Any reproduction or distribution of this document, in whole or in part, is prohibited unless permission is granted by the authors

# EE-559

# Deep Learning

## What’s on today?

- Vision transformer: from coherent text to coherent images
- Audio transformer: adapting transformers for sound understanding
- Audio-visual transformer: analyzing jointly audio & video data
- Vision and language models: fusing vision & language understanding
- Exercises: multimodal transformer

# Transformer

- generative sequence modeling
- universal self-supervised learning algorithm
- for 1-D sequences of any form

**Concepts**: Transformer as self-supervised learning algorithm, sequences of bytes, trained to maximize the likelihood (mode covering)

---

# Vision transformer

# Transformers for visual tasks

image generation

image completion

image classification

---

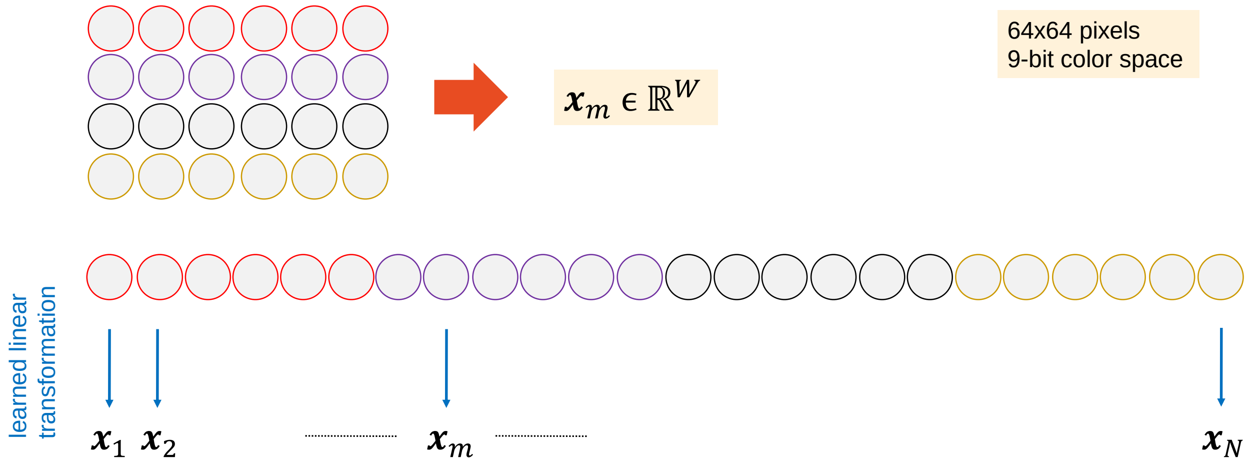
# From language to vision

domain differences

visual entities: large scale variations

high resolution of pixels (compared to words)

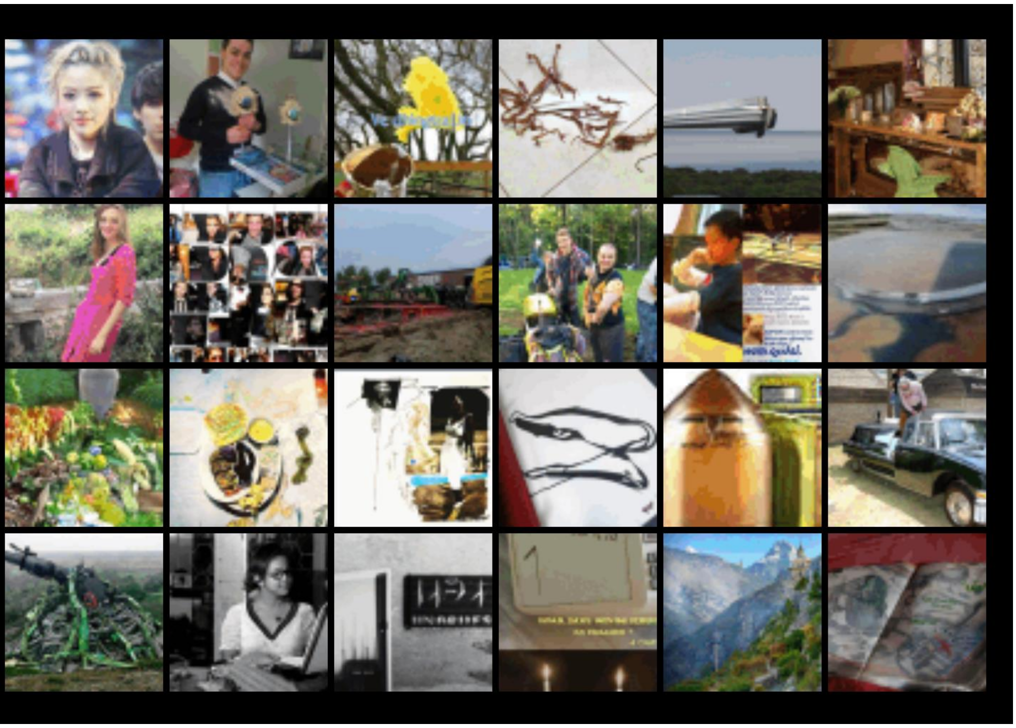
# imageGPT – transformer decoder



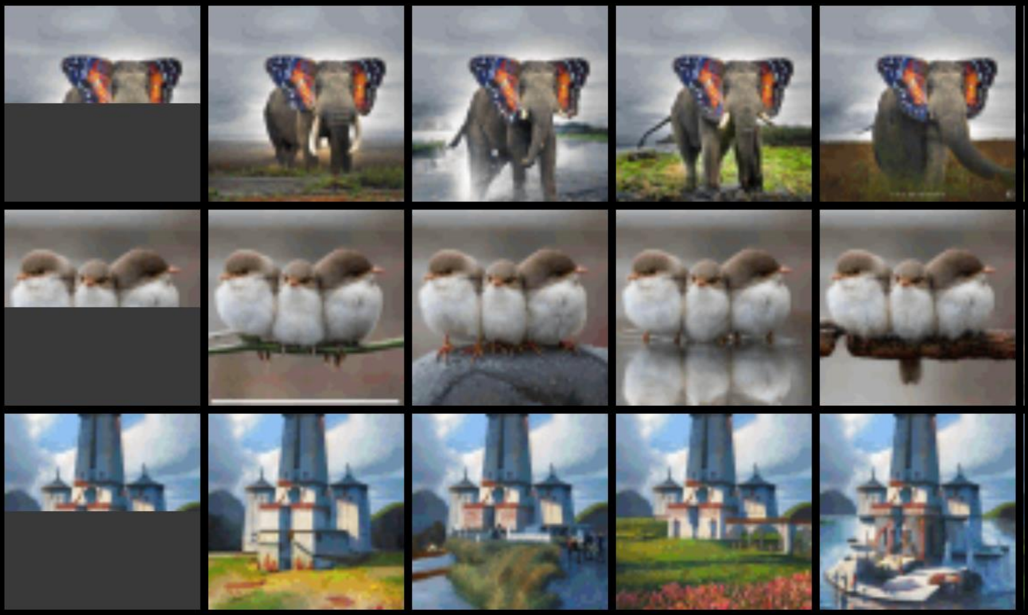
https://openai.com/research/image-gpt

---

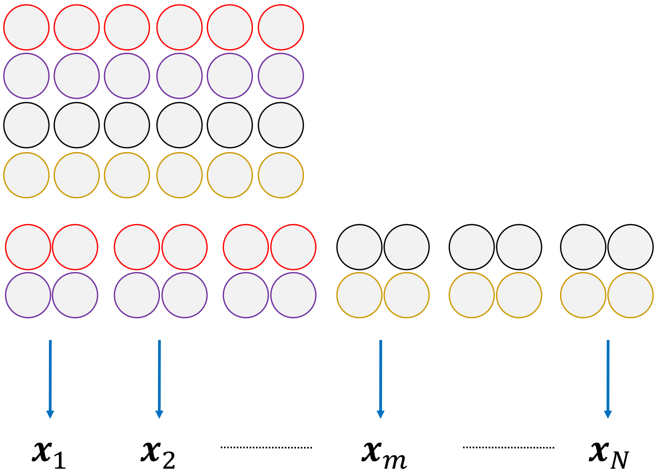
# Sample generated images



## Sample completed images



# Vision transformer



learned linear transformation

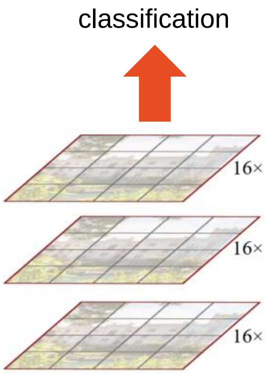
$\boldsymbol{p}_m \in \mathbb{R}^{P \times P}$

