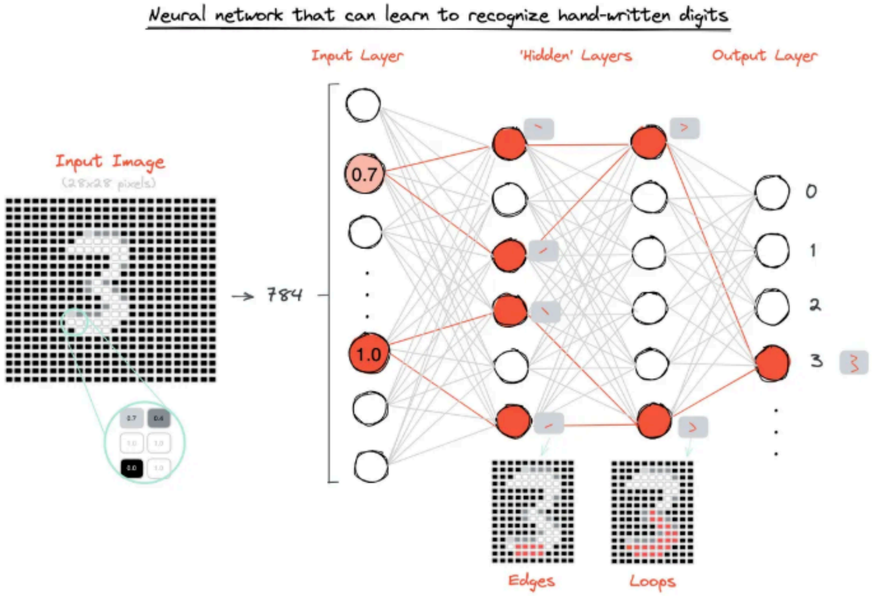
HELMHOLTZAI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

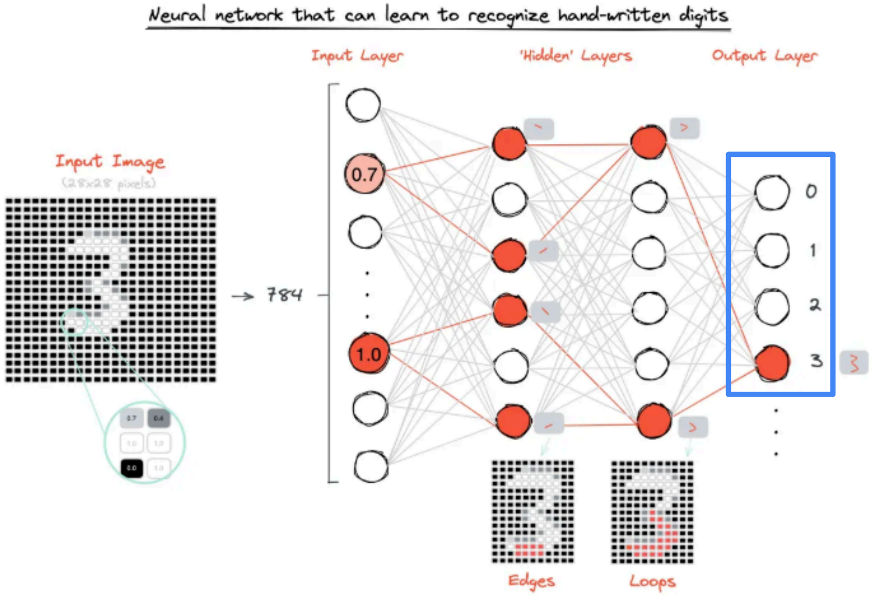
Introduction to Training Foundational Models with 4M

