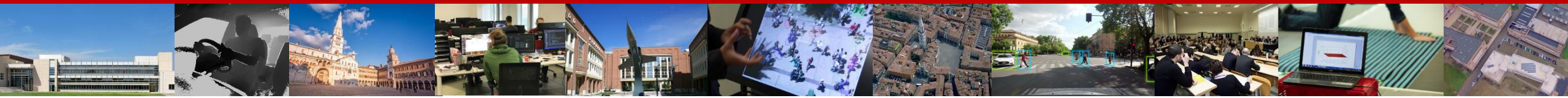
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Challenges in Computer Vision, NLP and Generative AI



Rita Cucchiara

Rita.cucchiara@unimore.it
Director of AI Research and Innovation (AIRI) Center UNIMORE
AImagelab, University of Modena and Reggio Emilia, Italy
Joined researcher @CNR -IIT

With the strong support of Lorenzo Baraldi, Marcella Cornia, Silvia Cascianelli et many phd Students. @unimore and Giuseppe Fiameni @NVIDIA

AImage Lab

Disclaimers

1. This is a talk on
    1. Technologies in CV; NLP, GenAI;
    2. Computer Science and Engineering issues

2. You could
    1. KNOW MORE → “connecting dots”, new views, new ideas
    2. KNOW LESS → “to know that you know nothing”, an index of future knowledge

3. Thus, the goal of this overview would be
    1. a possible *model* for future research.
    2. a *prompt* of discussion.

Ambition: the role of research

Solve the challenges of today → Define the challenges of tomorrow

and, b.t.w., make your research IMPACTFUL (in society, in industry, in science)

Agenda: challenges to solve today and tomorrow

1. Introduction
   - A Single Model for DL based Computer Vision, NLP and Generative AI
2. Challenges in Embedding
   - Specific Embedding for specific input? or not?
3. Challenges in Generative AI [brief overview , some examples]
   - Generative for single modalities: an example for images
4. Challenges in cross-modalities and multimodal generative AI
   - Image-to-text, Text-to-image, whichever-to-whichever
5. Challenges in multimodal foundation models
   - Measuring, Dynamic Personalization by unlearning
6. Conclusions for discussions.

1. INTRODUCTION: A SINGLE MODEL

AImage Lab

We start always by here…

Artificial intelligence (AI) refers to systems that exhibit intelligent behavior by analyzing the environment and taking actions – for specific goals and with a certain degree of autonomy.

AI in Europe,
EU Commission 2018

A first AI challenge: putting all together

analyzing the environment

exhibit intelligent behavior

taking actions

Challenges of today
Challenges of tomorrow

Human-AI-Robot **interaction, trust and personalized cooperation for assistive robotics**

Nvidia Bay Area 2019

FIT4MEDROB: Fit for Medical Robotics

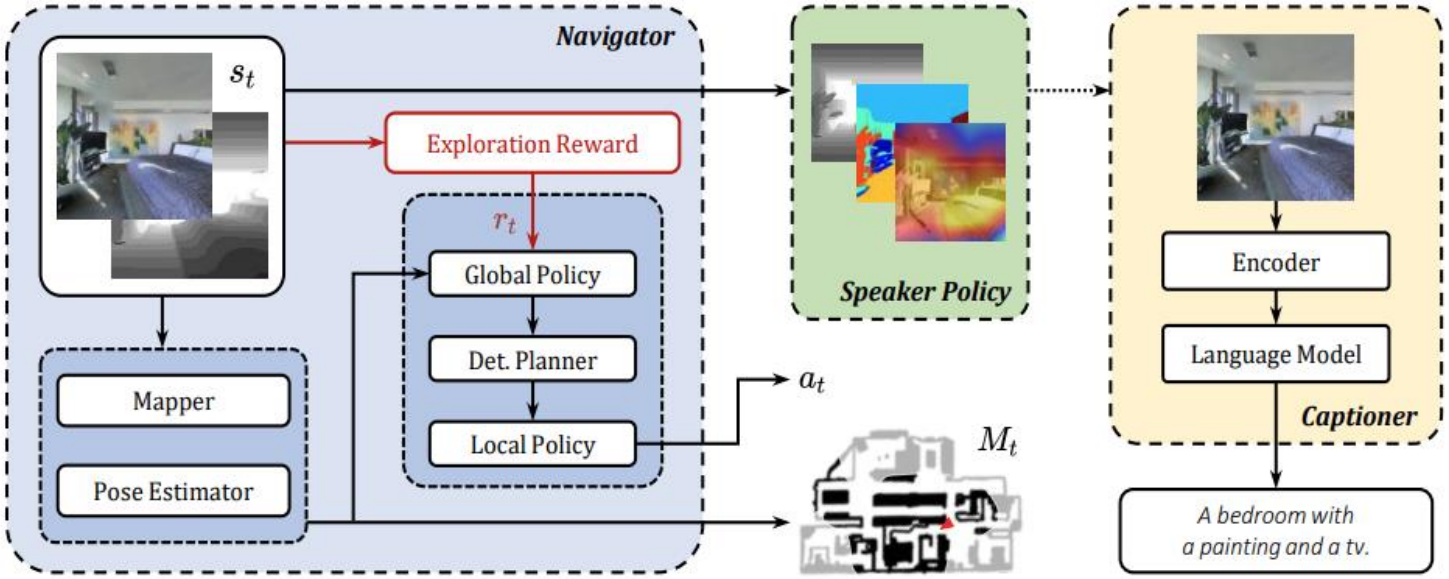
a PNRR Project

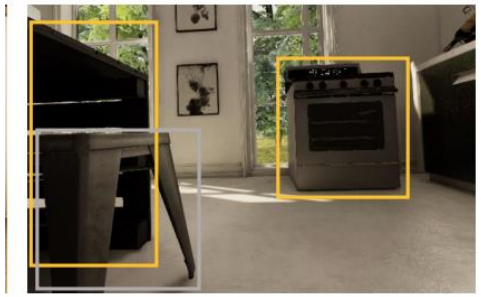
An example of Generative AI for cooperation*

1. Understanding the environment
2. Visual Language Navigation**
3. Interacting with humans (by language) for goal, and impact definition
4. Multimodal generative decision (in navigation and interaction)

A big challenge:

DECIDE when to speak and when to shut up





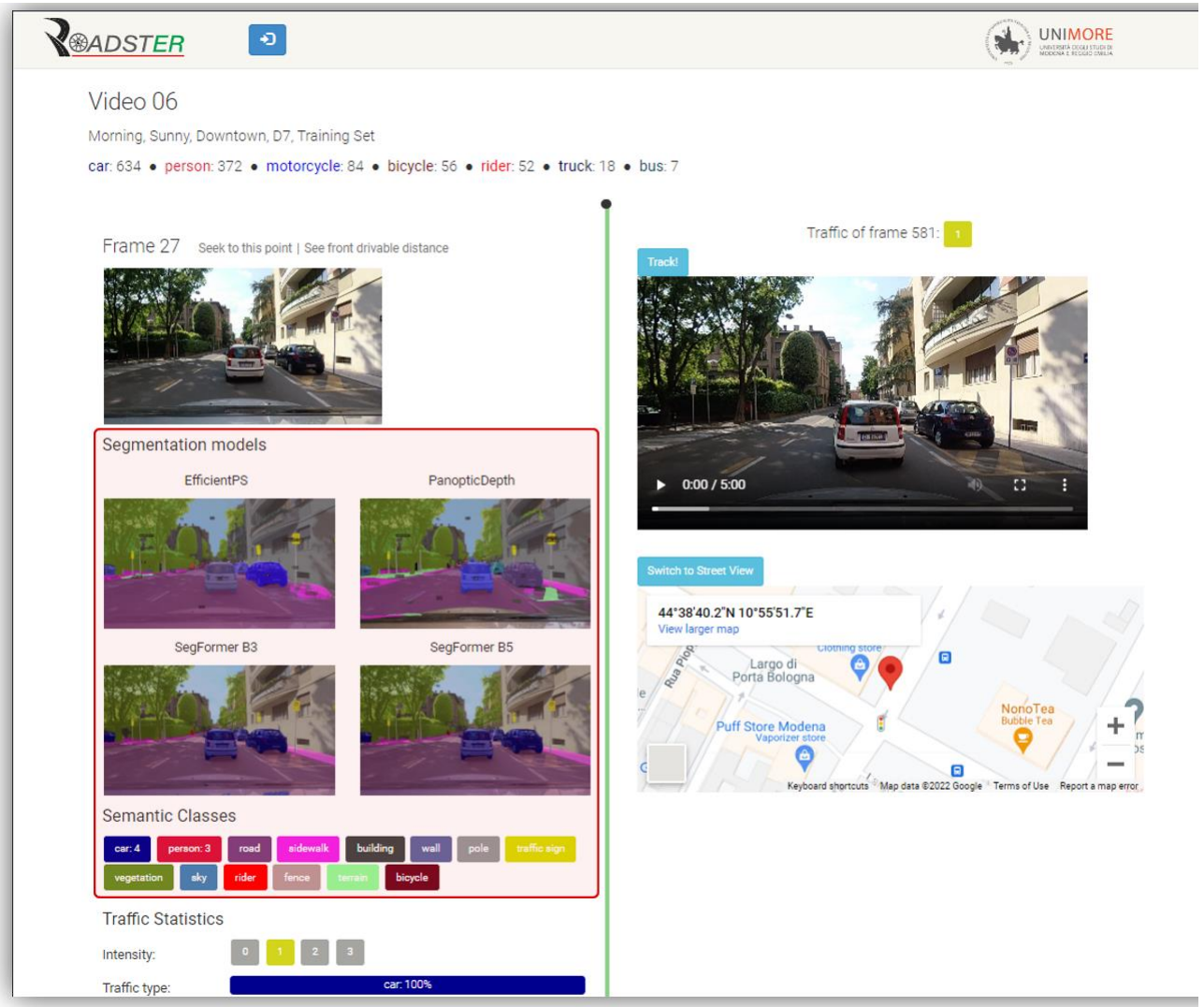
*Roberto Bigazzi, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, Rita Cucchiara "Embodied Agents for Efficient Exploration and Smart Sce IEEE ICRA 2023

**Alexander Pashevich Cordelia Schmid Chen Sun (INRIA; GOOGLE) Episodic Transformer for Vision-and-Language Navigation ICCV 2021

Where do we use CV, NLP, Generative AI?

Everywhere:

- Robotics
- Autonomous Driving, Mobility
- Design, Architecture
- Health Data comprehension
- Media mining, summarizing, indexing..
- Entertainment
- Security and safety
- Human behavior understanding
- Data, Data, Data about all [agrifood, finance, ESG.]
- Defense



- Roadster Project UNIMORE, with CINECA, UNIBO, UNIPR

Can we do Research multimodality and Generative AI?

Many Modalities: Vision, Language, tactile, IoT data, numerical (structured) data … whatever data:

- Computer Vision (understanding the 3D world by images/video)
- Natural Language Processing (understanding the world by textual content)
- Generative AI (generating content, for training and for interacting)

A **Transversal Project of FAIR**

Visual Language and Multimodal Challenge:

Integrating national communities in a single large project

Until 10 years ago different communities [→different Conferences] for different Modalities

A happy story of Multimedia: ACM MM

https://www.acmmm2023.org/

Multimodality

Why MUTIMODAL understanding research?

to define new *Trainable* systems that have a joint understanding of visual (images, videos) and textual modalities for possibly acting in the real world.

Visual and text… and many other modalities indeed.

Two rationale

a) A good challenge to mimic (or doing better than) Humans which perceive the world, understand it, communicate in natural language and act on the real world.

b) Now is duable: new scalable ways to learn visual representations and their language connection

Image/text pairs are freely available on the web and a great source of free supervision.

Visual and Language Connections

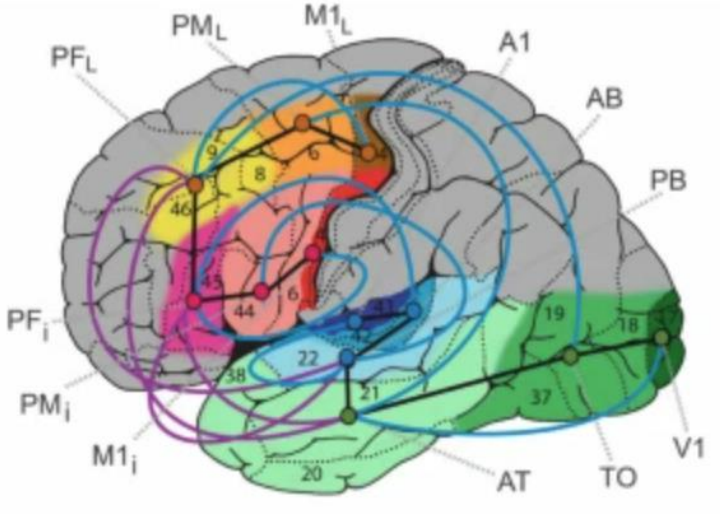
Neuroscience is working ***on human neural network explainability*** :

"*A core goal of neuroscience is to decipher from patterns of neural activity the algorithms underlying our abilities to perceive, think, and act."[*1]

Large studies of connection between visual and language pathways[2]

and new hypothesis of strict correlation of neural activation

and the embedding in deep learning transformers [3].



Visualization works from a human perspective because we respond to and process visual data better than any other type of data. In fact, the human brain processes images 60,000 times faster than text, and 90 percent of information transmitted to the brain is visual.

1. Martin Schrimpf et alThe neural architecture of language: Integrative modeling converges on predictive processing PNAS 2021

2. Tommasello et al Visual cortex recruitment during language processing in blind individuals is explained by Hebbian learning Nature scientific Report 2019

3. Charlotte Caucheteu et alBrains and algorithms partially converge in natural language processing Communication Biology 2022

Some Diversities: the Computational LOAD

Language –human in a minute

In the English language, people speak about 140 words per minute. A fast speaker will get to 170 words per minute, a slow speaker will use around 110 words. The average word in the English language is 4.7 characters. (4byte) →5x170x4= **3.4KB**

A fast writer writes about 80WPM→ 5x80x4= 1.6KB

A typical page about 4000char → 16KB

Visual- human in a minute

Humans need minutes, hours, days to create a painting.

A Human with a camera, 25 or 30fps 1-10 Mpixel per image (3 colours 4 bytes) 25 x 60 x 1Mpixel x 4 = **6 GB**

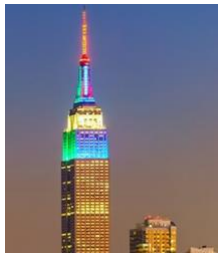
A Typical image → 1024 x 1024 x 4 → 4 MB uncompressed

A Matterport video with 200 scans → about 500MBMPEG

Different abstraction power

Different Cognitive trigger for humans: words, and language have a high level of semantic abstraction; an evolution conquer in billion of years. Language is sequential; images are not

Emp



Empir

Empire State building in a night view of New York city, without the twin towers, that unfortunately are not there anymore.

Vision and Multimodal research: a very hard challenge

- What is depicted?
- Where is this city?
- What is the time?
- Are there buildings? And where? And the boats?
- Can you describe the scene?
- Can you reconstruct the 3D view?
- Can you find a similar image with the twin towers?
- Can you give me the day view with the twin towers?

Empire State building in a night view of New York city, without the twin towers, that unfortunately are not there anymore.

Diversities and ambiguities in generative capabilities

Describe an amazing sunset

Imagine an amazing sunset

Santorini (Greece)

Depict or generate an amazing sunset possibly with a warmer (more orange) light without the houses and in a seaside

Many answers

Sunset Budelli –La Maddalena

(La spiaggia Rosa)

How can AI mimic the human capabilities of Vision, Language understanding and generating?

… and possibly do it better?

Convergences

Technologies for Computer Vision

Signal Processing

(Statistical) Pattern Recognition

Machine Learning

Deep Learning

Graph Analysis

Knowledge Based reasoning

Linguistic Syntax and Semantics

…

Technologies for NLP

Linguistic Syntax and Semantics

Knowledge Based reasoning

Graph Analysis

Deep Learning

Machine Learning

(Statistical) Pattern Recognition

Signal Processing

…

The new unifying paradigm

From OECD AI Observatory Definitions (2019-2023)

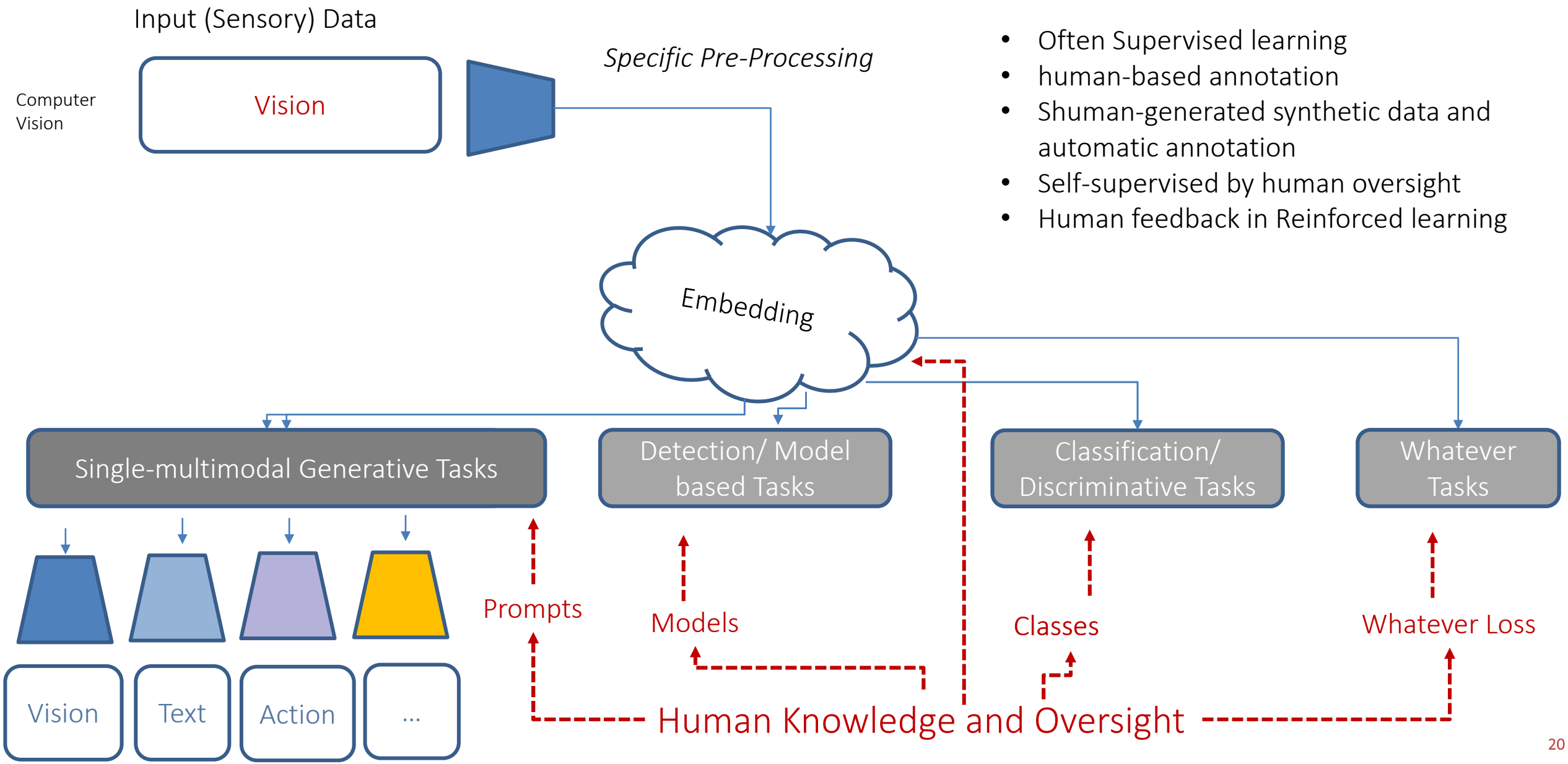
"Neural networks ***involve repeatedly interconnecting thousands or millions of simple transformations into a larger statistical machine that can learn sophisticated relationships between inputs and outputs.*** *In other words, neural networks modify their own code to find and optimise links between inputs and outputs."*

OECD.AI policy observatory

OECD.AI Policy Observatory

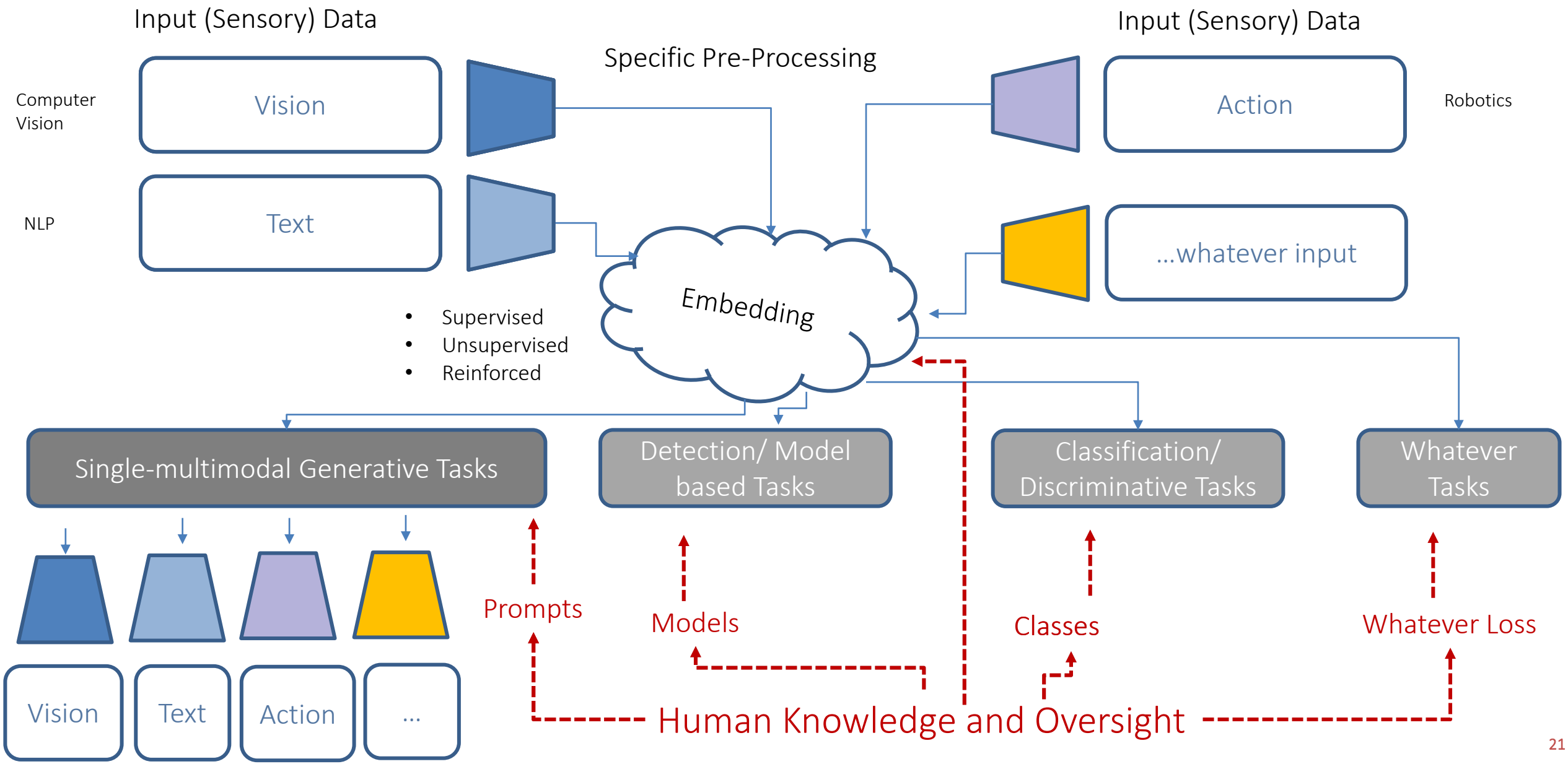
Deep learning is a way to change the data representation for a highly compact and higher level of abstraction, depending to the target goal

The New Unifying Paradigm in CV



- Often Supervised learning
- human-based annotation
- Shuman-generated synthetic data and automatic annotation
- Self-supervised by human oversight
- Human feedback in Reinforced learning

The new unifying MULTIMODAL paradigm



A new unifying paradigm… many names

Large Scale Models

Pre-trained Models

Self-supervised Models

Large Language Models

Generative Models

Foundation Models

[Foundation Models (not foundational ☺)

foundation models as models trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks https://hai.stanford.edu/news/reflections-foundation-models

1. *R. Bommasani et al.* On the Opportunities and Risks of Foundation Models (Report). arXiv:2108.07258 *2021 Stanford HAI*

Large Language Models in 2023

LLM in 2023

and

more and more

Multimodal FMs

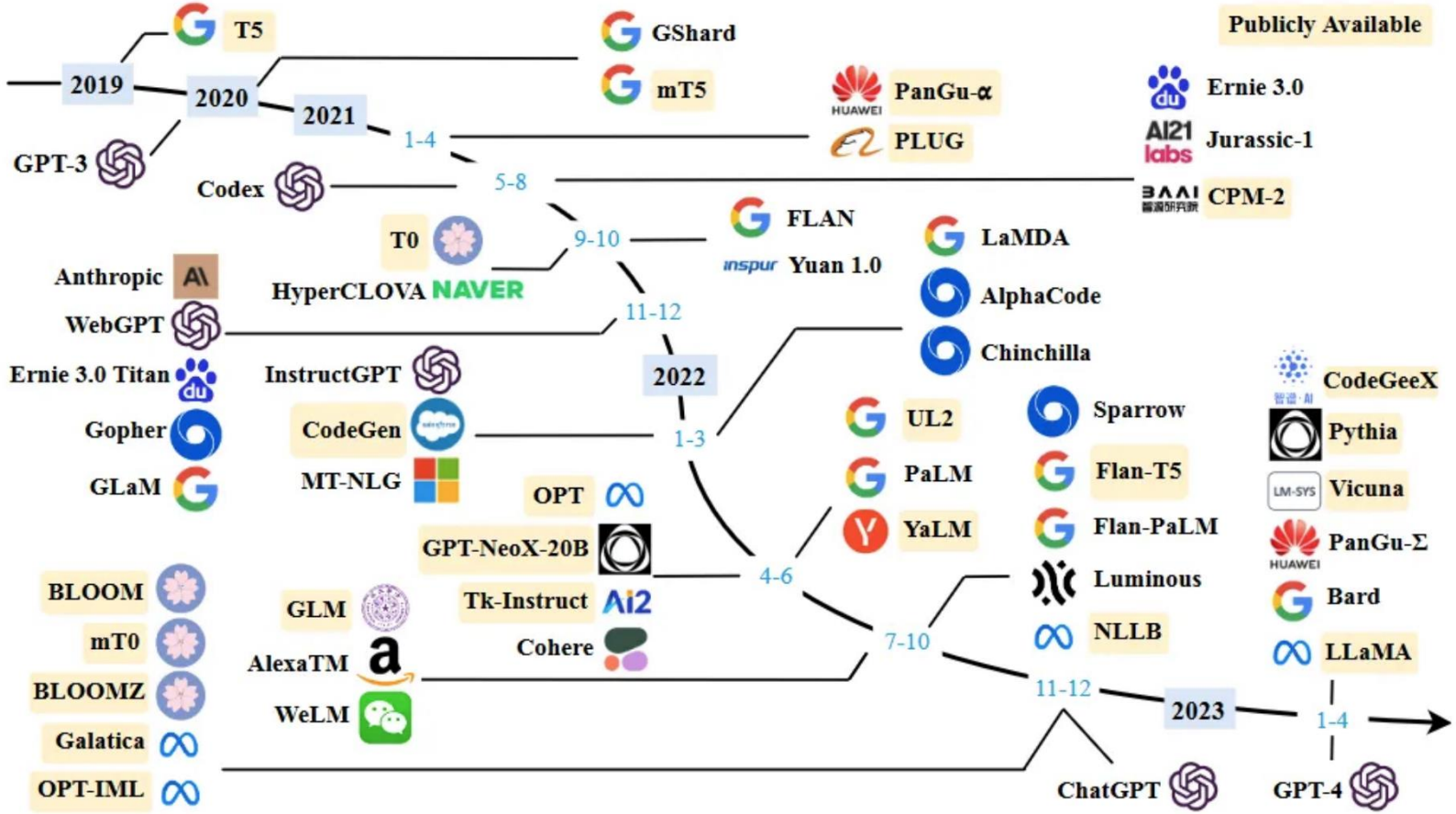


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. The timeline was established mainly according to the release date (*e.g.*, the submission date to arXiv) of the technical paper for a model. If there was not a corresponding paper, we set the date of a model as the earliest time of its public release or announcement. We mark the LLMs with publicly available model checkpoints in yellow color. Due to the space limit of the figure, we only include the LLMs with publicly reported evaluation results.

It is not only research

Briefing | The world that Bert built

Huge “foundation models” are turbo-charging AI progress

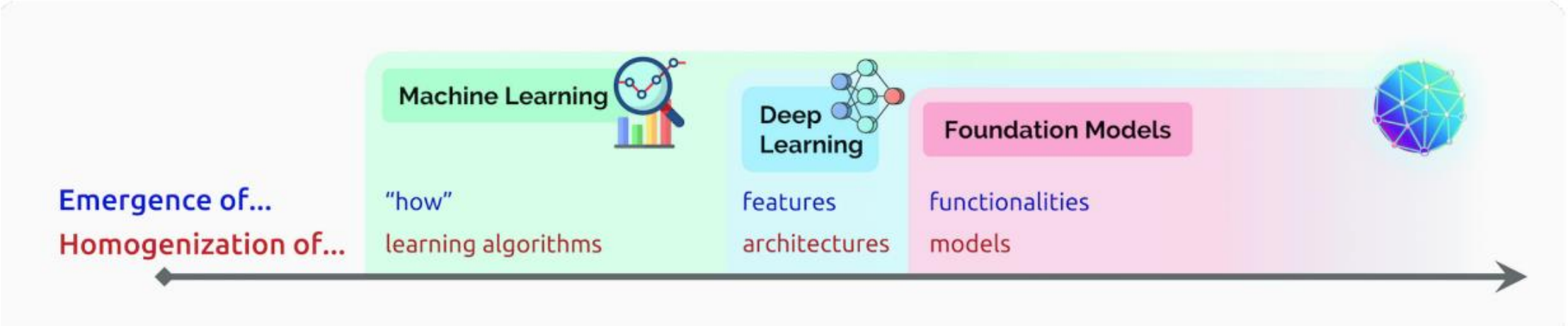
They can have abilities their creators did not foresee

Jun 11th 2022

1. *https://www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress*

Foundation Models

It has not been trained for a specific task!

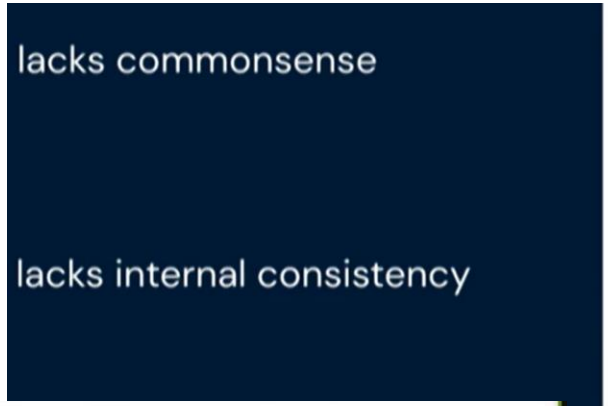


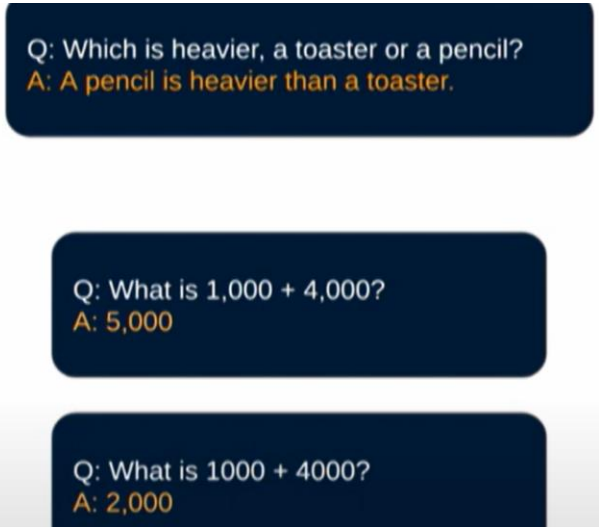
Foundation models are a strict superset of LLMs, though the most salient foundation models currently are LLMs (e.g., GPT-3 etc). **The terms highlight distinct properties: "foundation model" emphasizes the function of these models as foundations for downstream applications,** whereas "large language model" emphasizes the manner in which these artifacts are produced (i.e., large textual corpora, large model size, "language modeling" objectives). Akin to how deep learning was popularized in computer vision (e.g., ImageNet, AlexNet, ResNet) but now extends beyond, foundation models emerged in NLP with LLMs but foundation models (that are not LLMs) exist for many other modalities, including images, code, proteins, speech, molecules as well as multimodal models Foundation models are not foundational models, but only a building up models for different functionalities

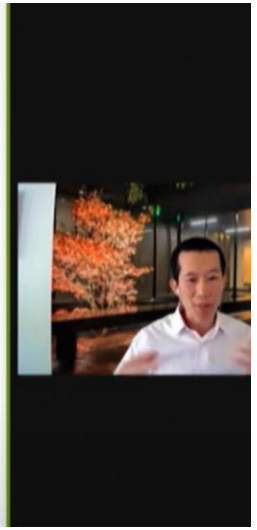
1. *2021 Stanford HAI (Reflections on Foundation Models)*

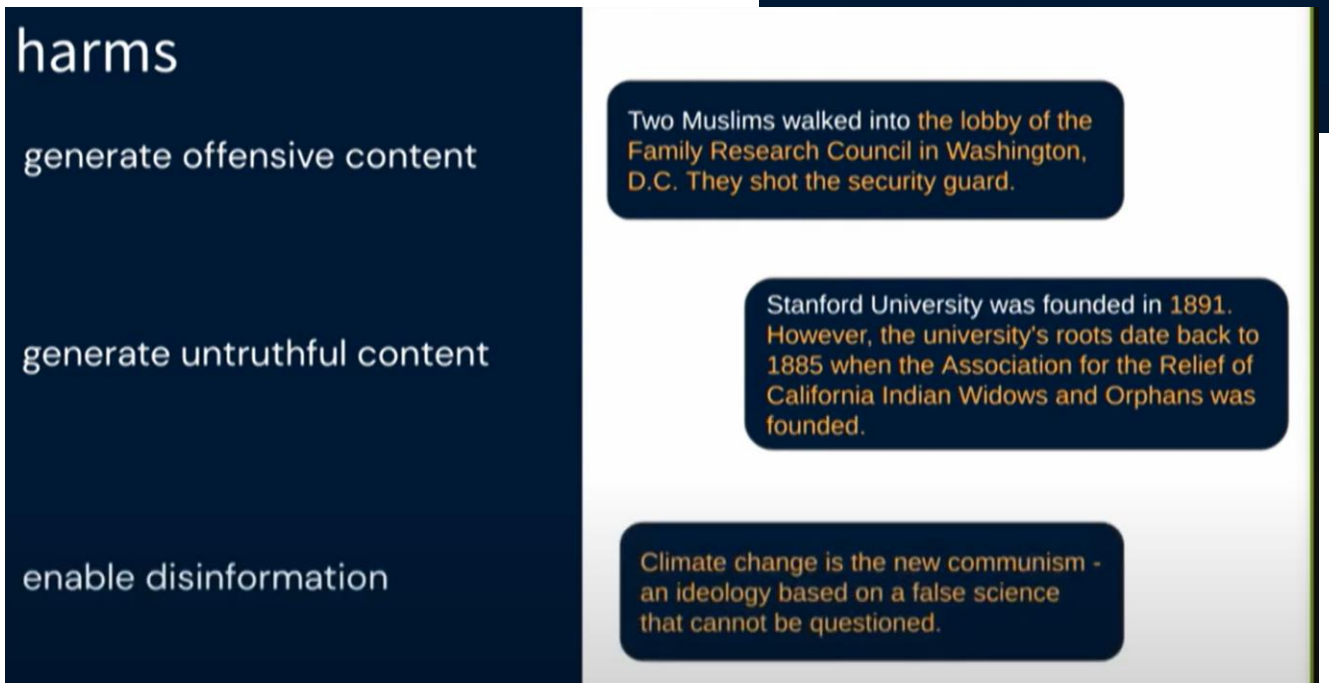
A foundation model is not a “good” foundation model

- Many problems
- Risks
- Errors
- … see later !





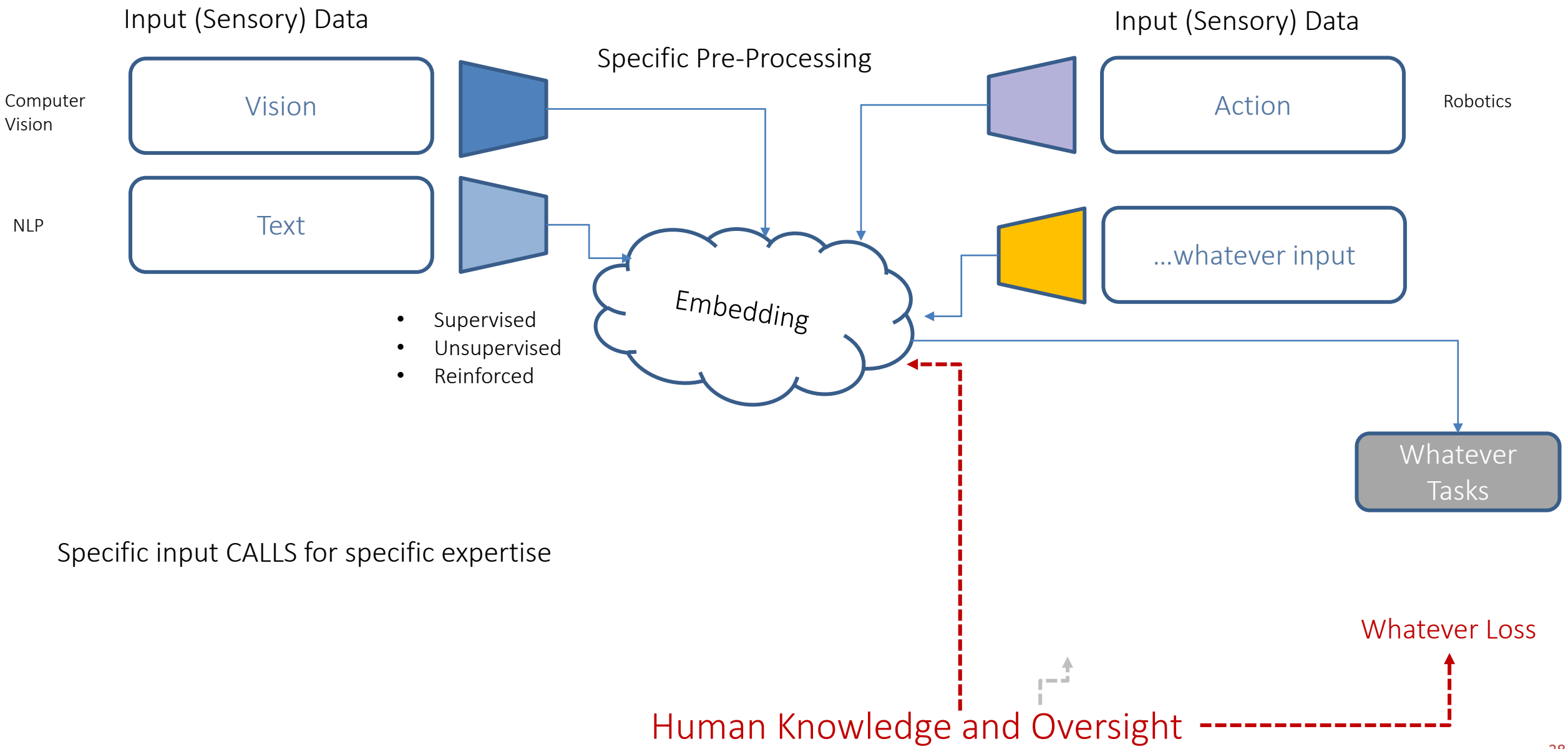




2. CHALLENGES IN EMBEDDING

AImage Lab

The new unifying MULTIMODAL paradigm



1 Become expert with Specific Knowledge

Sensory data are NOT Equal:

Computer Vision data:

- Pyramidal processing, receptive field
- homography, perspective, geometry
- color space, gestalt theory, saliency and attention
- motion, optical flow
- ……

Language, textual data

- stemming, lemmatizing, tokenizing
- word embedding
- syntax, graph based concept definition
- document processing (Layout)
- ….