$P = 16$

$\boldsymbol{x}_m \in \mathbb{R}^{W}$

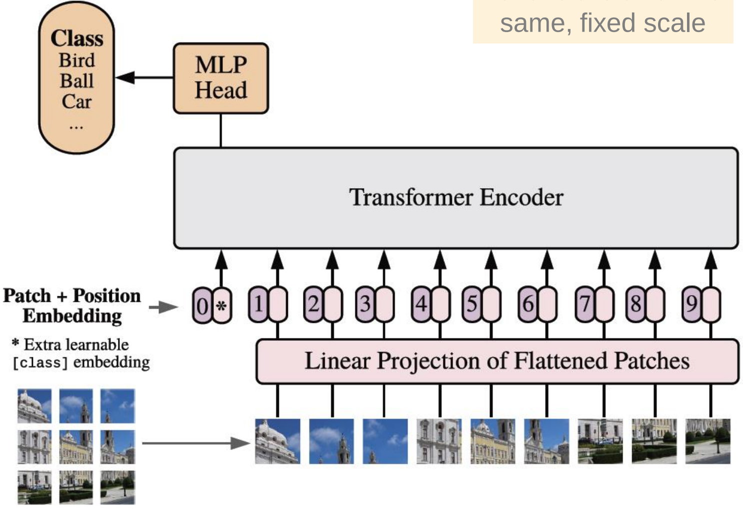
**Concepts**: Learned 1D positional encoding, single scale,
supervised training on 303,000,000 labelled images of 18,000 classes

# ViT



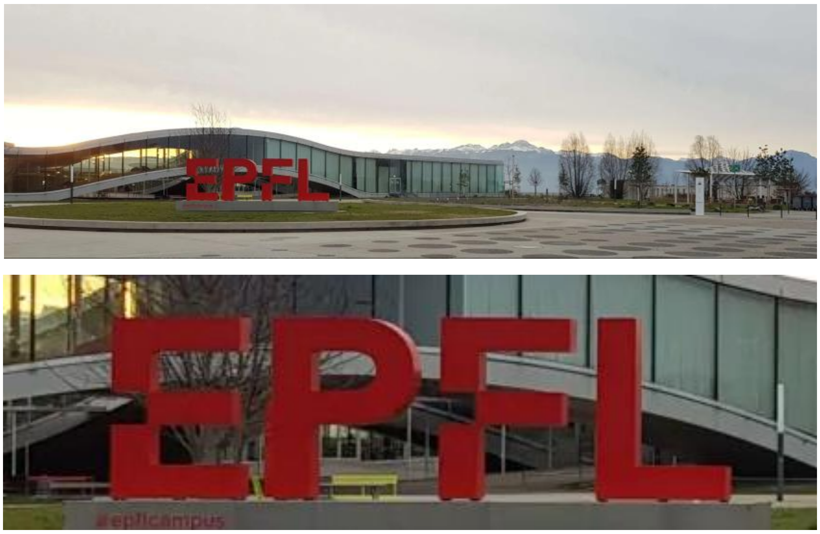
patch → token

Tokens are all at the same, fixed scale


Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

arXiv:2010.11929

# Scale



# Scale: hierarchical architecture

non-overlapping local windows

cross-window connection

model at various scales

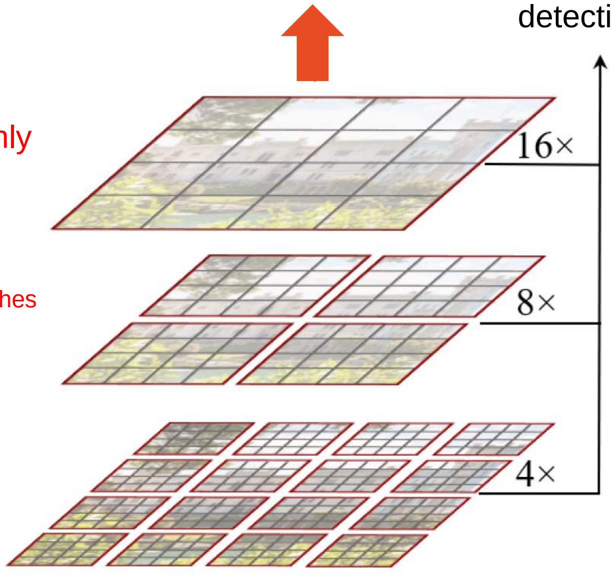
*limits self-attention computation*

---

# Swin

classification

segmentation, detection, ..



self-attention only within each local window

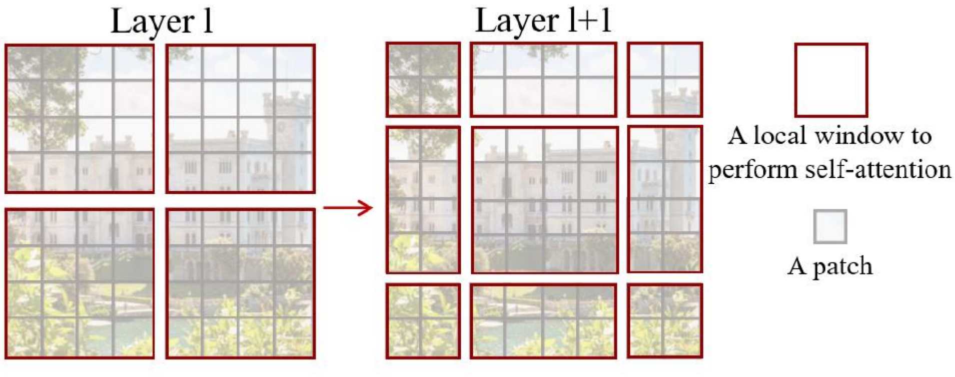
fixed number of patches in each window

complexity: linear to image size

hierarchical feature maps

arXiv:2103.14030

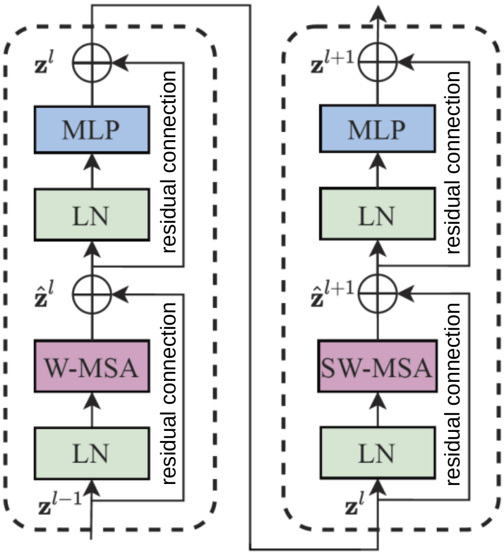
# Cross-window connection



shifted window approach

arXiv:2103.14030

# Consecutive Swin transformer blocks



*shift of the window partition between consecutive self-attention layers*

2-layer MLP with GELU nonlinearity

LayerNorm layer

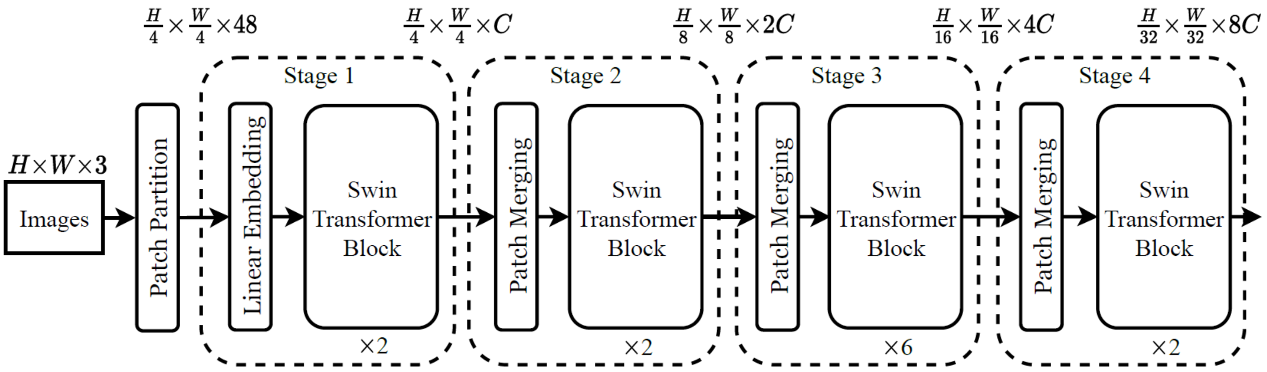
window based multi-head self-attention module

shifted window based multi-head self-attention module

arXiv:2103.14030

# Swin architecture

$H$: image height
$W$: image width
$C$: size of the embedding



*gradually merging neighboring patches in deeper transformer layers*

arXiv:2103.14030

---

# Audio transformer

# Transformers for audio tasks

- audio classification
- automatic speech recognition
- speech synthesis
- voice conversion