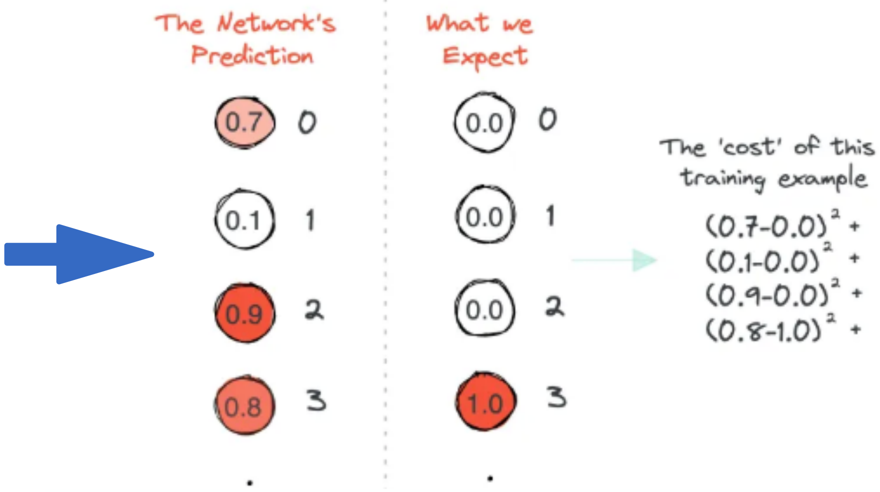
Gunjan Joshi
Helmholtz-Zentrum Dresden Rossendorf