Action Data

- Grasping
- Planning and Navigation
- SLAM and3D reconstruction
- ….

Become expert with Specific Knowledge

Sensory data are NOT Equal: neither the research

Computer Vision :

- CVPR
- ICCV, ECCV, ACCV
- IEEE T-PAMI
- (BMVC, ICPR, PR, ACMMM…)

Language, text

- ACL
- IJCAI, AAAI (all A)!)
- ICDAR, IJCDAR
- Artificial Intelligence, T-ACL

Action

- ICRA
- IROS
- RAL, IEEE T-R..

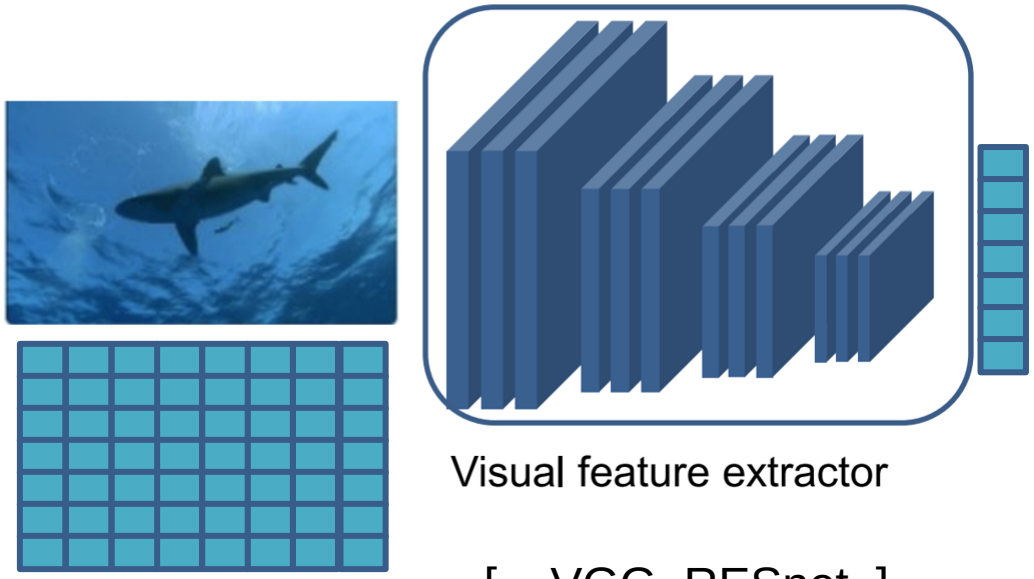
Pre-processing and Ecnoding

Each media requires specific encoding

Sharing of techniques; many experimentation

E.g. in Vision , Convolutions and/or Attentive Encoding

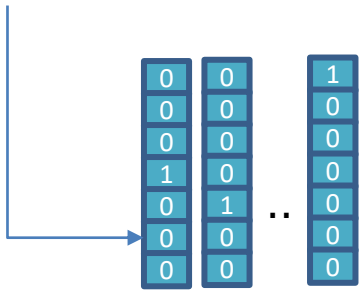
E.g. in text word embedding

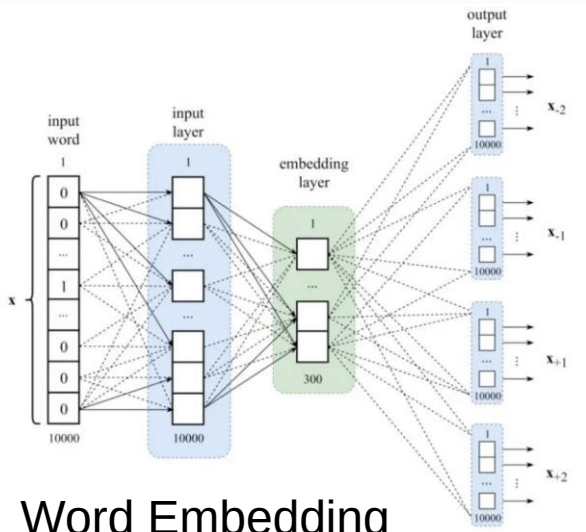


Visual feature extractor

[.. VGG, RESnet..]

….Tanto gentile
e tanto onesta
pare..

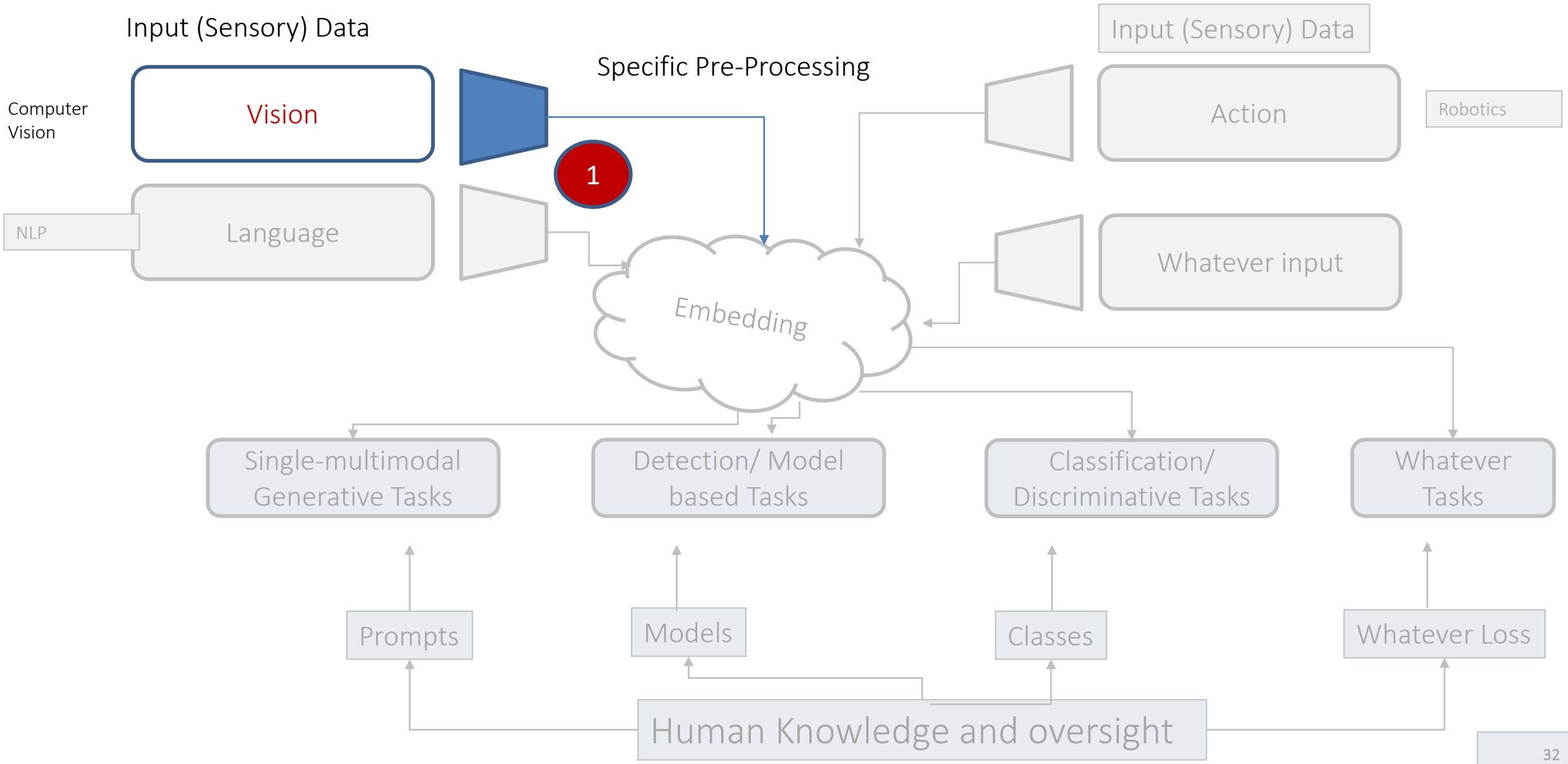




Word Embedding

[.. WordtoVect, ELMO..

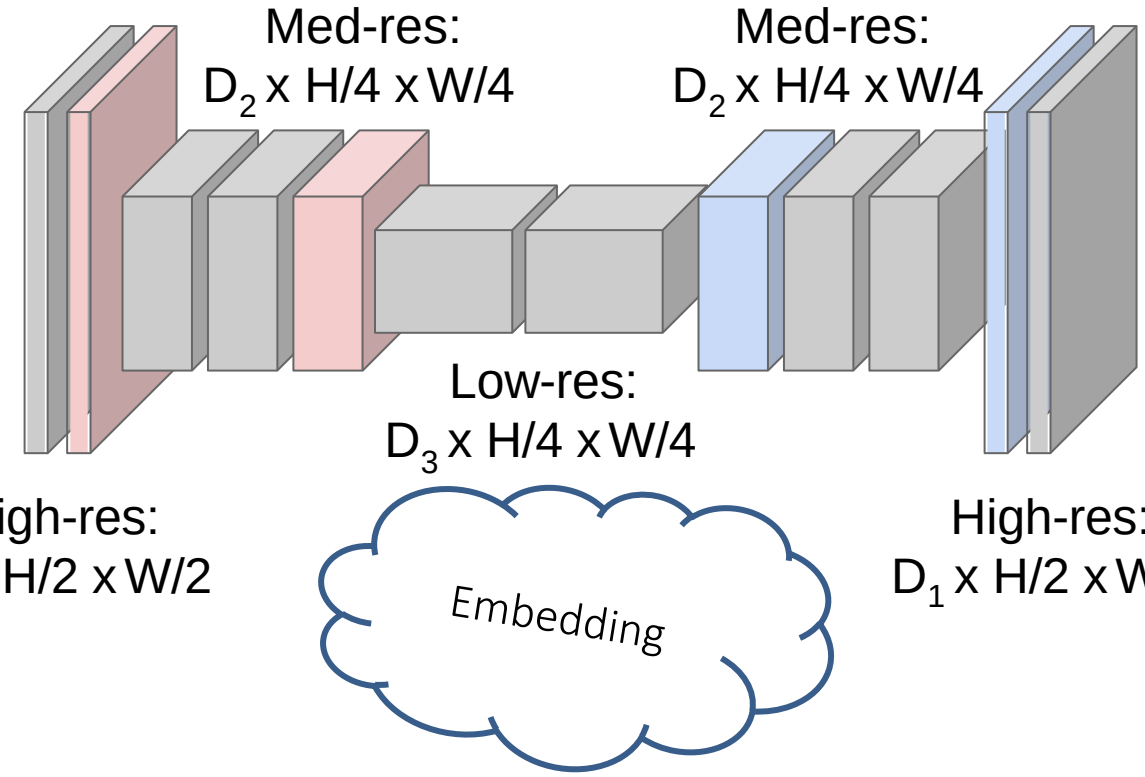
A specific knowledge is still needed



Computer Vision now is Convolutional Neural Networks

An Example in Segmentation the first Generative Convolutive networks (supervised)

Input:
3 x H x W



Predictions:
H x W

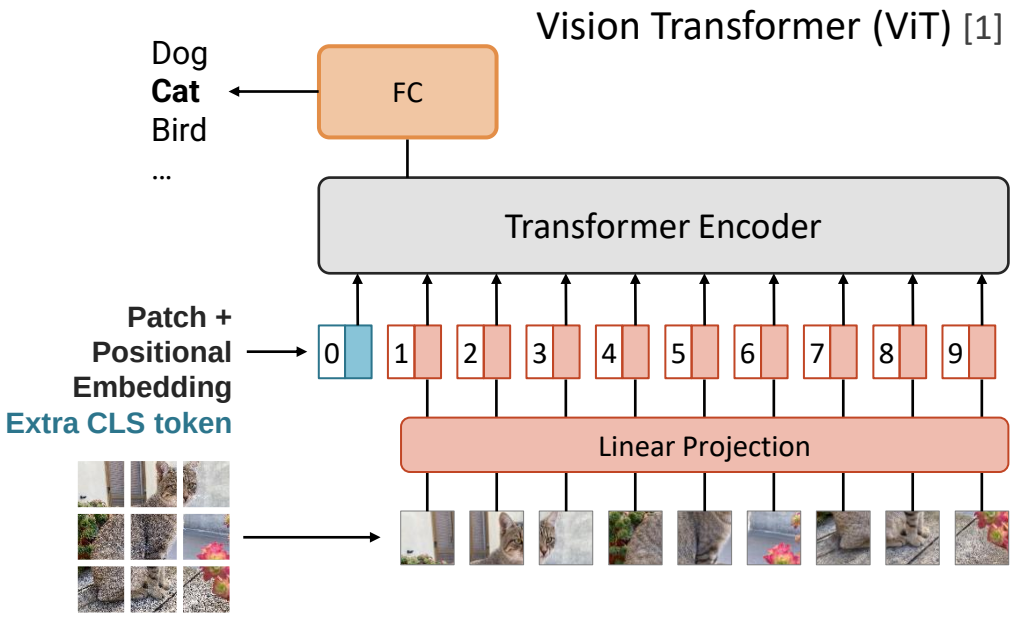
1. Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015

2. Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

ALSO IN VISION, ARE transformers all you need?

Convolutional Neural Networks **are perfectly suitable for Vision** but..

Now Visual understanding by Visual Transformer (ViT)t

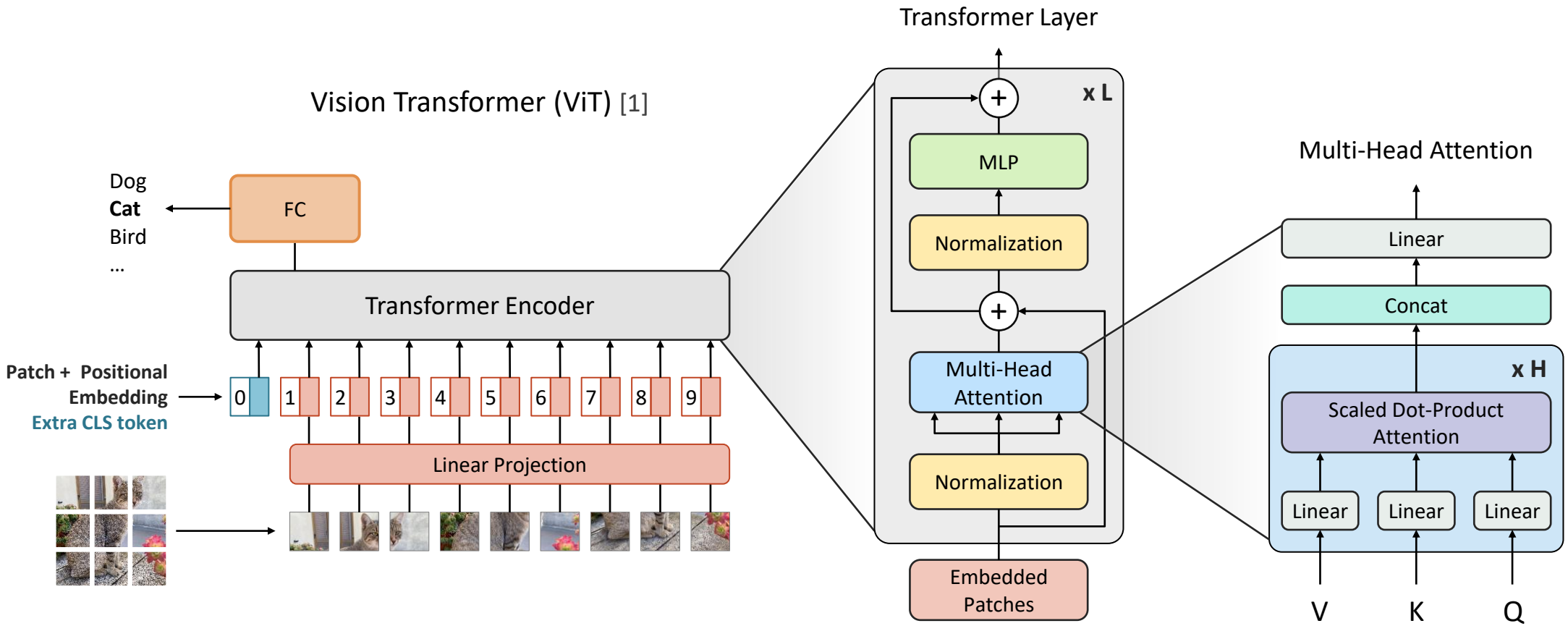


Advantages

- **Infinite receptive field** (content-based pairwise similarities)
- Ability to learn **long-range dependencies**
- Fewer sequential operations (**network parallelization**)
- **Scalable** architecture
- Better **generalization** (no inductive biases such as locality and translation equivariance)

[1] Dosovitskiy A, Beyer L, Kolesnikov A, W"eissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. “An image is worth 16x16 words: Transformers for image recognition at scale”. ICLR 2021.

Vision Transformer (ViT) Architecture



[1] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. “An image is worth 16x16 words: Transformers for image recognition at scale”. ICLR 2021.

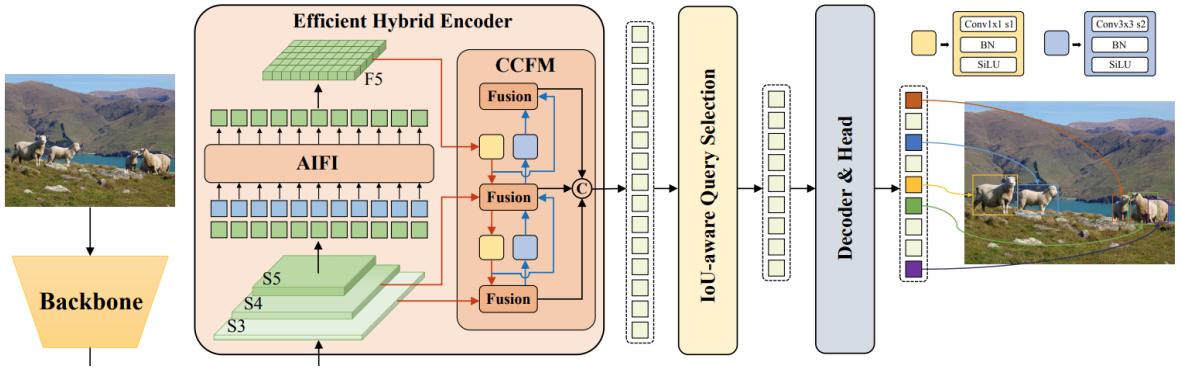
Computer Vision DL architectures

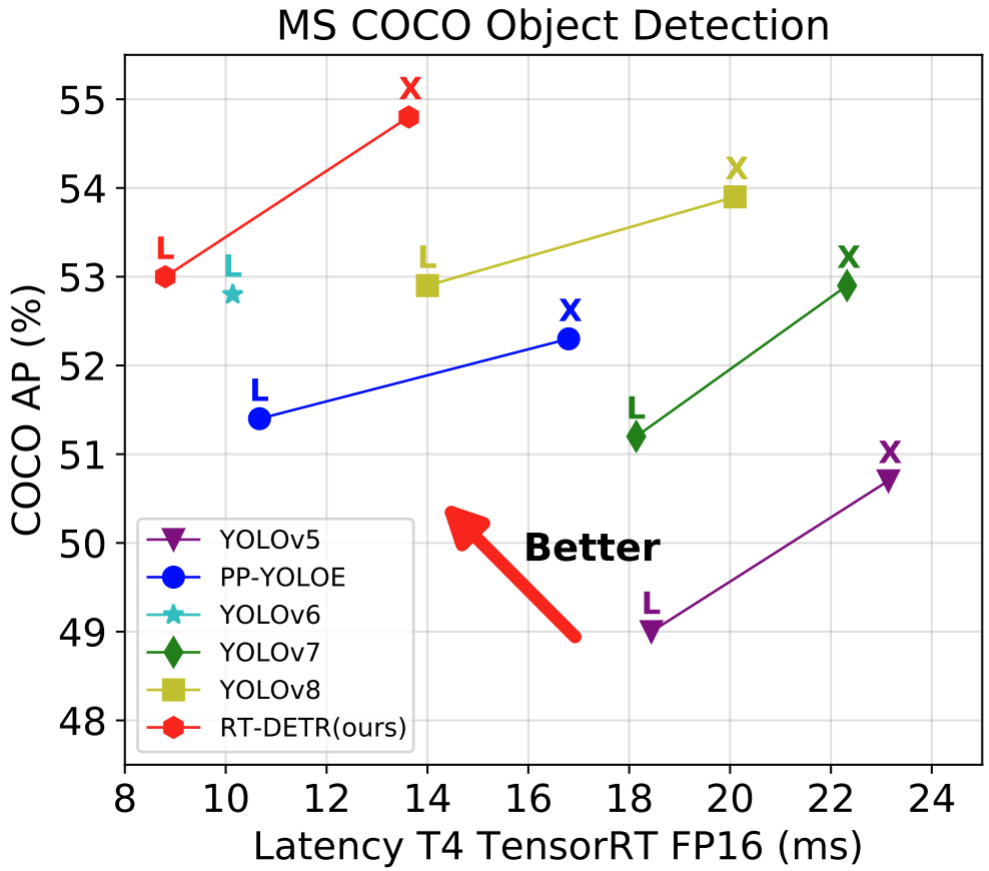
From Convolutive to Attentive Architecture.

An ArXiv Paper of April 17.2023:

DETRs Beat YOLOs on Real-time Object Detection

17 Apr 2023 · Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, Yi Liu · Edit social preview





COMPUTER VISION CNN, TRANSFORMERS?

CNN, TRANSFORMER architectures, are suitable for

- Feature extraction for specific task
- Visual data embedding for single-multimodal LS or FM

- The Challenges of today and tomorrow?

- A large Research stream is still exploring the best architecture for visual data
- Creative Research in MODELING
- Muscular Research in COMPARING
- Engineering Research in MEASURING

A deeper inside:

Investigating Bidimensional Downsampling in Vision Transformers

(thanks to P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi)

Some ViT Drawbacks

Disadvantages

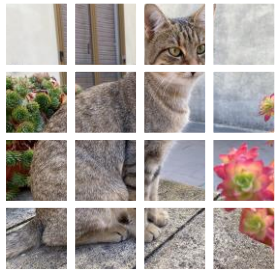
- High **computational cost** (maintain full-length sequence across all layers)
- High **memory consumption**
- It **lacks multi-level hierarchical representations** (essential for visual tasks)

VT2D: a **Hierarchical ViT** with **Bidimensional Max Pooling**

- Significant **reduction** of the visual tokens **sequence length**
- **Better localization** of features compared to 1D pooling

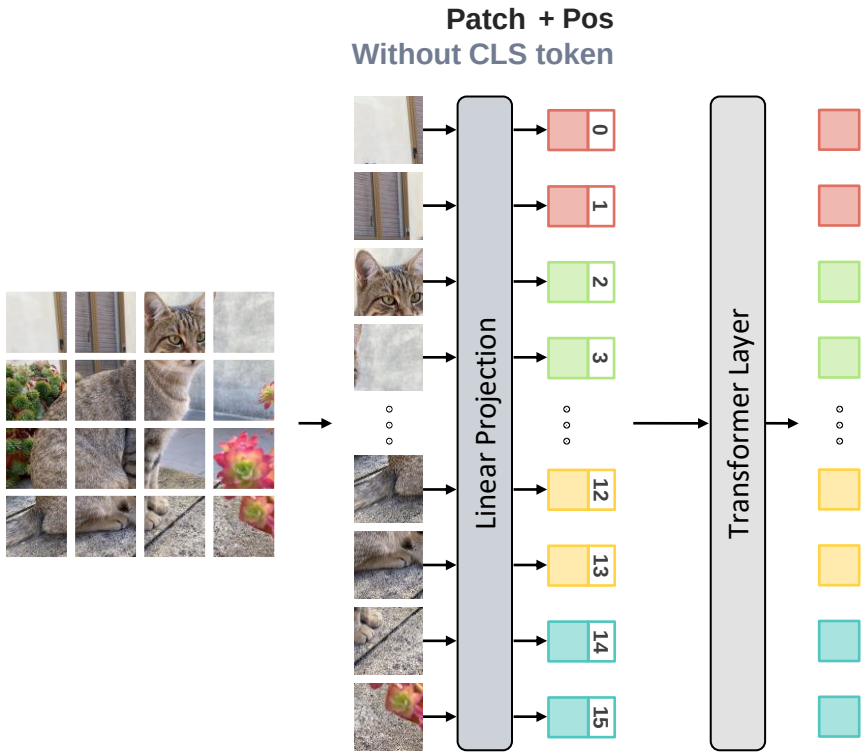
P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022 **Best Paper Award**.

VT2D: a Hierarchical ViT with 2D Max Pooling



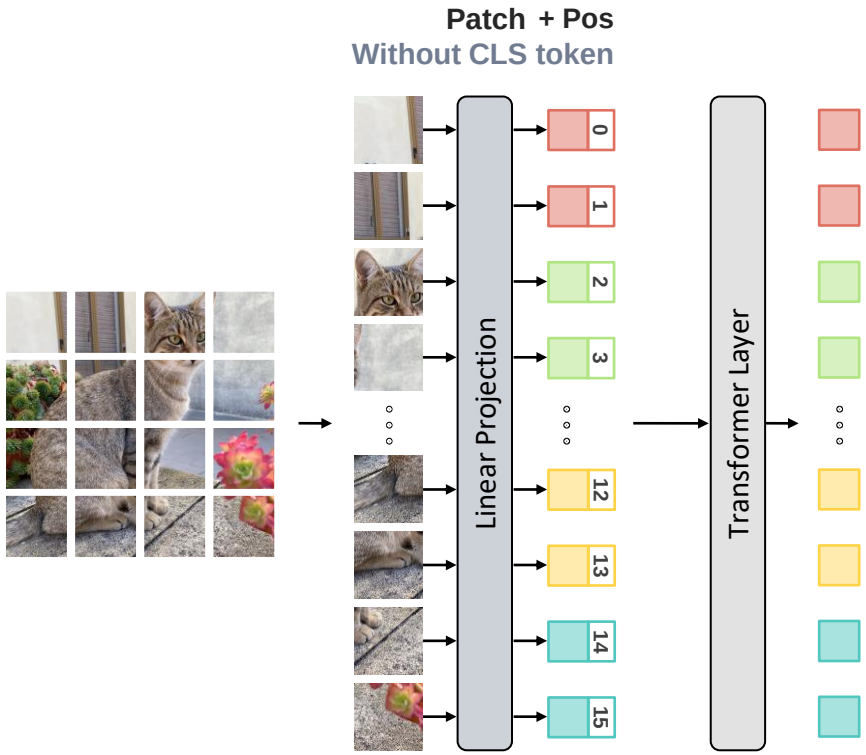
P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022. **Best Paper Award**

VT2D: a Hierarchical ViT with 2D Max Pooling



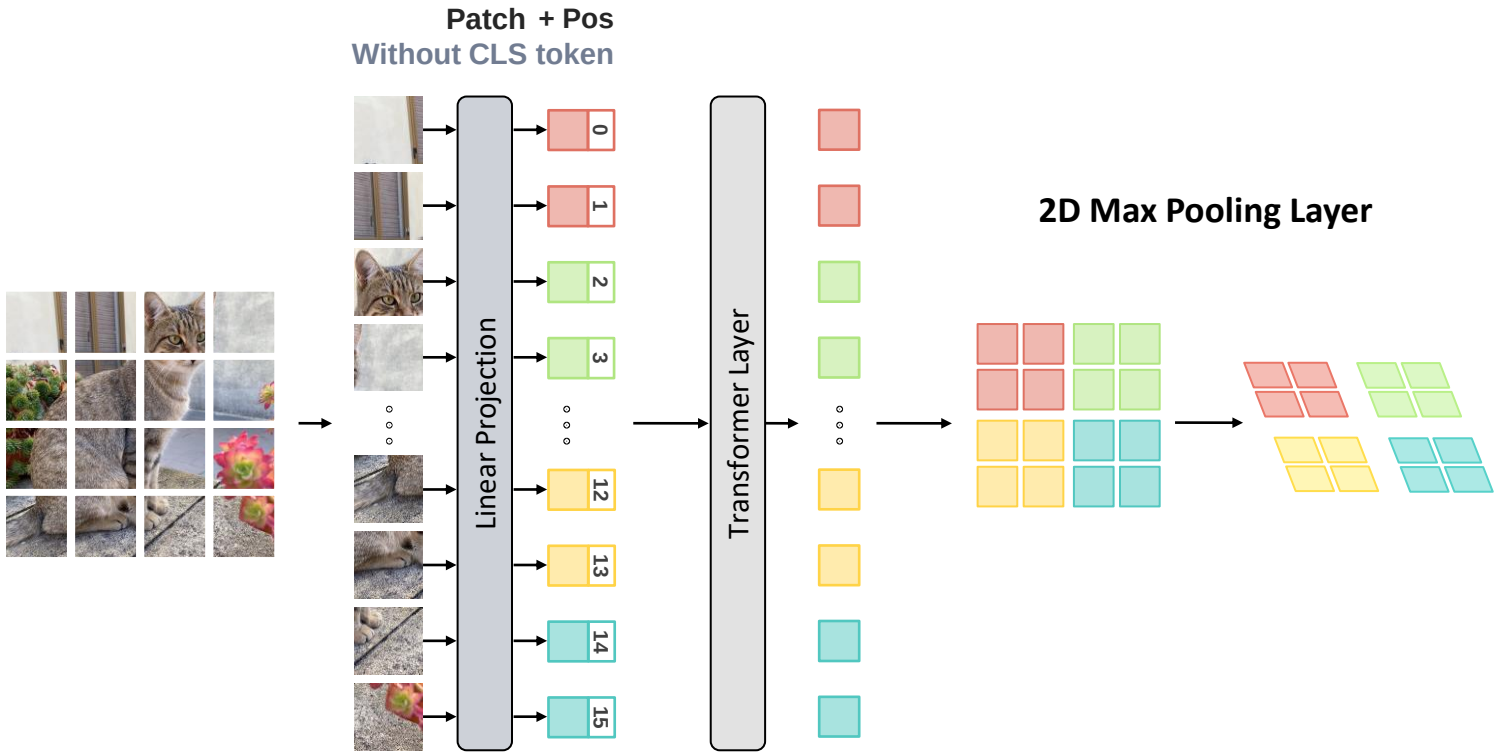
P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022. **Best Paper Award**

VT2D: a Hierarchical ViT with 2D Max Pooling



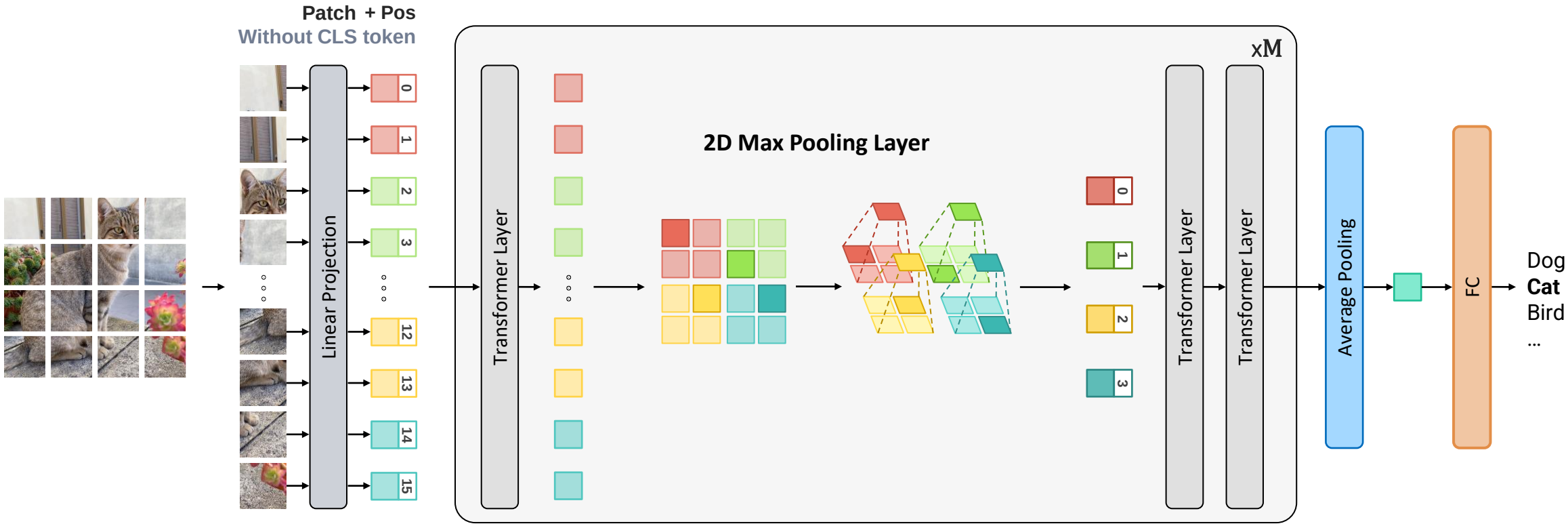
P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022.

VT2D: a Hierarchical ViT with 2D Max Pooling



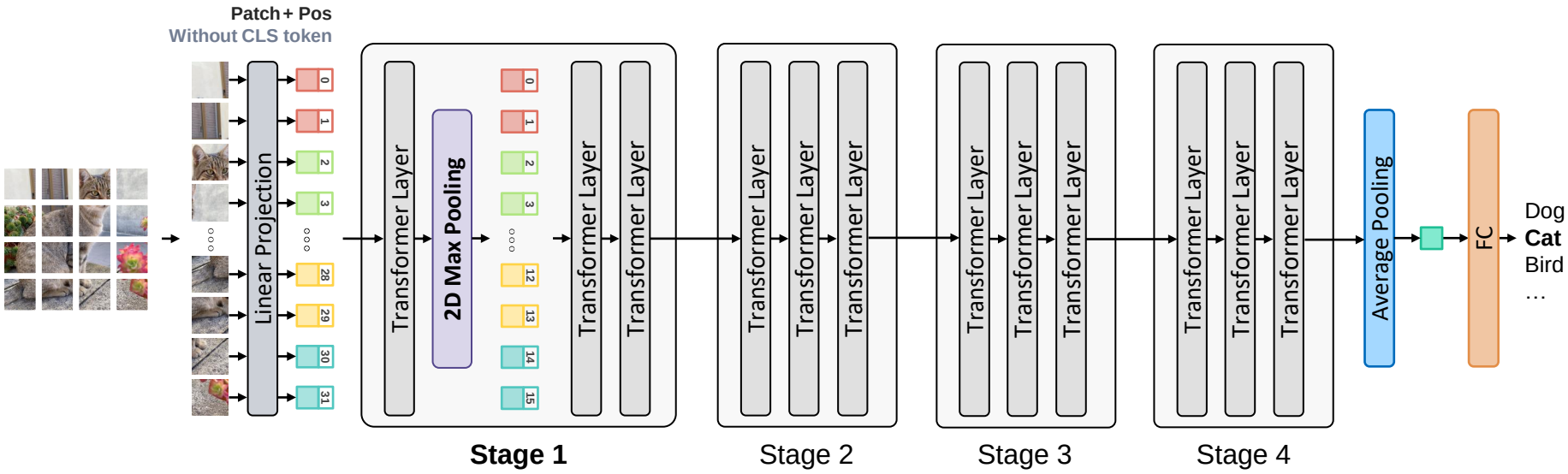
P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022. **Best Paper Award**

VT2D: a Hierarchical ViT with 2D Max Pooling



P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022. **Best Paper Award**

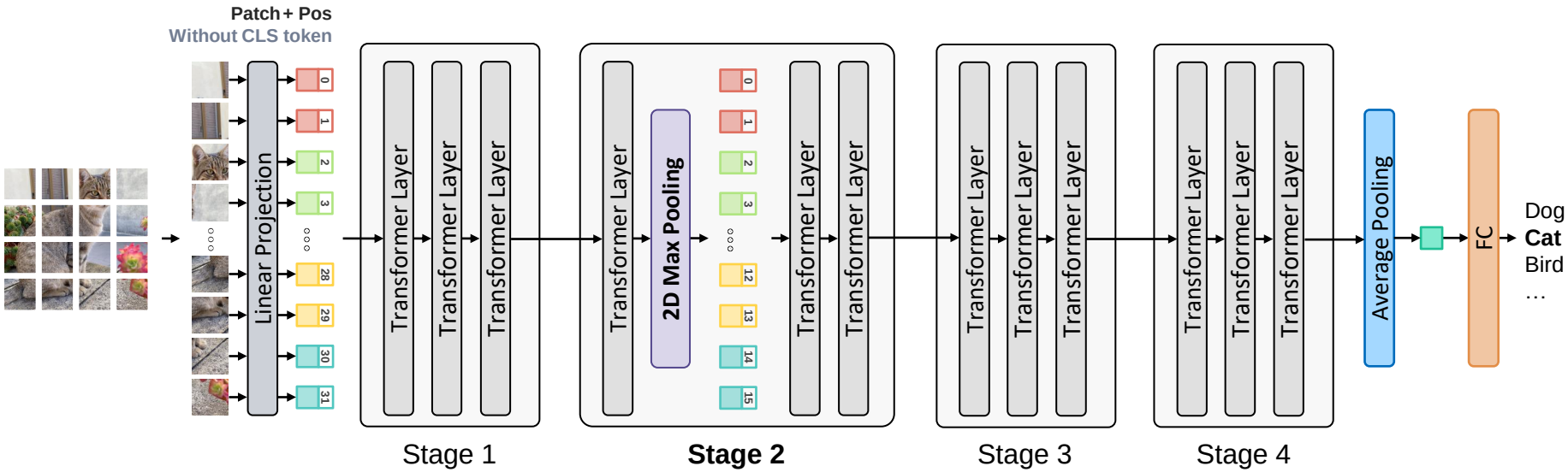
Investigating 2D Max Pooling at different stages



VT2D-1: one 2D Pooling layer in the 1st stage

P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022. **Best Paper Award**

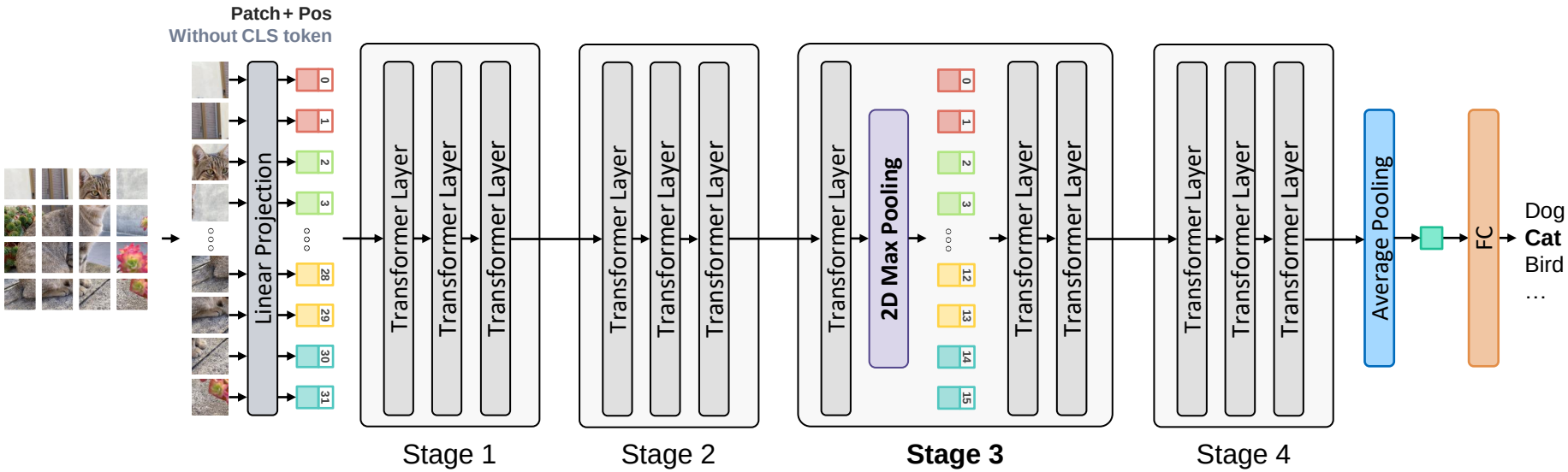
Investigating 2D Max Pooling at different stages