# Speech representation learning

- multiple sound units in each input utterance
- variable lengths of sound units
- no explicit segmentation of sound units
- non-lexical information of how it is delivered
- noise interleaving / overlapping with speech

*speaker identity, emotion, hesitation, interruptions*

*laughter, coughing, background sounds*

# Hidden-Unit BERT (HuBERT)

predict predetermined cluster assignments

learn representation of unmasked inputs

capture long-range temporal relations

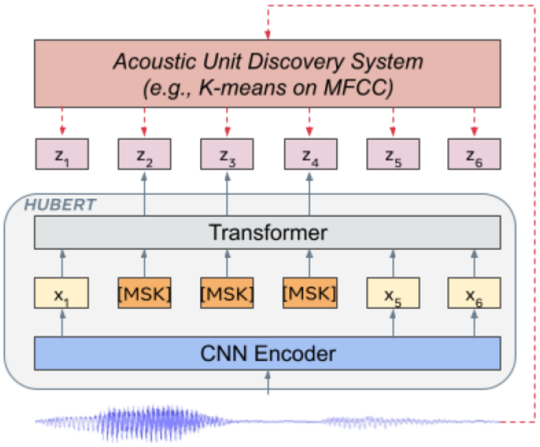
*self-supervised representation learning with access to speech-only data*

arXiv:2106.07447

# Hidden-Unit BERT (HuBERT)

learn a **combined acoustic & language model** over the continuous inputs

prediction loss over the masked regions only



offline clustering (*K*=100) provides aligned target labels

self-supervised speech representation learning

*K-means*: clustering

*MFCC*: Mel Freq. Cepstral Coefficients

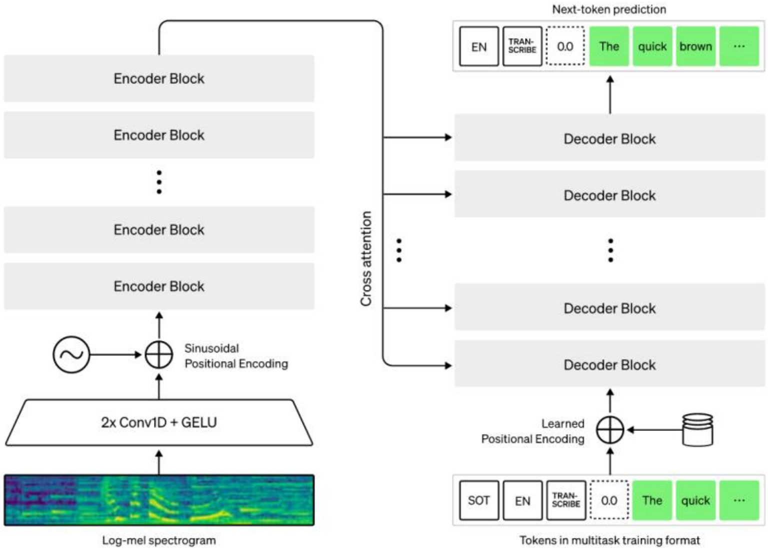
CNN: Convolutional Neural Network

arXiv:2106.07447

# Whisper

GELU: Gaussian Error Linear Unit

Mel Scale: sounds of equal relative distance sound to humans as they are equal in distance from one another



30-sec segments

audio conditional language model

arXiv:2212.04356

*“[…] problems such as getting stuck in repeat loops, not transcribing the first or last few words of an audio segment, or complete hallucination where the model will output a transcript entirely unrelated to the actual audio.”*

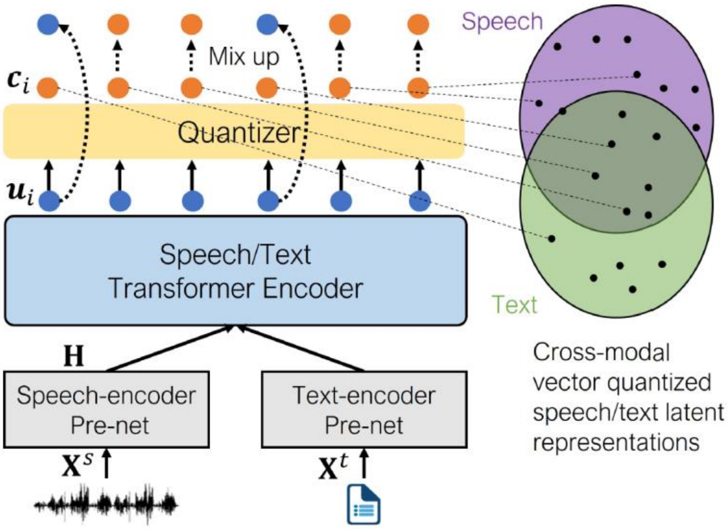
arXiv:2212.04356

# SpeechT5: joint pre-training

H: speech utterance
$u_i$: continuous representations
$c_i$: discrete representation

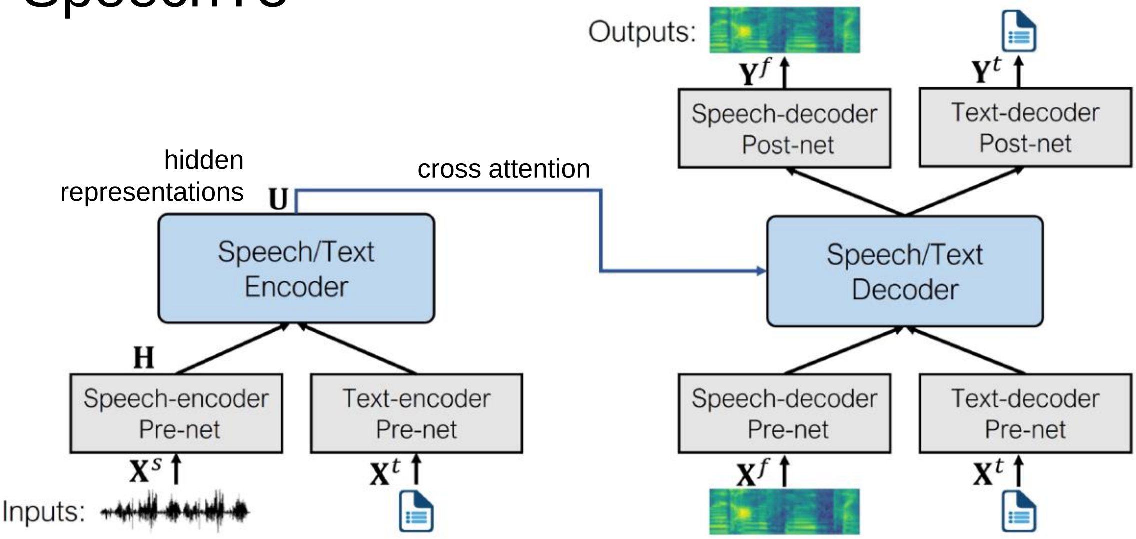
learns alignment between acoustic and textual representation

shares tokens across modalities


Speech
Mix up
$c_i$
Quantizer
$u_i$
Speech/Text Transformer Encoder
Text
H
Speech-encoder Pre-net
Text-encoder Pre-net
$X^s$
$X^t$
Cross-modal vector quantized speech/text latent representations

arXiv:2110.07205

# SpeechT5



hidden representations

cross attention

arXiv:2110.07205

slido

## Would you like to ask any questions about your mini-project?

ⓘ Start presenting to display the poll results on this slide.