“Plain Vanilla” Feed forward Neural Network

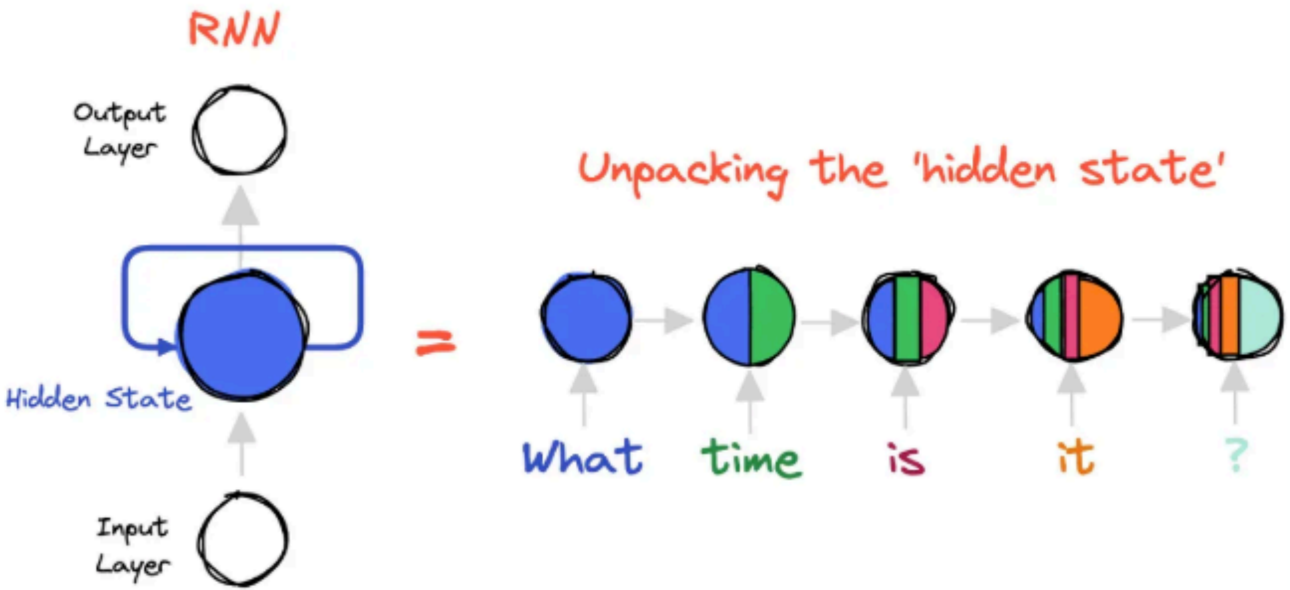


"Plain Vanilla" Feed forward Neural Network



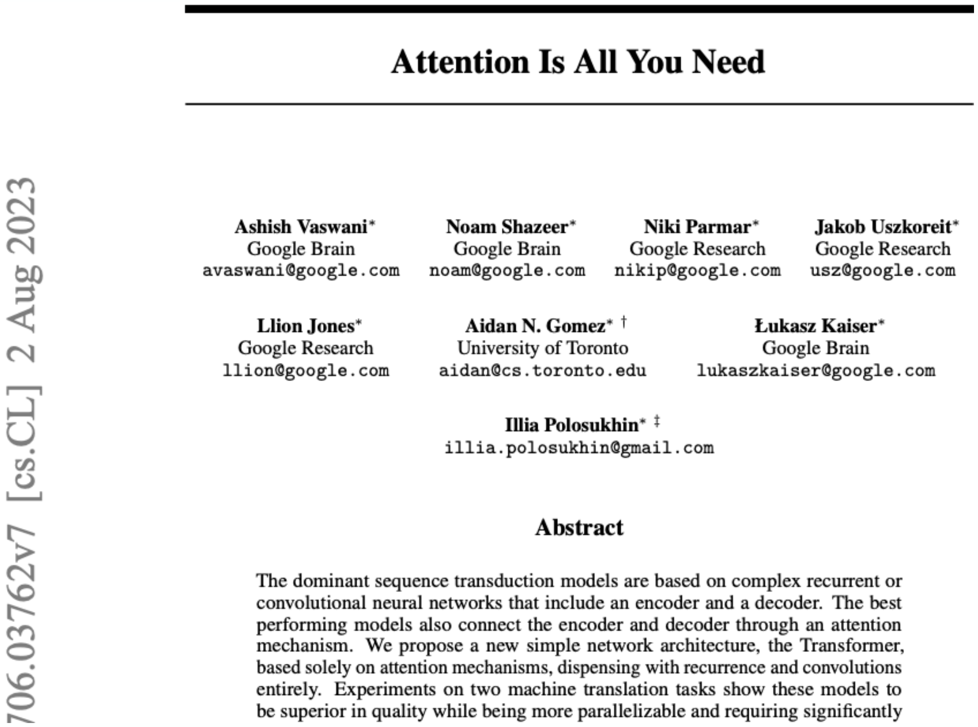


Recurrent Neural Networks



[The A.I Hacker - Illustrated Guide to Recurrent Neural Networks]

Transformers



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

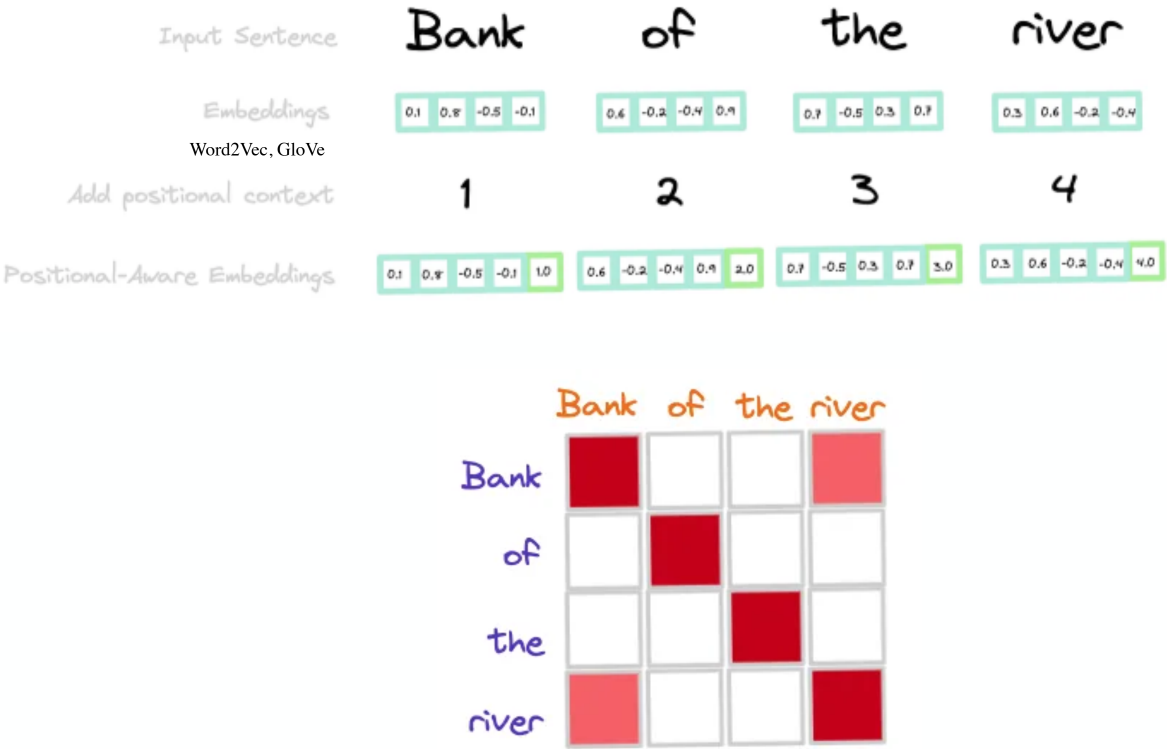
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly

706.03762v7 [cs.CL] 2 Aug 2023

Transformers brought two key innovations from its predecessor (RNNs)

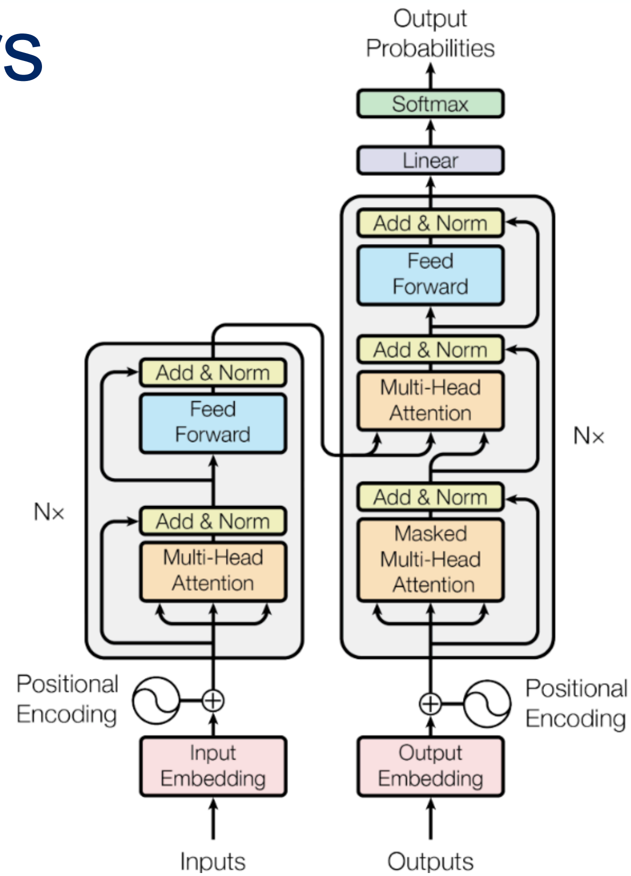
- **Positional Encodings**
- **Self-Attention**