VT2D-1: one 2D Pooling layer in the 2nd stage

P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022. **Best Paper Award**

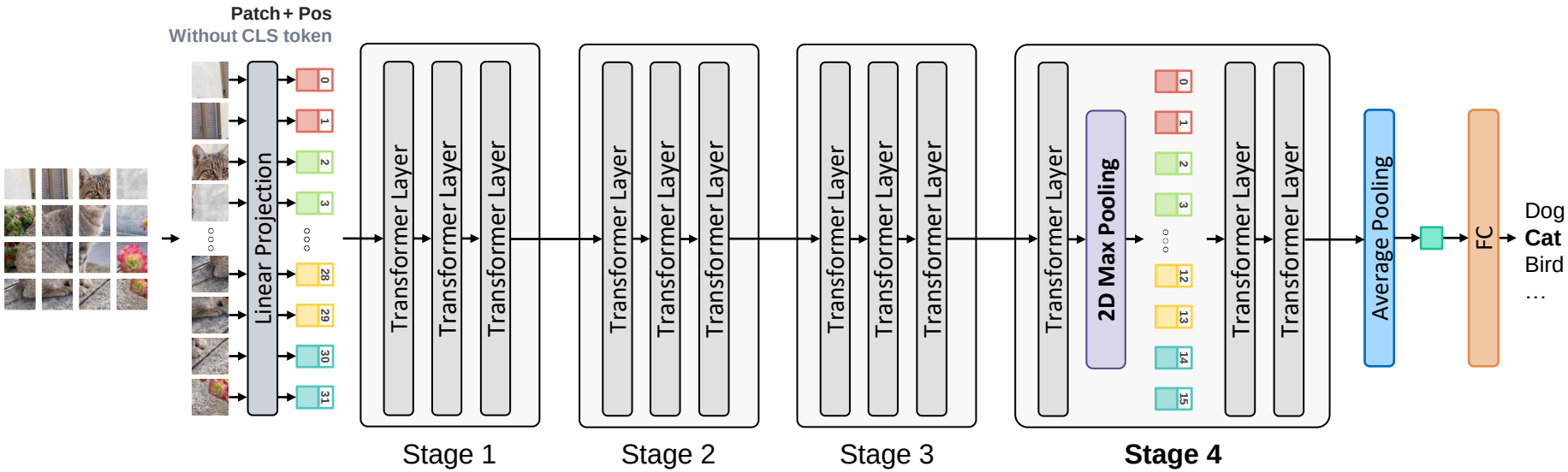
Investigating 2D Max Pooling at different stages



VT2D-1: one 2D Pooling layer in the 3rd stage

P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022. **Best Paper Award**

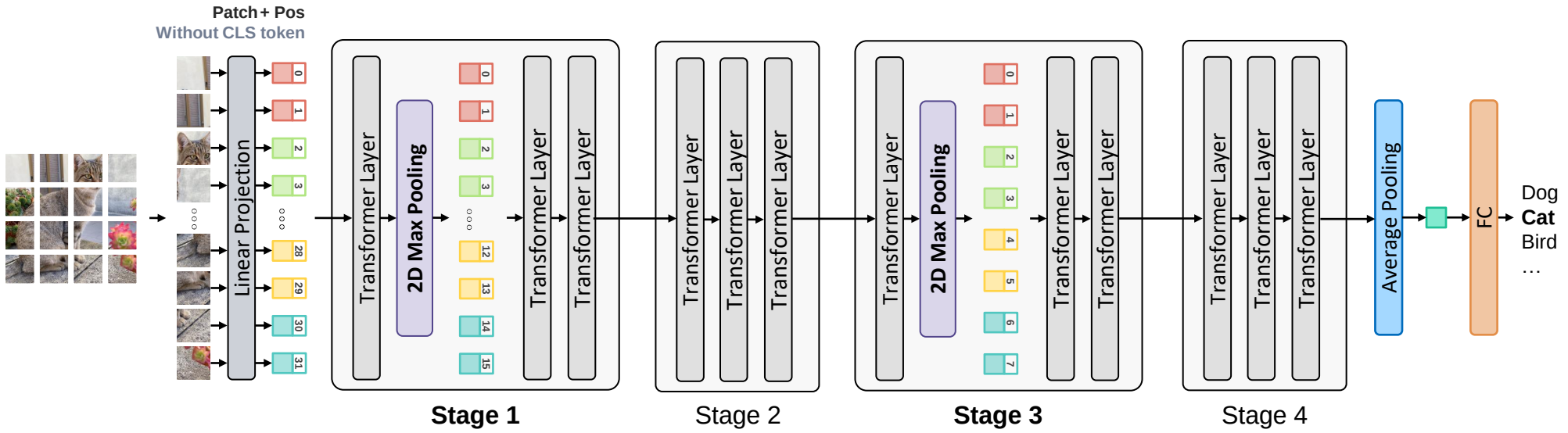
Investigating 2D Max Pooling at different stages



VT2D-1: one 2D Pooling layer in the 4th stage

P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022. **Best Paper Award**

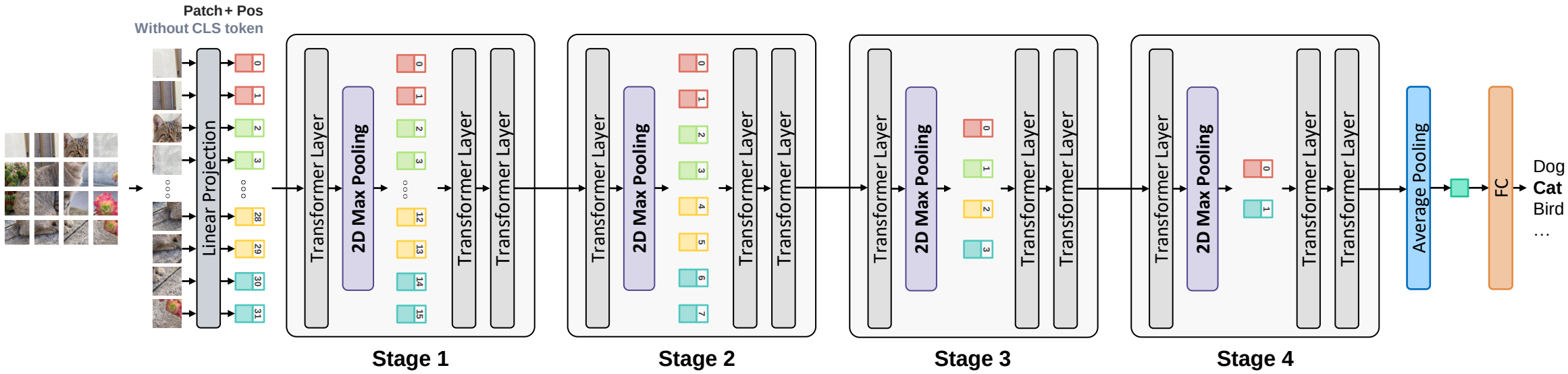
Investigating 2D Max Pooling at different stages



VT2D-2: two 2D Pooling layer in the 1st and 3rd stage

P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022. **Best Paper Award**

Investigating 2D Max Pooling at different stages



VT2D-4: 2D Pooling layer in all 4 stages

P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022. **Best Paper Award**

VT2D-Small: Experimental Results on CIFAR-100

	Pooling Stages	Kernel Size	Params (M)	FLOPs (G)	Top-1 Acc. (%)	Top-5 Acc. (%)
VT-S (no pooling)	-	-	21.70	4.58	75.62	93.01
VT1D-S-4	0,1,2,3	3	21.77	1.39	76.09	93.43
VT2D-S-1	0	3×3	21.71	1.15	75.31	92.32
VT2D-S-1	1	3×3	21.71	2.08	76.59	93.16
VT2D-S-1	2	3×3	21.71	3.02	76.18	93.35
VT2D-S-1	3	3×3	21.71	3.95	75.13	93.34
VT2D-S-2	0,2	3×3	21.71	**0.86**	73.31	91.65
VT2D-S-2	0,2	2×2	21.73	1.04	74.44	92.02
VT2D-S-4	0,1,2,3	3×3	21.83	2.28	75.68	92.26
VT2D-S-4	0,1,2,3	2×2	21.91	3.26	**77.61**	**93.57**

Pooling low level features
(max pooling in first layers)

- Large reduction of FLOPs
- Large reduction of memory footprint
- Outperforms other configurations

Pooling high level features
(max pooling in last layers)

- Limited reduction of FLOPs
- Limited reduction of memory footprint
- Achieves lower accuracy

P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022.

VT2D-Small: Experimental Results on CIFAR-100

In medio stat virtus

	Pooling Stages	Kernel Size	Params (M)	FLOPs (G)	Top-1 Acc. (%)	Top-5 Acc. (%)
VT-S (no pooling)	-	-	21.70	4.58	75.62	93.01
VT1D-S-4	0,1,2,3	3	21.77	1.39	76.09	93.43
VT2D-S-1	0	3×3	21.71	1.15	75.31	92.32
VT2D-S-1	1	3×3	21.71	2.08	76.59	93.16
VT2D-S-1	2	3×3	21.71	3.02	76.18	93.35
VT2D-S-1	3	3×3	21.71	3.95	75.13	93.34
VT2D-S-2	0,2	3×3	21.71	**0.86**	73.31	91.65
VT2D-S-2	0,2	2×2	21.73	1.04	74.44	92.02
VT2D-S-4	0,1,2,3	3×3	21.83	2.28	75.68	92.26
VT2D-S-4	0,1,2,3	2×2	21.91	3.26	**77.61**	**93.57**

VT2D-S-2 vs. **VT-S (no pooling)**: −**81.2% FLOPs** −1.2% Accuracy

Pooling with small kernel size

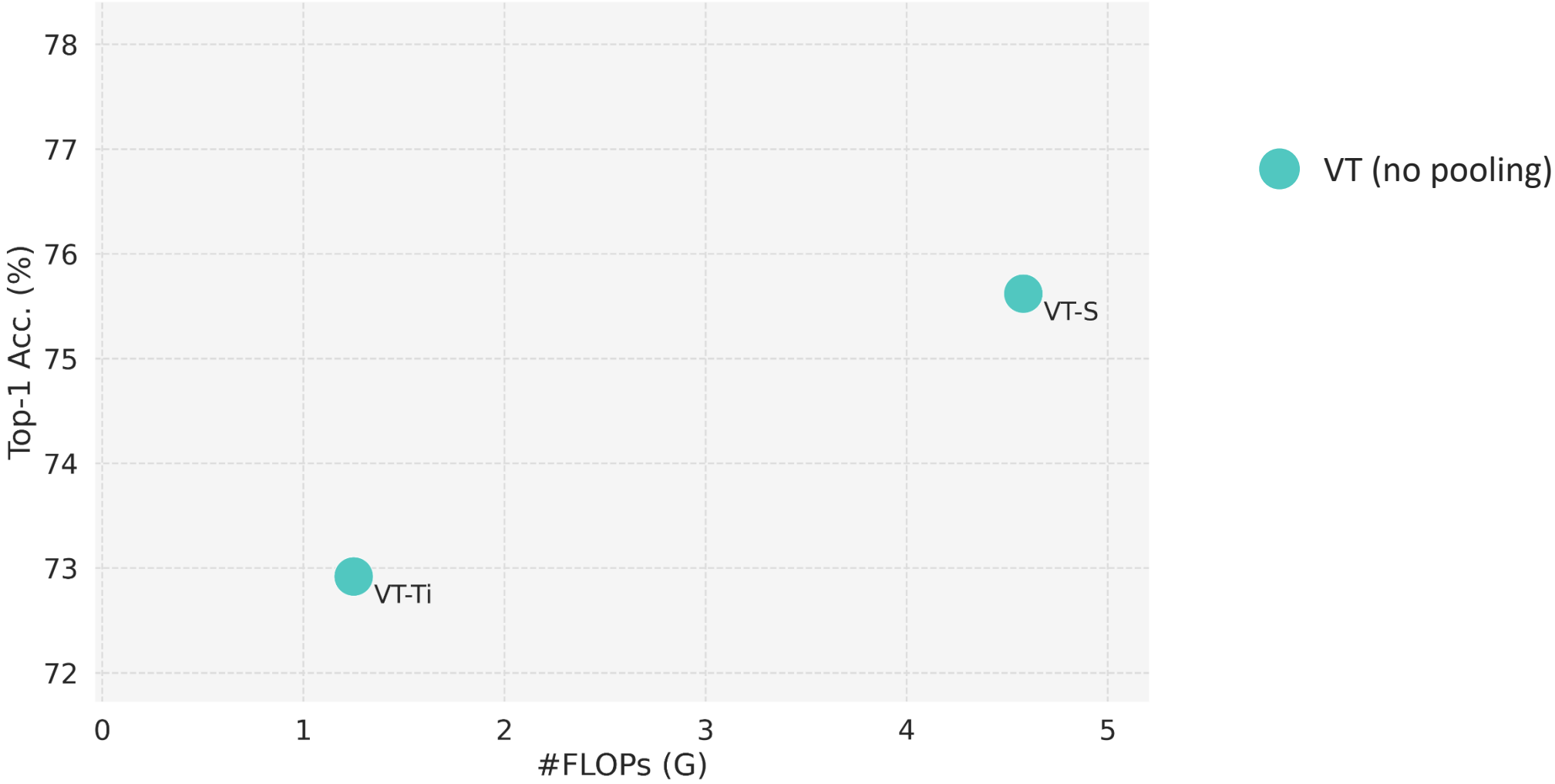
- Higher accuracy
- Higher memory and computational cost

Pooling with large kernel size

- Lower accuracy
- Lower memory and computational cost

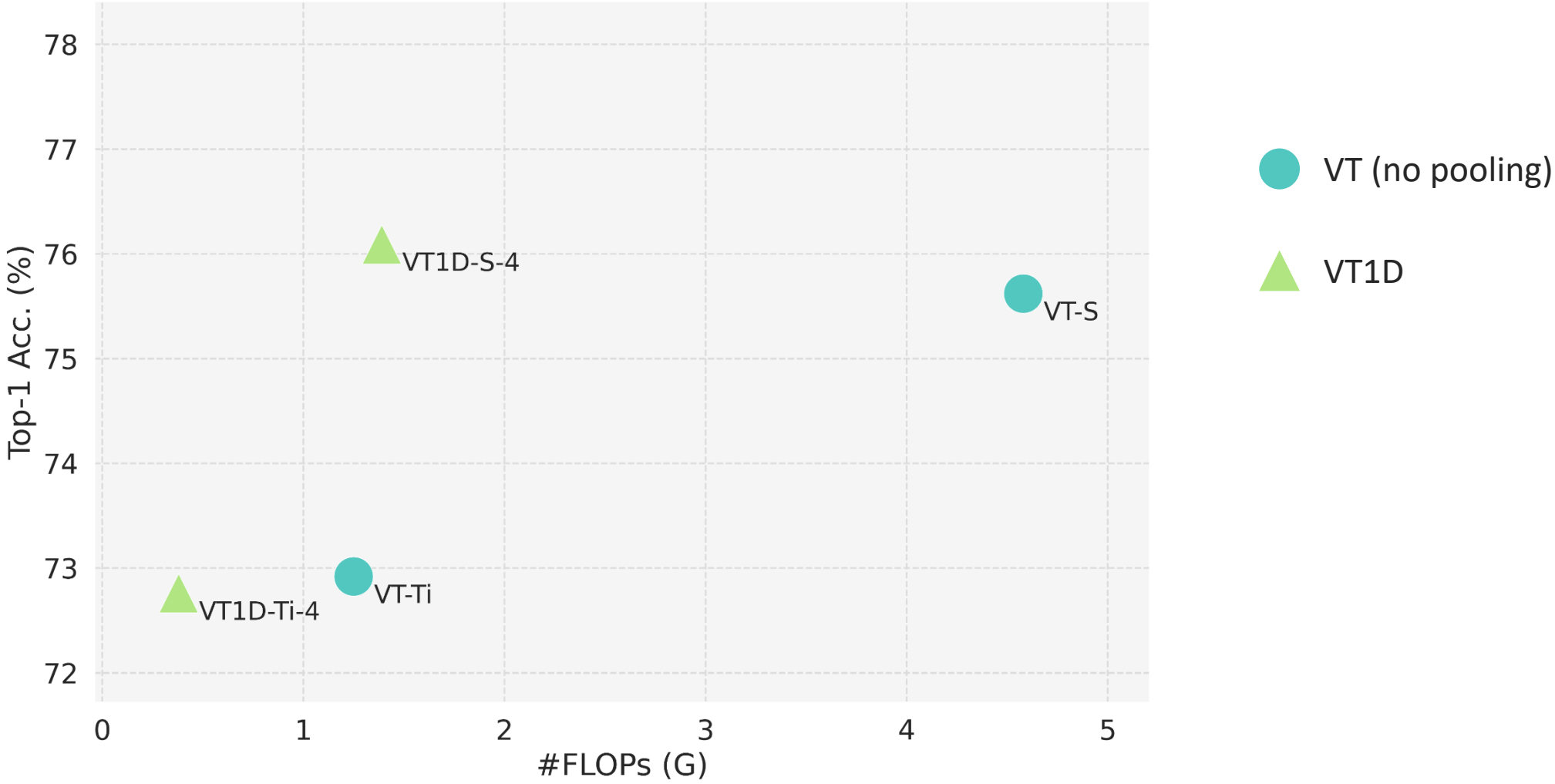
P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022.

Performance Comparison: Accuracy vs. FLOPs



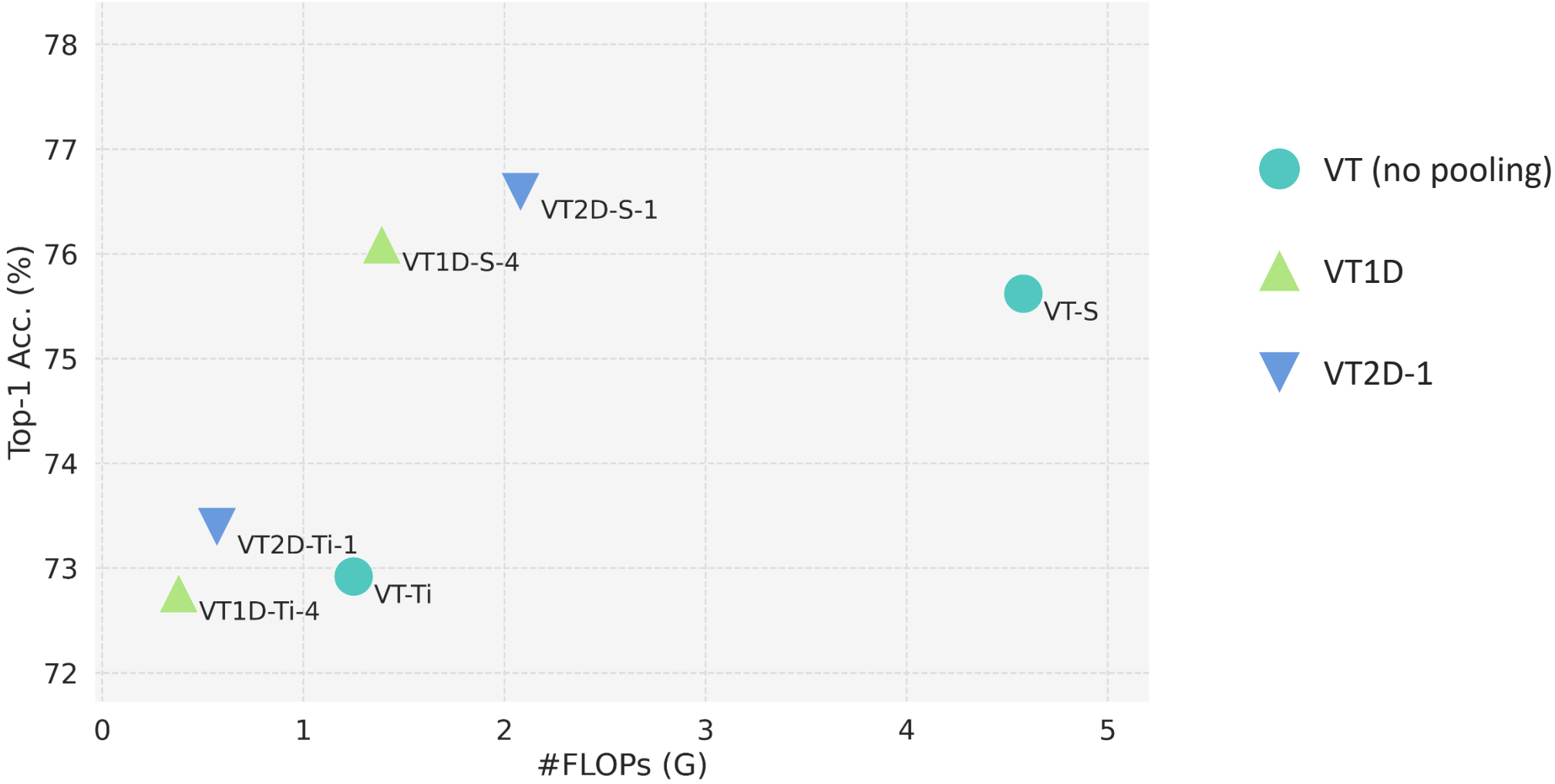
P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022.

Performance Comparison: Accuracy vs. FLOPs



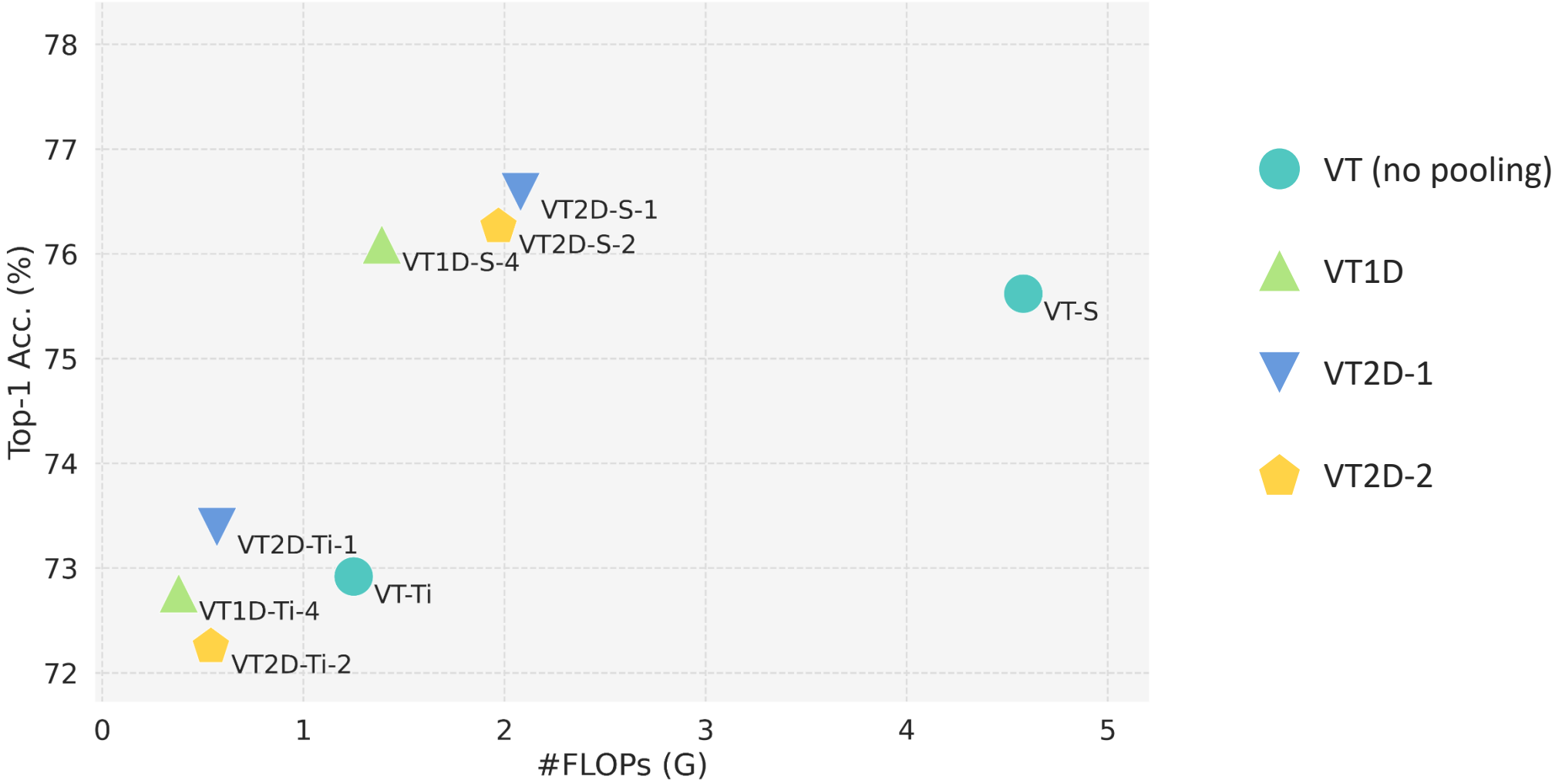
P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022.

Performance Comparison: Accuracy vs. FLOPs



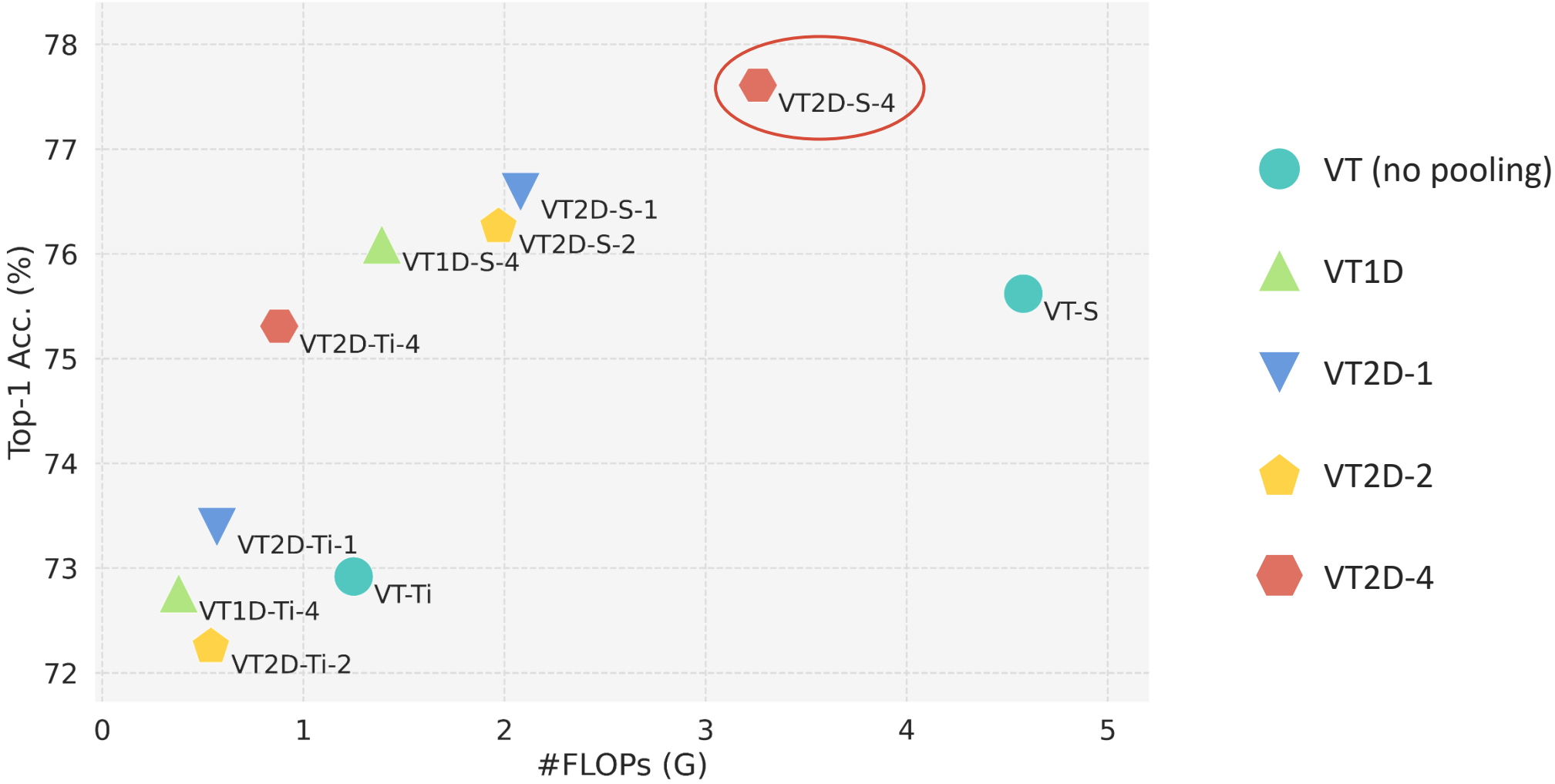
P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022.

Performance Comparison: Accuracy vs. FLOPs



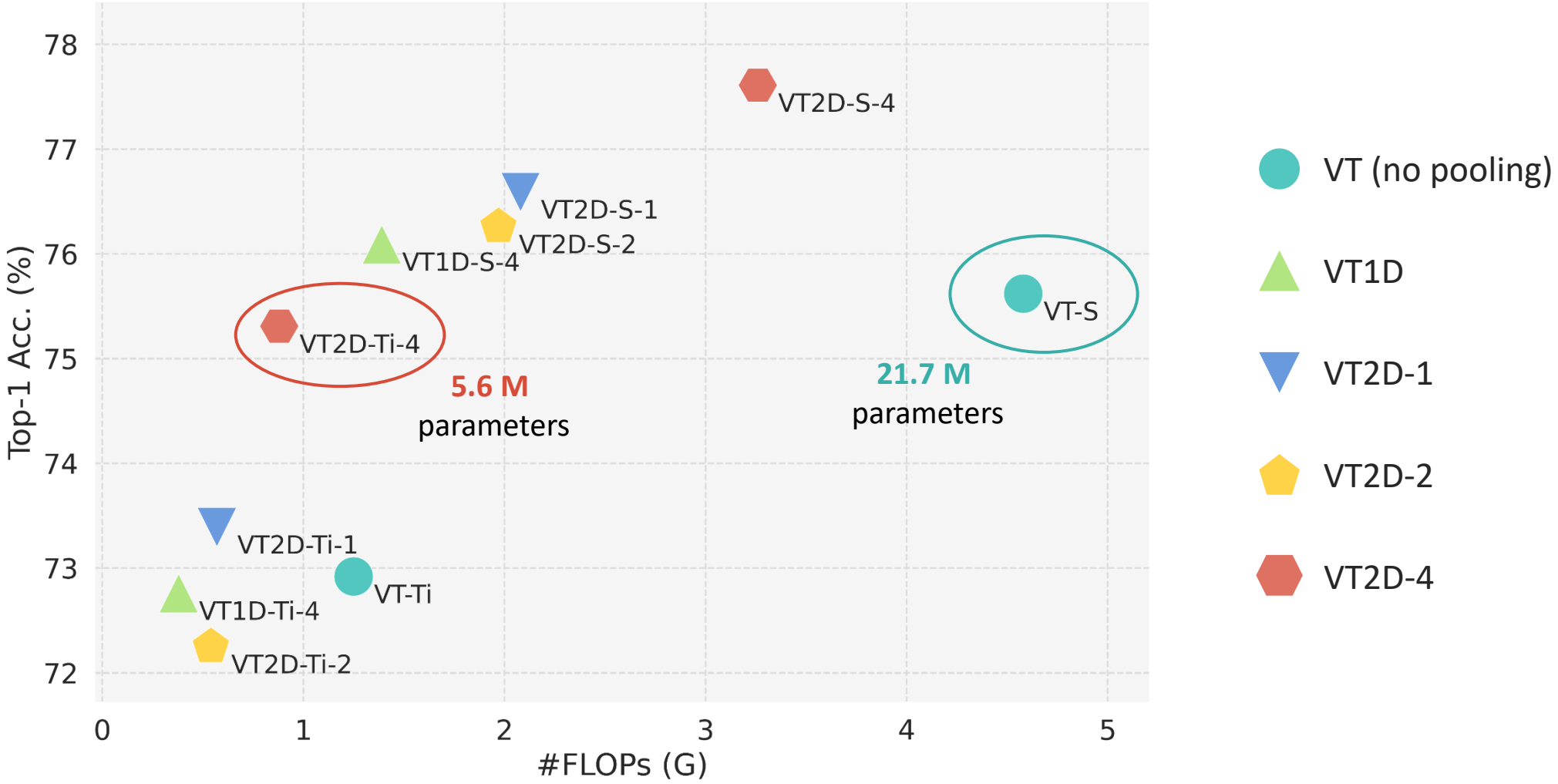
P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022.

Performance Comparison: Accuracy vs. FLOPs



P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022.

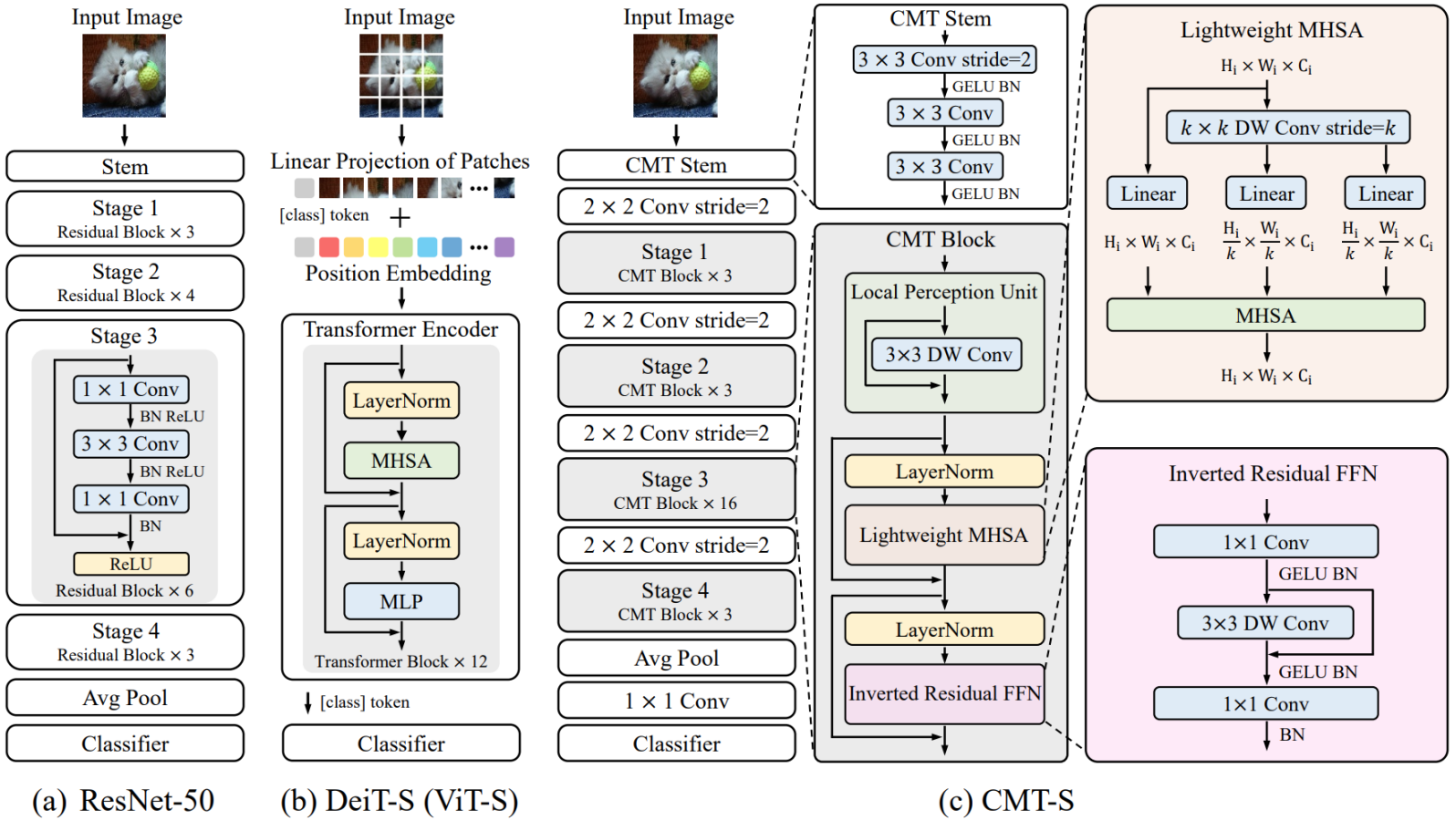
Performance Comparison: Accuracy vs. FLOPs



P. Bruno, R. Amoroso, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Investigating Bidimensional Downsampling in Vision Transformer Models”. ICIAP 2022.

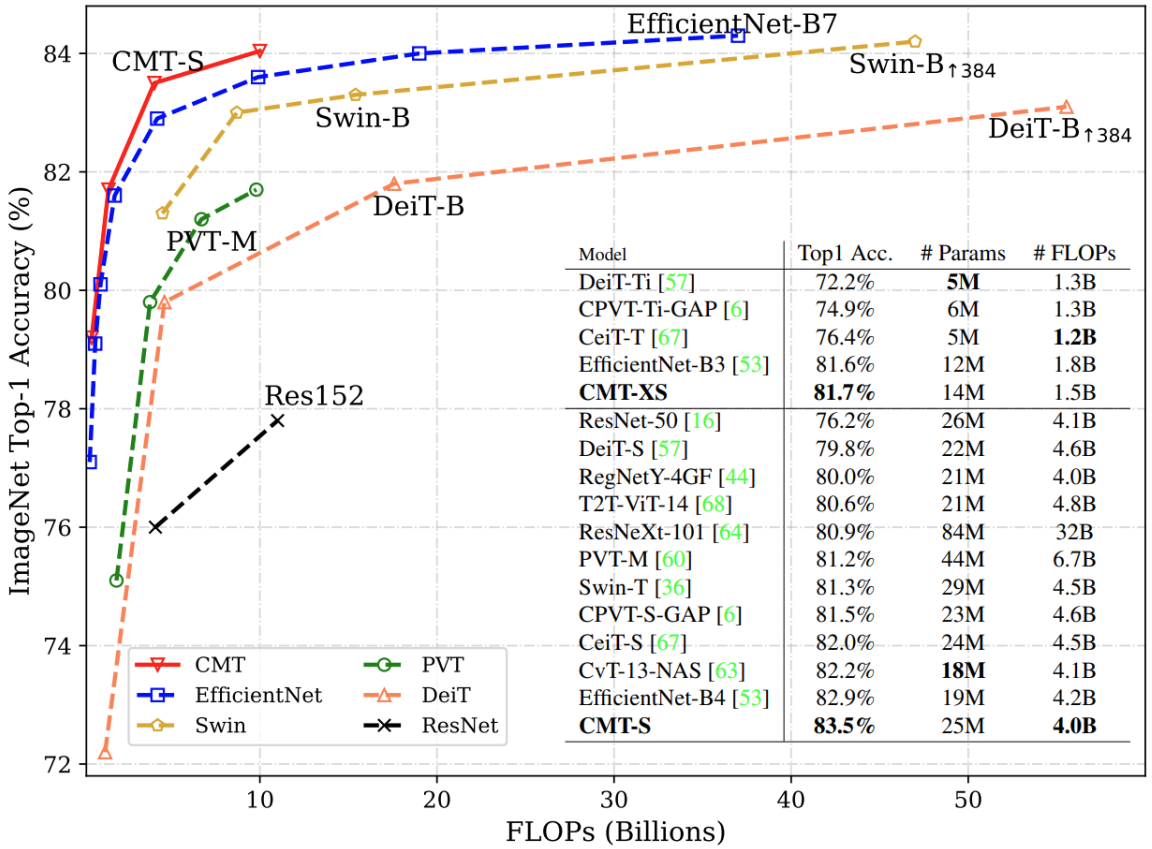
Many explorations of ViT+CNN…

Many architectures try to find the best of convolutive and attention-based architecture



1. J. Guo et al "CMT: Convolutional Neural Networks Meet Vision Transformers» CVPR 2022

CMT (CNNs meet transformers) architecture for visual



Model	Top1 Acc.	# Params	# FLOPs
DeiT-Ti [57]	72.2%	**5M**	1.3B
CPVT-Ti-GAP [6]	74.9%	6M	1.3B
CeiT-T [67]	76.4%	5M	**1.2B**
EfficientNet-B3 [53]	81.6%	12M	1.8B
CMT-XS	**81.7%**	14M	1.5B
ResNet-50 [16]	76.2%	26M	4.1B
DeiT-S [57]	79.8%	22M	4.6B
RegNetY-4GF [44]	80.0%	21M	4.0B
T2T-ViT-14 [68]	80.6%	21M	4.8B
ResNeXt-101 [64]	80.9%	84M	32B
PVT-M [60]	81.2%	44M	6.7B
Swin-T [36]	81.3%	29M	4.5B
CPVT-S-GAP [6]	81.5%	23M	4.6B
CeiT-S [67]	82.0%	24M	4.5B
CvT-13-NAS [63]	82.2%	**18M**	4.1B
EfficientNet-B4 [53]	82.9%	19M	4.2B
CMT-S	**83.5%**	25M	**4.0B**

(a) ImageNet Accuracy vs. FLOPs

Table 2. **ImageNet Results of CMT**. CNNs and transformers with similar accuracy are grouped together for comparison. The propose CMTs consistently outperform other methods with less computational cost.