# Audio-visual transformer

# Transformers for audio-visual tasks

segmentation

localization

question answering

# Audio-visual fusion

frozen vision transformer
+
with trainable latent audio-visual adapters

(inserted into each transformer layer)

bi-directional adapter
to
transfer information from audio to visual tokens

(and from visual to audio tokens)

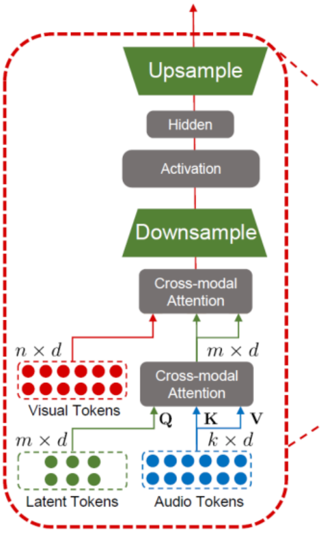
arXiv:2212.07983

# Adapter: four high-level components

| latent tokens | cross-modal attention operation | cross-modal attention operation | discriminative representation |
|---|---|---|---|
| learning compressed audio or visual representation | compresses all the tokens from one modality into the latent tokens | fusion between the latent tokens of one modality and the tokens from another modality | fused tokens fed into module that computes a discriminative audio-visual representation |

arXiv:2212.07983

# Audio to visual adapter



fused tokens: fed to an adapter

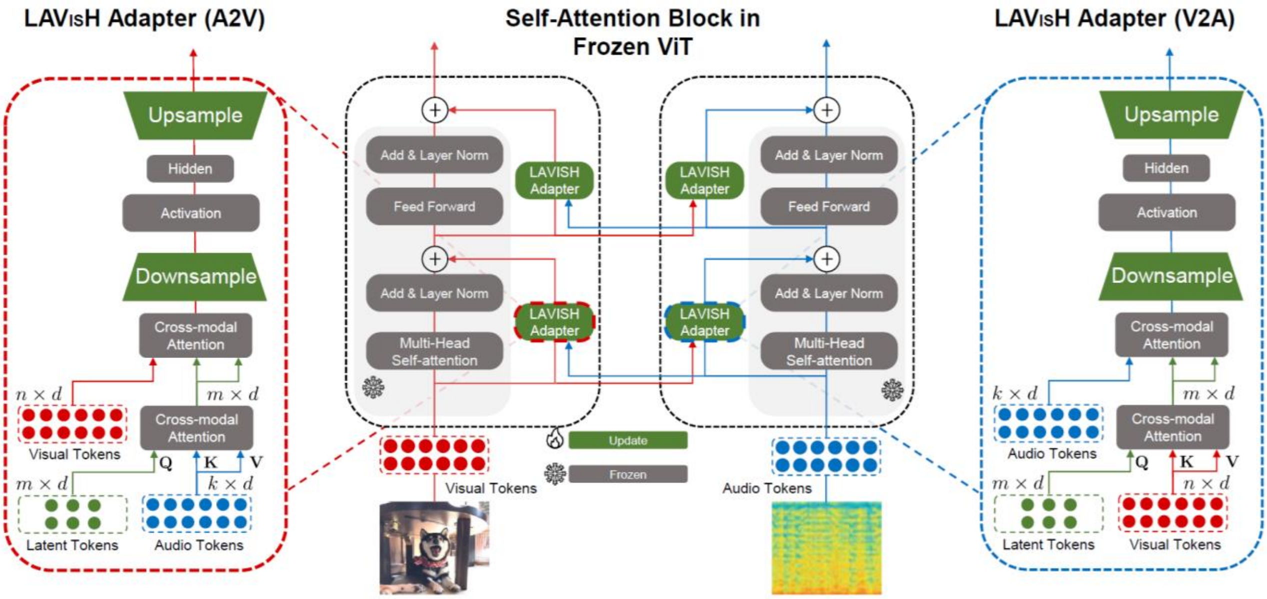
adapter module: *computes a discriminative audio-visual representation*

audio-visual fusion between latent tokens of one modality & tokens from another modality

compresses all the tokens from one modality into the latent tokens

learning compressed audio or visual representation

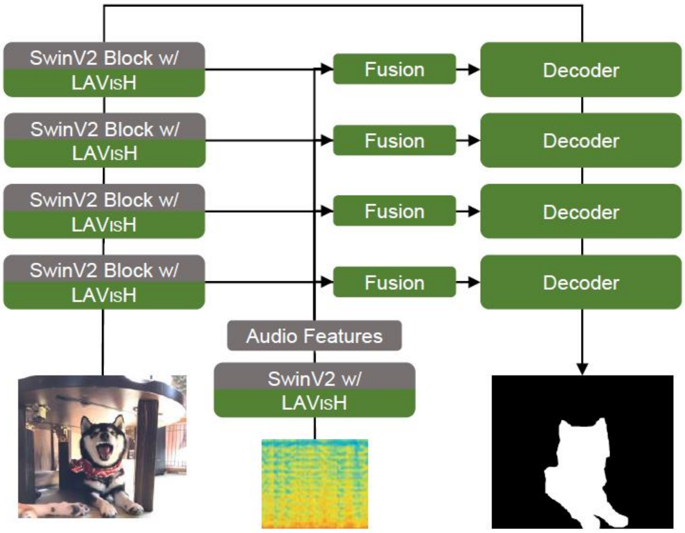
arXiv:2212.07983



LAVisH: latent audio-visual hybrid

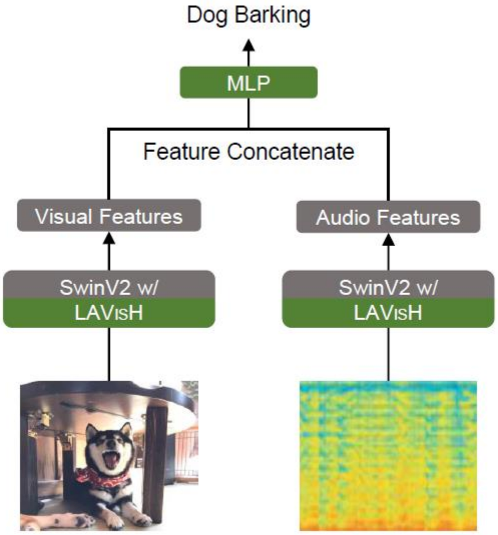
arXiv:2212.07983

# Audio-visual segmentation



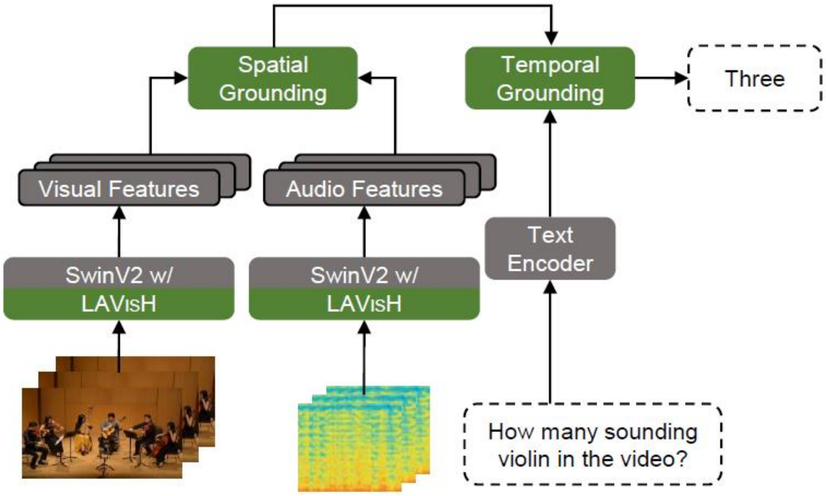
arXiv:2212.07983

# Audio-visual event localization



arXiv:2212.07983

# Audio-visual question answering



arXiv:2212.07983

# Vision and language models

---

## CM3Leon

image tokenizer

256 × 256 image
1024 tokens from
vocabulary of 8192

text tokenizer

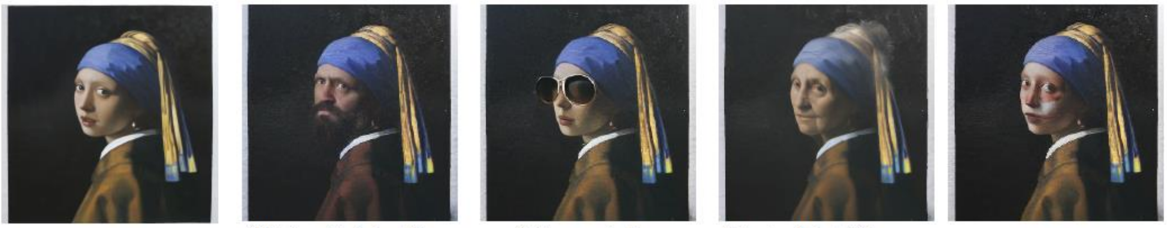
vocabulary size of 56320
special token <break>
to indicate a transition
between modalities

