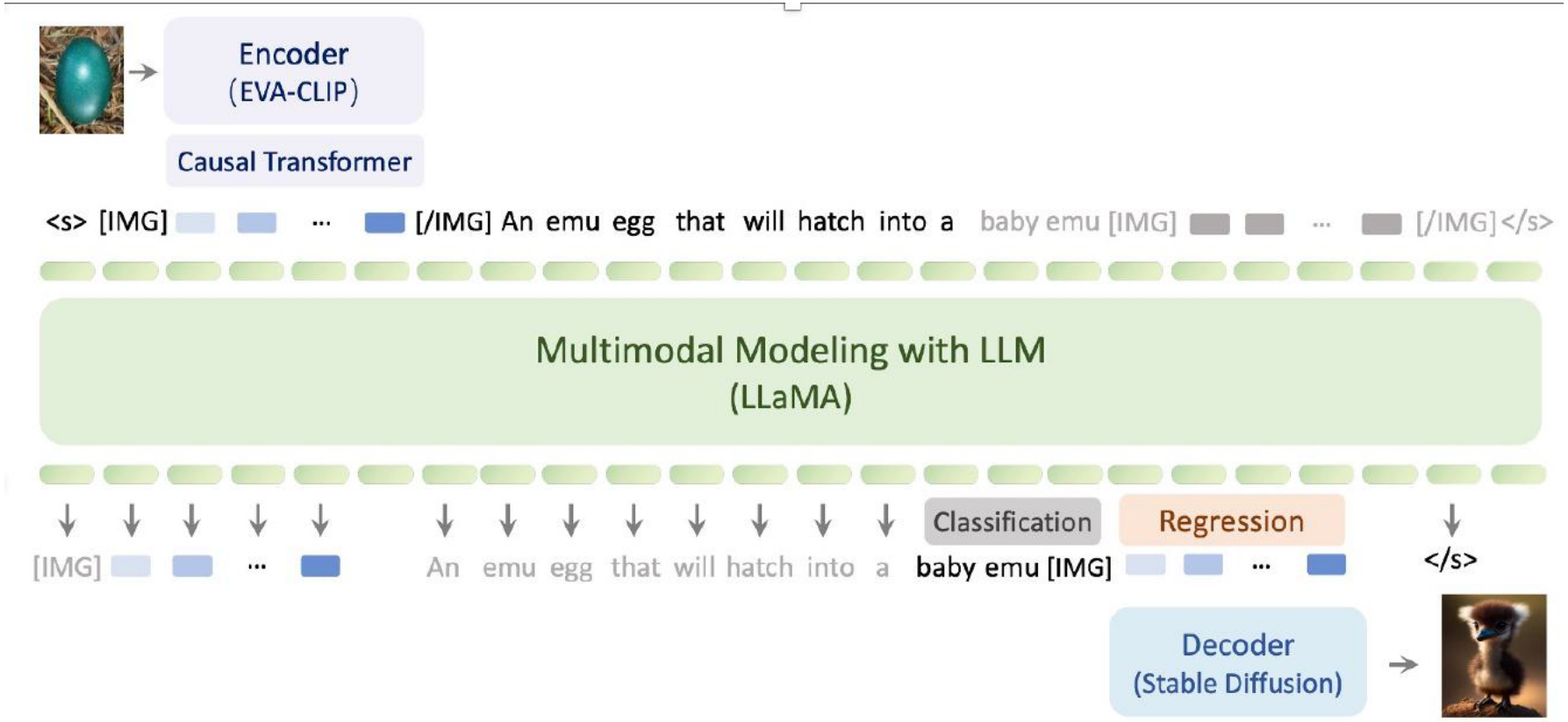
Emu2: Motivation

- Multimodal tasks encompass anything involving understanding and generation in single or multiple modalities
- Previous multimodal systems largely rely on designing task-specific architecture and collecting a sizeable supervised training set
- But humans can solve a new task in context, i.e., with only a few demonstrations or simple instructions
- This paper demonstrate that a scaled-up multimodal generative pretrained model (37B parameters) can harness similar in-context learning abilities