Model	Top-1 Acc.	Top-5 Acc.	Throughput	# Params	Resolution	# FLOPs	Ratio
CPVT-Ti-GAP [6]	74.9%	-	-	6M	224^2	1.3B	2.6×
DenseNet-169 [22]	76.2%	93.2%	-	14M	224^2	3.5B	7×
EfficientNet-B1 [53]	79.1%	94.4%	-	7.8M	240^2	0.7B	1.2×
CMT-Ti	**79.1%**	**94.5%**	1323.5	9.5M	160^2	**0.6B**	**1**×
ResNet-50 [16]	76.2%	92.9%	-	25.6M	224^2	4.1B	2.7×
CoaT-Lite Mini [65]	78.9%	-	-	11M	224^2	2.0B	1.3×
DeiT-S [57]	79.8%	-	940.4	22M	224^2	4.6B	3.1×
EfficientNet-B3 [53]	81.6%	95.7%	732.1	12M	300^2	1.8B	1.2×
CMT-XS	**81.8%**	**95.8%**	857.4	15.2M	192^2	**1.5B**	**1**×
ResNeXt-101-64x4d [64]	80.9%	95.6%	-	84M	224^2	32B	8×
T2T-ViT-19 [68]	81.2%	-	-	39.0M	224^2	8.0B	2×
PVT-M [60]	81.2%	-	528.1	44.2M	224^2	6.7B	1.7×
Swin-T [36]	81.3%	-	755.2	29M	224^2	4.5B	1.1×
CPVT-S-GAP [6]	81.5%	-	-	23M	224^2	4.6B	1.2×
RegNetY-8GF [44]	81.7%	-	591.6	39.2M	224^2	8.0B	2×
CeiT-S [67]	82.0%	95.9%	-	24.2M	224^2	4.5B	1.1×
EfficientNet-B4 [53]	82.9%	96.4%	349.4	19M	380^2	4.2B	1×
Twins-SVT-B [5]	83.1%	-	-	56.0M	224^2	8.3B	2.1×
CMT-S	**83.5%**	**96.6%**	562.5	25.1M	224^2	**4.0B**	**1**×
ViT-B/16$_{\uparrow 384}$ [10]	77.9%	-	85.9	55.5M	384^2	77.9B	8.4×
TNT-B [14]	82.8%	96.3%	-	65.6M	224^2	14.1B	1.5×
DeiT-B$_{\uparrow 384}$ [57]	83.1%	-	85.9	85.8M	384^2	55.6B	6.0×
CvT-21$_{\uparrow 384}$ [63]	83.3%	-	-	31.5M	384^2	24.9B	2.7×
Swin-B [36]	83.3%	-	88M	224^2	278.1	15.4B	1.5×
Twins-SVT-L [5]	83.3%	-	288.0	99.2M	224^2	14.8B	1.7×
CeiT-S$_{\uparrow 384}$ [67]	83.3%	96.5%	-	24.2M	384^2	12.9B	1.4×
BoTNet-S1-128 [48]	83.5%	96.5%	-	75.1M	256^2	19.3B	2.1×
EfficientNetV2-S [54]	83.9%	-	-	22M	224^2	8.8B	1×
EfficientNet-B6 [53]	84.0%	96.8%	96.9	43M	528^2	19.2B	2.0×
CMT-B	**84.5%**	**96.9%**	285.4	45.7M	256^2	**9.3B**	**1**×
EfficientNet-B7 [53]	84.3%	97.0%	55.1	66M	600^2	37B	1.9×
CMT-L	**84.8%**	**97.1%**	150.4	74.7M	288^2	**19.5B**	**1**×

1. J. Guo et al "CMT: Convolutional Neural Networks Meet Vision Transformers» CVPR 2022

Encoding multimodality together

Many transformer .based encoders

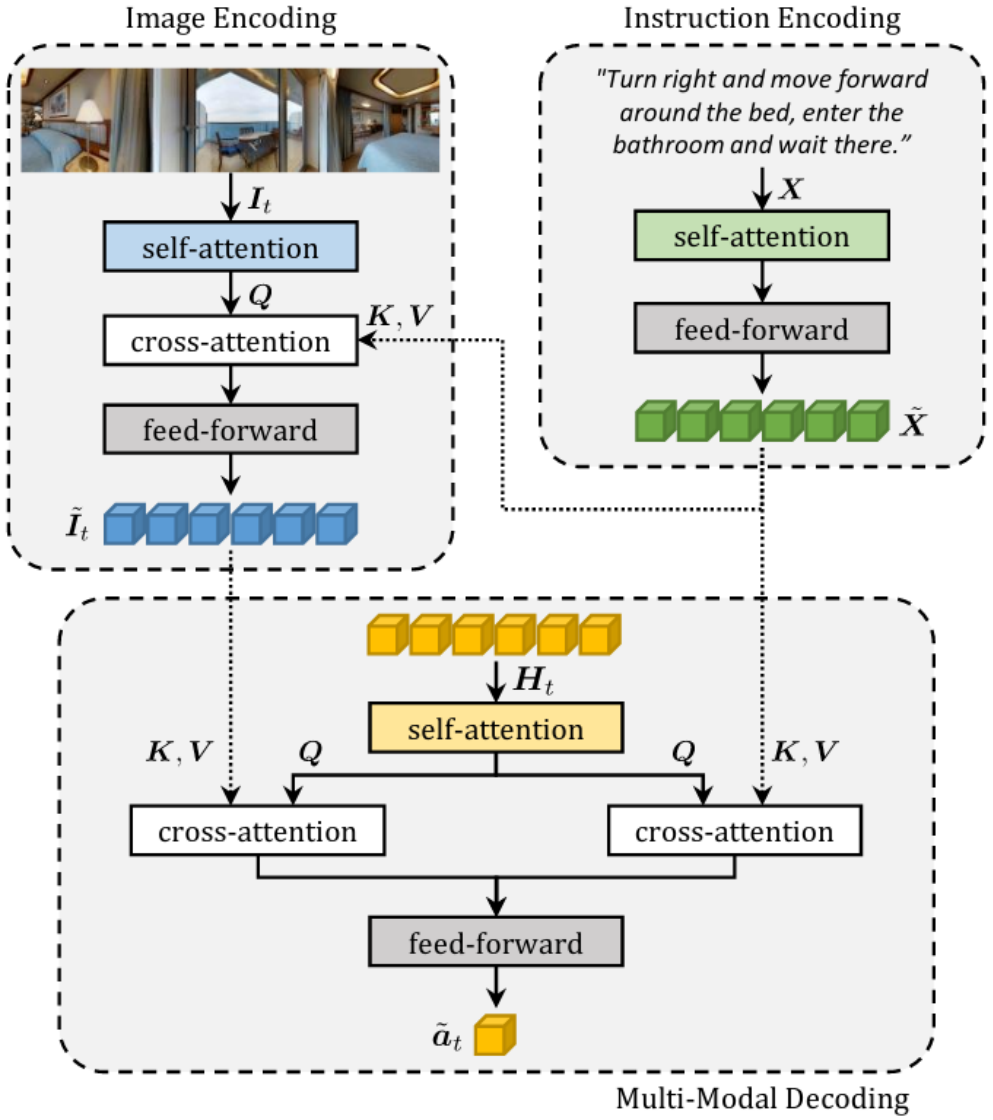
Self-attention for single modality

Cross-attention for two modalities

Take-at-home-message:

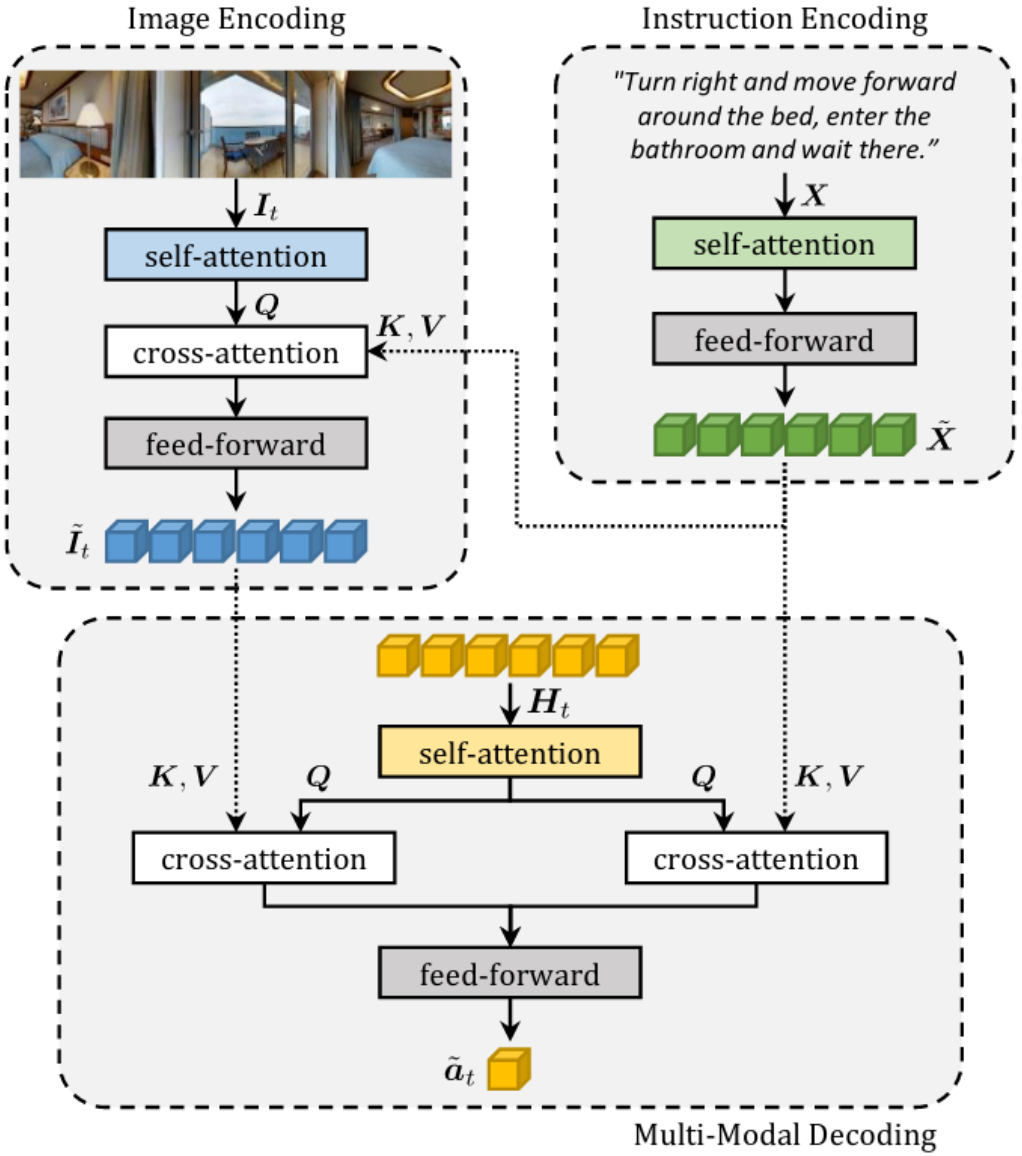
Multimodality now can be done “easily”

Think about it!



Landi, L Baraldi, M Cornia, M Corsini, R Cucchiara Multimodal attention networks for low-level vision-and-language navigation – CVIU Journal , 2021

Perceive-Transform-Act



Based on multi-head attention*

$$\mathrm{MH}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Concat}(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_h)\boldsymbol{W}^O$$

with:

$$\boldsymbol{h}_i = \mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V)$$

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V}$$

In self-attention layers, the Keys, Queries, and Values (K, Q, V) come from the same source sequence.

After each block, a residual connection, followed by layer normalization are added.

* as (*Vaswani et. al, NeurIPS 2017*):

AImage Lab

Instruction:

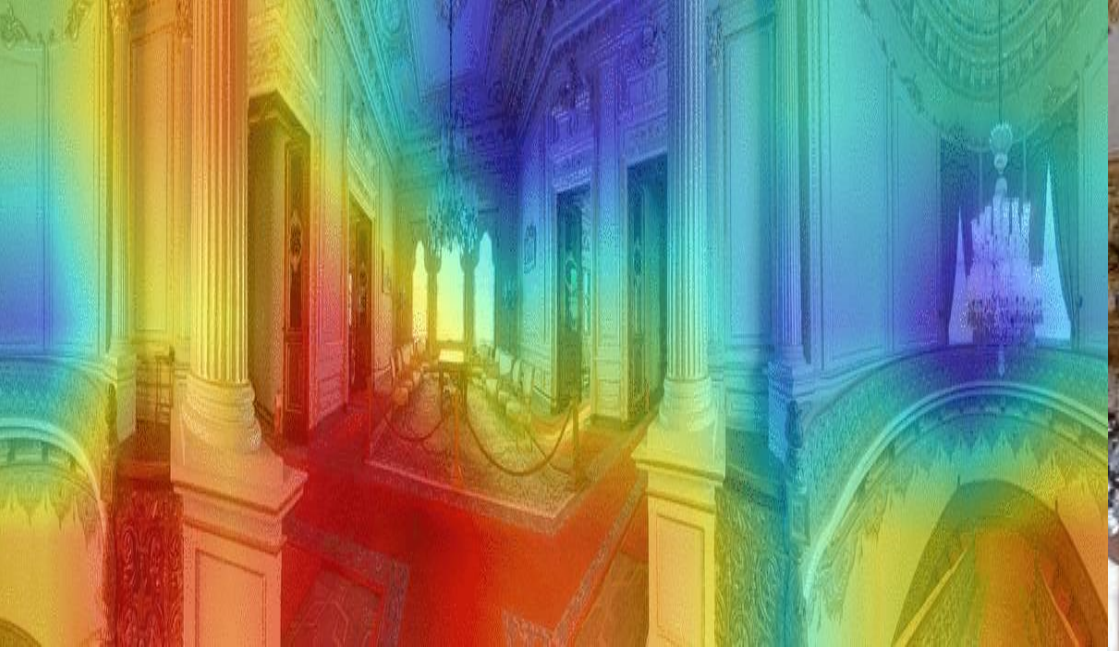
Walk up the stairs. Turn right at the top of the stairs and walk along the red ropes. Walk through the open doorway straight ahead along the red carpet. Walk through that hallway into the room with couches and a marble coffee table.

Legend: L *left* R *right* F *forward* E *end episode*

Instruction:

Walk up the stairs. Turn right at the top of the stairs and walk along the red ropes. Walk through the open doorway straight ahead along the red carpet. Walk through that hallway into the room with couches and a marble coffee table.

Legend: L *left* R *right* F *forward* E *end episode*



Instruction:

Walk up the stairs. Turn right at the top of the stairs and walk along the red ropes. Walk through the open doorway straight ahead along the red carpet. Walk through that hallway into the room with couches and a marble coffee table.

Legend: L *left* R *right* F *forward* E *end episode*

Instruction:

Walk up the stairs. Turn right at the top of the stairs and walk along the red ropes. Walk through the open doorway straight ahead along the red carpet. Walk through that hallway into the room with couches and a marble coffee table.

Legend: L *left* R *right* F *forward* E *end episode*

Instruction:

Walk up the stairs. Turn right at the top of the stairs and walk along the red ropes. Walk through the open doorway straight ahead along the red carpet.
Walk through that hallway into the room with couches and a marble coffee table.

Legend: L *left* R *right* F *forward* E *end episode*

Instruction:

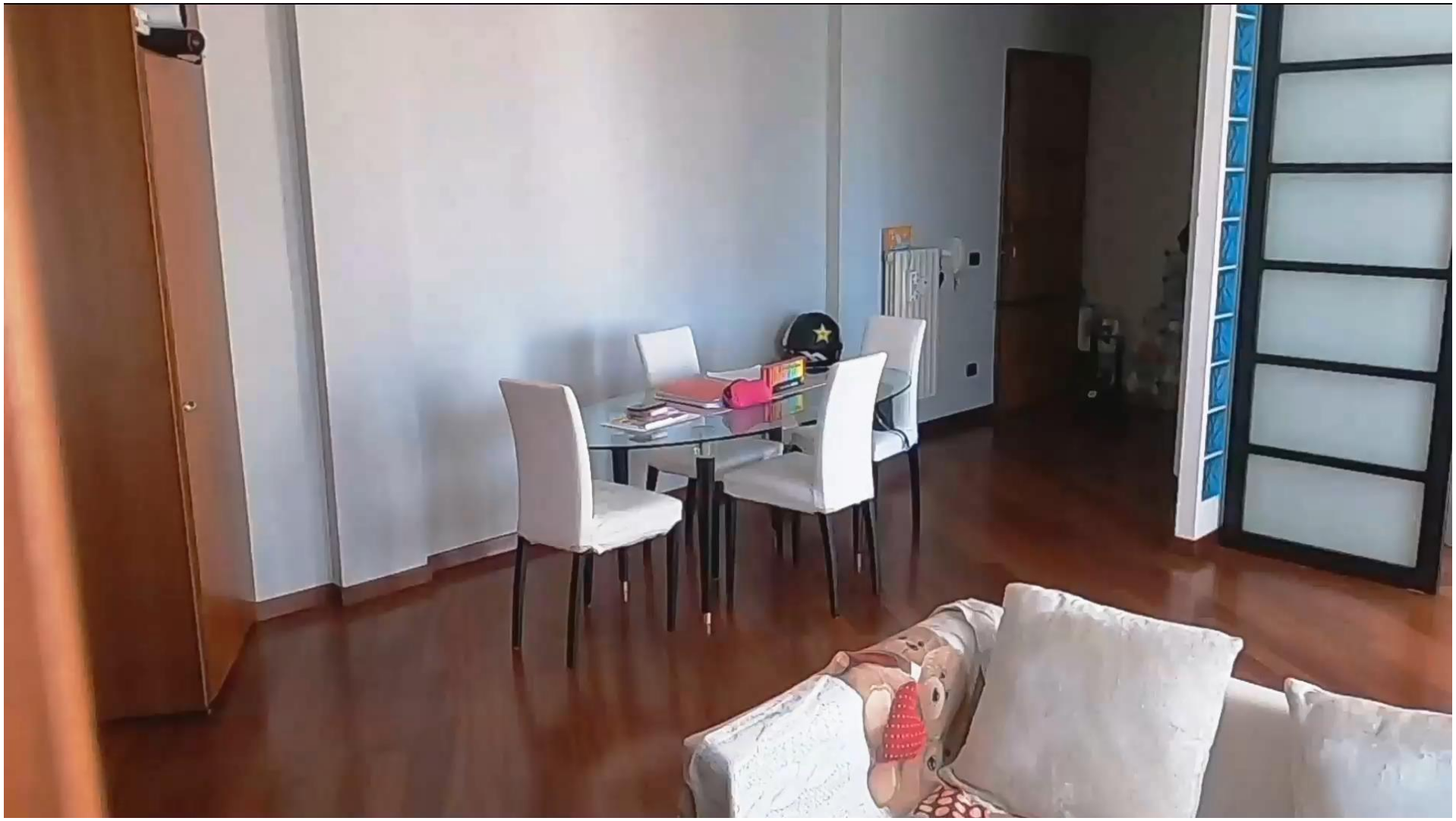
Walk up the stairs. Turn right at the top of the stairs and walk along the red ropes. Walk through the open doorway straight ahead along the red carpet. Walk through that hallway into the room with couches and a marble coffee table.

Legend: L *left* R *right* F *forward* E *end episode*

PTA is working

Not only simulation

PTA+ explainability



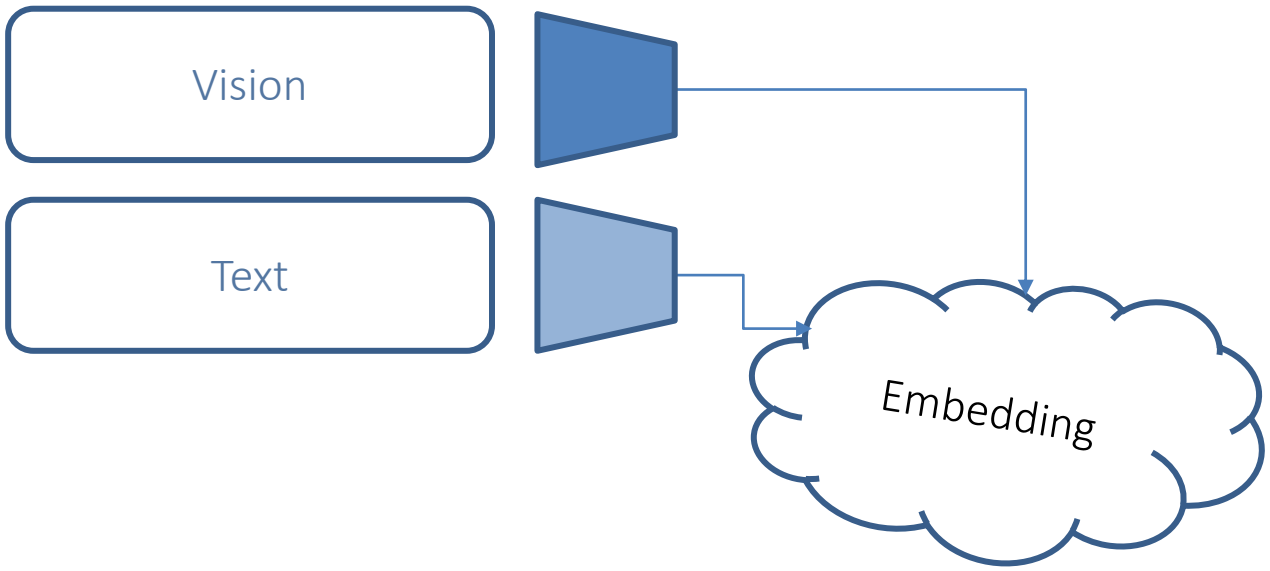
1. *R Bigazzi, F Landi, S Cascianelli, L Baraldi, M Cornia, R Cucchiara Focus on impact: indoor exploration with intrinsic motivation IEEE Robotics and Automation Letters 7 (2), 2985-2992*

Subscribe

AImage Lab

Embedding visual and language: the CLIP Power

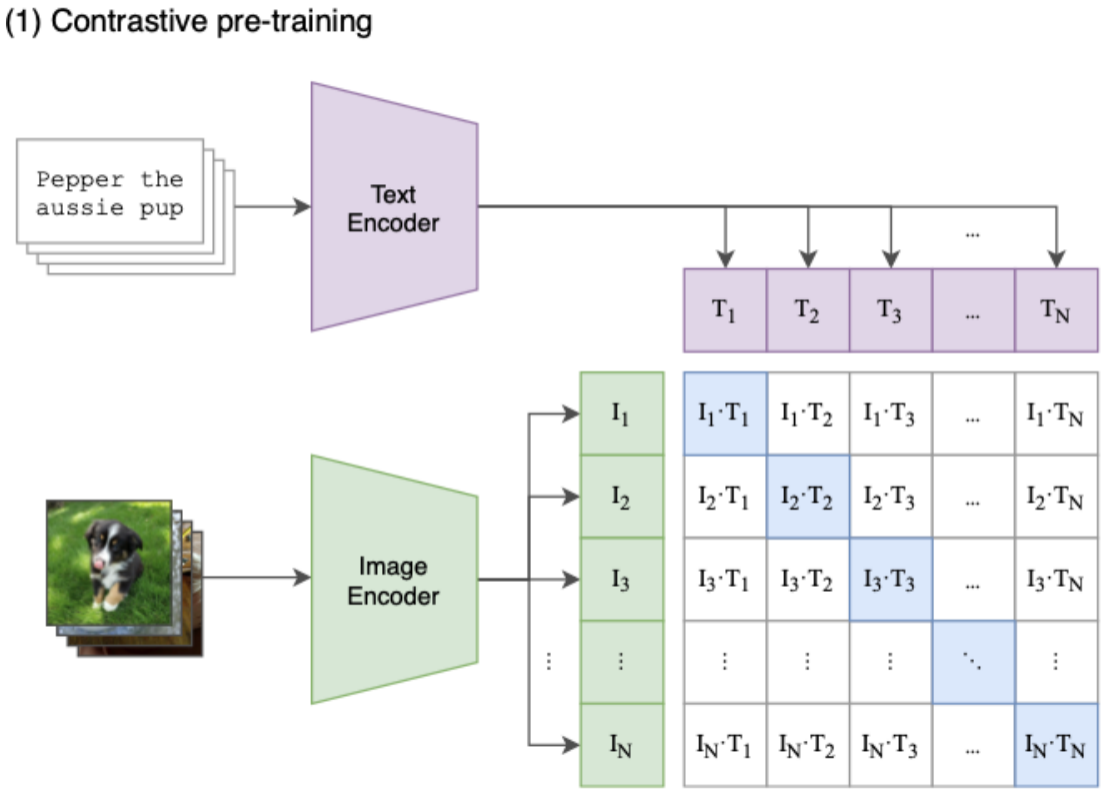
Vision-language models have demonstrated impressive capabilities in challenging tasks such as image captioning, image generation, and visual question answering. Typically, they consist of three key elements: **an image encoder**, **a text encoder**, and **a strategy to fuse information from the two encoders**.



A fantastic idea: create a **discriminative embedding by contrastive learning**: **CLIP** , a true foundation model.

CLIP

2021: OpenAI CLIP (Contrastive Language–Image Pre-training). The input to CLIP is 400 million image-text pairs crawled from the internet. It encodes text using Transforms, encodes images using Vision Transformers, and applies contrastive learning to train the model. Contrastive training matches correct image and text pairs using cosine similarity



This is a basic module for many research

- Retrieval
- Generative AI (e.g. with Sig
- Multimodal Discriminative tasks
- …

1. https://arxiv.org/abs/2103.00020

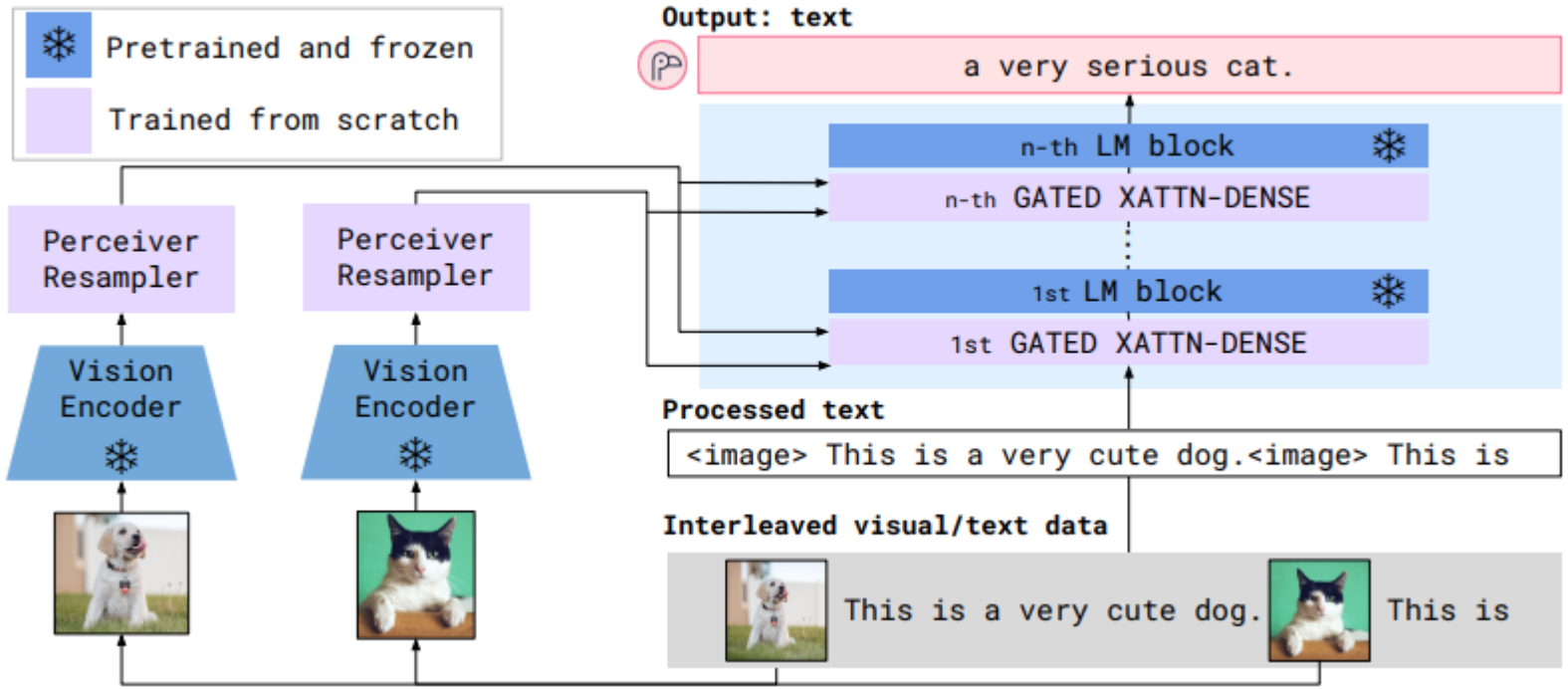
FLAMINGO

2022: DeepMind group of Visual Language Models; Flamingo.

They have two parts: a vision model that can understand visual scenes, and a language model that helps with reasoning. The models use their pre-training knowledge to work together.

Flamingo models can also take high-quality images or videos thanks to a Perceiver architecture (that can analyze a large number of visual input features and produce a small number of visual tokens

The Flamingo-80B, the biggest version with 80 billion parameters, set a new record in few-shot learning for many tasks that involve understanding language, images, and videos.



1. https://arxiv.org/abs/2204.14198

Other multimodal FMs

- Microsoft FLORENCE (2022) only images
- Microsoft Kosmos-1, a multimodal model that can perceive different modalities, learn context and follow instructions. Uses a Transformer-based causal language model. Used for generative images, VQ%A
- Google's PaLM-E is an embodied multimodal model: different embodiments, including internet-scale language, vision, and visual-language domains. The biggest PaLM-E model, PaLM-E-562B, has 562 billion parameters with new tasks: telling jokes based on an image or doing robot tasks such as perceiving, talking, and planning.
- OpenAI's GPT-4 is a large multimodal model capable of processing image and text inputs and producing text outputs. It scored 90th percentile on a simulated bar exam and 99th percentile (with vision) on Biology Olympiad.
- a long long list…….

- And thus what can we do in such a research???

What can we do in multimodal research?

A lot.

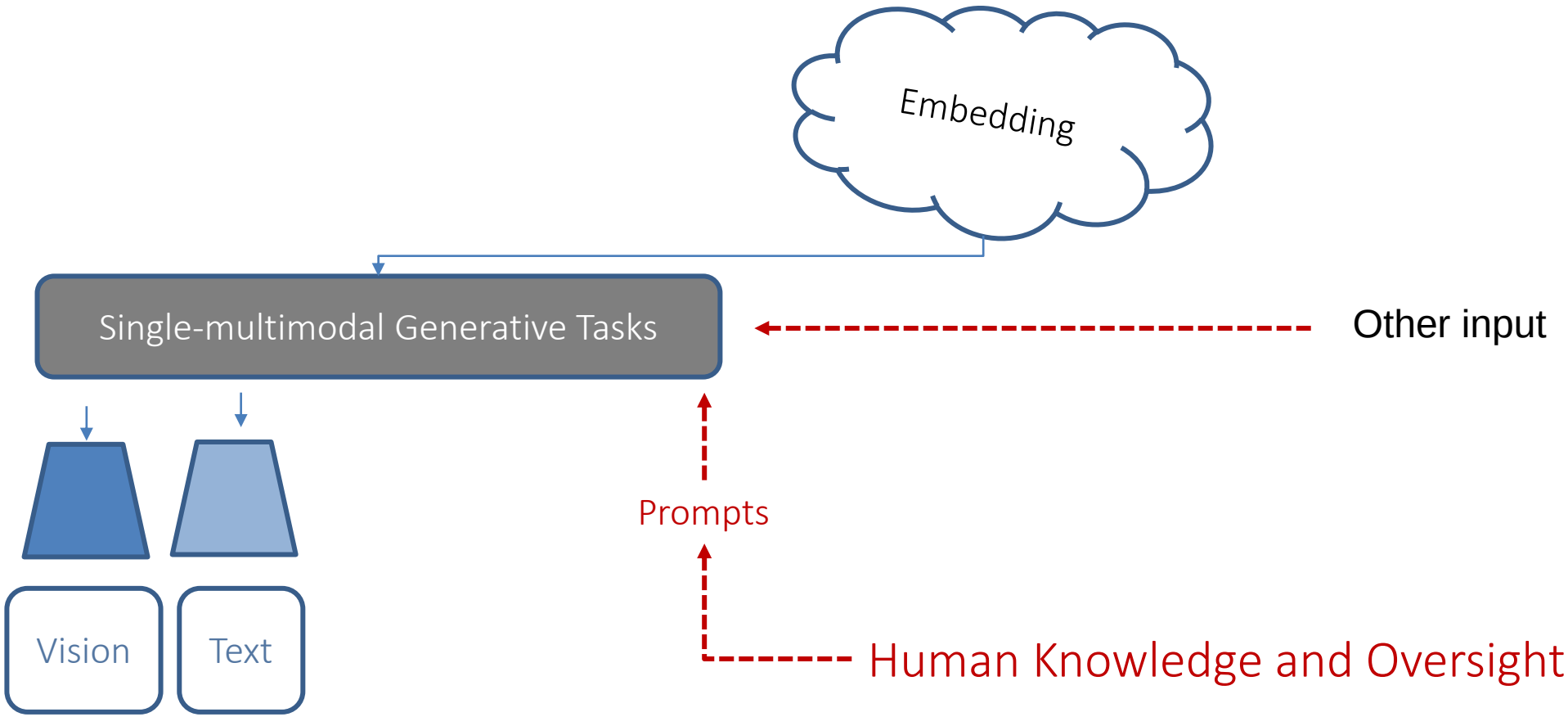
Standing on the shoulder of giants.

Use their models. Solve open challenges. CREATE new challenges.