Transformers



Transformers

Encoder



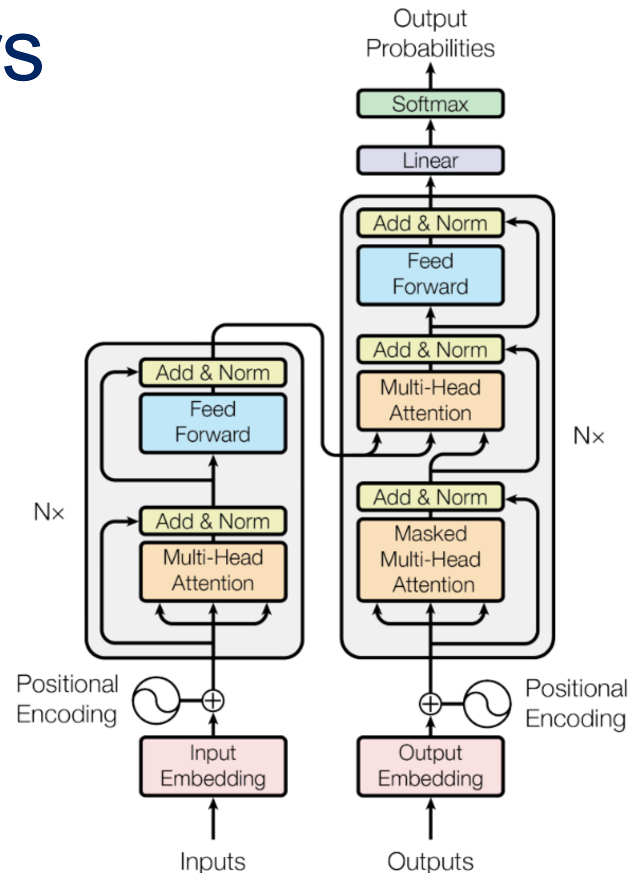
Decoder

https://arxiv.org/pdf/1706.03762

Transformers

BERT

Encoder



GPT

Decoder

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of

Transformers

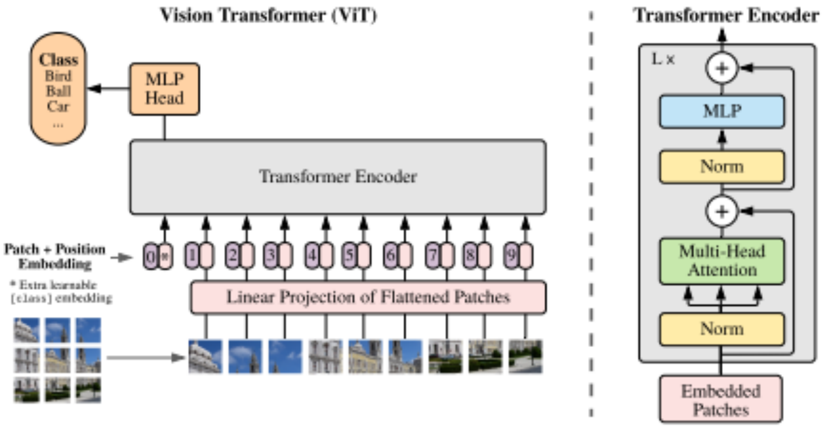
AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

Treat an image like a **sequence of patch tokens**, just like words in a sentence and use the same transformer architecture from NLP.



Transformers

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

Jiasen Lu[1], Dhruv Batra[1,2], Devi Parikh[1,2], Stefan Lee[1,3]
[1]Georgia Institute of Technology, [2]Facebook AI Research, [3]Oregon State University

Abstract

We present ViLBERT (short for Vision-and-Language BERT), a model for learning task-agnostic joint representations of image content and natural language. We extend the popular BERT architecture to a multi-modal two-stream model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers. We pretrain our model through two proxy tasks on the large, automatically collected Conceptual Captions dataset and then transfer it to multiple established vision-and-language tasks – visual question answering, visual commonsense reasoning, referring expressions, and caption-based image retrieval – by making only minor additions to the base architecture. We observe significant improvements across tasks compared to existing task-specific models – achieving state-of-the-art on all four tasks. Our work represents a shift away from learning groundings between vision and language only as part of task training and towards treating visual grounding as a pretrainable and transferable capability.

1 Introduction

"... spend the summer linking a camera to a computer and getting the computer to describe what it saw."

Marvin Minsky on the goal of a 1966 undergraduate summer research project [1]

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1] Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-

Flamingo: a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac[*,‡] Jeff Donahue[*] Pauline Luc[*] Antoine Miech[*]
Iain Barr[†] Yana Hasson[†] Karel Lenc[†] Arthur Mensch[†] Katie Millican[†]
Malcolm Reynolds[†] Roman Ring[†] Eliza Rutherford[†] Serkan Cabi Tengda Han
Zhitao Gong Sina Samangooei Marianne Monteiro Jacob Menick
Sebastian Borgeaud Andrew Brock Aida Nematzadeh Sahand Sharifzadeh
Mikolaj Binkowski Ricardo Barreira Oriol Vinyals Andrew Zisserman
Karen Simonyan[*,‡]

[*] Equal contributions, ordered alphabetically, [†] Equal contributions, ordered alphabetically, [‡] Equal senior contributions

DeepMind

HELMHOLTZ AI Artificial Intelligence Cooperation Unit

Foundation Model

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

1.1.1 Naming.

We introduce the term *foundation models* to fill a void in describing the paradigm shift we are witnessing; we briefly recount some of our reasoning for this decision. Existing terms (e.g., *pretrained model, self-supervised model*) partially capture the technical dimension of these models, but fail to capture the significance of the paradigm shift in an accessible manner for those beyond machine learning. In particular, foundation model designates a model class that are distinctive in their sociological impact and how they have conferred a broad shift in AI research and deployment. In contrast, forms of pretraining and self-supervision that technically foreshadowed foundation models fail to clarify the shift in practices we hope to highlight.