arXiv:2309.02591

# CM3Leon: enabled tasks

- text-guided image editing
- image-to-image grounded generation
- image-to-text tasks
- conditional text generation

# Text-guided image editing



Input

“What would she look like as a bearded man?”

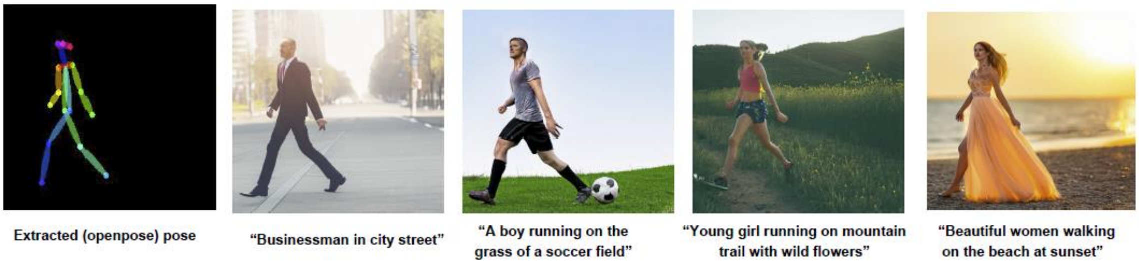
“Put on a pair of sunglasses”

“she should look 100 years old”

“Apply face paint”

arXiv:2309.02591

# Image-to-image grounded generation



Extracted (openpose) pose

“Businessman in city street”

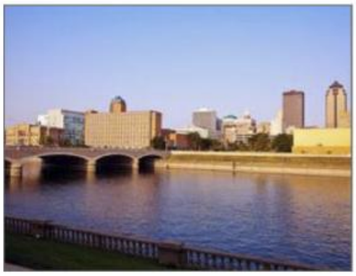
“A boy running on the grass of a soccer field”

“Young girl running on mountain trail with wild flowers”

“Beautiful women walking on the beach at sunset”

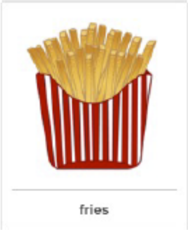
arXiv:2309.02591

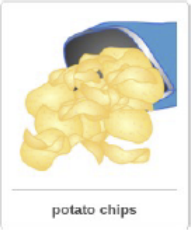
# Image-to-text tasks



| | |
|---|---|
| Caption: Describe the given image | A beautiful view of a city from across a river. |
| Long Caption: Describe the given image in very detail | A view of tall buildings in a city. The photo is taken from a park across a river. We can see a bridge over the river. |
| VQA: Question: what time of the day is the photo taken? | Sunset time |
| Reasoning: Question: Does this passage describe the weather or the climate? Context: Figure: Des Moines. The temperature recorded …Please explain your answer. | Weather. Because the atmosphere is the layer of air that surrounds Earth. Both weather and climate tell you about the atmosphere. … |

# Conditional text generation





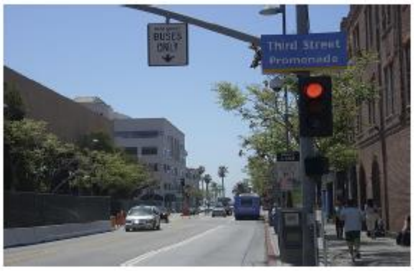
Instruction: Question: Which property do these two objects have in common?
Context: Select the best answer.
Options: (A) shiny (B) salty
Answer: Let's think step-by-step:

CM3Leon Output: Think about each object. Potato chips have a salty taste. The pretzel is salty. So the answer is (B).

arXiv:2309.02591

---

# Conditional text generation



Instruction: Describe all the objects in the given image in very detail.

CM3Leon Output: A street sign is on a metal pole. The sign is blue with white writing. There is a red light on the pole. The sky is a bright blue. There are cars parked on the street. There are trees on the side of the street. There are buildings on the side of the street.

arXiv:2309.02591

# Natural language as a training signal

*easier to scale*

supervision contained in natural language

*learning perception*

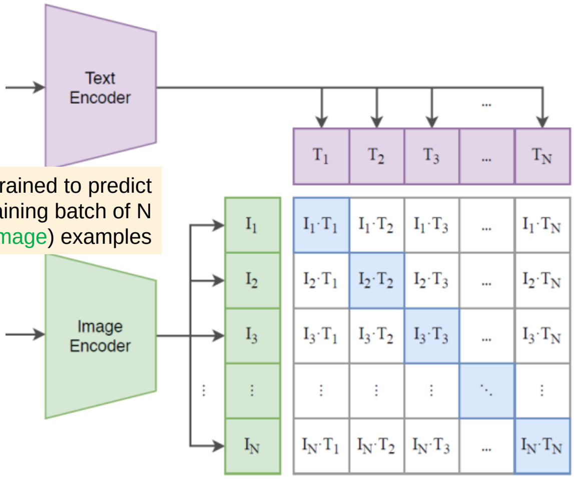
*enables zero-shot transfer*

Contrastive Language-Image Pre-training (**CLIP**)

arXiv:2103.00020

---

# Contrastive language-image pre-training



|  | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ |
| $I_2$ | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ |
| $I_3$ | $I_3 \cdot T_1$ | $I_3 \cdot T_2$ | $I_3 \cdot T_3$ | ... | $I_3 \cdot T_N$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $I_N$ | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ |

jointly trained to predict pairings of a training batch of N (text, image) examples

**multi-modal embedding space**

**maximize** the cosine similarity of the embeddings of the N correct pairs

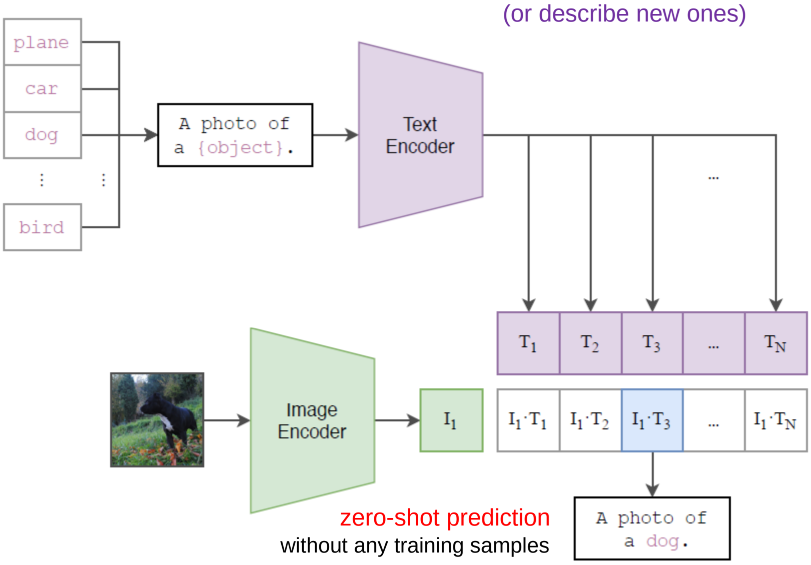
**minimizing** cosine similarity of the embeddings of the NxN - N incorrect pairs