3. CHALLENGES IN GENERATIVE AI

After the suitable embedding…

The big journey in Generative al



Generative ML is not NEW

Generative ML has a long story *

1. *Gaussian Mixture Models (GMM),*
2. *Hidden Markov Models (HMM),*
3. *Latent Dirichlet Allocation (LDA),*
4. *Boltzmann Machines (BM)……*

… e.g. in tracking:

zt observation of the world at time t (or frame k)

xt the status of the model at the time t (or frame k)

Probabilistic tracking of object in video

With first order Markov assumptions

GENERATIVE ML given x to can generate z

The prior today at state k,
is the prediction yesterday at state k-1

$$p(x_k \mid z_{1:k-1}) = \int f(x_k \mid x_{k-1}) p(x_{k-1} \mid z_{1:k-1}) dx_{k-1}$$

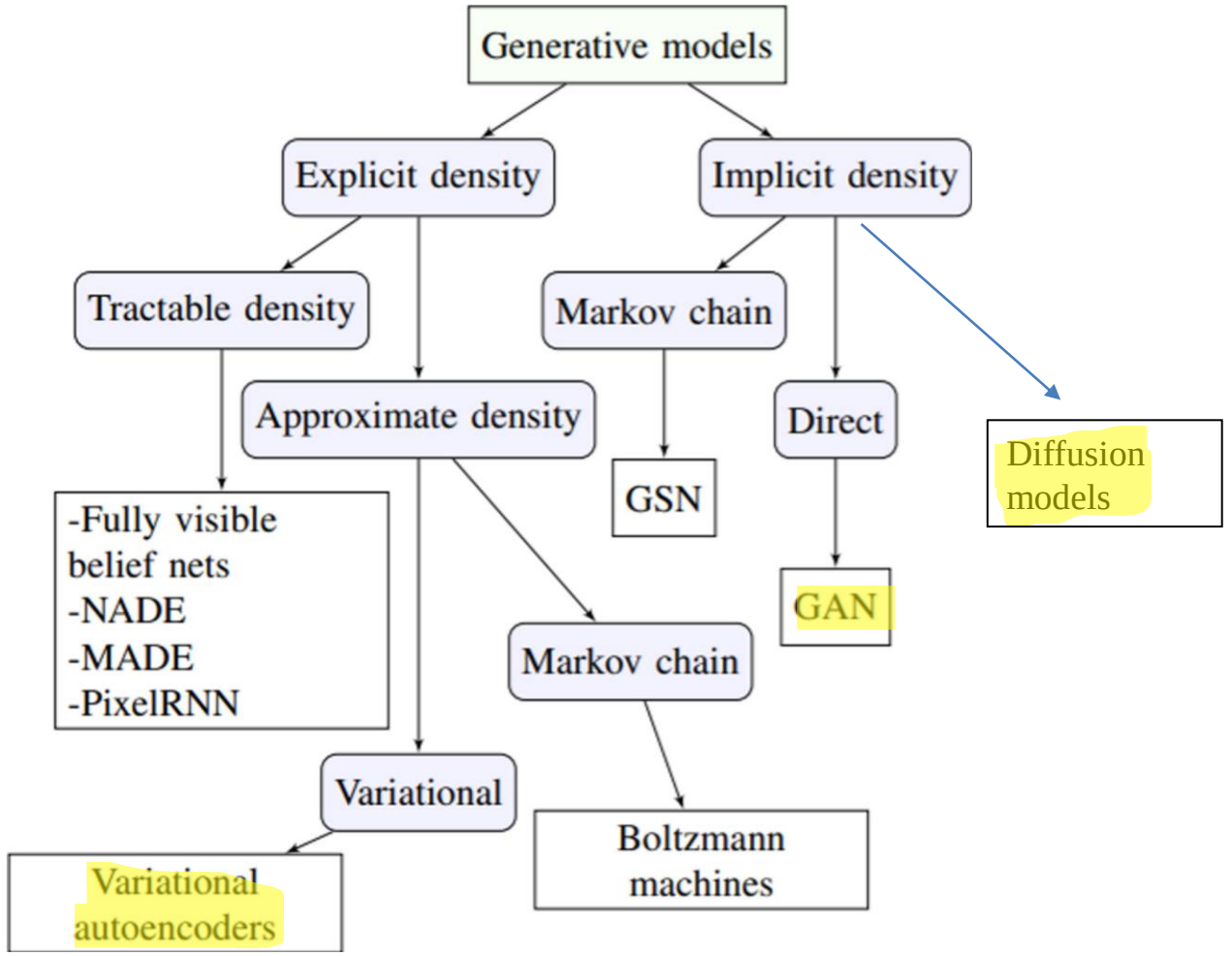
$$p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{x}_{t-1}) \propto p(\mathbf{z}_t|\mathbf{x}_t)\ p(\mathbf{x}_t|\mathbf{x}_{t-1})$$

1. *Harshvardhan GM A et al " comprehensive survey and analysis of generative models in machine learning" Computer Science Review 2020*

Generative ML

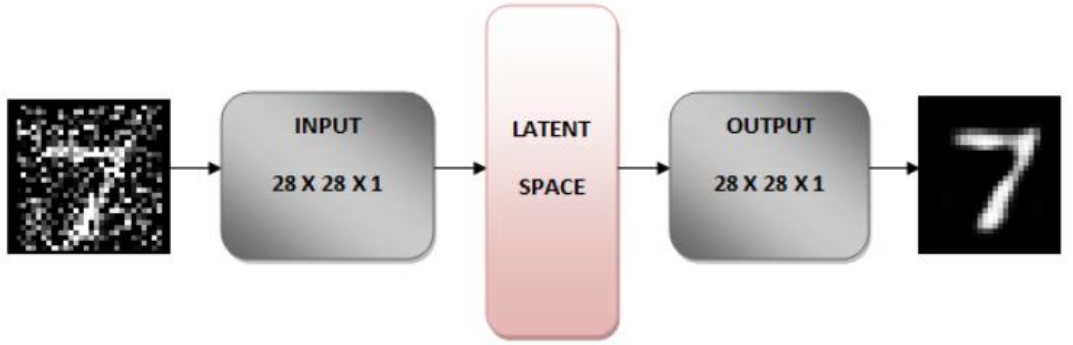
After the Deep Learning era

- Autoencoders
- VAE
- GANS
- Diffusion Models
- ..Overfitting learning (e.g. NeRF)

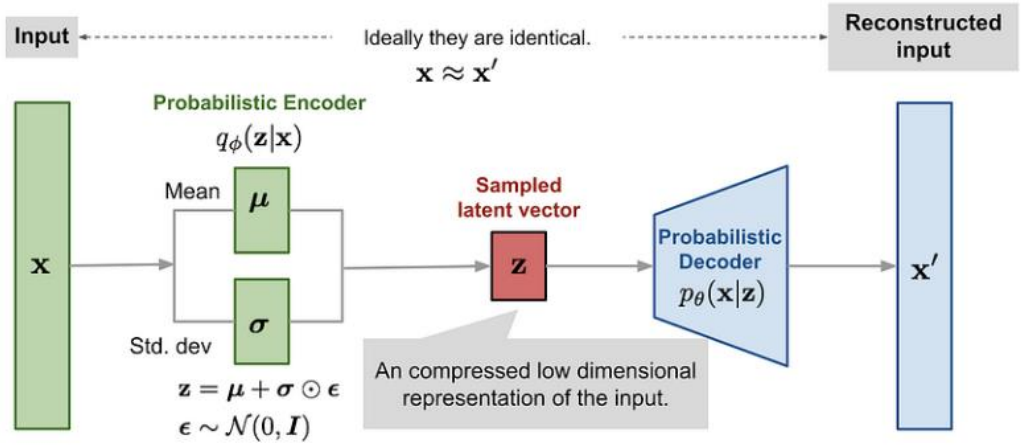


Generative al

From Simple Autoencoders



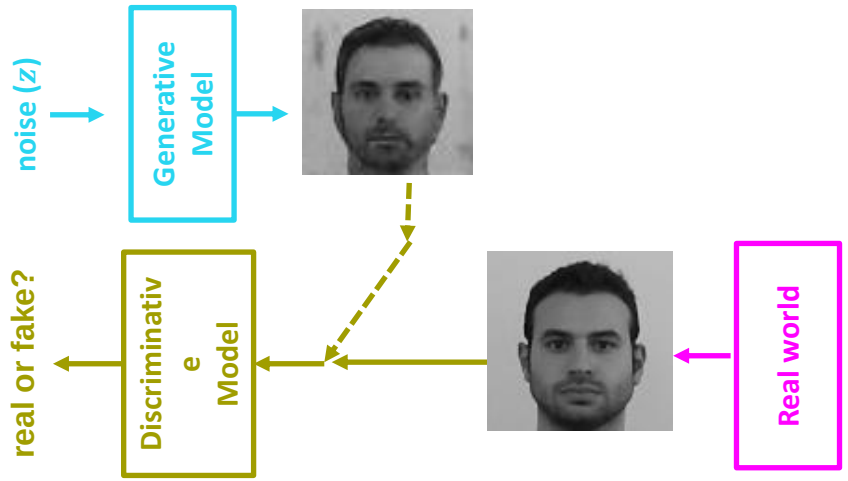
To Variational Autoencoders



Gans arrived

To **Generative Adversarial Networks**

- A Generative autoencoder enriched in training by a Discriminator Architecture trained with Adversarial examples by the Generator



Generative adversarial networks are based on a game theoretic scenario in which the generator network must compete against an adversary. The generator network directly produces samples. Its adversary, the discriminator network, attempts to distinguish between samples drawn from the training data and samples drawn from the generator.

(Deep Learning Y Goodfellow, Y. Bengio, A. Courville 2016, the Bible)

- Objective

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim Pdata(x)}[\log D(x)] + \mathbb{E}_{z\sim pz(z)}[\log\left(1 - D\left(G(z)\right)\right)]$$

Maximize D — Value of — Expectation — fake

Minimize G — Distribution of real data — Prob. Of D(*real*) — Distribution of generated samples — Prob. Of D(*fake*)

1. *The GAN Loss Y. Goodfellow Nips 2014*

GAN family

A large family

Conditioned GAN

Cycle GAN

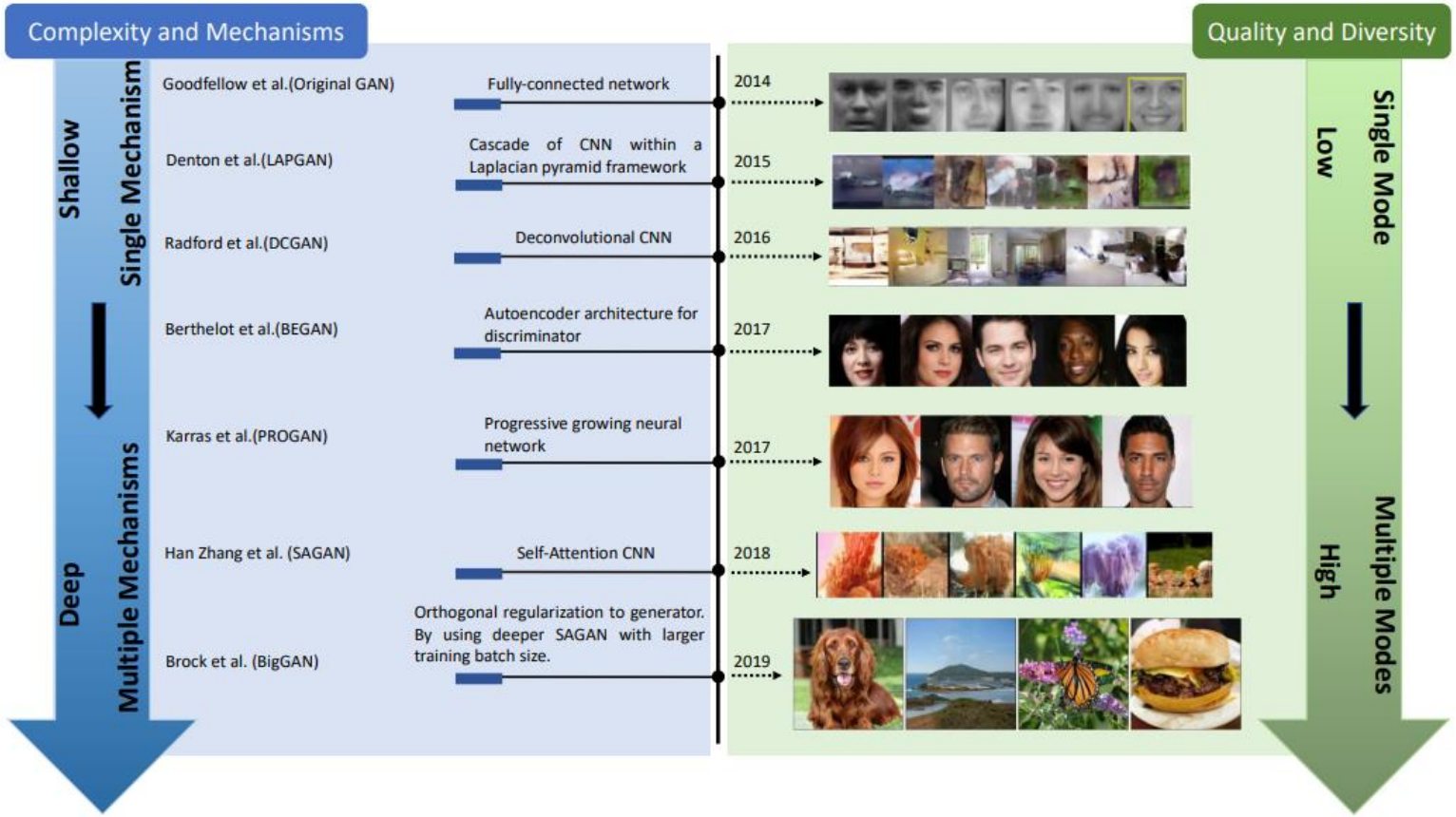
….

A challenge in GAN

How make them general enough

and also perfectly controllable?

How use them to solve new challenges?



See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/349189619

Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy

Article in ACM Computing Surveys · February 2021

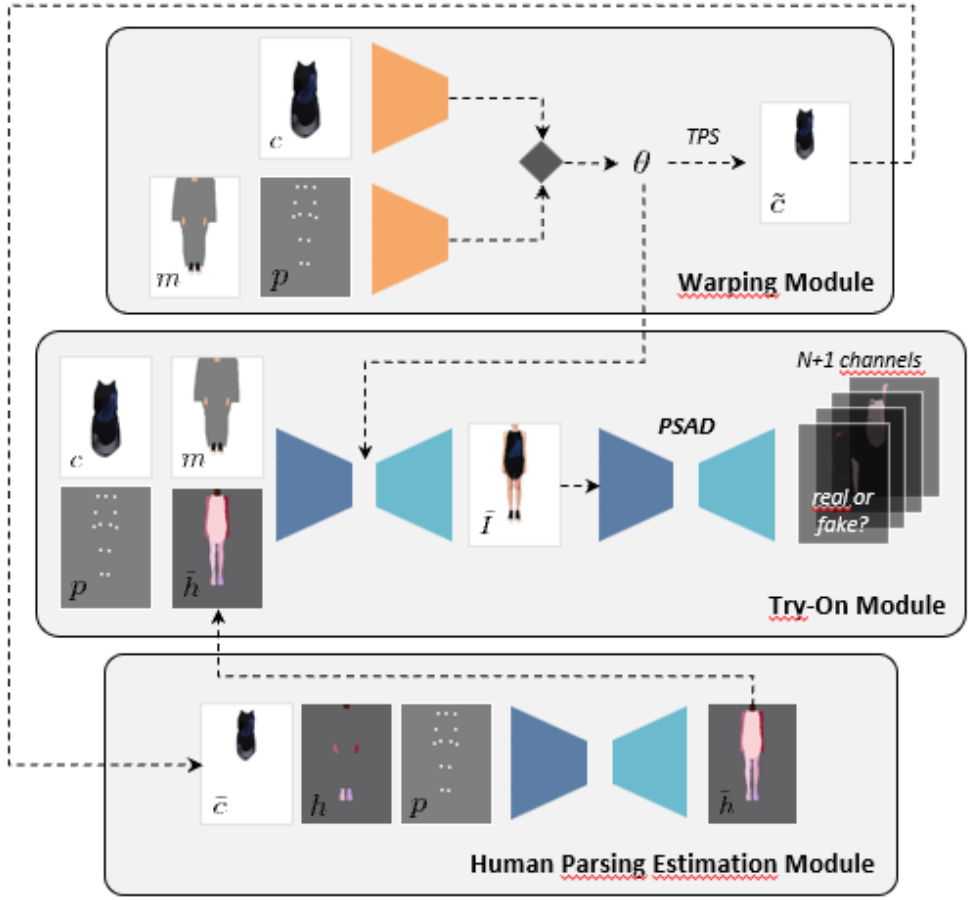
DOI: 10.1145/3439723

1. *https://arxiv.org/pdf/1906.01529.pdf*

Just an example…

DressCode the largest garment dataset for Vton (thanks to YOOX NET-a-PORTER)

Monomodal generative AI



https://ailb-web.ing.unimore.it/dress-code/demo

Davide Morelli, Matteo Fincat, Marcella Cornia Federico Landi Fabio Cesari Rita Cucchiara Dress Code: High-Resolution i-category Virtual Try-On ECCV 2022

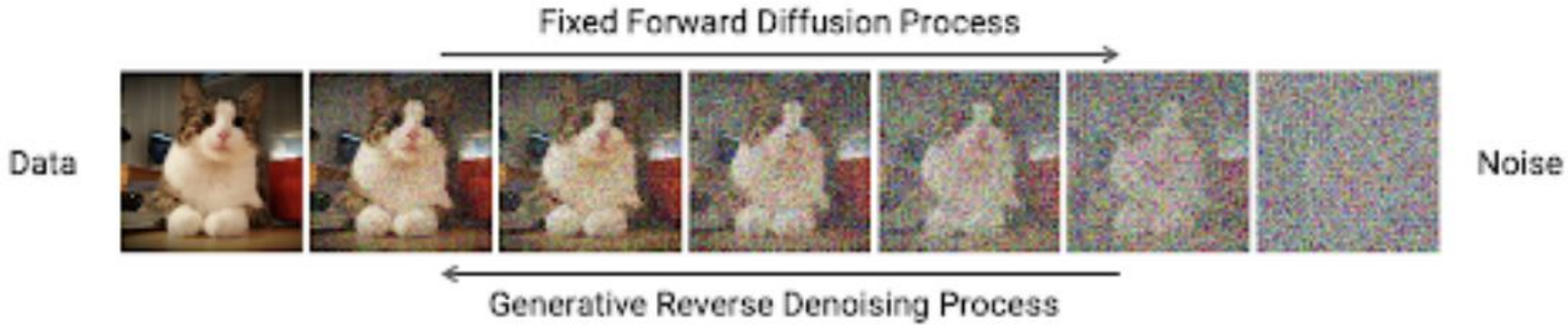
Diffusion models

Diffusion Models , also known as **denoising diffusion models** or **score-based generative models**, demonstrate surprisingly high sample quality, often outperforming generative adversarial networks, strong mode coverage and sample diversity.

Diffusion models consist of two processes:

forward diffusion

parametrized reverse



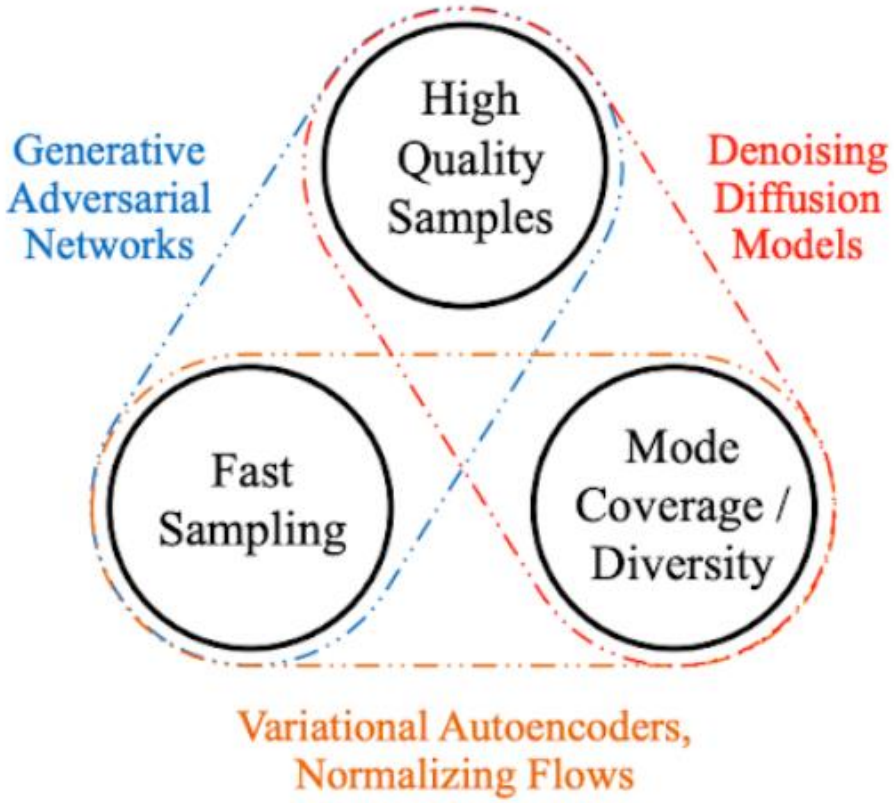
A small drawback: need learning thousands of diffusion Enormously.

A big challenge: how to make them more efficient…*

1. *M Ning, E Sangineto, A Porrello, S Calderara, R Cucchiara Input Perturbation Reduces Exposure Bias in Diffusion Models ICLR 2023*

The Generative Trilemma

Diffusion GANS



Published as a conference paper at ICLR 2022

TACKLING THE GENERATIVE LEARNING TRILEMMA WITH DENOISING DIFFUSION GANS

Zhisheng Xiao*
The University of Chicago
zxiao@uchicago.edu

Karsten Kreis
NVIDIA
kkreis@nvidia.com

Arash Vahdat
NVIDIA
avahdat@nvidia.com

https://arxiv.org/pdf/2112.07804.pdf

4. CHALLENGES IN CROSS-MODALITIES

AImage Lab

Cross-modality Generative AI

PRIN Creative PRIN (Sapienza, UNIMORE, UniTN)

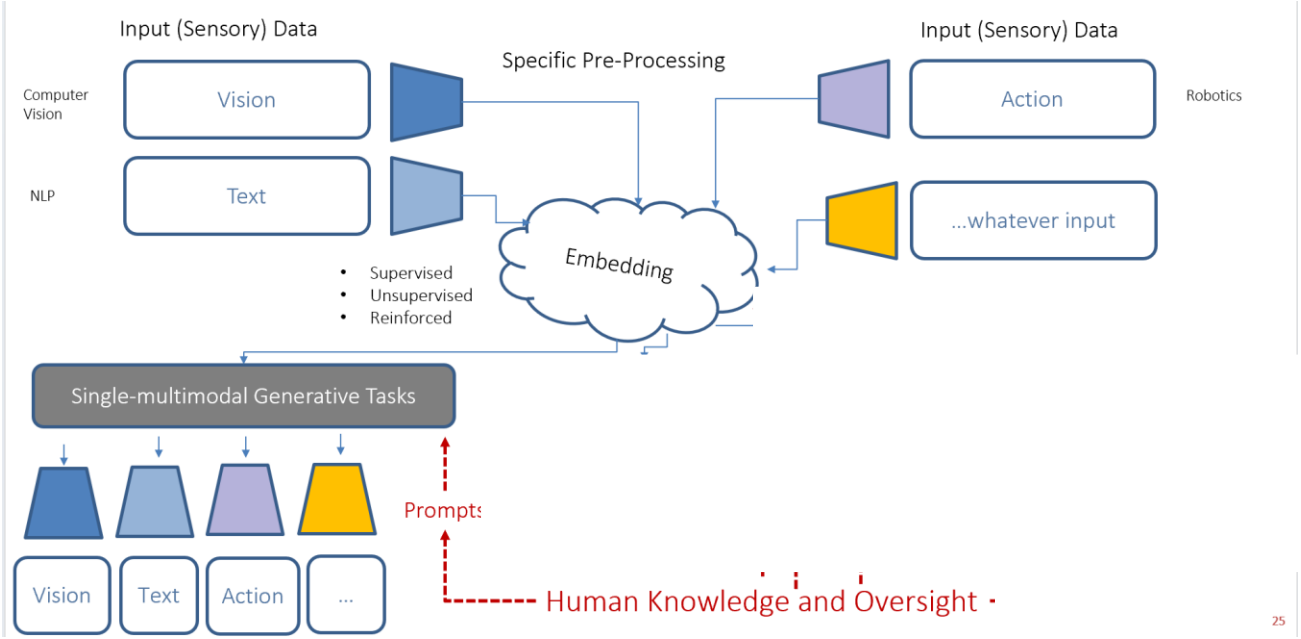
TP VLMC – FAIR

Many challenges in cross modalities

CREATIVE

CRoss-modal understanding and gEnerATIon of Visual and tExtual content

- Image-to-text
- Text-to-images
- whatever-to-whatever
- More whatever –to-whatever



Stable diffusion Models

Great ideas come from mixing knowledge!

Putting together Embedding and Generative diffusion Models

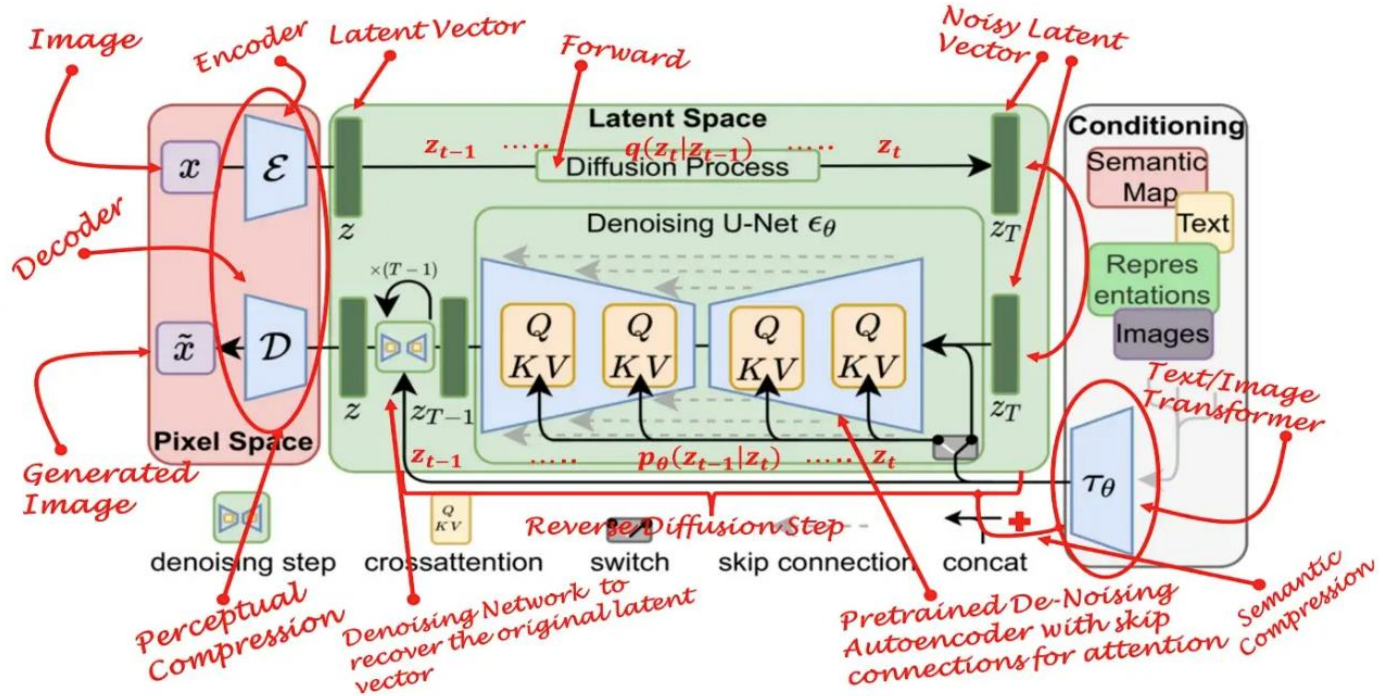


Figure 1: Latent Diffusion Model (Base Diagram:[3], Concept-Map Overlay: Author)

a Time-to-market of less than one year

Dall-E
Dall-”2
Midjourney
etc etc

Whati s the problem?

1. *R Rombach, A Blattmann, D Lorenz, P Esser, B Ommer “High-resolution image synthesis with latent diffusion models” CVPR 2022 1400 cit in 1 yeae*

The challenge of prompt based diffusion models

Prompt-based diffusion model

multimodal prompts : **image target editing**

The biggest challenge is that ……..they work well!.

What about transparency, interpretability etc?

How can we detect fake images?

Should they detect by

a) syntax, perceptive, probabilistic model of the signature of diffusion?

b) Semantc, by the lack of pertinency? Entropy, probability of surprise? (and what about normal anomalies?

c) Should by defined by human-like decision process?

Spot the fake

'The giraffe is standing alone in the wilderness.'

a giraffe standing in the middle of a field

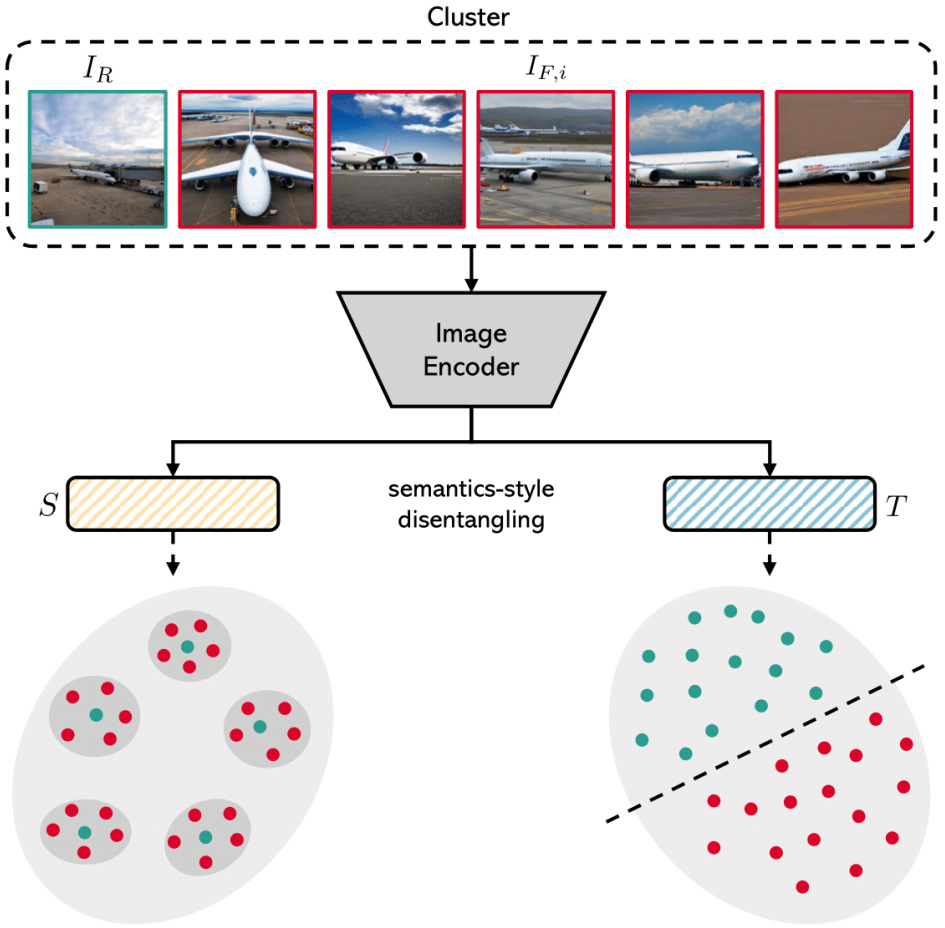
'A woman wearing a coat is standing in the snow near monuments while holding an umbrella.

a woman walking in the snow with an umbrella"

New brave ideas

Disentanging semantic and style for Fake detection

Backbone	Dataset	Validation Set		Test Set	
		In-Cluster Accuracy	Overall Accuracy	In-Cluster Accuracy	Overall Accuracy
RN50	ImageNet	78.15	95.94	78.62	95.97
ViT-B/32	ImageNet	68.27	91.46	68.06	93.70
CLIP RN50	OpenAI	96.33	99.38	96.72	99.44
CLIP ViT-B/32	OpenAI	95.79	99.27	95.49	99.22
OpenCLIP ViT-B/32	LAION-400M	90.72	98.41	91.57	98.55
OpenCLIP ViT-B/32	LAION-2B	97.93	99.65	98.06	99.66



elsa
European Lighthouse on Secure and Safe AI

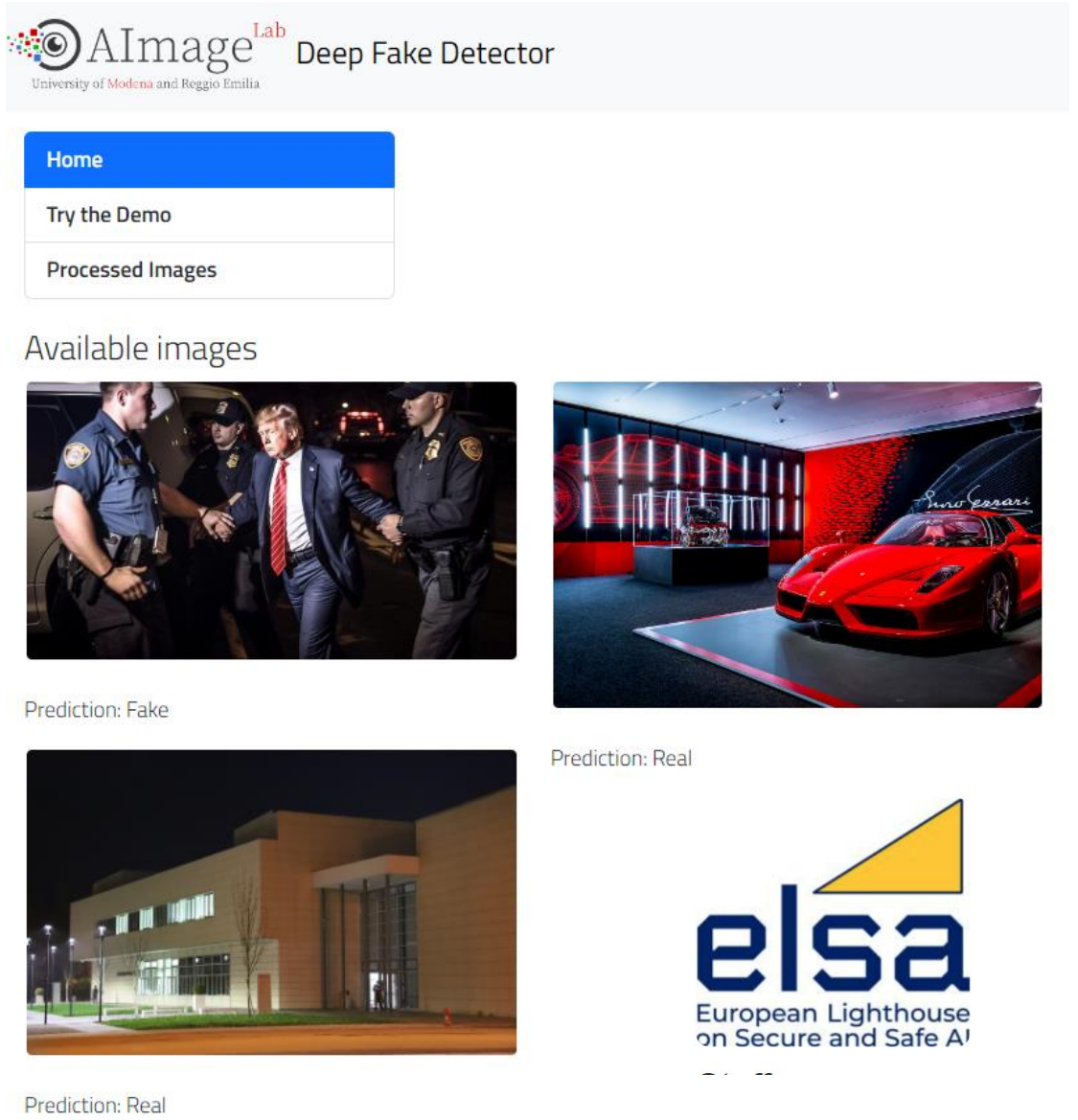
1. R. Amoroso, .D,Morelli L. Baraldi, A.DelBimbo, L.Baraldi, R.Cucchiara. " Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images“ under review at ACM TOMM.

Deep fake detection performance

Backbone	Dataset	Validation Set		Test Set	
		In-Cluster Accuracy	Overall Accuracy	In-Cluster Accuracy	Overall Accuracy
RN50	ImageNet	78.15	95.94	78.62	95.97
ViT-B/32	ImageNet	68.27	91.46	68.06	93.70
CLIP RN50	OpenAI WIT	96.33	99.38	96.72	99.44
CLIP ViT-B/32	OpenAI WIT	95.79	99.27	95.49	99.22
OpenCLIP ViT-B/32	LAION-400M	90.72	98.41	91.57	98.55
OpenCLIP ViT-B/32	LAION-2B	97.93	99.65	98.06	99.66

- High accuracy on all detectors [1] up to 96.6%
- What patterns do detectors learn?
  - **Style**
  - Semantic
  - Bias in generated data

  - Other activities for Fake signature detection



1. R. Amoroso, .M.Cornia A.DelBimbo, L.Baraldi, R.Cucchiara. " Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images“ under review at ACM TOMM.

A next research challenge ICCV 2023

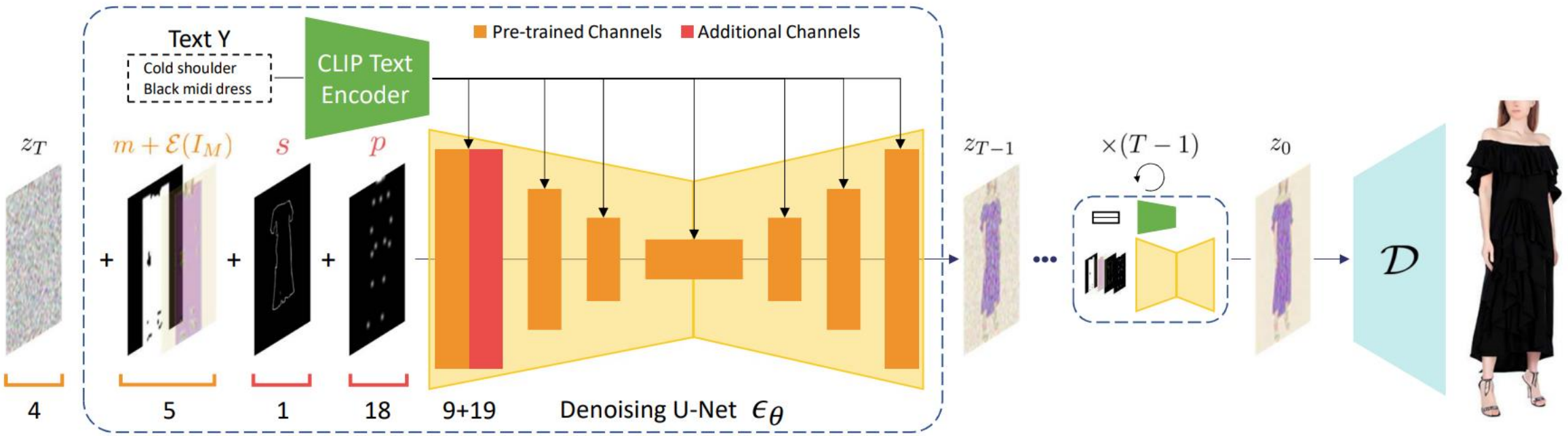
https://ailb-web.ing.unimore.it/dfad2023/

1. *Ask Lrenzo Baraldi, Federico Cocchi*

Multimodal Garment Designer (MGD)

Idea: enhancing Stable Diffusion to work with multiple modalities (*i.e.,* text, human pose, garment sketches)

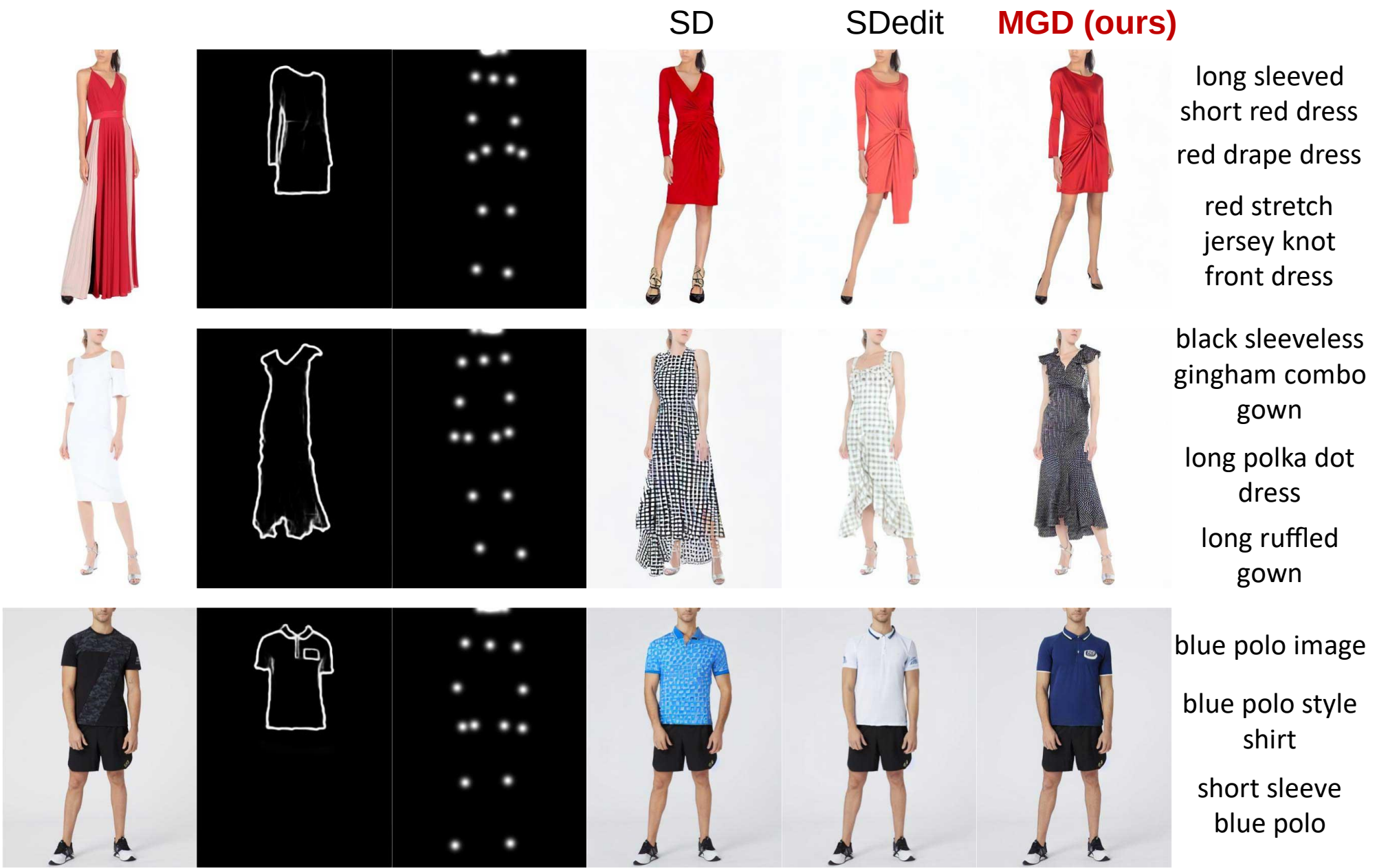
- We extend the denoising U-Net to take additional channels as input.
- This strategy makes it possible to exploit the Stable Diffusion pre-trained weights, giving the model the ability to follow multimodal prompts while preserving the model's characteristics.