Emu [Previous Version]: Architecture

- **Emu's Model Architecture**

  Visual Encoder + **Causal Transformer** + Multimodal Modeling + Visual Decoder

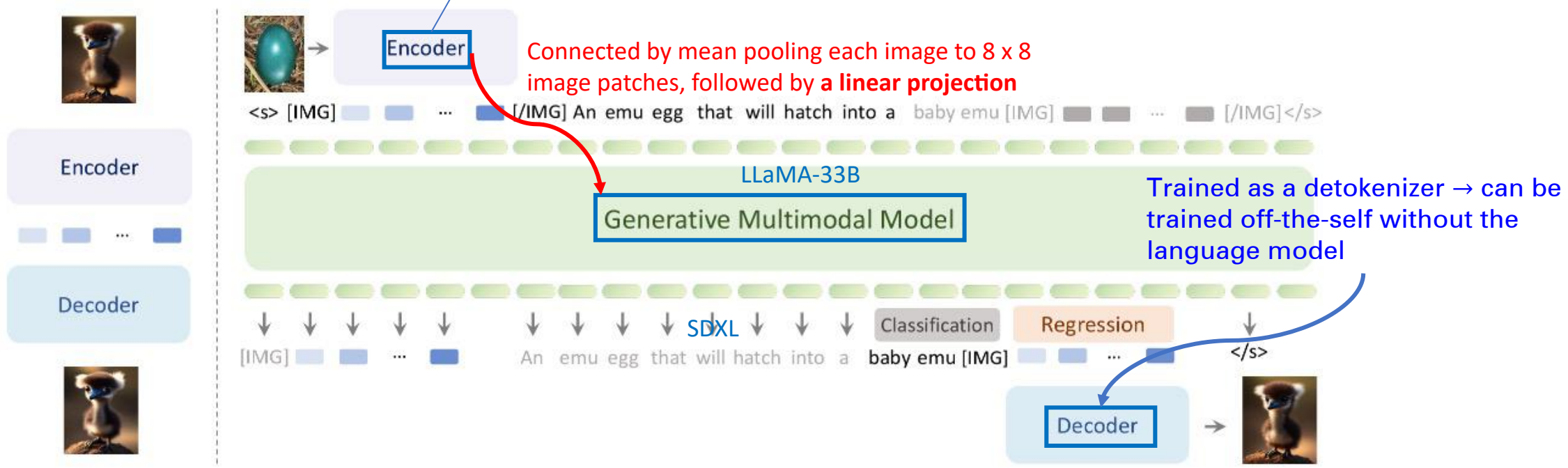


In Emu, for each training sample, the multimodal modeling LLM is used to generate N visual embeddings in an autoregressive manner to feed into image decoder as the condition of image generation training

Emu2: Objective & Architecture

- Unified autoregressive objective:
  Predict-the-next-multimodal-element (either visual embeddings or textual tokens)

- Model Architecture
  Visual Encoder + autoregressive Multimodal Modeling + Visual Decoder



Emu2: Training Objective

- Recall the training objective is: **Predict-the-next-multimodal-element**

  - Given an unlabeled web-scale corpora D consisting of interleaved multimodal sequences $x = (x_1, x_2, \ldots, x_n)$

  - First convert all continuous 2D signals into 1D latent embeddings sequence $u = (u_1, u_2, \ldots, u_m)$, where u_i can be either a discrete text token, or a visual embedding

  - Goal is to approximate the likelihood of the web-scale corpora p(x) with p(u)

$$\max_{\theta} \sum_{u \in D} \sum_{i=1}^{|u|} \log P(u_i | u_1, \ldots, u_{i-1}; \theta) \approx p(x)$$

- Two types of losses:

  - For discrete text tokens: cross-entropy loss

  - For continuous visual embeddings: $\ell 2$ regression loss

Emu2: Pretraining

- **Training:**

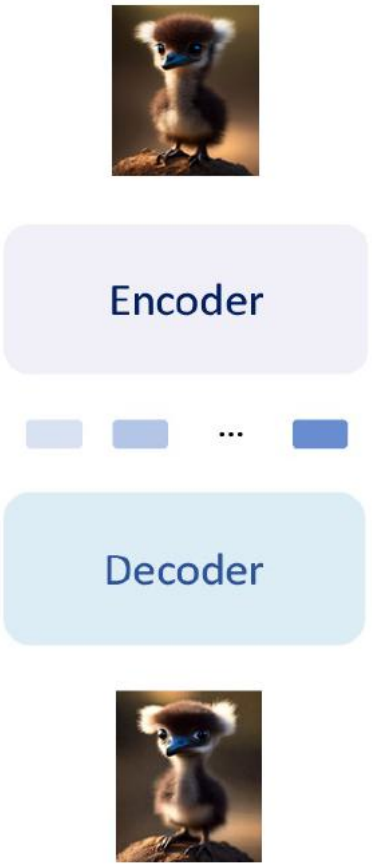
1. Pretrained on image-text and video-text pair data with only captioning loss on the text tokens