dataset of 400 million (image, text) pairs collected from the internet

arXiv:2103.00020

# Testing

natural language used to reference
learned visual concepts
(or describe new ones)



learned text encoder
synthesizes a zero-shot linear
classifier by embedding
the names or descriptions
of target classes

zero-shot prediction
without any training samples

arXiv:2103.00020

# VLM for few-shot learning

*few-shot prompting*

sequences
of arbitrarily
interleaved
visual &
textual data

*output: free-form text*

**Flamingo**

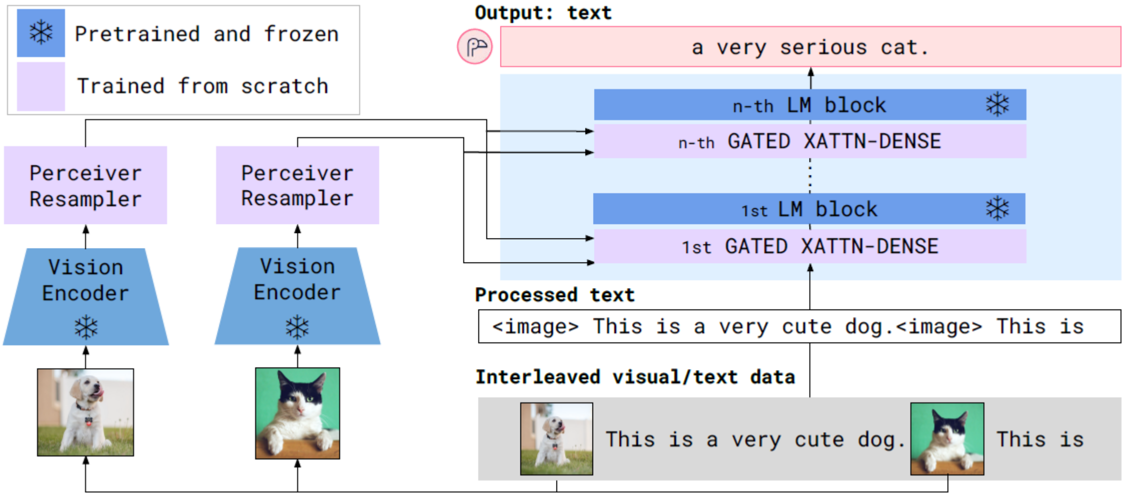
*visual tokens used to condition the frozen language model*

VLM: Vision and Language Model

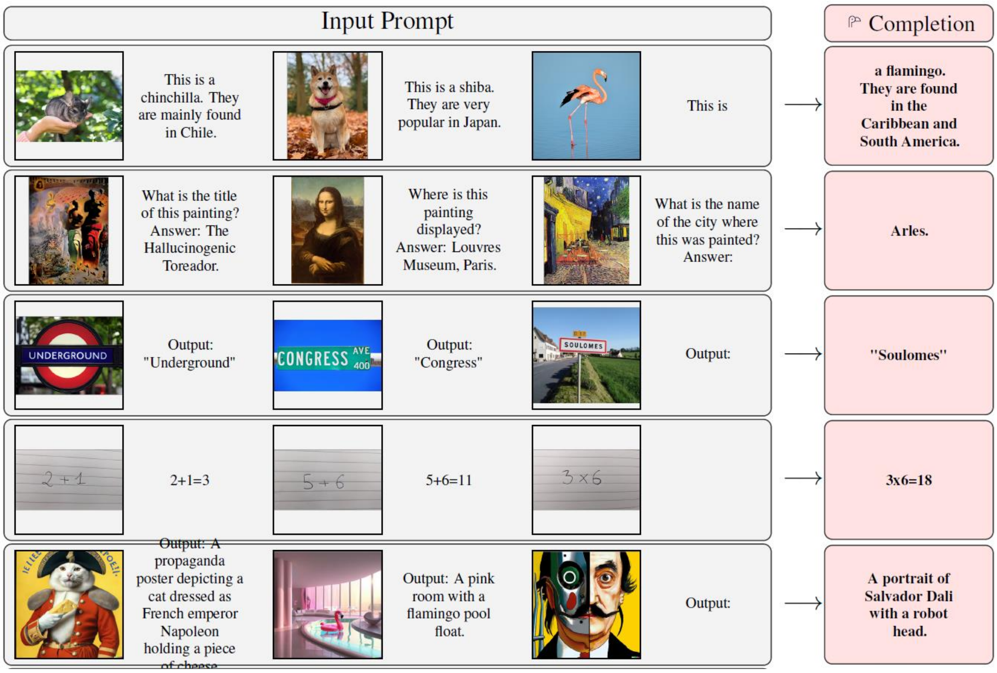
arXiv:2204.14198

# Architecture overview

visual tokens condition the frozen language model

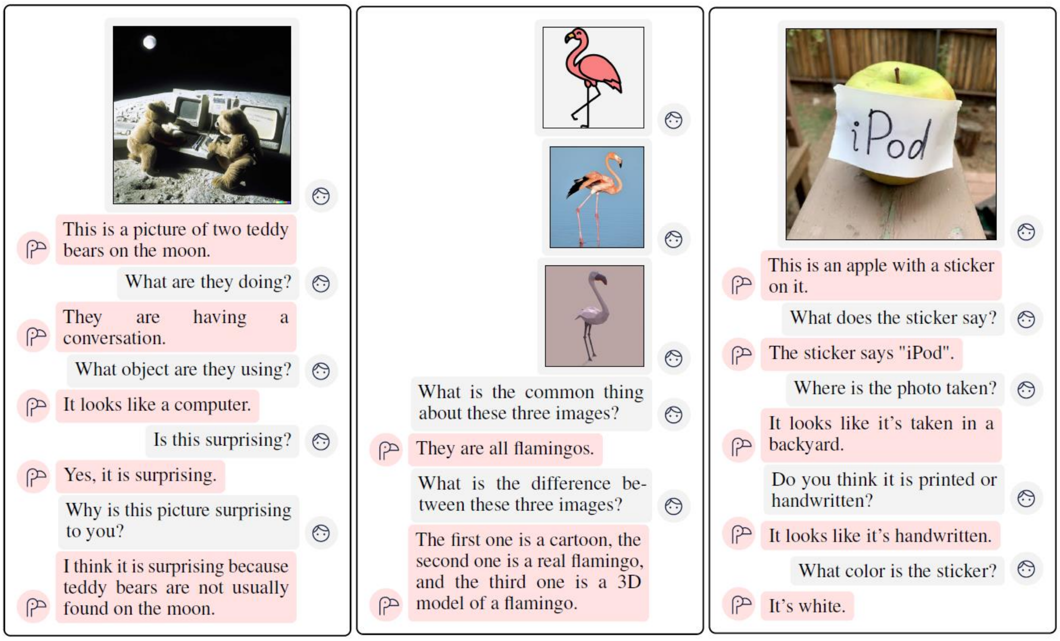


arXiv:2204.14198



| Input Prompt | | | Completion |
|---|---|---|---|
| This is a chinchilla. They are mainly found in Chile. | This is a shiba. They are very popular in Japan. | This is | **a flamingo. They are found in the Caribbean and South America.** |
| What is the title of this painting? Answer: The Hallucinogenic Toreador. | Where is this painting displayed? Answer: Louvres Museum, Paris. | What is the name of the city where this was painted? Answer: | **Arles.** |
| Output: "Underground" | Output: "Congress" | Output: | **"Soulomes"** |
| 2+1=3 | 5+6=11 | | **3x6=18** |
| Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese. | Output: A pink room with a flamingo pool float. | Output: | **A portrait of Salvador Dali with a robot head.** |

arXiv:2204.14198



This is a picture of two teddy bears on the moon.

What are they doing?

They are having a conversation.

What object are they using?

It looks like a computer.

Is this surprising?

Yes, it is surprising.

Why is this picture surprising to you?

I think it is surprising because teddy bears are not usually found on the moon.

What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.

This is an apple with a sticker on it.

What does the sticker say?

The sticker says "iPod".

Where is the photo taken?

It looks like it's taken in a backyard.

Do you think it is printed or handwritten?