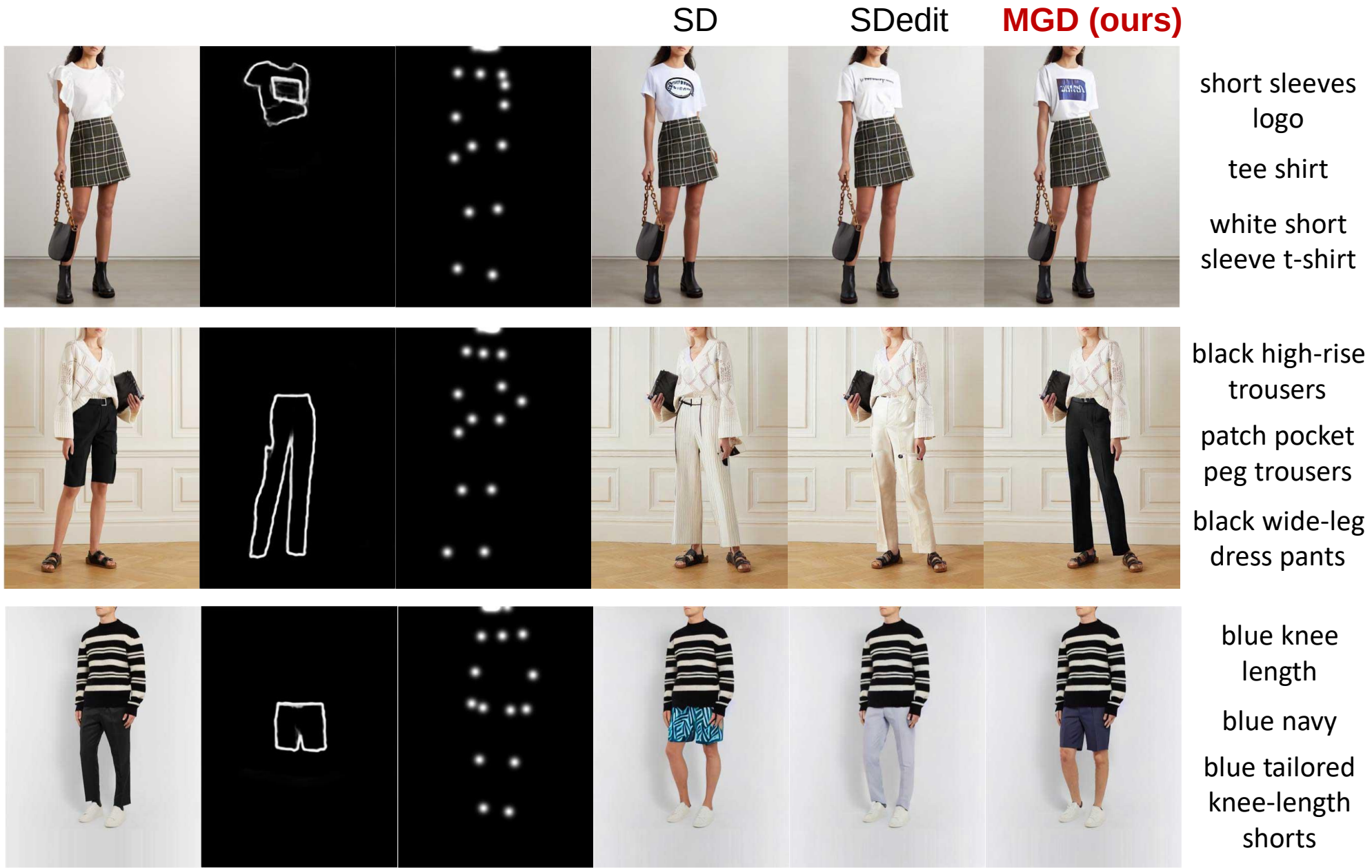
A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, R. Cucchiara "Multimodal Garment Designer: Human-Centric Latent Diffusions Models for Fashion Image Editing"
Under Review, 2023

AImage Lab

Qualitative Results



Qualitative Results



AImage Lab

Qualitative Results

AImage Lab

Qualitative Results

An enormus work under the table…

1. Definition of a very large multimodal Dataset
2. Definition of a platform for human-in-the-loop learning (from noun chunck to sentence
3. Definition a new architecture and a new approach
4. Definition of tests
5. Doing everything with tens of thousands of hours of training (and large use of CINECA GPUS)

Take at home message: Research needs time, effort, human critical mass… and good ideas/preparation

Take at home message 2: you must enjoy in.

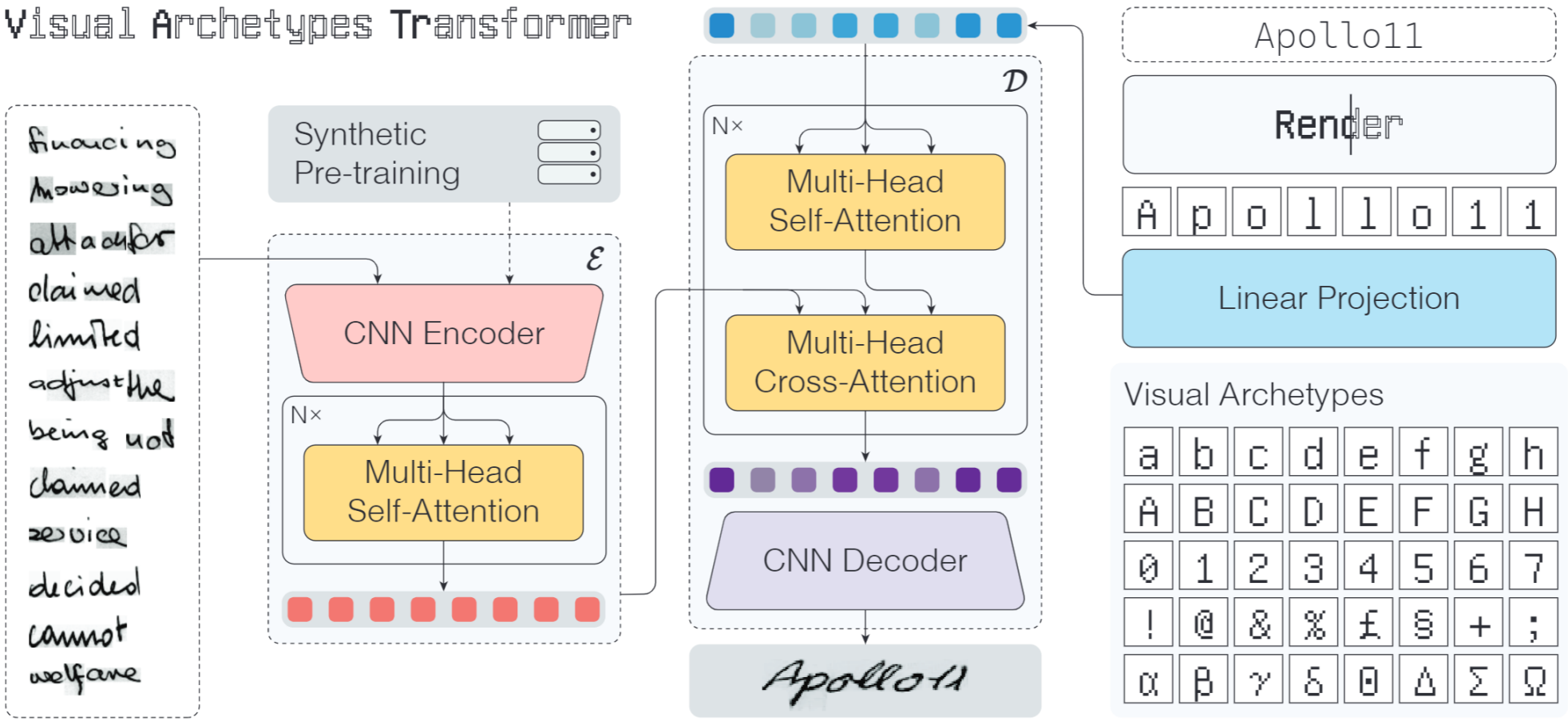
→ D. Morelli si Going in Internship to Amazon for 6 Month

→ We have contact with Gucci and Armani for improving the model.

→ Thanks Phd School in Pisa

Generating personalized images by handwritten promt

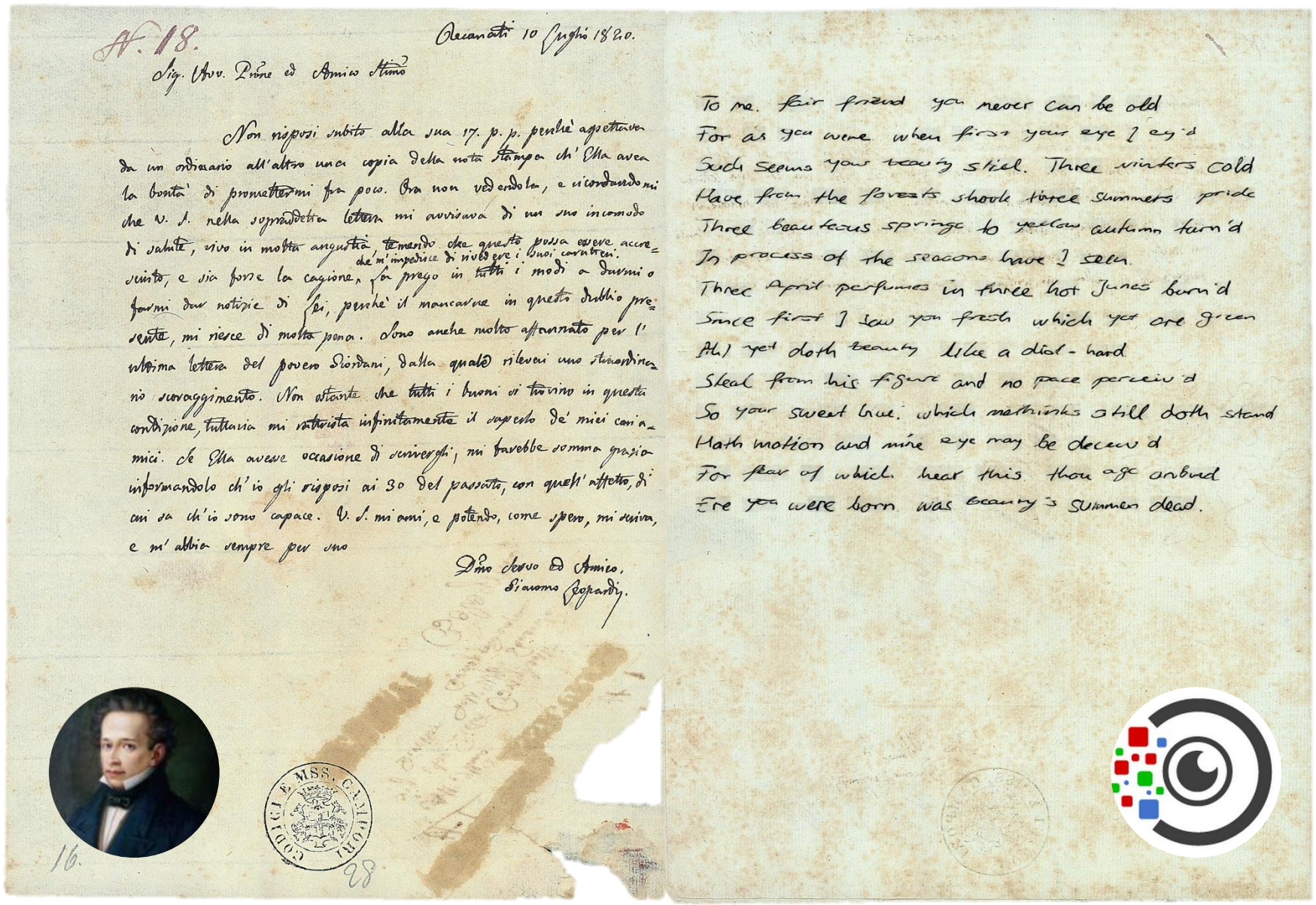
New research I generative AI for handwritten text generation



1. Pippi, V., Cascianelli, S., Cucchiara, R. Handwritten Text Generation from Visual Archetypes. CVPR (2023)

AImage Lab

Automatic Few-Shot HTG with style



N. 18.

Recanati 10 Luglio 1820.

Sig. Avv. Pirone ed Amico Stimo

Non risposi subito alla sua 17. p. p. perchè aspettava da un ordinario all'altro una copia della nota stampa ch' Ella avea la bontà di promettermi fra poco. Ora non vedendola, e ricordandomi che V. S. nella sopraddetta lettera mi avvisava di un suo incomodo di salute, vivo in molta angustia, temendo che questo possa essere accresciuto, e sia forse la cagione, che m'impedisce di rivedere i suoi caratteri. Per prego in tutti i modi a darmi o farmi dar notizie di Lei, perchè il mancarne in questo dubbio presente, mi riesce di molta pena. Sono anche molto affannato per l'ultima lettera del povero Giordani, dalla quale rilevai uno straordinario scoraggimento. Non ostante che tutti i buoni si trovino in questa condizione, tuttavia mi rattrista infinitamente il sapere de' miei carissimi. Se Ella avesse occasione di scrivergli, mi farebbe somma grazia informandolo ch' io gli risposi ai 30 del passato, con quell' affetto, di cui sa ch' io sono capace. V. S. mi ami, e potendo, come spero, mi desia, e m' abbia sempre per suo

Divo devoto ed Amico.
Giacomo Leopardi.

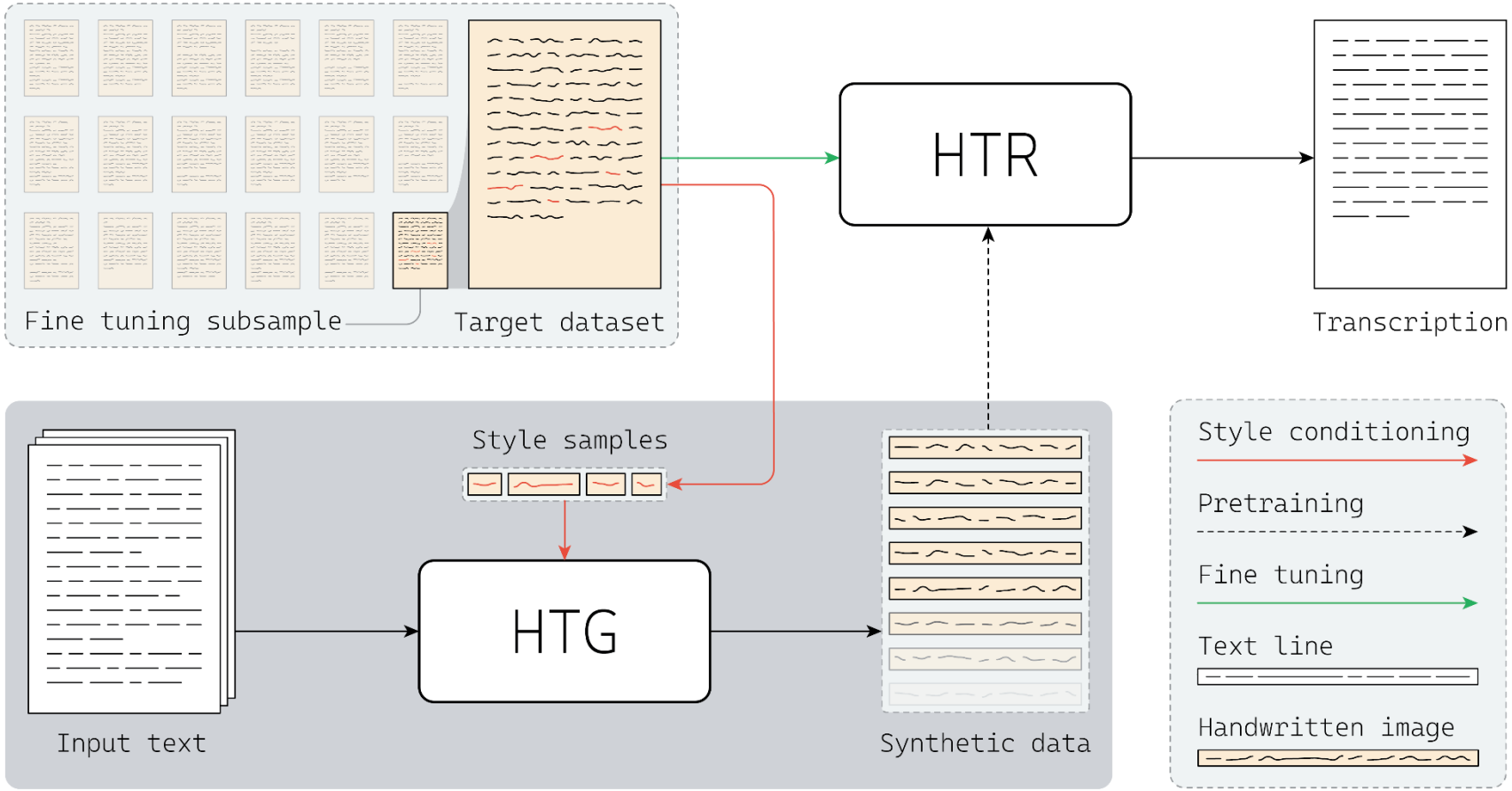
16.

28

To me. fair friend you never can be old
For as you were when first your eye I eyed
Such seems your beauty still. Three winters cold
Have from the forests shook three summers pride
Three beauteous springs to yellow autumn turn'd
In process of the seasons have I seen
Three April perfumes in three hot Junes burn'd
Since first I saw you fresh which yet are green
Ah! yet doth beauty like a dial-hand
Steal from his figure and no pace perceived
So your sweet hue. which methinks still doth stand
Hath motion and mine eye may be deceived
For fear of which hear this thou age unbred
Ere you were born was beauty's summer dead.

And if that isn't enough...

Neural Nets are powerful tools, but are also **data greedy**. To obtain better performance, we need specific training data that allow the model to adapt to new domains



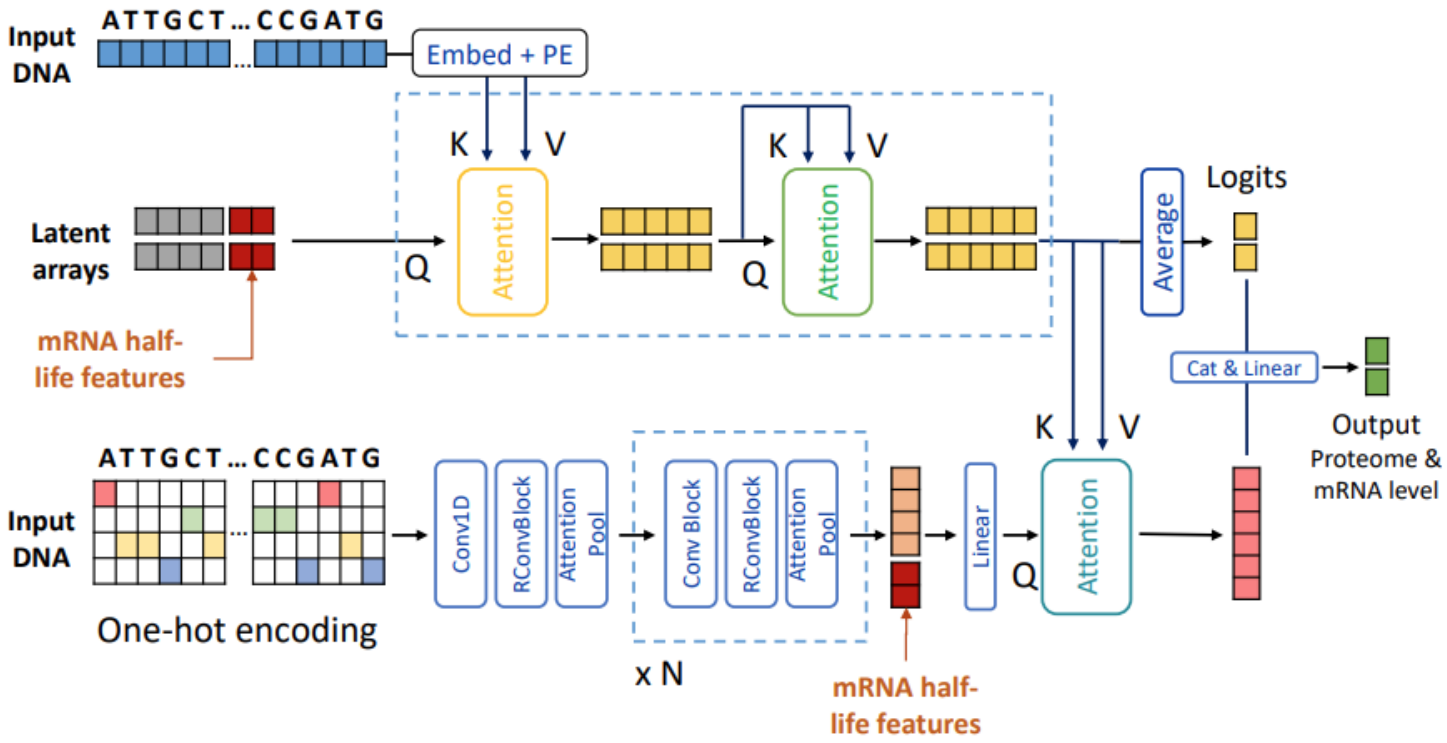
Pippi, V., Cascianelli, S., Kermorvant, C., Cucchiara, R. *How to Choose Pretrained Handwriting Recognition Models for Single Writer Fine-Tuning.* ICDAR (2023)

Generative Multimodality is the same with genoma

The power of the model is that you can apply everywhere

A generative Model for Protein and mRNA level from DNA encoding for Genoma Amalysis

EU DECIDER PRoject*



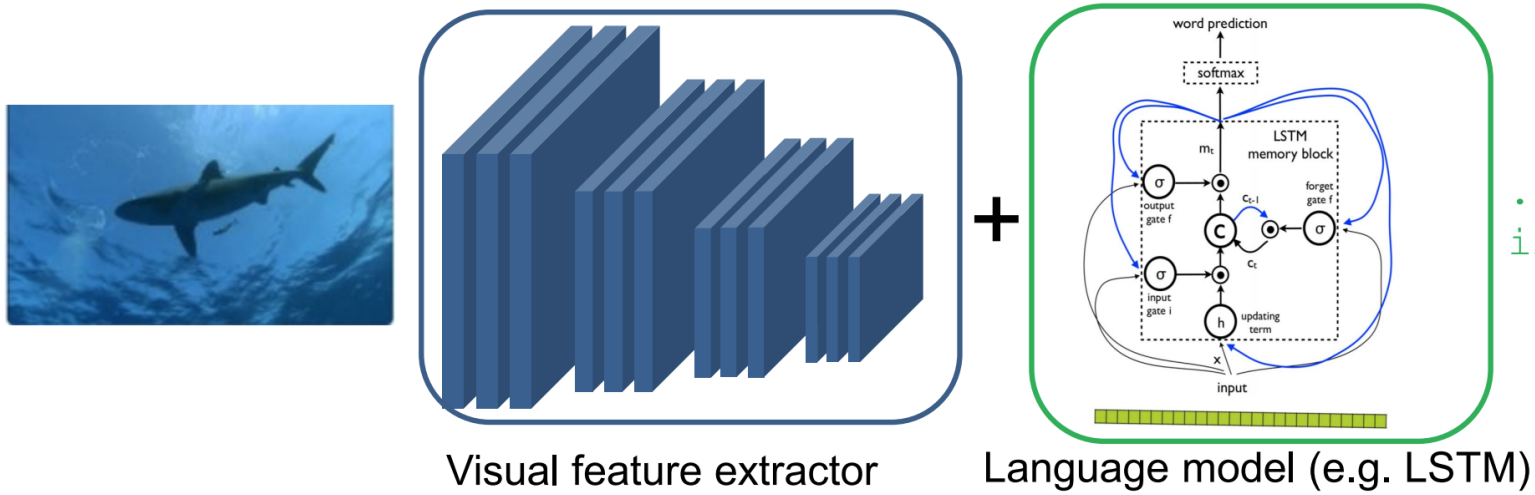
Towards cross-modality generative

A large research activity: **From visual –to-text (image captioning)**

Since 2015 [1,..3] joining visual feature extractors and language models

VISUAL DATA UNDERSTANDING

NATURAL LANGUAGE GENERATION

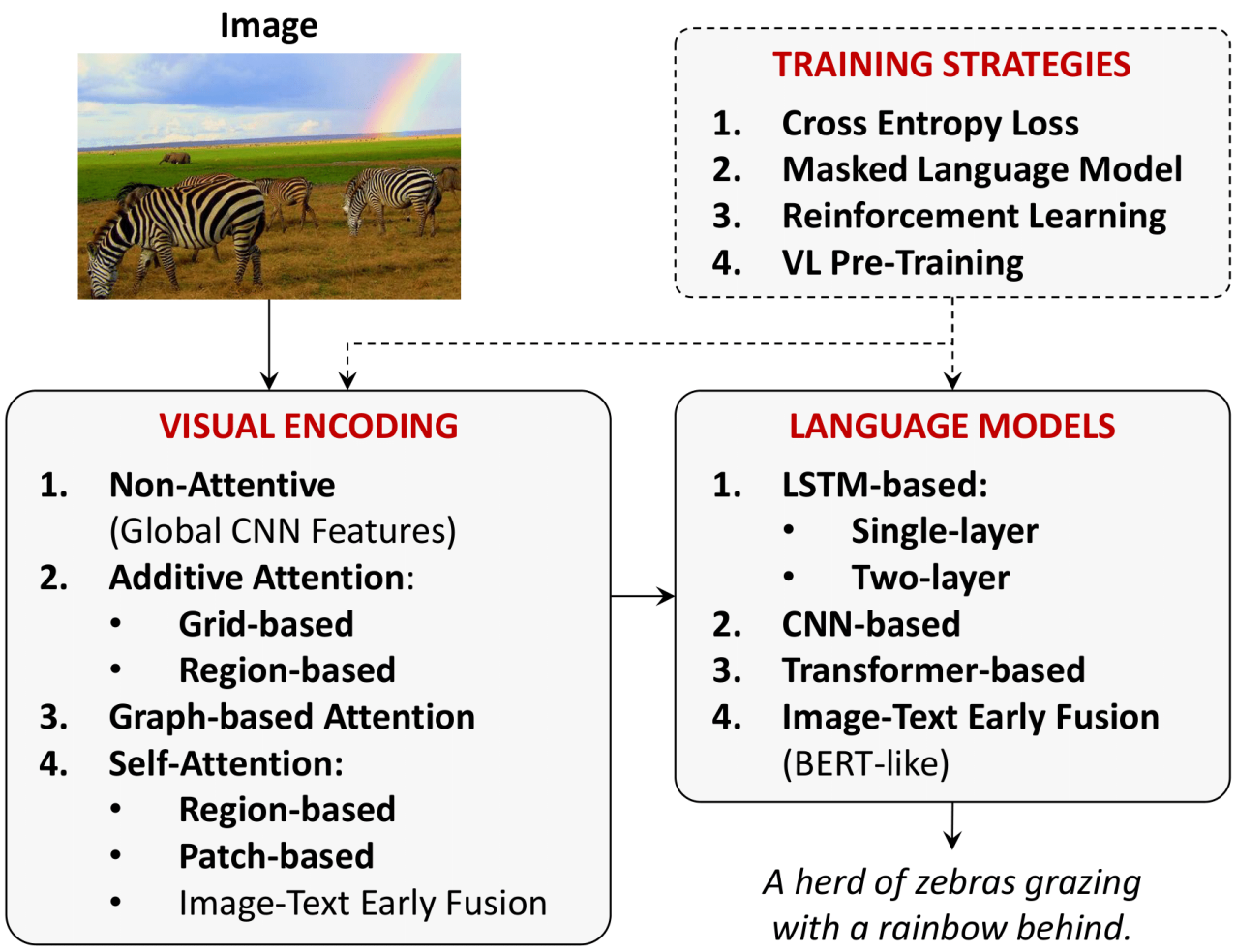


..a white shark swims in the ocean water..

- Mainly supervised (by human annotation)
- Trained by visual and language models

1. Karpathy, A., & Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *CVPR 2015*.
2. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. Show and tell: A neural image caption generator. In *CVPR 2015*.
3. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR 2015*.

The Image Captioning Journey



IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

From Show to Tell: A Survey on Deep Learning-based Image Captioning

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara

Abstract—Connecting Vision and Language plays an essential role in Generative Intelligence. For this reason, large research efforts have been devoted to image captioning, *i.e.* describing images with syntactically and semantically meaningful sentences. Starting from 2015 the task has generally been addressed with pipelines composed of a visual encoder and a language model for text generation. During these years, both components have evolved considerably through the exploitation of object regions, attributes, the introduction of multi-modal connections, fully-attentive approaches, and BERT-like early-fusion strategies. However, regardless of the impressive results, research in image captioning has not reached a conclusive answer yet. This work aims at providing a comprehensive overview of image captioning approaches, from visual encoding and text generation to training strategies, datasets, and evaluation metrics. In this respect, we quantitatively compare many relevant state-of-the-art approaches to identify the most impactful technical innovations in architectures and training strategies. Moreover, many variants of the problem and its open challenges are discussed. The final goal of this work is to serve as a tool for understanding the existing literature and highlighting the future directions for a research area where Computer Vision and Natural Language Processing can find an optimal synergy.

Index Terms—Image Captioning, Vision-and-Language, Deep Learning, Survey.

1 Introduction

Image captioning is the task of describing the visual content of an image in natural language, employing a visual understanding system and a language model capable of generating meaningful and syntactically correct sentences. Neuroscience research has clarified the link between human vision and language generation only in the last few years [1]. Similarly, in Artificial Intelligence, the design of architectures capable of processing images and generating language is a very recent matter. The goal of these research efforts is to find the most effective pipeline to process an input image, represent its content, and transform that into a sequence of words by generating connections between visual and textual elements while maintaining the fluency of language.

The early-proposed approaches to image captioning have entailed description retrieval [2], [3], [4], [5], [6], [7] or template filling and hand-crafted natural language generation techniques [8], [9], [10], [11], [12], [13], [14], [15]. While these have been treated in other surveys [16], [17], [18], image captioning is currently based on the usage of deep learning-based generative models. In its standard configuration, the task is an image-to-sequence problem whose inputs are pixels. These inputs are encoded as one or multiple feature vectors in the visual encoding step, which prepares the input for a second generative step, called the language model. This produces a sequence of words or subwords decoded according to a given vocabulary.

In these few years, the research community has improved model design considerably: from the first deep learning-based proposals adopting Recurrent Neural Networks (RNNs) fed with global image descriptors, methods have been enriched with attentive approaches and reinforcement learning up to the breakthroughs of Transformers and self-attention and single-stream BERT-like approaches. At the same time, the Computer Vision and Natural Language Processing (NLP) communities have addressed the challenge of building proper evaluation protocols and metrics to compare results with human-generated ground-truths. However, despite the investigation and improvements achieved in these years, image captioning is still far from being considered a solved task.

Several domain-specific proposals and variants of the task have also been investigated to accommodate for different user needs and descriptions styles. According to [19], [20], indeed, image captions can be perceptual, when focusing on low-level visual attributes; non-visual, when reporting implicit and contextual information; conceptual, when describing the actual visual content (*e.g.* visual entities and their relations). While the latter is commonly recognized as the target of the image captioning task, this definition encompasses descriptions focusing on different aspects and at various levels of detail (*e.g.* including attributes or not, mentioning named entities or high-level concepts only, describing salient parts only, or also finer details).

With the aim of providing a testament to the journey that captioning has taken so far, and with that of encouraging novel ideas, we trace a holistic overview of techniques, models, and task variants developed in the last years. Furthermore, we review datasets and evaluation metrics and perform quantitative comparisons of the main approaches. Finally, we discuss open challenges and future directions.

• *M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, and R. Cucchiara are with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy.*
E-mail: {matteo.stefanini, marcella.cornia, lorenzo.baraldi, silvia.cascianelli, rita.cucchiara}@unimore.it.
• *G. Fiameni is with NVIDIA AI Technology Centre, Italy.*
E-mail: gfiameni@nvidia.com

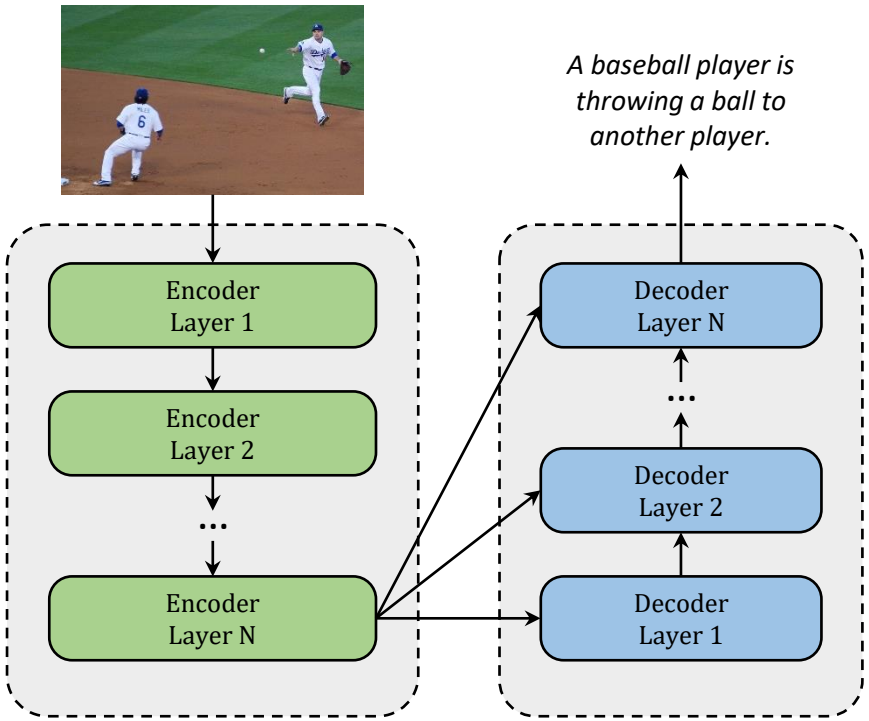
Manuscript received July, 2021; revised November, 2021.

1. M Stefanini, M Cornia, L Baraldi, S Cascianelli, G Fiameni, R Cucchiara From show to tell: a survey on deep learning-based image captioning IEEE Transactions on Pattern Analysis and Machine Intelligence 2022

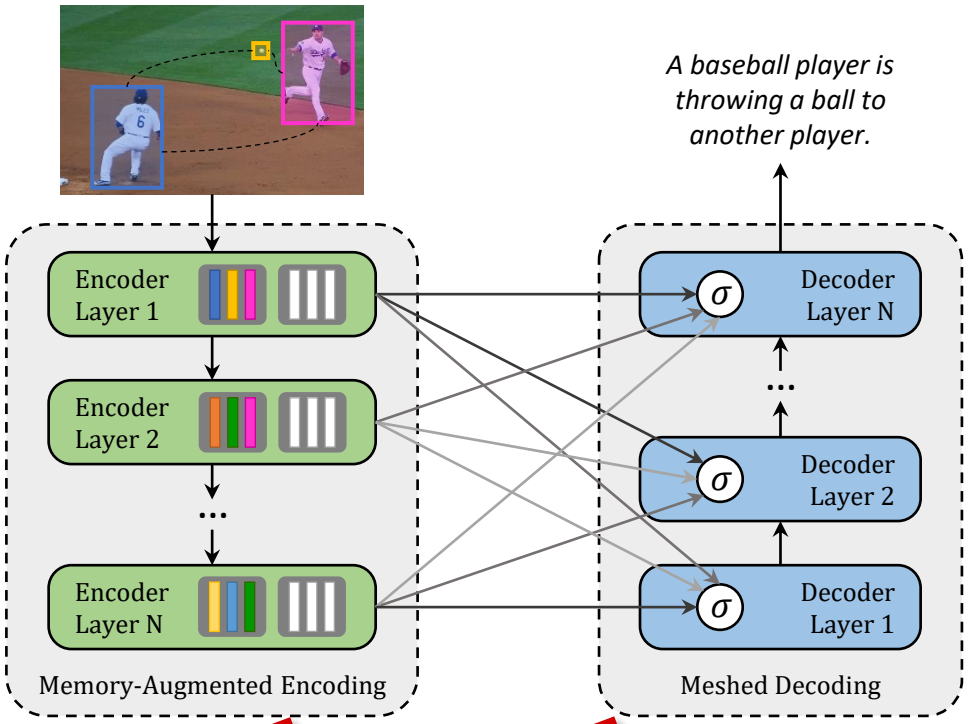
Meshed-Memory Transformer

A jump over the shoulder of giants…… let us modeling new ideas

Original Transformer



M² Transformer



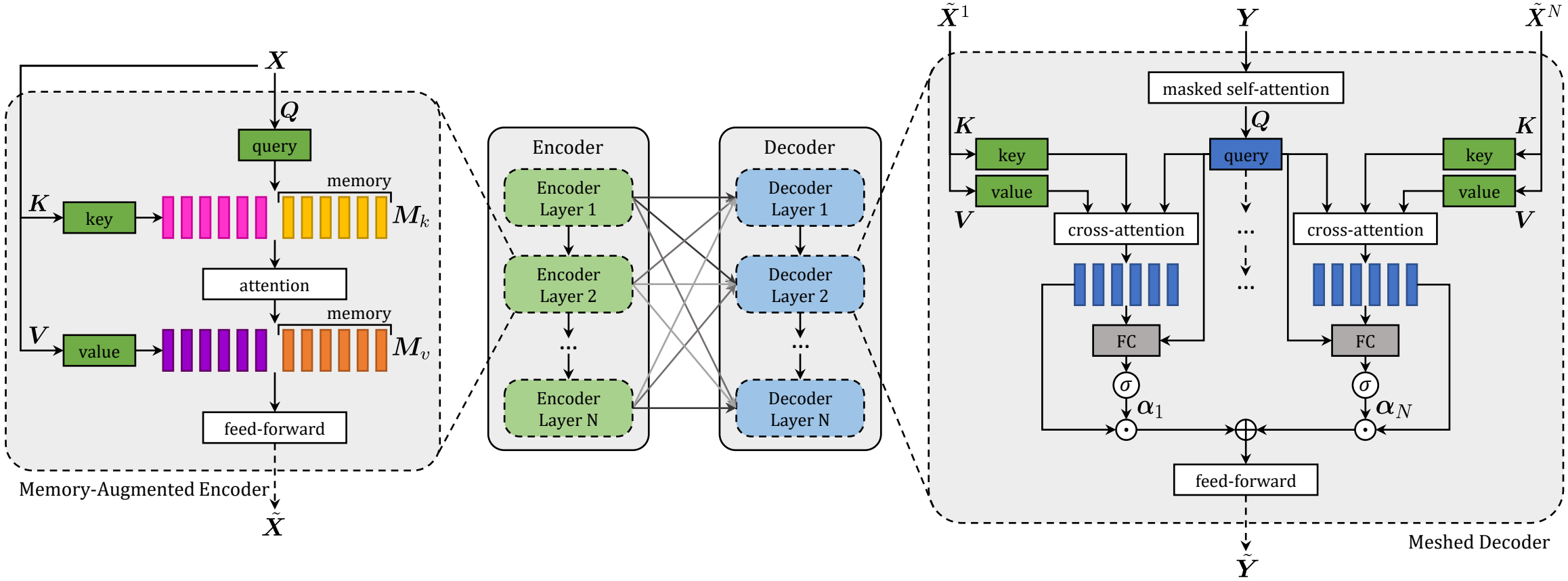
Relationships between image regions are modeled via **persistent memory vectors**.

Encoder and decoder layers are connected in **a mesh-like structure**.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In CVPR 2020.

Meshed-Memory Transformer

- In our encoder, the set of keys and values is extended **with learnable vectors** that can encode a **priori information**.
- A mesh connectivity is operated through a **learnable gating mechanism** which modulates the contribution of each encoder layer during cross attention.



Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In CVPR 2020.