2. Freeze the Visual Encoder and only optimize the linear projection layer and Multimodal Modeling with both text classification loss and image regression loss

- **Visual Decoding**

- Visual Decoder is trained to directly decode visual embeddings generated by the Visual Encoder into image

- Can be trained off-the-shelf without the language model



Emu2: Two Variants

• **Emu2** can be efficiently aligned to follow specific task instructions by fine-tuning the base model with conversational data to yield **Emu2-Chat**

Multimodal tokens Supervised by cross-entropy loss during training

• Instruction-Following Chat Training Objective:

- Two special tokens: [USER] and [ASSISTANT] to help organize different data types as bellow

<Sys.Msg.> [USER]: <Instruction> [ASSISTANT]: <Answer>

System message that varies between 2 major task categories: academic-task-oriented & multimodal chat

• **Emu2-Gen:** capable of accepting a mix of text, locations and images as conditions, and generating images that are ground in the specified text or subject

• **Controllable Visual Generation Training Objective:**

- use the same unified generative pretraining objective
- coordinates of each object is represented in image form by drawing the bounding box at its specified location on a black image

Emu2: Quantitative Results

Model	Shot	VQAv2	OKVQA	VizWiz	TextVQA	Hateful Memes
Kosmos-1 (1.6B)	0	51.0	-	29.2	-	-
	4	51.8	-	35.3	-	-
	8	51.4	-	39.0	-	-
Flamingo (9B)	0*	51.8	44.7	28.8	31.8	57.0
	4	56.3	49.3	34.9	33.6	62.7
	8	58.0	50.0	39.4	33.6	63.9
	16	59.4	50.8	43.0	33.5	64.5
Flamingo (80B)	0*	56.3	50.6	31.6	35.0	46.4
	4	63.1	57.4	39.6	36.5	<u>68.6</u>
	8	65.6	57.5	44.8	37.3	**70.0**
	16	66.8	**57.8**	48.4	37.6	**70.0**
IDEFICS (80B)	0*	60.0	45.2	36.0	30.9	60.6
	4	63.6	52.4	40.4	34.4	57.8
	8	64.8	55.1	46.1	35.7	58.2
	16	65.4	56.8	48.3	36.3	57.8
Emu (14B)	0*	52.9	42.8	34.4	-	-
	4	58.4	-	41.3	-	-
	8	59.0	-	43.9	-	-
	16	-	-	-	-	-
Emu2 (37B)	0	33.3	26.7	40.4	26.2	52.2
	4	67.0	53.2	54.6	48.2	62.4
	8	<u>67.8</u>	54.1	<u>54.7</u>	<u>49.3</u>	65.8
	16	**68.8**	<u>57.1</u>	**57.0**	**50.3**	66.0

Table 1. Zero-shot and few-shot evaluations of **Emu2**. 0* denotes text two-shot and image zero-shot results following Flamingo [5]. The best results are in **bold** and the second best are <u>underlined</u>.