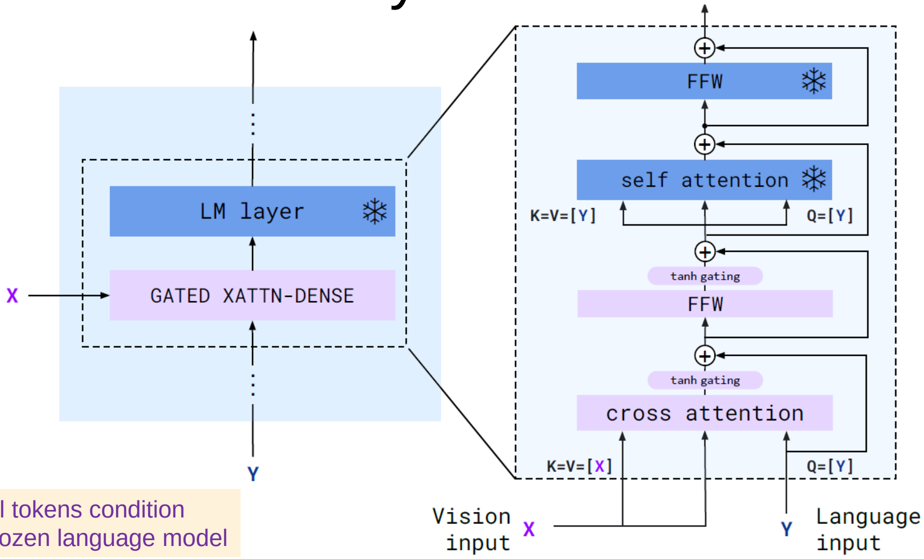
It looks like it's handwritten.

What color is the sticker?

It's white.

arXiv:2204.14198

# Cross-attention layers

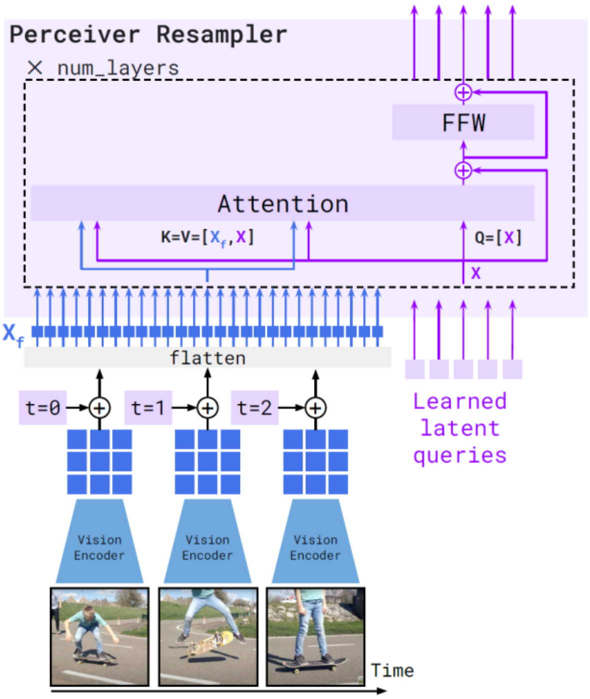


visual tokens condition
the frozen language model

arXiv:2204.14198

*“Future work is required to better understand*
*the effect of these added layers*
*on the optimization dynamics*
*and on the model itself.”*

arXiv:2204.14198



number of output tokens
=
number of learnt latent queries

cross-attending to
the flattened
visual features

learnt temporal
position encoding

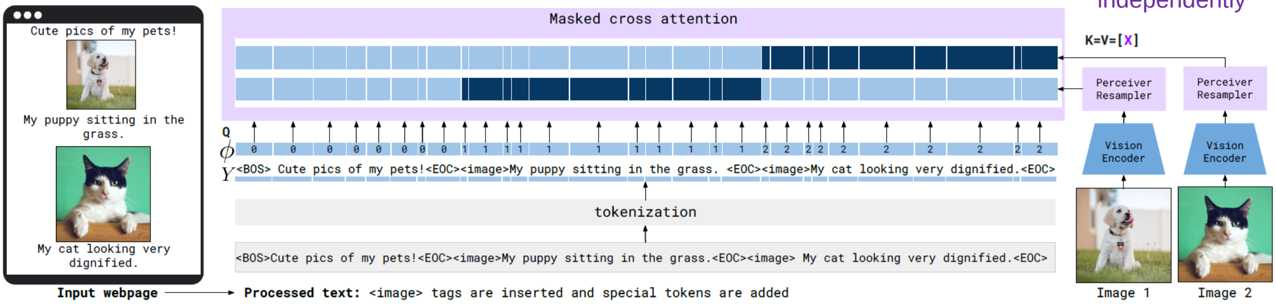
no explicit
spatial grid
position
encodings

arXiv:2204.14198

# Text interleaved with images/videos

selective cross-attention
dark blue: visible
light blue: masked

visual tokens from images processed independently


Cute pics of my pets!
My puppy sitting in the grass.
My cat looking very dignified.
Input webpage
Masked cross attention
K=V=[X]
Perceiver Resampler
Perceiver Resampler
Vision Encoder
Vision Encoder
Image 1
Image 2
Q
$\phi$
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
$Y$
<BOS> Cute pics of my pets!<EOC><image>My puppy sitting in the grass. <EOC><image>My cat looking very dignified.<EOC>
tokenization
<BOS>Cute pics of my pets!<EOC><image>My puppy sitting in the grass.<EOC><image> My cat looking very dignified.<EOC>
Processed text: <image> tags are inserted and special tokens are added

<image> : tag inserted at the location of the visual data in the text
<BOS> : beginning of sequence
<EOC> : end of chunk

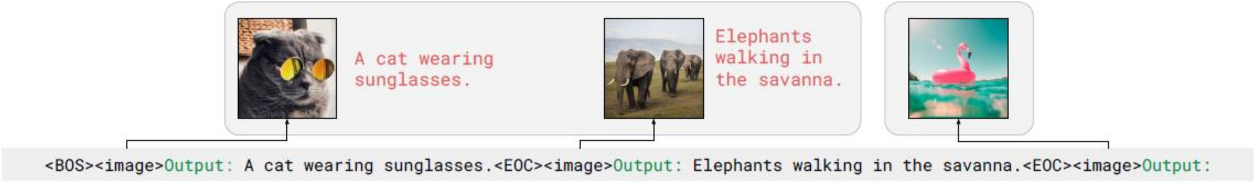
arXiv:2204.14198

# Vision-to-text task

**input**: vision
**output**: text

task-specific few-shot examples
(support examples)

query


A cat wearing sunglasses.
Elephants walking in the savanna.
<BOS><image>Output: A cat wearing sunglasses.<EOC><image>Output: Elephants walking in the savanna.<EOC><image>Output:

processed prompt

prepending `Output:` to the expected response

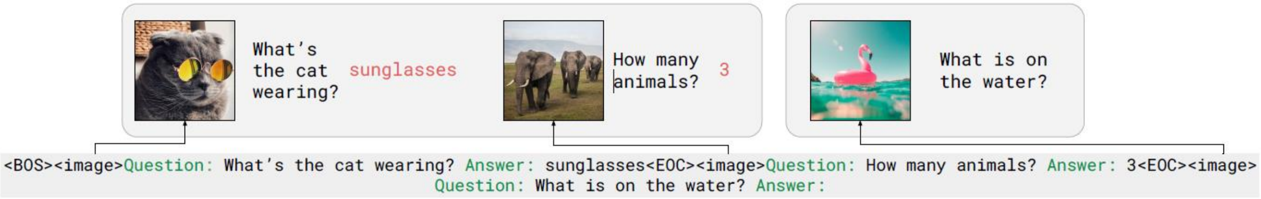
arXiv:2204.14198

# Visual question answering task

**input**: vision + text
**output**: text

task-specific few-shot examples
(support examples)

query



`<BOS><image>Question: What's the cat wearing? Answer: sunglasses<EOC><image>Question: How many animals? Answer: 3<EOC><image> Question: What is on the water? Answer:`