Meshed-Memory Transformer

		BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE		CIDEr	
		c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
CVPR 2017	SCST [33]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
CVPR 2018	Up-Down [4]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
ICCV 2019	RDN [18]	80.2	95.3	-	-	-	-	37.3	69.5	28.1	37.8	57.4	73.3	121.2	125.2
ECCV 2018	RFNet [15]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
ECCV 2018	GCN-LSTM [48]	80.8	95.9	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
CVPR 2019	SGAE [46]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
ICCV 2019	ETA [24]	81.2	95.0	65.5	89.0	50.9	80.4	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
ICCV 2019	AoANet [14]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
ICCV 2019	GCN-LSTM+HIP [49]	**81.6**	95.9	66.2	90.4	51.5	81.6	39.3	71.0	28.8	38.1	59.0	74.1	127.9	130.2
	$\mathcal{M}^2$ Transformer	**81.6**	**96.0**	**66.4**	**90.8**	**51.8**	**82.7**	**39.7**	**72.8**	**29.4**	**39.0**	**59.2**	**74.8**	**129.3**	**132.1**

→ At the beginning of 2020, our model reached the first place in the COCO leaderboard.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In CVPR 2020.

M^2 Transformer: Results

Ground-truth: A truck parked near a tall pile of hay.

Transformer: A truck is parked in the grass in a field.

M^2 Transformer: A green truck parked next to a pile of hay.

Ground-truth: A cat looking at his reflection in the mirror.

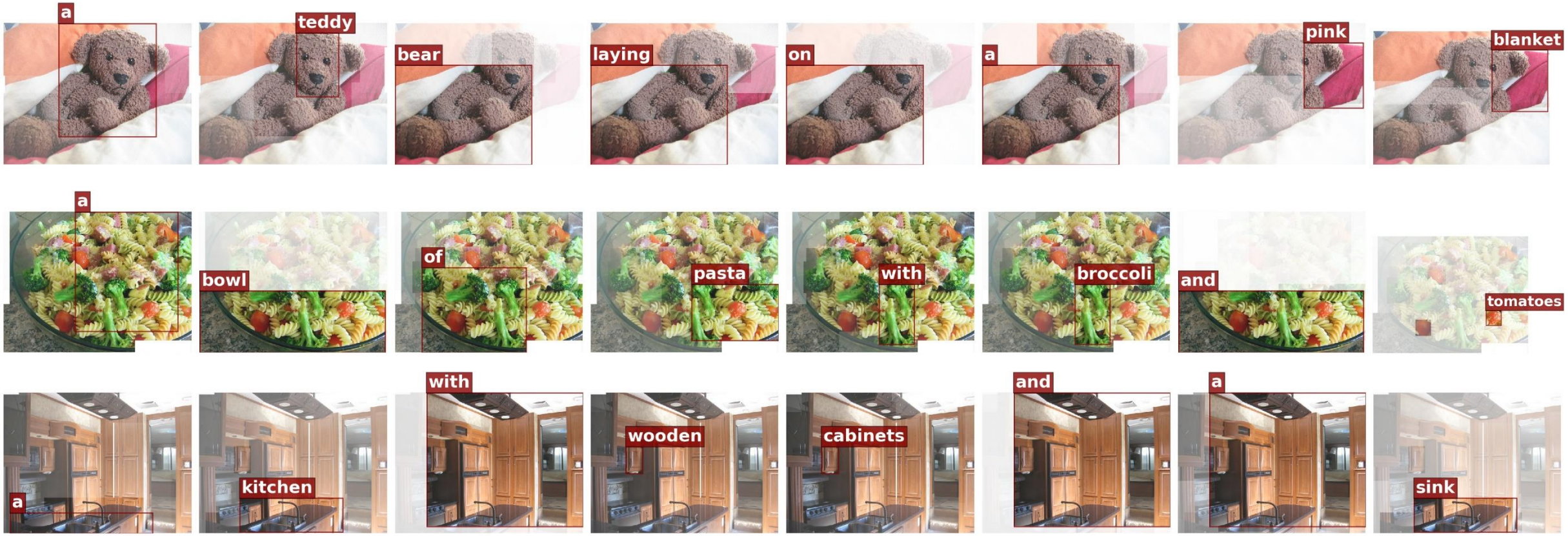
Transformer: A cat sitting in a window sill looking out.

M^2 Transformer: A cat looking at its reflection in a mirror.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-Memory Transformer for Image Captioning." CVPR 2020.

M^2 Transformer: Attention Visualizations

- A bit of explainability: To visualize the attended image regions, we employ the **Integrated Gradients method** which approximates the integral of gradients with respect to the input.



Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-Memory Transformer for Image Captioning." CVPR 2020.

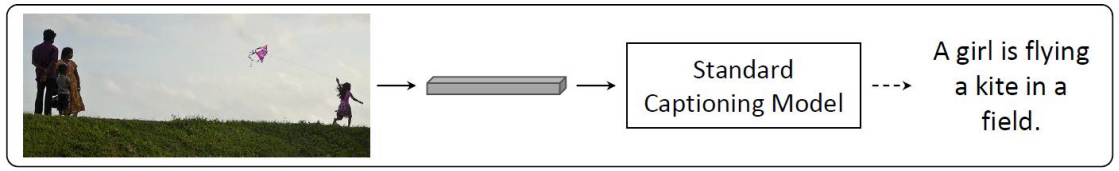
AImage Lab

Explainability through Captioning

A new challenge: Controllable Image Captioning?

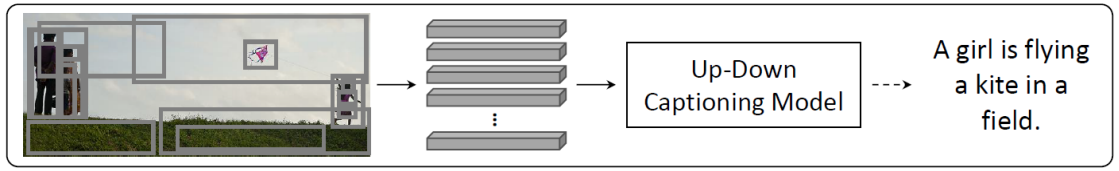
Early captioning approaches:

- Global image feature vector



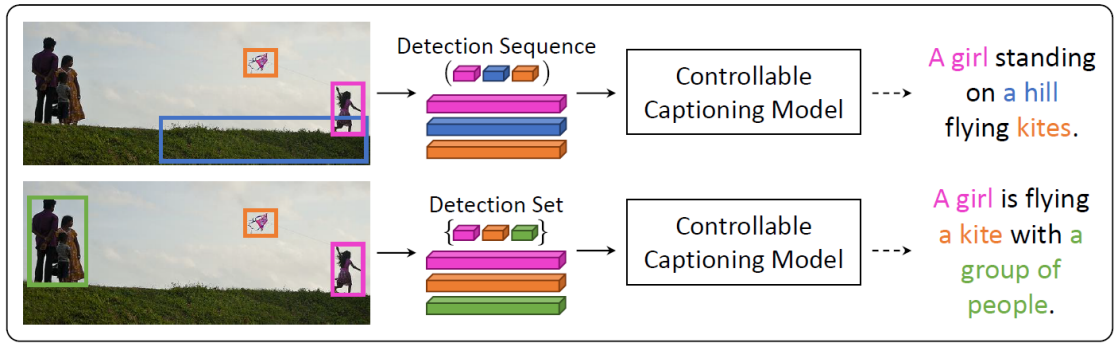
Attention-based approaches:

- Weakly interpretable (through attention)
- Not controllable.
  - We can't decide which regions get processed
  - No control over the generation process.



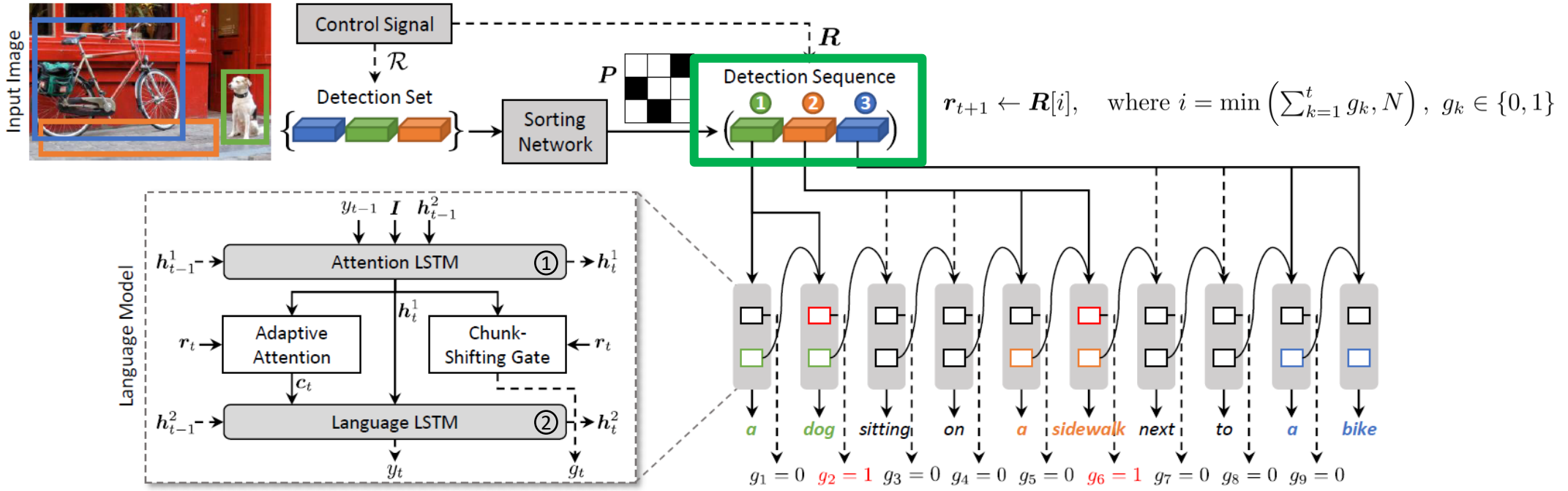
Show, control and tell

- Controllable via regions
  - A sequence (ordered)
  - A set (unordered)



Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

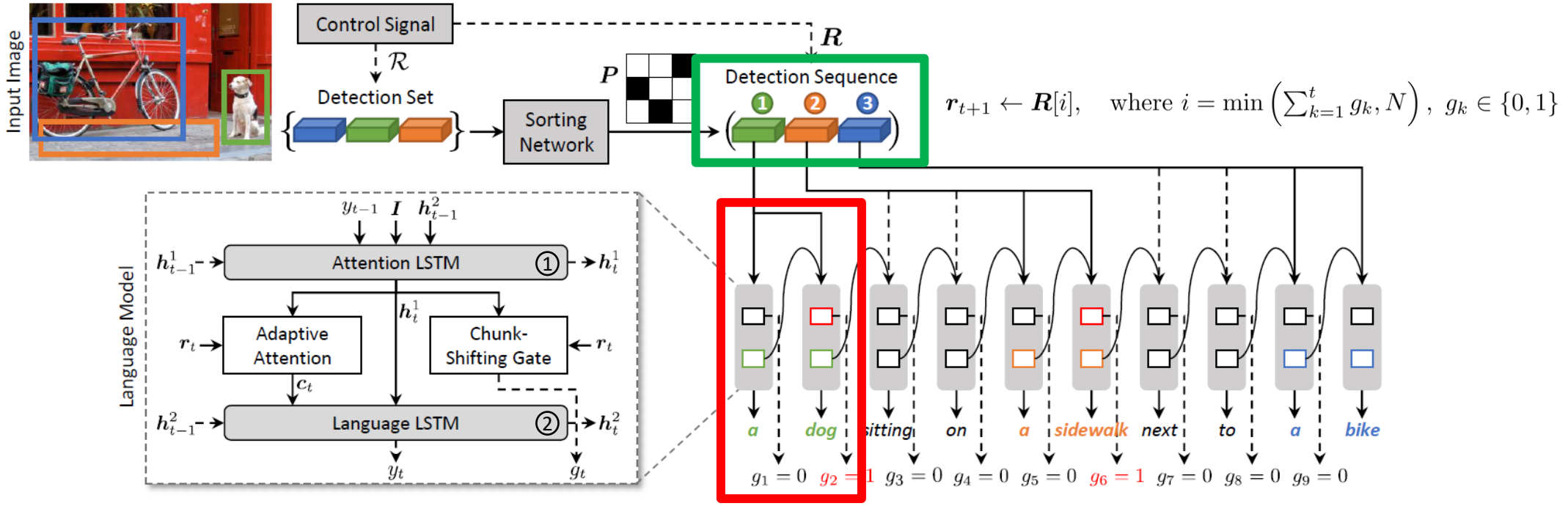
Controllable Captions with LTMs (2019)



- Language model takes as input **a sequence of regions**
- Switches between one region and the next one via a learned chunk-shifting gate
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

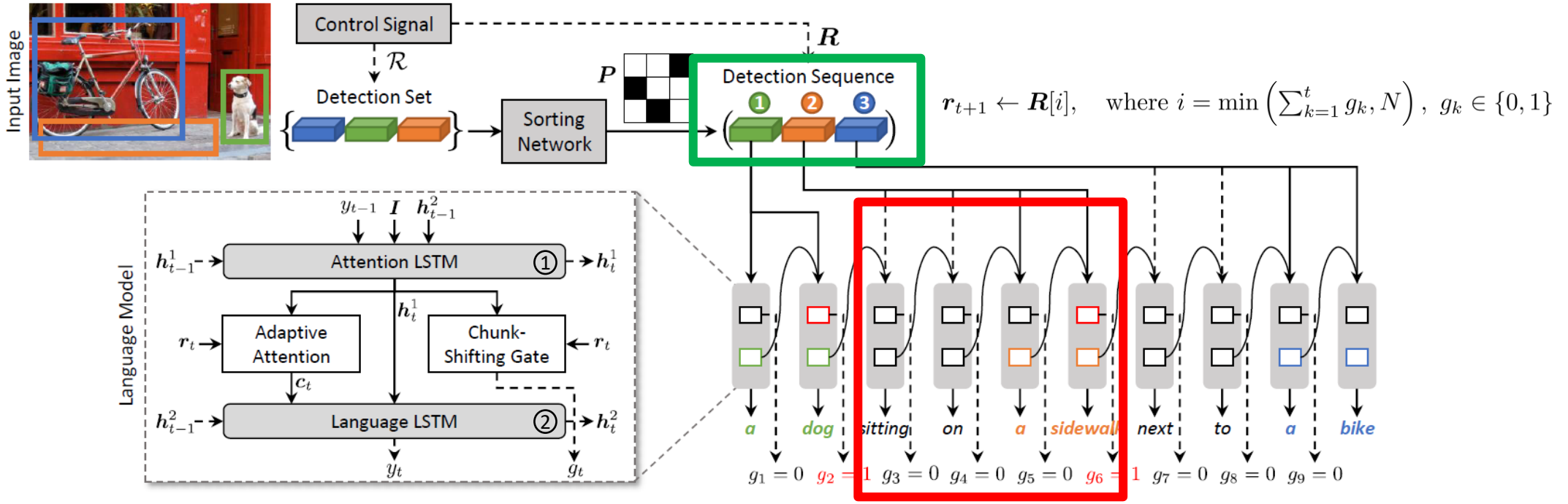
Generating Controllable Captions



- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

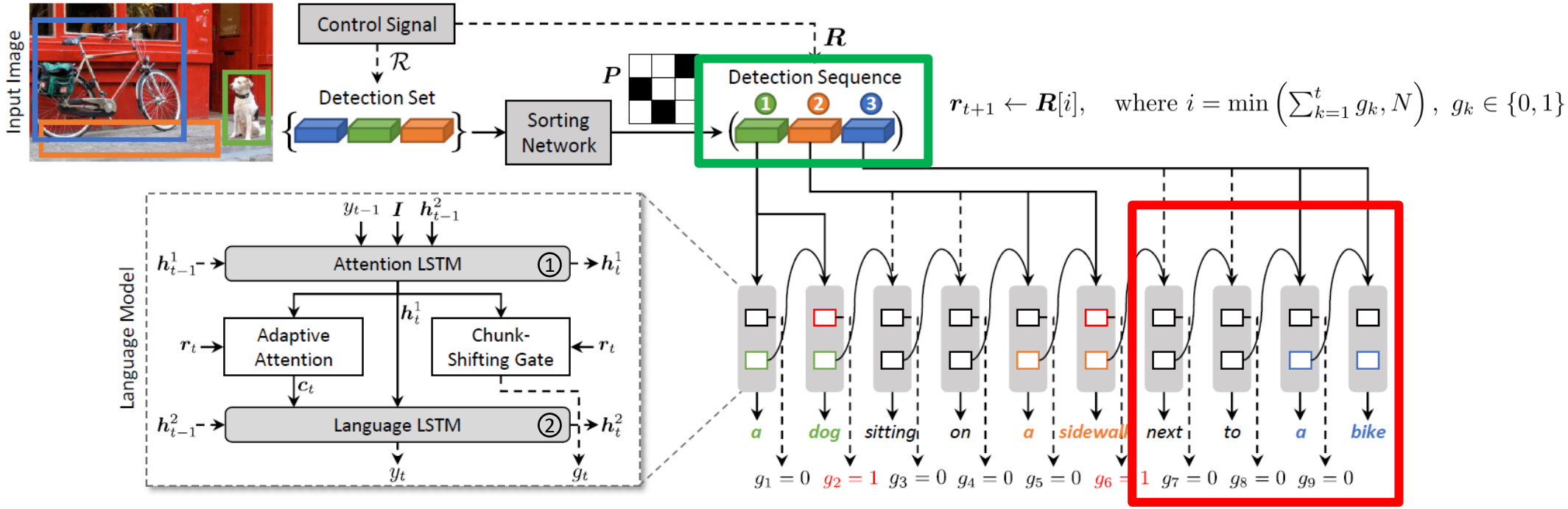
Generating Controllable Captions



- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Generating Controllable Captions

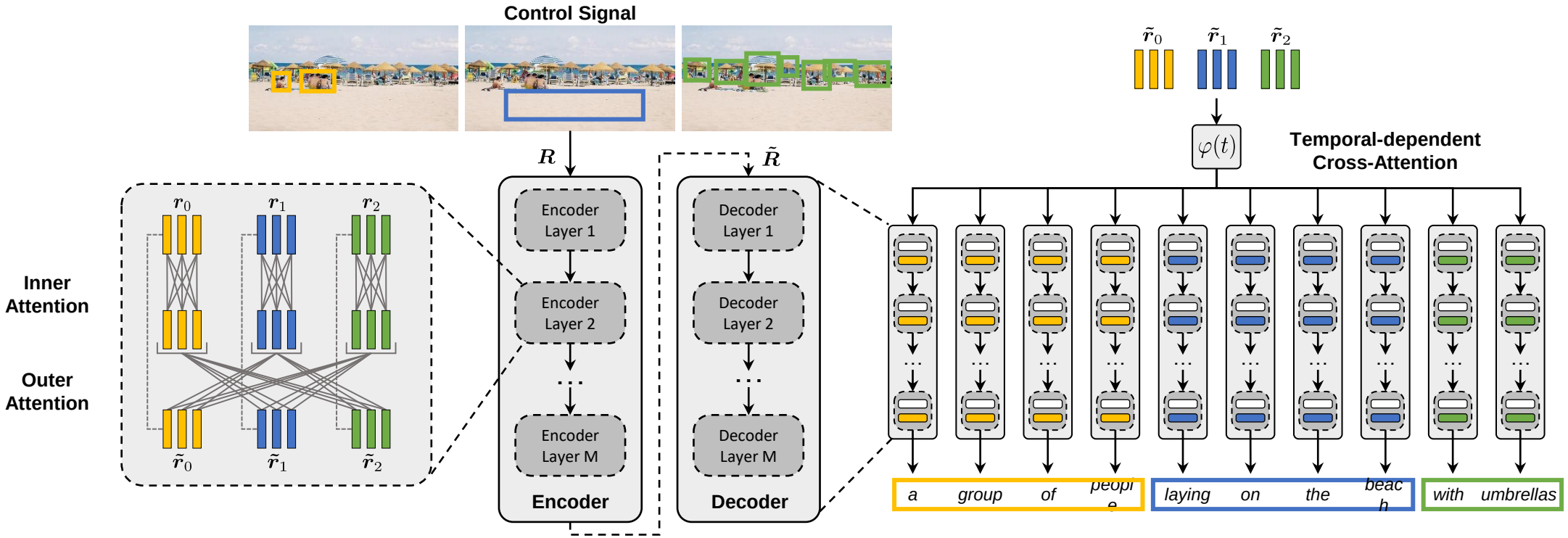


- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllable Captions with transFormers (2021)

Now a similar approach with Transformer

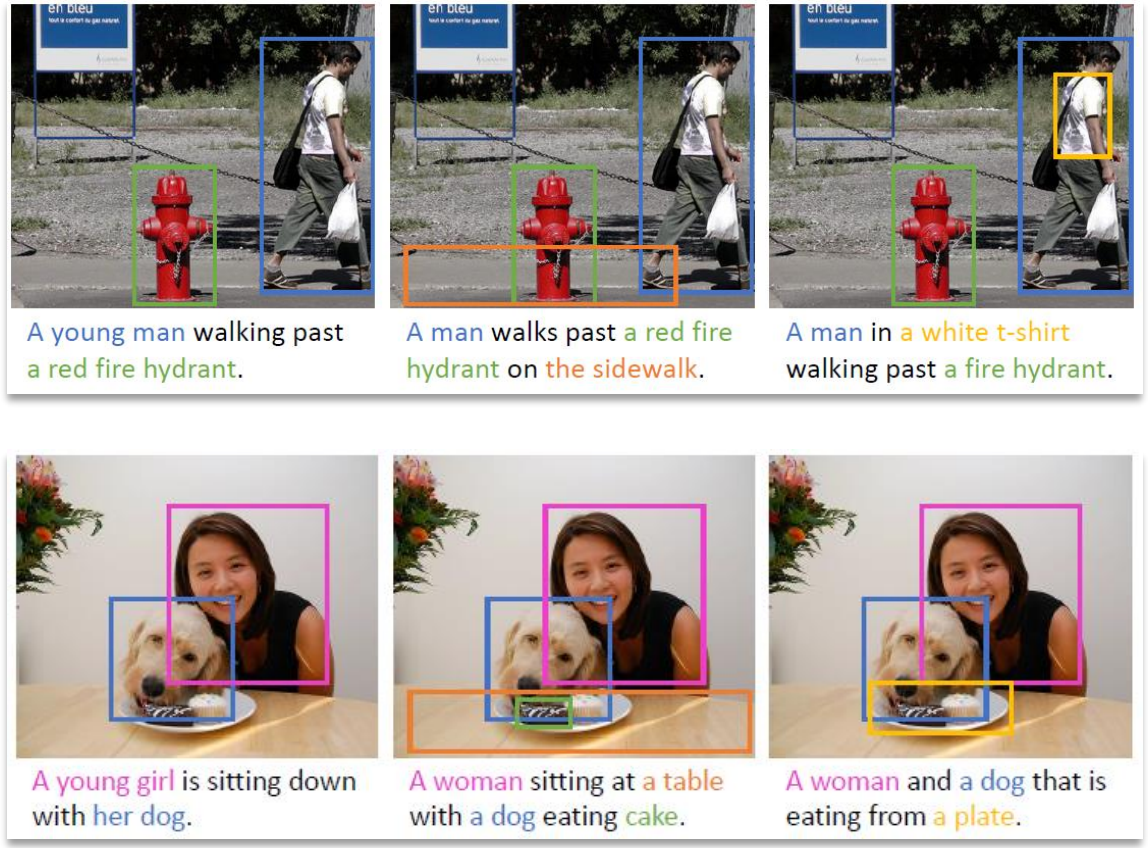


Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Fully-Attentive Iterative Networks for Region-Controlled Image and Video Captioning." Under Review.

NEW COCO Entities Dataset

For training and evaluation, we collect COCO-Entities
→ more than 120,000 images

- COCO with noun chunks associated to regions
- Semi-automatically annotated



Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllable Captioning Results

Controllability via a sequence of regions

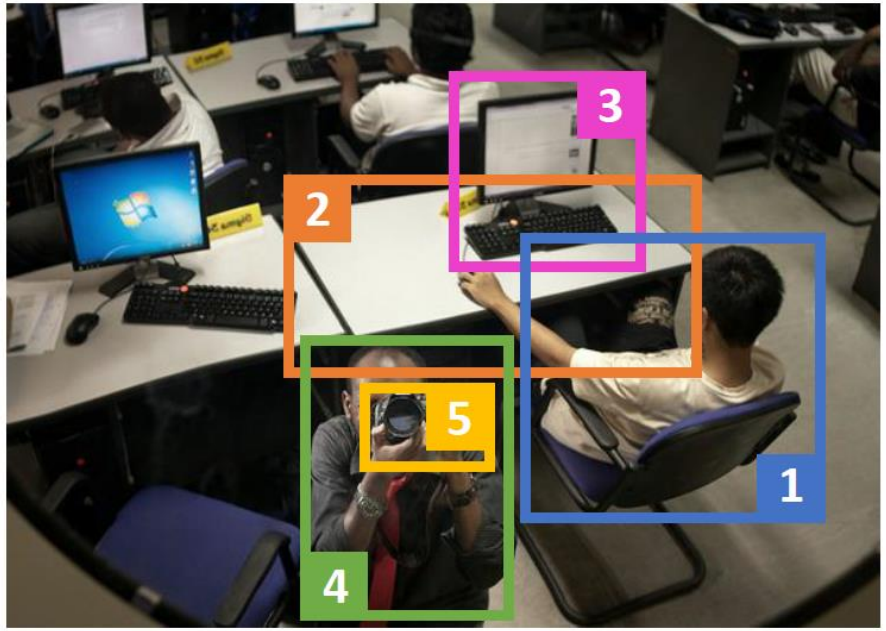
Method	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW
Neural Baby Talk	12.9	19.2	40.4	120.2	29.5	0.305	-	-	-	-	-	-	-	-	-	-	-	-
Up-Down	12.9	19.3	40.0	119.9	29.3	0.296	14.2	20.0	42.1	133.9	30.0	0.310	-	-	-	-	-	-
$\mathcal{M}^2$ Transformer	13.9	20.2	41.5	130.3	31.0	0.314	16.5	22.1	44.5	154.5	33.4	0.345	-	-	-	-	-	-
Oscar	14.0	21.9	42.0	134.9	31.8	0.301	16.5	22.1	45.4	155.8	34.4	0.334	-	-	-	-	-	-
LSTM-based																		
Controllable Up-Down	17.3	23.0	46.7	161.0	39.1	0.396	17.4	22.9	47.1	168.5	39.0	0.397	17.9	23.6	48.2	171.3	40.7	0.443
Show, Control and Tell	**20.9**	**24.4**	**52.5**	**193.0**	**45.3**	**0.508**	**22.5**	**25.6**	**55.1**	**210.1**	**48.1**	**0.615**	**22.3**	**25.6**	**55.3**	**209.7**	**48.5**	**0.649**
Transformer-based																		
Controllable Transformer	18.9	24.3	47.9	177.5	41.1	0.431	20.2	25.0	49.6	194.0	42.8	0.464	20.3	25.0	49.7	194.0	42.8	0.468
Show, Control and Tell	**25.0**	**27.4**	**56.2**	**225.0**	**49.4**	**0.623**	**26.1**	**28.0**	**57.8**	**238.6**	**50.2**	**0.671**	**26.5**	**28.0**	**58.1**	**243.4**	**51.1**	**0.683**

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Fully-Attentive Iterative Networks for Region-Controlled Image and Video Captioning." Under Review.

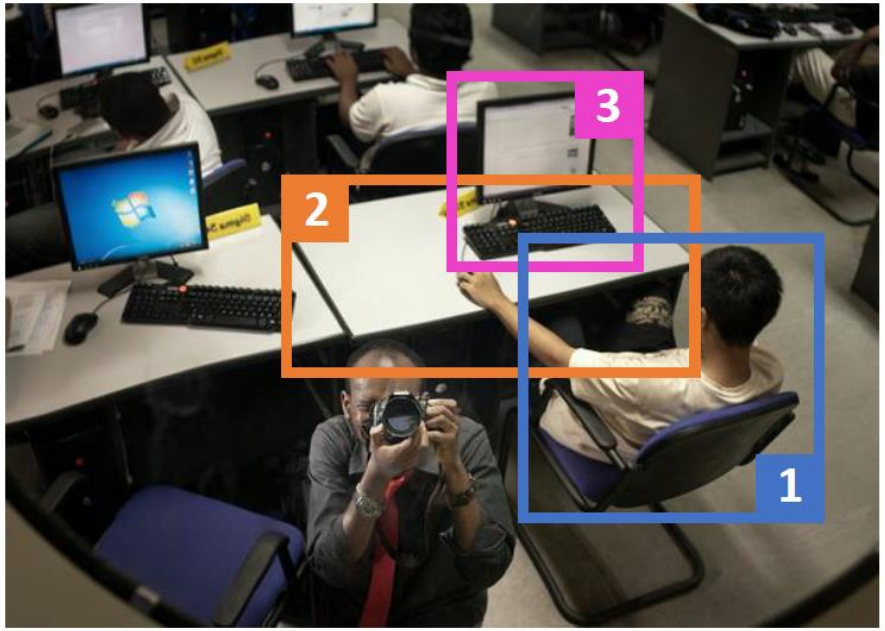
Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllability via a Sequence of Regions

Results when controlling with a sequence of regions



A man sitting at a desk with a computer and a man holding a camera.

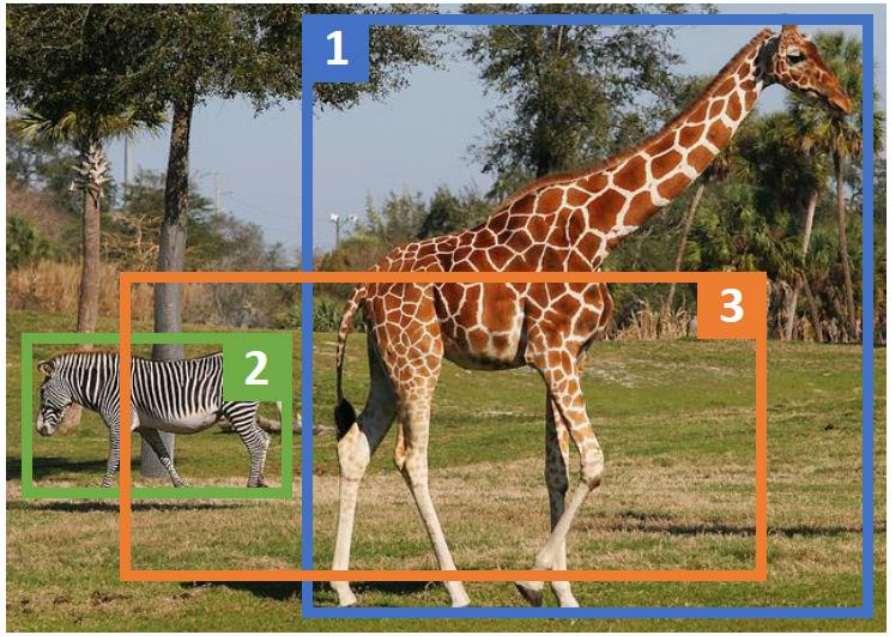


A man sitting at a desk with a computer.

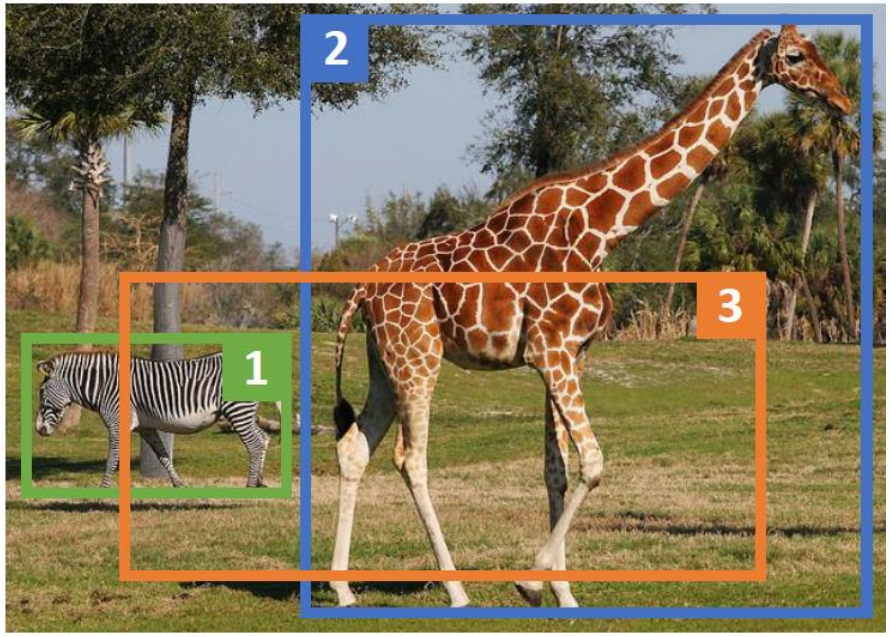
Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllability via a Sequence of Regions

Results when controlling with a sequence of regions



A giraffe standing in front of a zebra in a field.

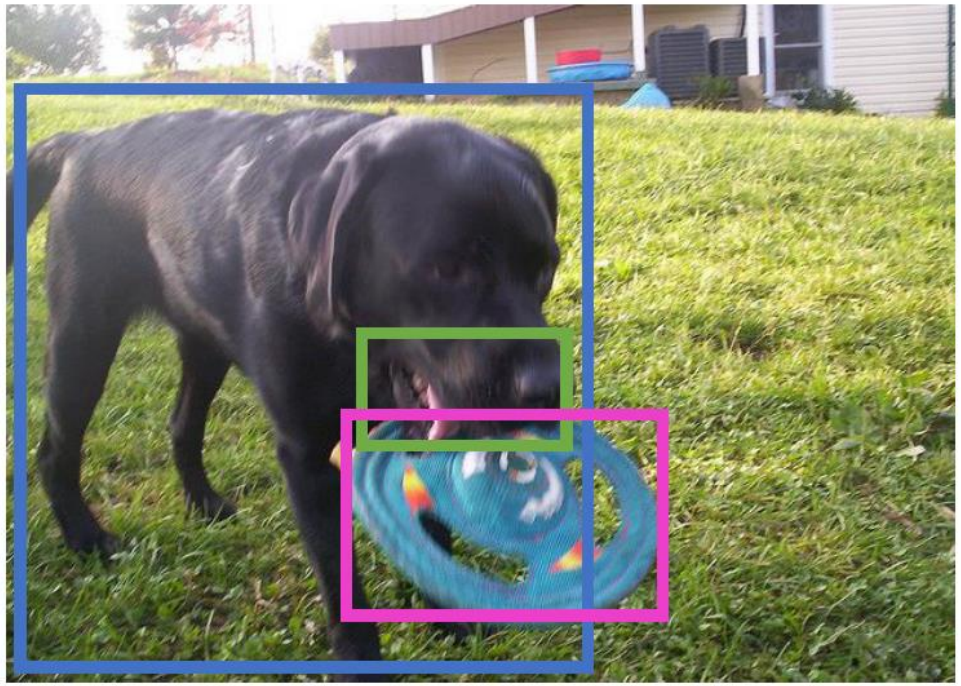


A zebra standing next to a giraffe in a field.

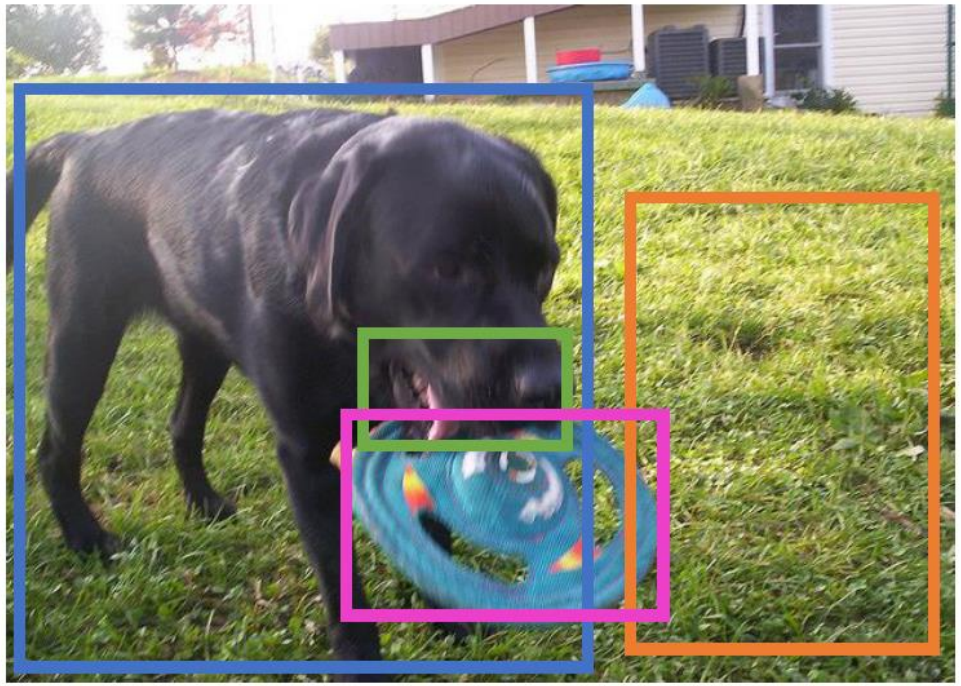
Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllability via a Set of Regions

Results when controlling with a set of regions



A dog holding a frisbee in its mouth.



A dog standing in the grass with a frisbee in its mouth.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllability via a Set of Regions

Results when controlling with a set of regions

A man in a black jacket skiing down a hill.

A man on skis down a snow covered slope.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

5. CHALLENGES IN MULTIMODAL FOUNDATION MODELS

A future challenge 1 : the measure

Large scale multimodal models revolutionized CV; NLP, Generative aI etc

Many images, text etc can be generated

HOW CAN WE EVALUATE THE GOODNESS OF GENERATIVE AI?

The problem of EVALUATION

Captioning, text description , VQA…

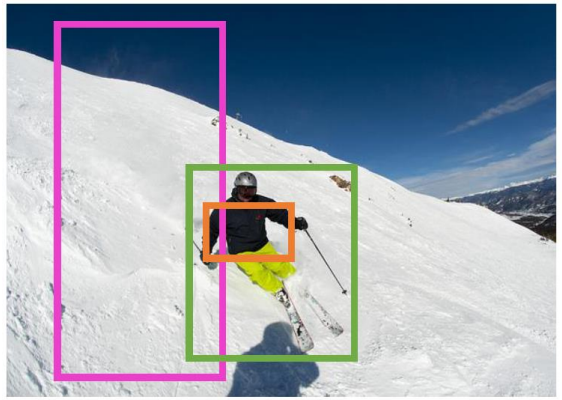
but how to evaluate them?

- Some Linguistic Measures
- Some Captioning Measures (CiDer)
- Some Captioning Measues based on FM (Clip-S)
- Human-Feedback

Now, focus on:

- **New evaluation metrics (CVPR 2023)**

What is better??



A man in a black jacket skiing down a hill.

A man on skis down a snow covered slope.

What is better??

Standard Captioner:
A group of people riding skateboards in a field.

Universal Captioner:
*A group of people riding **segways** in a field.*

Standard Captioner:
A tall building sitting in the middle of a body of water.

Universal Captioner:
*An aerial view of the **Burj Al Arab** in **Dubai**.*

Standard Captioner:
A woman with blonde hair is posing for a picture.

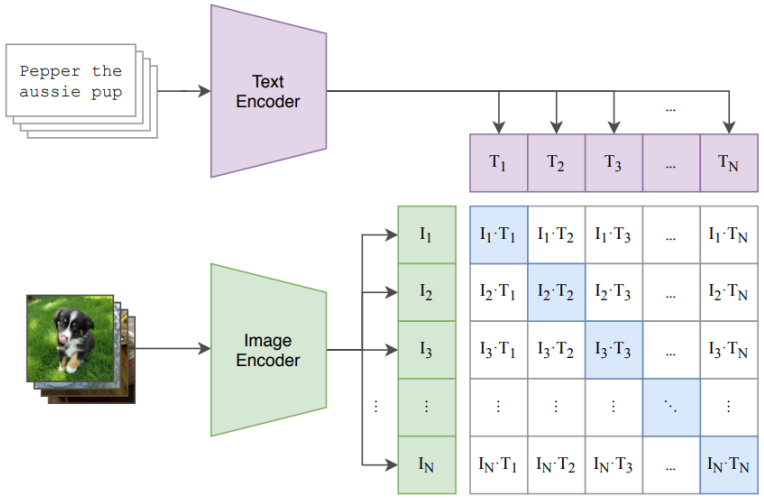
Universal Captioner:
*A picture of **Marilyn Monroe** with a red lipstick.*

MeasuRes

Is so difficult to measure captions

and often GT captions (e.g. in Coco) have no sense..