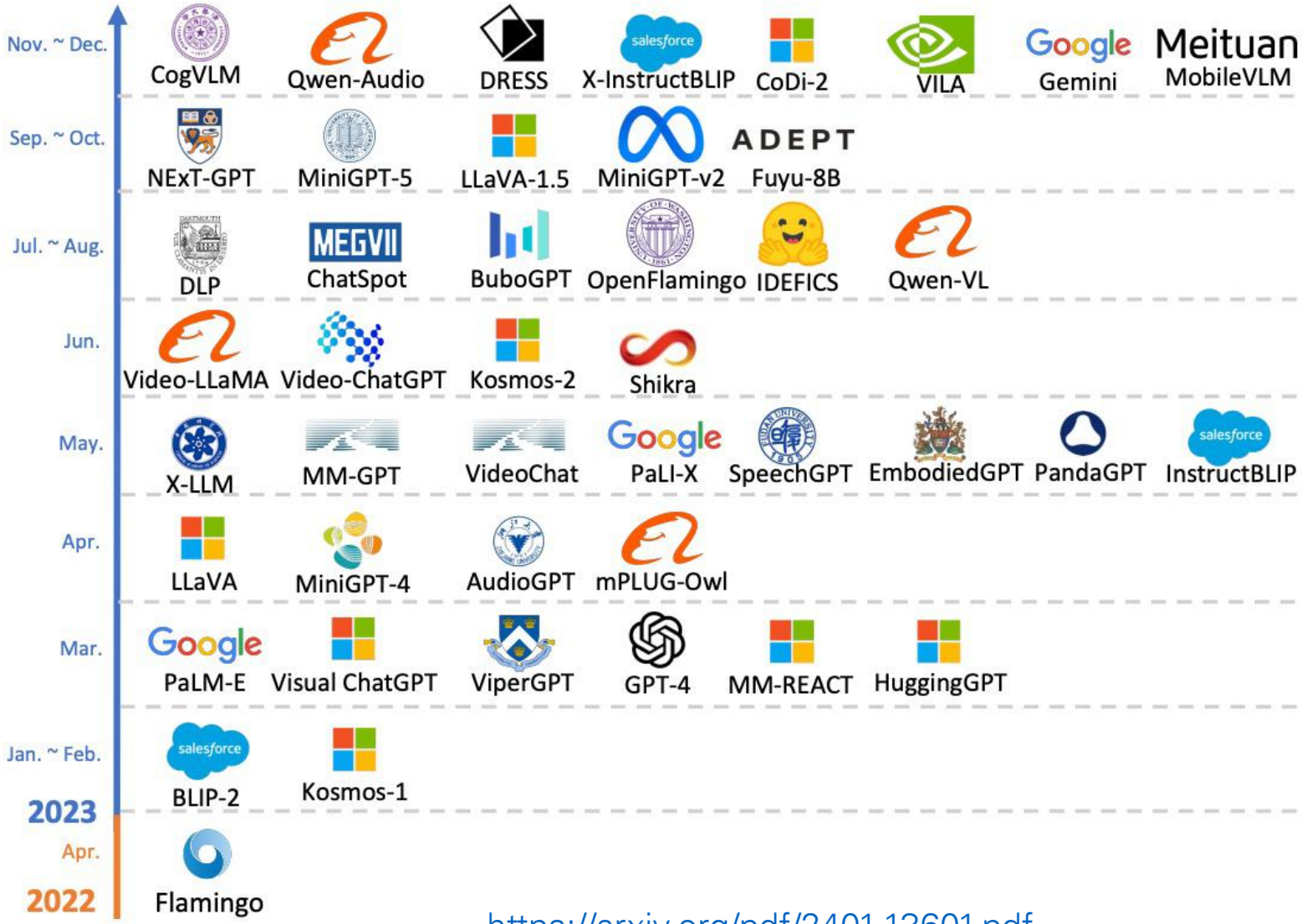
Outperforms Flamingo-80B and IDEFICS-80B under all few-shot settings with a much smaller model scale

Emu2: Controllable Visual Generation

Figure 4. Visualization of **Emu2-Gen**'s controllable generation capability. The model is capable of accepting a mix of text, locations and images as input, and generating images in context. The presented examples include text- and subject-grounded generation, stylization, multi-entity composition, subject-driven editing, and text-to-image generation.

Models	CLIP-I ↑	CLIP-T ↑
unimodal generation models		
MUSE [14]	-	**0.320**
Imagen [65]	-	0.270
DALL-E 2 † [62]	-	0.314
DALL-E 3 † [10]	-	**0.320**
SDv1.5 [63]	0.667	0.302
SDXL [59]	0.674	0.310
multimodal generation models		
GILL [38]	0.684	-
SEED [28]	0.682	-
Emu [72]	0.656	0.286
Emu2-Gen	**0.686**	0.297

Great survey paper on multimodal LLMs



https://arxiv.org/pdf/2401.13601.pdf