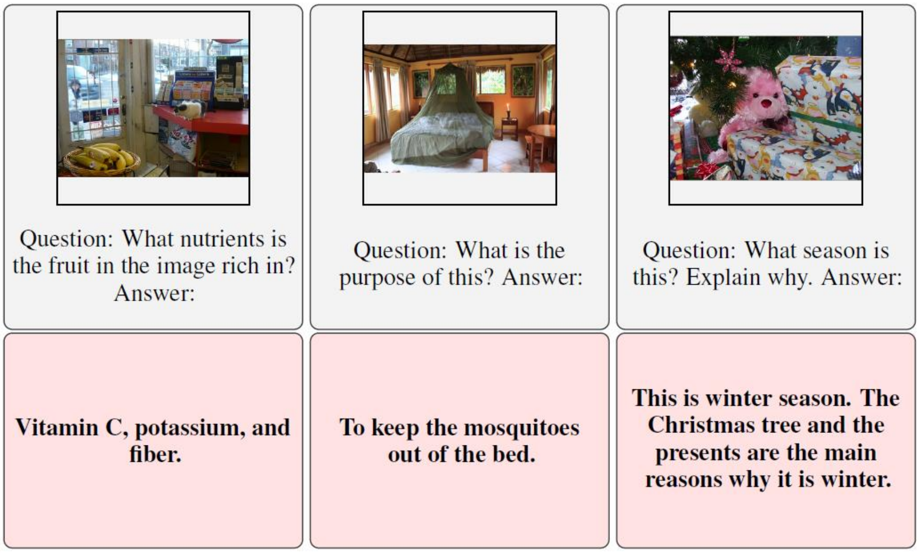
processed prompt

prompting in the format "`Question:` {question} `Answer:` {answer}"

arXiv:2204.14198

# Simple form of interaction

**input**: single image + text prompt
(question or start of a caption)
**output**: text



Question: What nutrients is the fruit in the image rich in? Answer:

**Vitamin C, potassium, and fiber.**

Question: What is the purpose of this? Answer:

**To keep the mosquitoes out of the bed.**

Question: What season is this? Explain why. Answer:

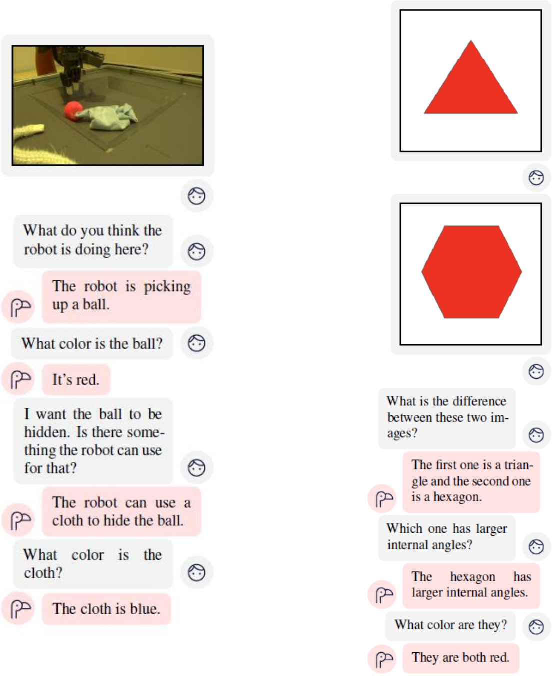
**This is winter season. The Christmas tree and the presents are the main reasons why it is winter.**

input prompt

completion

arXiv:2204.14198

# Dialogue



What do you think the robot is doing here?

The robot is picking up a ball.

What color is the ball?

It's red.

I want the ball to be hidden. Is there something the robot can use for that?

The robot can use a cloth to hide the ball.

What color is the cloth?

The cloth is blue.

What is the difference between these two images?

The first one is a triangle and the second one is a hexagon.

Which one has larger internal angles?

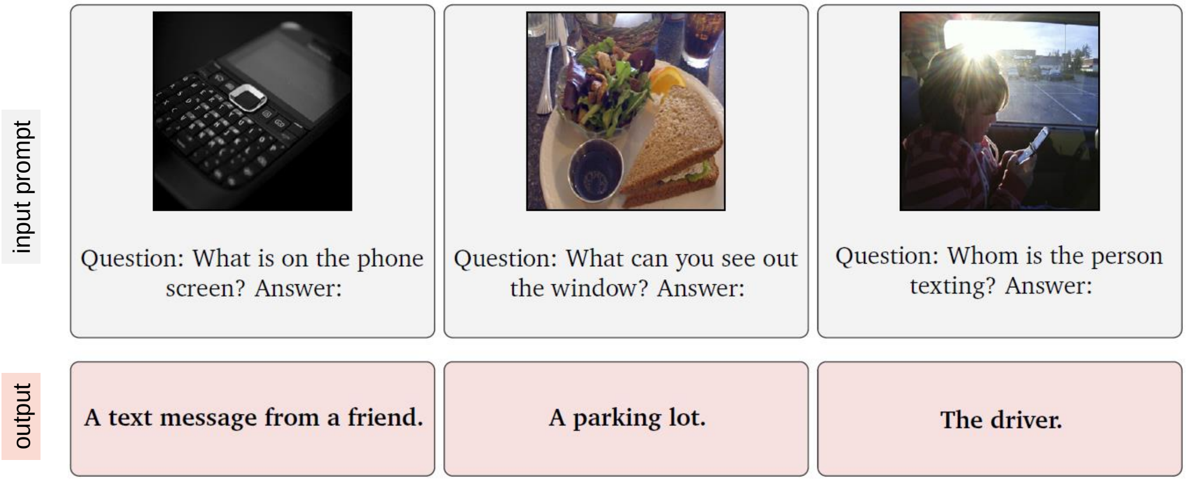
The hexagon has larger internal angles.

What color are they?

They are both red.

arXiv:2204.14198

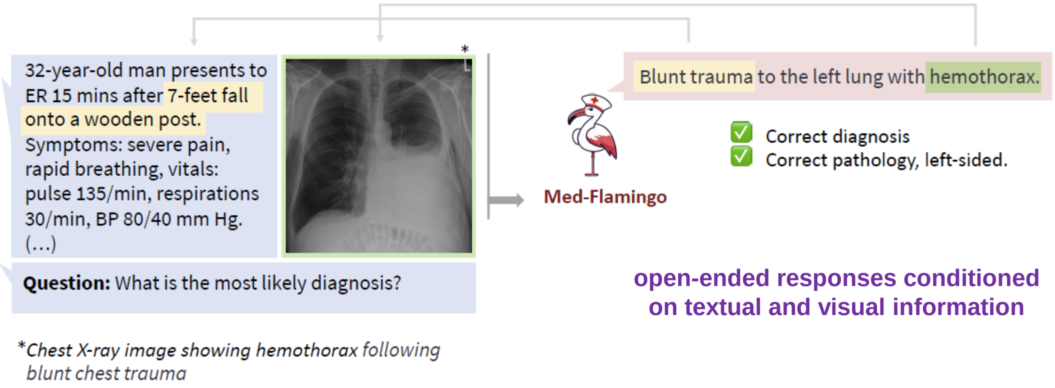
# Hallucinations



| | | | |
|---|---|---|---|
| input prompt | Question: What is on the phone screen? Answer: | Question: What can you see out the window? Answer: | Question: Whom is the person texting? Answer: |
| output | **A text message from a friend.** | **A parking lot.** | **The driver.** |

arXiv:2204.14198

# Medical generative vision-language model



32-year-old man presents to ER 15 mins after 7-feet fall onto a wooden post. Symptoms: severe pain, rapid breathing, vitals: pulse 135/min, respirations 30/min, BP 80/40 mm Hg. (…)

**Question:** What is the most likely diagnosis?

*Chest X-ray image showing hemothorax following blunt chest trauma

Blunt trauma to the left lung with hemothorax.

✓ Correct diagnosis
✓ Correct pathology, left-sided.

open-ended responses conditioned on textual and visual information

in-context learning

arXiv:2307.15189

# Exercises

# Today’s exercise

**Practice.**

You will become familiar with **multimodal transformer:** text with categorical and numerical features for classification

# What did we learn today?

- Vision transformer
- Audio transformer
- Audio-visual transformer
- Vision and language models
- Exercises

# EE-559

# Deep Learning

andrea.cavallaro@epfl.ch