Existing metrics for image-text correspondence are either only based on **(few) human references** or multi-modal embeddings trained on **noisy data**.



A silver bicycle is parked in a living room.

METEOR	CIDEr	CLIP-S
23.1	68.6	0.686

A silver bicycle leaning up against a kitchen table and chairs.

METEOR	CIDEr	CLIP-S
32.4	63.7	0.637

PAC-S a NEW MEASURE

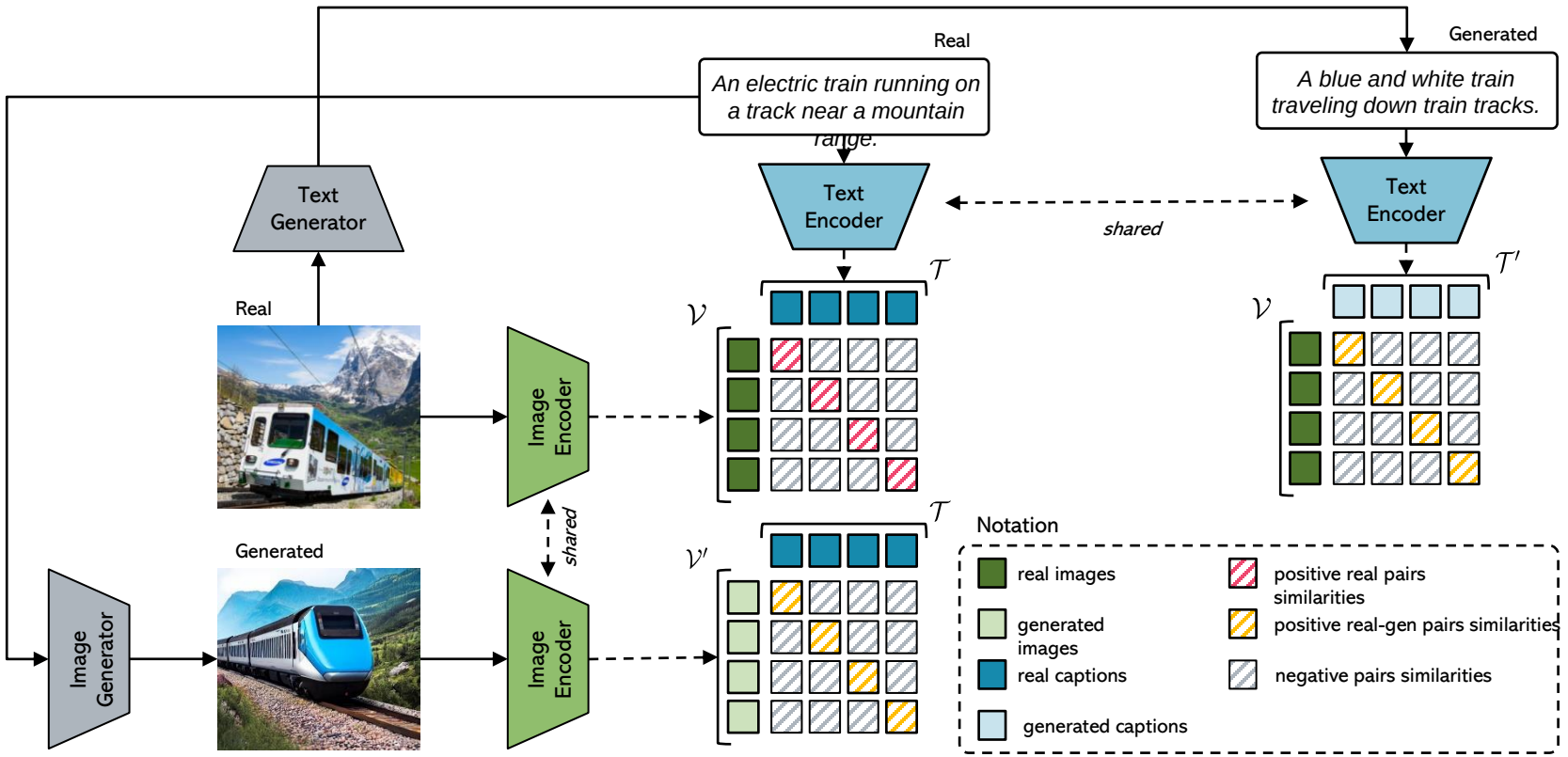
PAC-S: A new metric for evaluating Image-text correspondence

The metric outperforms previous reference-free and reference-based metrics in terms of ***correlation with human judgment***.

	METEOR	CIDEr	CLIP-S	PAC-S
A yellow bus passes through an intersection.	42.7	167.0	0.816	0.836
A yellow bus is traveling down a city street just past an intersection.	33.9	94.5	0.813	0.844

S. Sarto, M. Barraco, M. Cornia, L. Baraldi, R. Cucchiara "Positive-Augmented Constrastive Learning for Image and Video Captioning Evaluation«, CVPR 2023

Positive-Augmented Contrastive Learning



- ***Dual-encoder architecture*** comparing the visual and textual inputs via cosine similarity
- Usage of ***synthetic generators*** of both visual and textual data (Stable Diffusion[1] and BLIP[2], respectively)

Fine-tuning on human annotated data by taking into account ***contrastive relationship*** between real and generated matching image-caption pairs.

1. *Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.*
2. *Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In ICML, 2022.*

Image Captioning Correlation with Human Judgment

PAC score achieves the **best correlation with human judgment** and accuracy on all the considered image datasets, demonstrating its ***effectiveness*** compared to previously proposed metrics.

	Flickr8k-Expert		Flickr8k-CF	
	Kendall τ_b	Kendall τ_c	Kendall τ_b	Kendall τ_c
BLEU-1	32.2	32.3	17.9	9.3
BLEU-4	30.6	30.8	16.9	8.7
ROUGE	31.1	32.3	19.9	10.3
METEOR	41.5	41.8	22.2	11.5
CIDEr	43.6	43.9	24.6	12.7
SPICE	51.7	44.9	24.4	12.0
BERT-S	-	39.2	22.8	-
LEIC	46.6	-	29.5	-
BERT-S++	-	46.7	-	-
UMIC	-	46.8	-	-
TIGEr	-	49.3	-	-
ViLBERTScore	-	50.1	-	-
MID	-	54.9	37.3	-
CLIP-S	51.1	51.2	34.4	17.7
PAC-S	**53.9** (+2.8)	**54.3** (+3.1)	**36.0** (+1.6)	**18.6** (+0.9)
RefCLIP-S	52.6	53.0	36.4	18.8
RefPAC-S	**55.4** (+2.8)	**55.8** (+2.8)	**37.6** (+1.2)	**19.5** (+0.7)

	Composite	
	Kendall τ_b	Kendall τ_c
BLEU-1	29.0	31.3
BLEU-4	28.3	30.6
ROUGE	30.0	32.4
METEOR	36.0	38.9
CIDEr	34.9	37.7
SPICE	38.8	40.3
BERT-S	-	30.1
BERT-S++	-	44.9
TIGEr	-	45.4
ViLBERTScore	-	52.4
FAIEr	-	51.4
CLIP-S	49.8	53.8
PAC-S	**51.5** (+1.7)	**55.7** (+1.9)
RefCLIP-S	51.2	55.4
RefPAC-S	**52.8** (+1.6)	**57.1** (+1.7)

	Pascal-50S				
	HC	HI	HM	MM	Mean
length	51.7	52.3	63.6	49.6	54.3
BLEU-1	64.6	95.2	91.2	60.7	77.9
BLEU-4	60.3	93.1	85.7	57.0	74.0
ROUGE	63.9	95.0	92.3	60.9	78.0
METEOR	66.0	97.7	94.0	66.6	81.1
CIDEr	66.5	97.9	90.7	65.2	80.1
BERT-S	65.4	96.2	93.3	61.4	79.1
BERT-S++	65.4	98.1	96.4	60.3	80.1
TIGEr	56.0	99.8	92.8	74.2	80.7
ViLBERTScore	49.9	99.6	93.1	75.8	79.6
FAIEr	59.7	99.9	92.7	73.4	81.4
MID	67.0	99.7	97.4	76.8	85.2
CLIP-S	55.9	**99.3**	96.5	72.0	80.9
PAC-S	**60.6** (+4.7)	**99.3** (+0.0)	**96.9** (+0.4)	**72.9** (+0.9)	**82.4** (+1.5)
RefCLIP-S	64.9	**99.5**	95.5	73.3	83.3
RefPAC-S	**68.2** (+3.3)	**99.5** (+0.0)	**95.6** (+0.1)	**75.9** (+2.6)	**84.8** (+1.5)

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. JAIR, 47:853–899, 2013

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In CVPR, 2015

examples

	CLIP-S	PAC-S
A person tries to catch a **ball** on a beach.	**0.781**	0.798
A person tries to catch a **frisbee** on a beach.	0.759	**0.828**

	CLIP-S	PAC-S
A **baby horse** is seen standing in between another elephant's legs.	**0.782**	0.793
A **baby elephant** is seen standing in between another elephant's legs.	0.769	**0.820**

Visual Concept evaluation by querying

Idea: A very brave challenge: emulating humans in making questions,,, and finding answers..

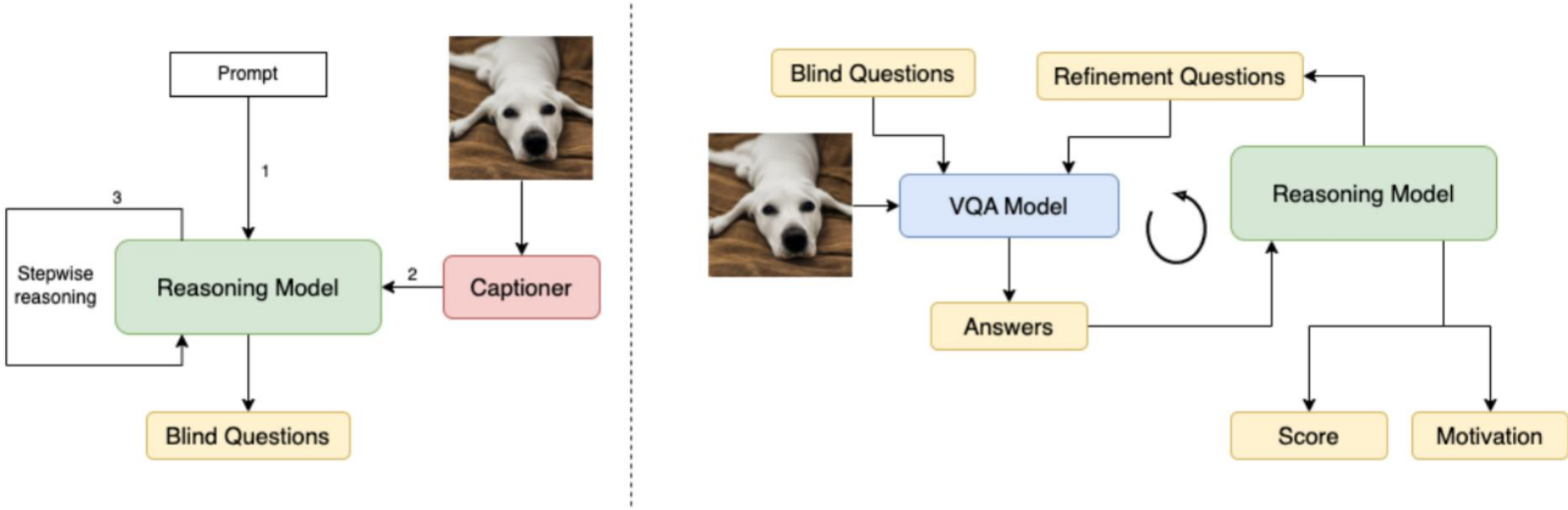


Figure 2. Visual Concept Evaluation Pipeline

Ask Lorenzo Baraldi : UNIMORE & UNITN

A future challenge : unlearning pre-trained models

Unlearning: away to catastrophically forget something the user ask to forget.

Something not possible for humans

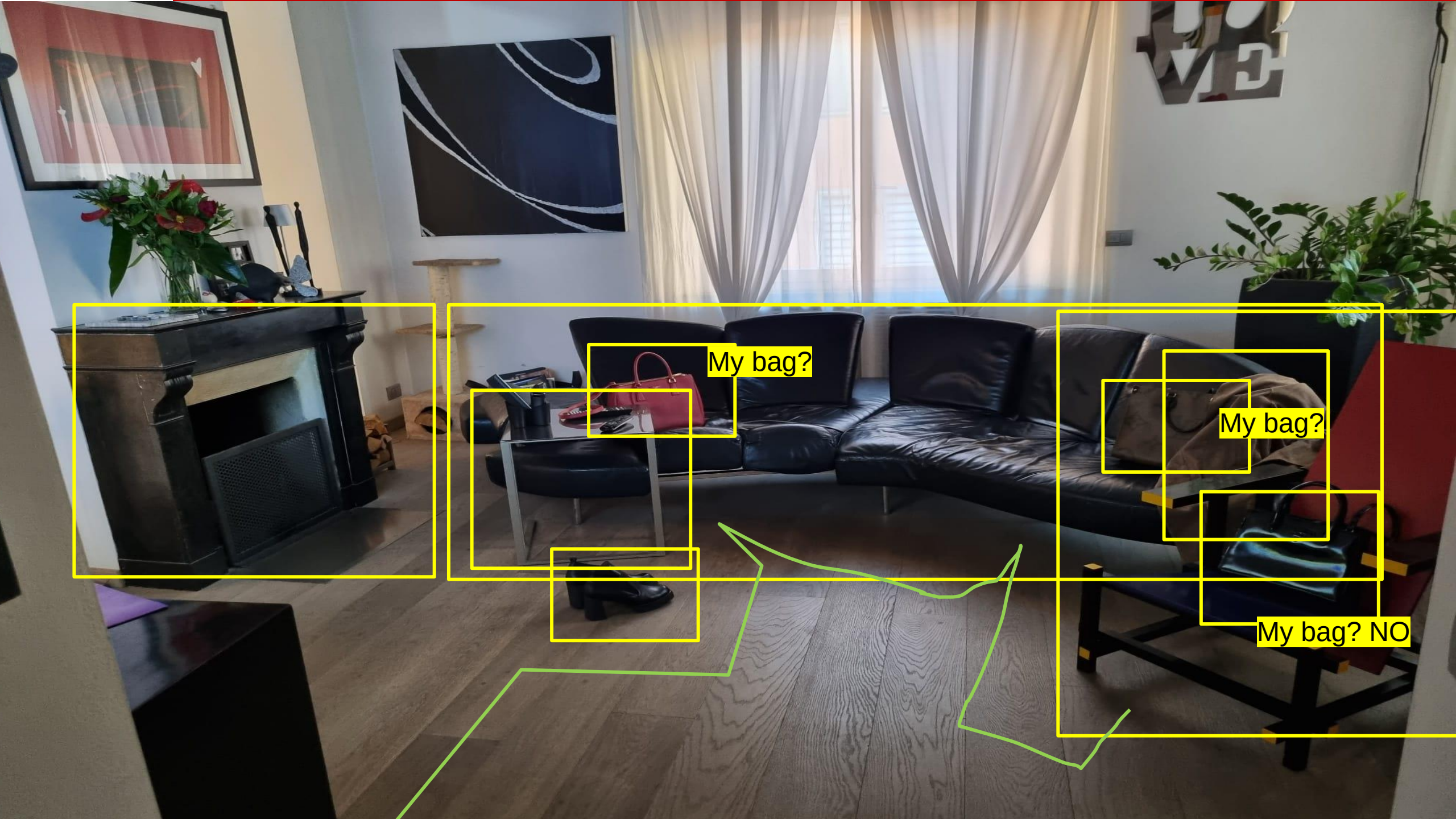
Useful for fairness, privacy, trustworthy and for PERSONALIZATION

We could need PLASTICITY

We could need UNLEARNING

Similar to extinction learning or inhibitory learning in humans (Kia Nobre)

Please Robot give me my BAG!


My bag?
My bag?
My bag? NO

Unlearning as hyper personalization

Let's suppose the robot can recognize different classes of objects (classification, detection, search and retrieval)

classes

Hand-bag

Gym-bag

School-bag

Shop-bag

Unlearning as hyper personalization

“My bag”

classes

Hand-bag

Gym-bag

~~School-bag~~

~~Shop-bag~~

AImage Lab

After Unlearning

This is your hand-bag

This is your gym-bag

How can a system do it?

This is **NOT** yours

This is **NOT** yours

Filtering

This is your hand-bag

This is your gym-bag

This is **NOT** yours

This is **NOT** yours

1) Filter the output

I recognize that this is a work-bag but I do not tell you….

Unlearning

This is your hand-bag

This is **NOT** yours

This is **NOT** yours

This is your work-bag

2) Unlearn a single class

? I do not recognize it as a known object
(e.g. I confuse it with another class,
possibly with a low confidence)

Multiple Class Unlearning

This is your hand-bag

This is **NOT** yours

This is **NOT** yours

This is your work-bag

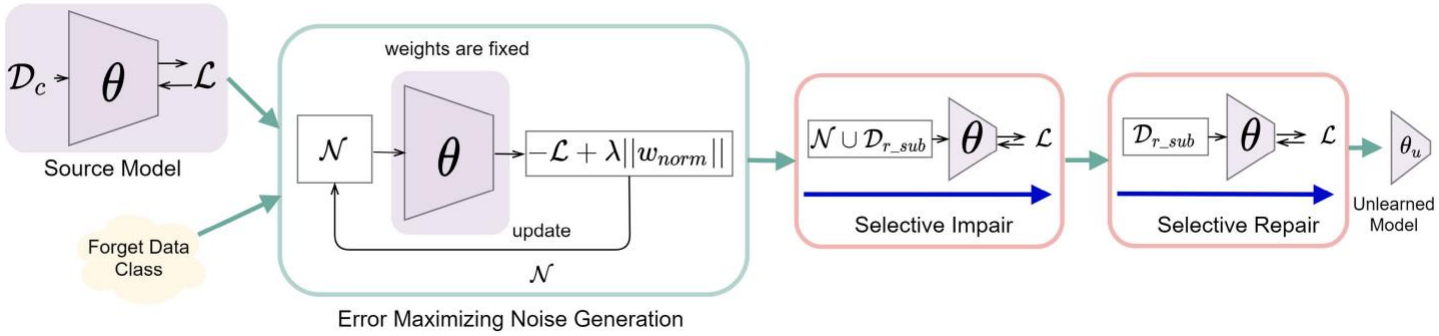
2) Unlearn more classes

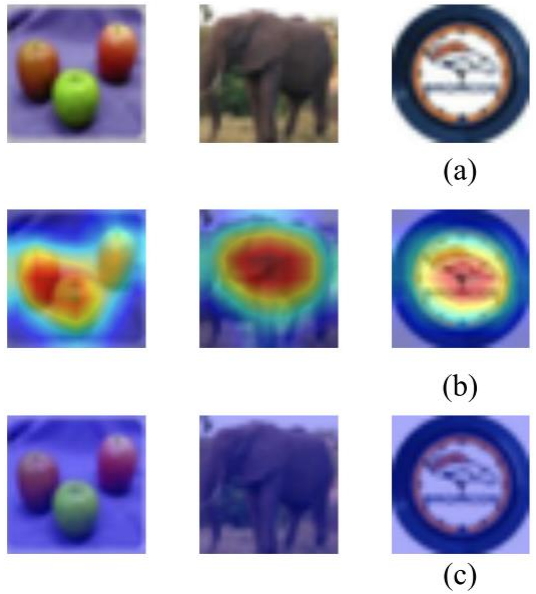
? *I do not recognize them as a known object*
(e.g. I confuse the with another class, possibly with a low confidence)

Unlearning in vision: two main families

1. One aims at making **the model unlearn by destroying its performance on the subject of the unlearning, and splitting its probability among all the other classes**[1]

- e.g. learning a noise matrix to deteriorate the model's performances)





2. Another realizes unlearning **by removing some classes and shifting their probabilities to the second most likely**[2]

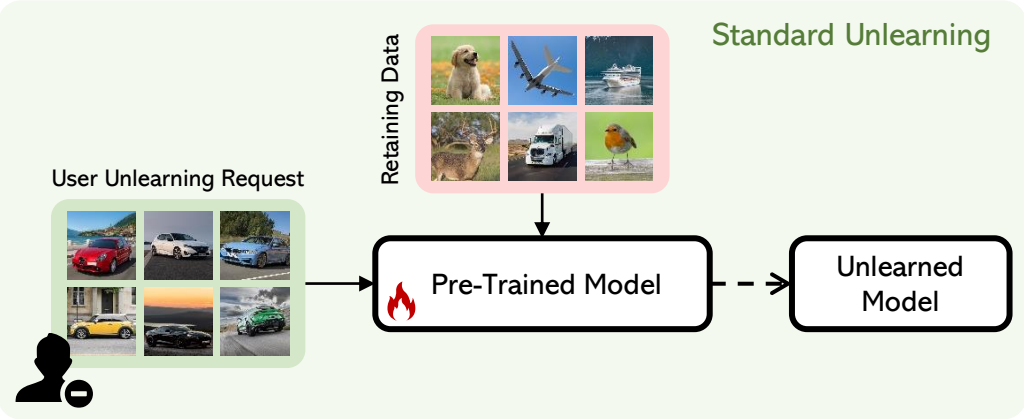
[1] A. K. Tarun, et al. «Fast Yet Effective Machine Unlearning». arXiv preprint arXiv:2111.08947 (2021).

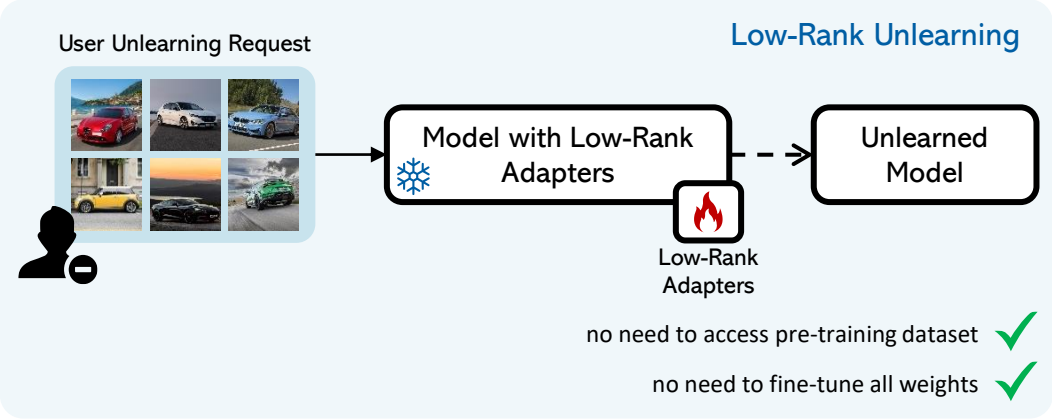
[2] S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Multi-Class Explainable Unlearning for Image Classification via Weight Filtering», **under review**

Unlearning Pre-trained models

No retaining data available?

- a strategy to realize unlearning **without** either accessing the **retaining data** or creating **hand-crafted proxies**
- It only requires access to **some images** of the classes that are to be unlearned
- It does not even require that those images come from the original dataset: we provide experiments with **random images**, downloaded from the **web**

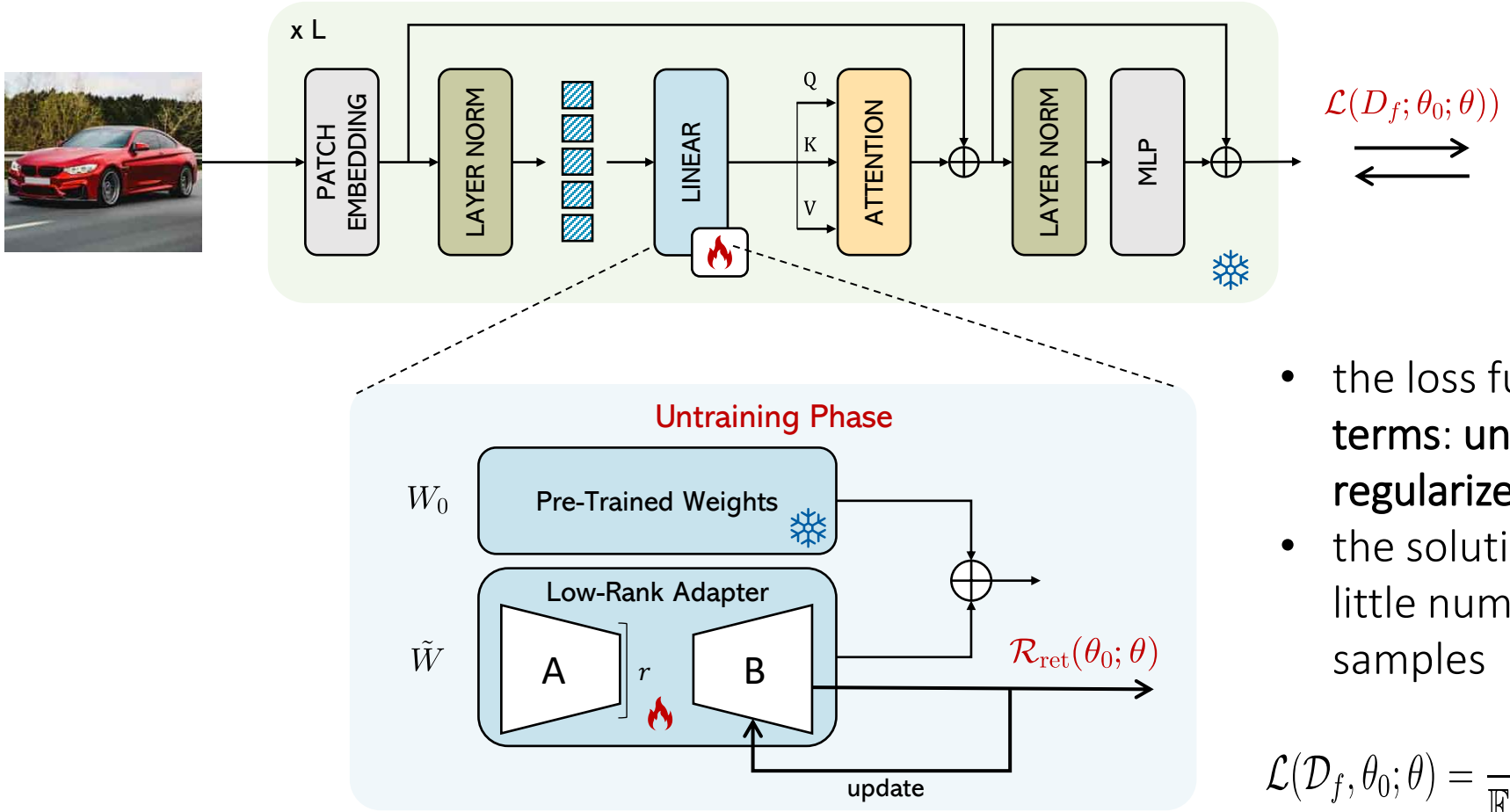




S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **under review**

The Model

unlearning architecture We inject a **trainable low-rank** decomposition into the linear layer producing the query, key and value vectors



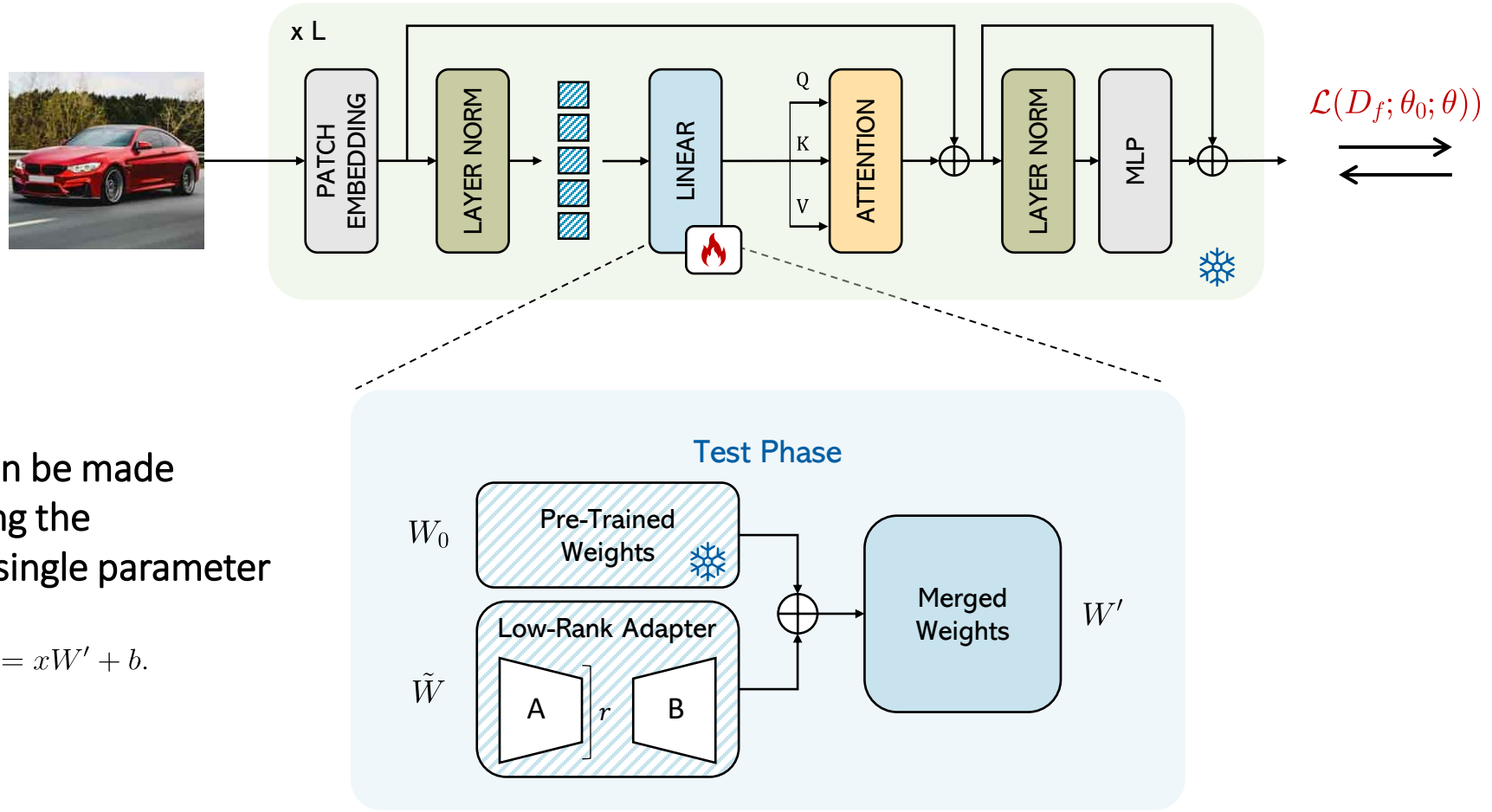
- the loss function is composed of **two terms**: **unlearning factor** and a **regularizer**
- the solution is **extremely fast**, given the little number of required untraining samples

$$\mathcal{L}(\mathcal{D}_f,\theta_0;\theta)=\frac{1}{\mathbb{E}_{\mathbf{x},\mathbf{y}\in\mathcal{D}_f}\mathcal{L}_{\text{CE}}(g_{\theta'}(\mathbf{x}),\mathbf{y};\theta)}+\lambda\|\text{vec}(B)\|_1.$$

S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **under review**

The Model

Our unlearning architecture in Test Phase



During the evaluation, ***W*** can be made inaccessible by just collapsing the decomposition, back into a single parameter matrix.

$$W' \leftarrow W_0 + BA, \;\; f(x) = xW' + b.$$

S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **under review**

Experimental results on CIFAR-10 and CIFAR-20

- We achieve **comparable results** to approaches using the **retaining data**
- The use of **LoRa** layer, combined with our loss, **performs better** than other baselines, in the **same setting**

			ViT-T		ViT-S		Swin-S	
		$\mathcal{D}_r$	Acc_r [%] ↑	Acc_f [%] ↓	Acc_r [%] ↑	Acc_f [%] ↓	Acc_r [%] ↑	Acc_f [%] ↓
CIFAR-10	Original model	-	82.0	82.0	84.0	84.0	89.8	89.8
	Retrained model	✓	80.9	0.0	85.4	0.0	88.8	0.0
	Fine-tuned model	✓	80.2	7.9	81.3	3.0	85.0	2.3
	Random labels [17]	✓	83.0	0.0	85.1	0.0	88.9	0.0
	Negative gradient [14]	✓	84.4	0.0	85.8	0.0	88.9	0.0
	Negative gradient w/ L_1 regularization	✗	80.8	0.3	82.2	1.0	85.4	2.1
	Negative gradient w/ low-rank	✗	80.9	**0.1**	82.5	0.9	85.4	1.8
	Bounded loss w/ L_1 regularization	✗	81.2	**0.1**	82.3	**0.8**	85.5	1.4
	Bounded loss w/ low-rank (Ours)	✗	**81.9**	0.1	**83.5**	**0.8**	**86.0**	**0.8**
CIFAR-20	Original model	-	67.0	67.0	71.9	71.9	74.4	74.4
	Retrained model	✓	64.2	0.0	69.7	0.0	72.7	0.0
	Fine-tuned model	✓	64.5	8.2	67.2	8.6	68.3	4.6
	Random labels [17]	✓	66.2	0.0	70.8	0.0	73.2	0.0
	Negative gradient [14]	✓	67.6	0.0	71.4	0.0	72.2	0.0
	Negative gradient w/ L_1 regularization	✗	62.9	1.1	68.0	1.2	67.9	3.8
	Negative gradient w/ low-rank	✗	63.0	1.0	67.8	1.0	67.9	3.8
	Bounded loss w/ L_1 regularization	✗	63.1	1.2	67.9	0.8	68.0	3.7
	Bounded loss w/ low-rank (Ours)	✗	**63.5**	**0.9**	**68.2**	**0.8**	**68.2**	**3.4**

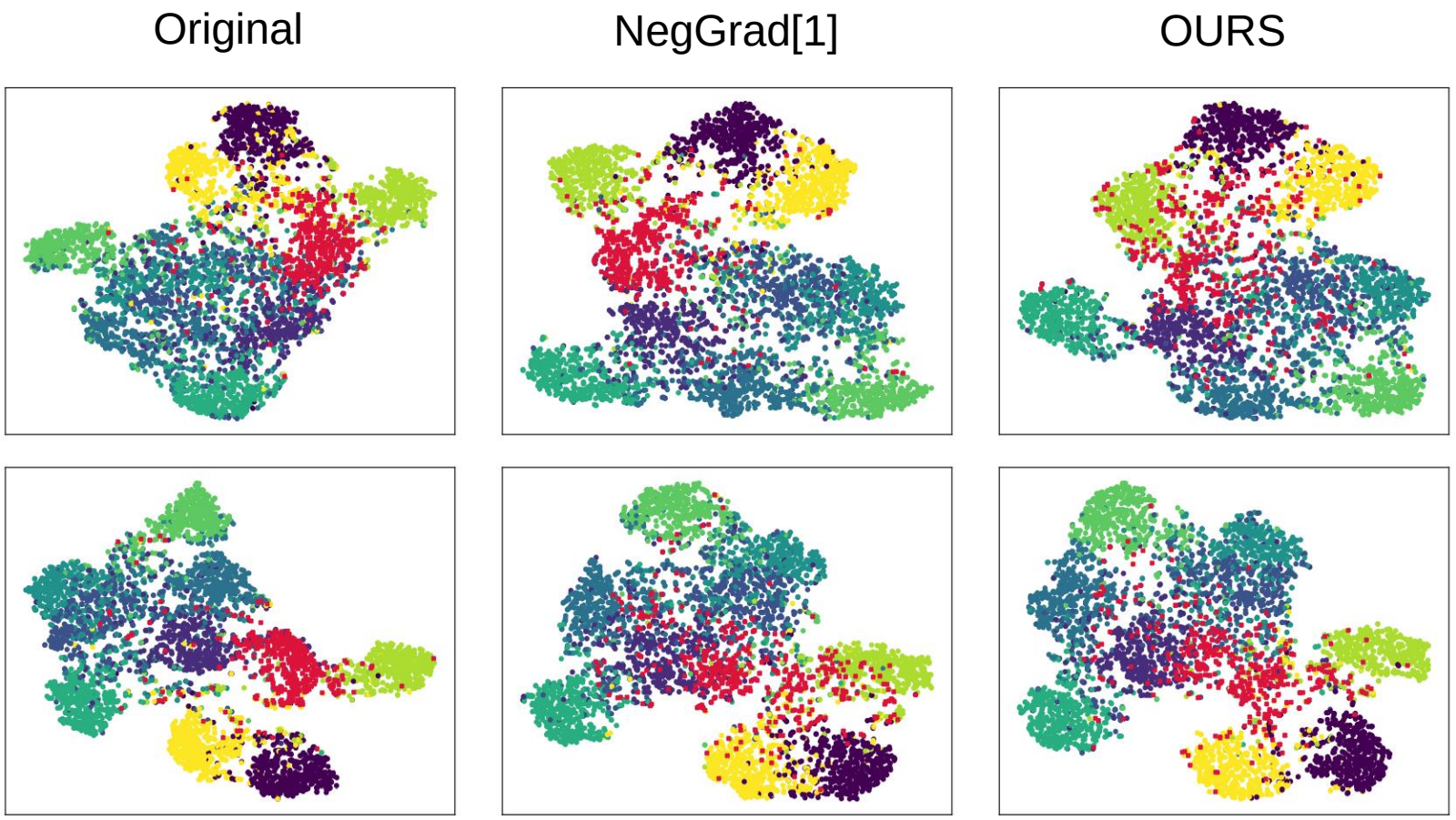
S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **under review**

Results

Experimental results on CIFAR-10 and CIFAR-20

Visualizing the **embedding space** with the **T-SNE** algorithm,

The **low-rank unlearning** brings the embedding of unlearned samples toward **other classes**



[1] A. Golatkar, A. Achille, and S. Soatto. 2020. «Eternal sunshine of the spotless net: Selective forgetting in deep networks». In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **under review**

6. CONCLUSIONS FOR DISCUSSIONS

AImage Lab

Conclusions

- Computer Vision and Pattern Recognition; Natural Language Processing and Computational Linguistics; Discriminative and Generative Machine Learning are converging in single multimodal models
- Multimodal models could be large scale foundation model (or not) or adapted, distilled and created with modular elements
- Now research in cross-modal and multimodal generative AI is absolutely the challenge of today: many applications (e g industry, fashion,media…)
- Many challenges in all sub-tasks of the unifying model
- Many challenges for the future (evaluating as human can do, changing the pre-trained hypothesis, hyper-personalization..just to same someone)

And it is not enough…robustness, accuracy, human oversight, are not enough to cope with trustworthiness.. we need an ethic-by-design-AI .

Conclusions

Get on the shoulders of giants…

and jump far!

Thanks

Thanks to Aimagelab researchers.

http://Aimagelab.unimore.it

Rita.cucchiara@unimore.it

https://www.ellis.unimore.it/

Thanks to the AIMAGELAB colleagues (Costantino Grana, Roberto Vezzani, Elisa Ficarra, Simone Calderara, Lorenzo Baraldi, Marcella Cornia, Enver Sangineto, Vittorio Cuculo, Silvia Cascianelli, Angelo Porello…)

Thanks to all Phd students!

Thanks to NVIDIA, and CINECA.

thanks Dino Pedreschi and all Staff for your effort

in the National Phd School in Ai for Society