https://arxiv.org/abs/2108.07258

Foundation Model

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

2108.07258v3 [cs.LG] 12 Jul 2022

1.1.1 Naming.

We introduce the term *foundation models* to fill a void in describing the paradigm shift we are witnessing; we briefly recount some of our reasoning for this decision. Existing terms (e.g., *pretrained model, self-supervised model*) partially capture the technical dimension of these models, but fail to capture the significance of the paradigm shift in an accessible manner for those beyond machine learning. In particular, foundation model designates a model class that are distinctive in their sociological impact and how they have conferred a broad shift in AI research and deployment. In contrast, forms of pretraining and self-supervision that technically foreshadowed foundation models fail to clarify the shift in practices we hope to highlight.

Stanford defines foundation models as:

*“**Models** trained on **broad data** (generally using **self supervision** at scale) that can be **adapted** (fine-tuned) to a wide range of downstream tasks”*

Foundation Model

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

2108.07258v3 [cs.LG] 12 Jul 2022

1.1.1 Naming.

We introduce the term *foundation models* to fill a void in describing the paradigm shift we are witnessing; we briefly recount some of our reasoning for this decision. Existing terms (e.g., *pretrained model, self-supervised model*) partially capture the technical dimension of these models, but fail to capture the significance of the paradigm shift in an accessible manner for those beyond machine learning. In particular, foundation model designates a model class that are distinctive in their sociological impact and how they have conferred a broad shift in AI research and deployment. In contrast, forms of pretraining and self-supervision that technically foreshadowed foundation models fail to clarify the shift in practices we hope to highlight.

Stanford defines foundation models as:

*"**Models** trained on **broad data** (generally using **self supervision** at scale) that can be **adapted** (fine-tuned) to a wide range of downstream tasks"*

Foundation Model = (Large corpus of (unlabeled) data + Scale + SSL) → Transfer learning capability

4M: Massively Multimodal Masked Modeling

4M: Massively Multimodal Masked Modeling

David Mizrahi[1,2*] Roman Bachmann[1*] Oğuzhan Fatih Kar[1]
Teresa Yeo[1] Mingfei Gao[2] Afshin Dehghan[2] Amir Zamir[1]
[1]Swiss Federal Institute of Technology Lausanne (EPFL) [2]Apple

https://4m.epfl.ch

Abstract

Current machine learning models for vision are often highly specialized and limited to a single modality and task. In contrast, recent large language models exhibit a wide range of capabilities, hinting at a possibility for similarly versatile models in computer vision. In this paper, we take a step in this direction and propose a multimodal training scheme called 4M. It consists of training a **single unified Transformer encoder-decoder** using a **masked modeling objective** across a **wide range of input/output modalities** – including text, images, geometric, and semantic modalities, as well as neural network feature maps. 4M achieves **scalability** by unifying the representation space of all modalities through mapping them into discrete tokens and performing multimodal masked modeling on a small randomized subset of tokens.

4M leads to models that exhibit several key capabilities: (1) they can perform a diverse set of vision tasks out of the box, (2) they excel when fine-tuned for unseen downstream tasks or new input modalities, and (3) they can function as a generative

4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities

Roman Bachmann[1†*] Oğuzhan Fatih Kar[1*] David Mizrahi[2†*] Ali Garjani[1]
Mingfei Gao[2] David Griffiths[2] Jiaming Hu[2] Afshin Dehghan[2] Amir Zamir[1]
[1]Swiss Federal Institute of Technology Lausanne (EPFL) [2]Apple

https://4m.epfl.ch

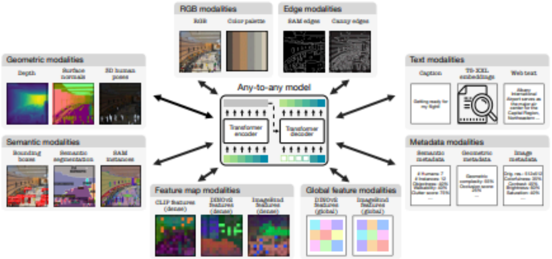
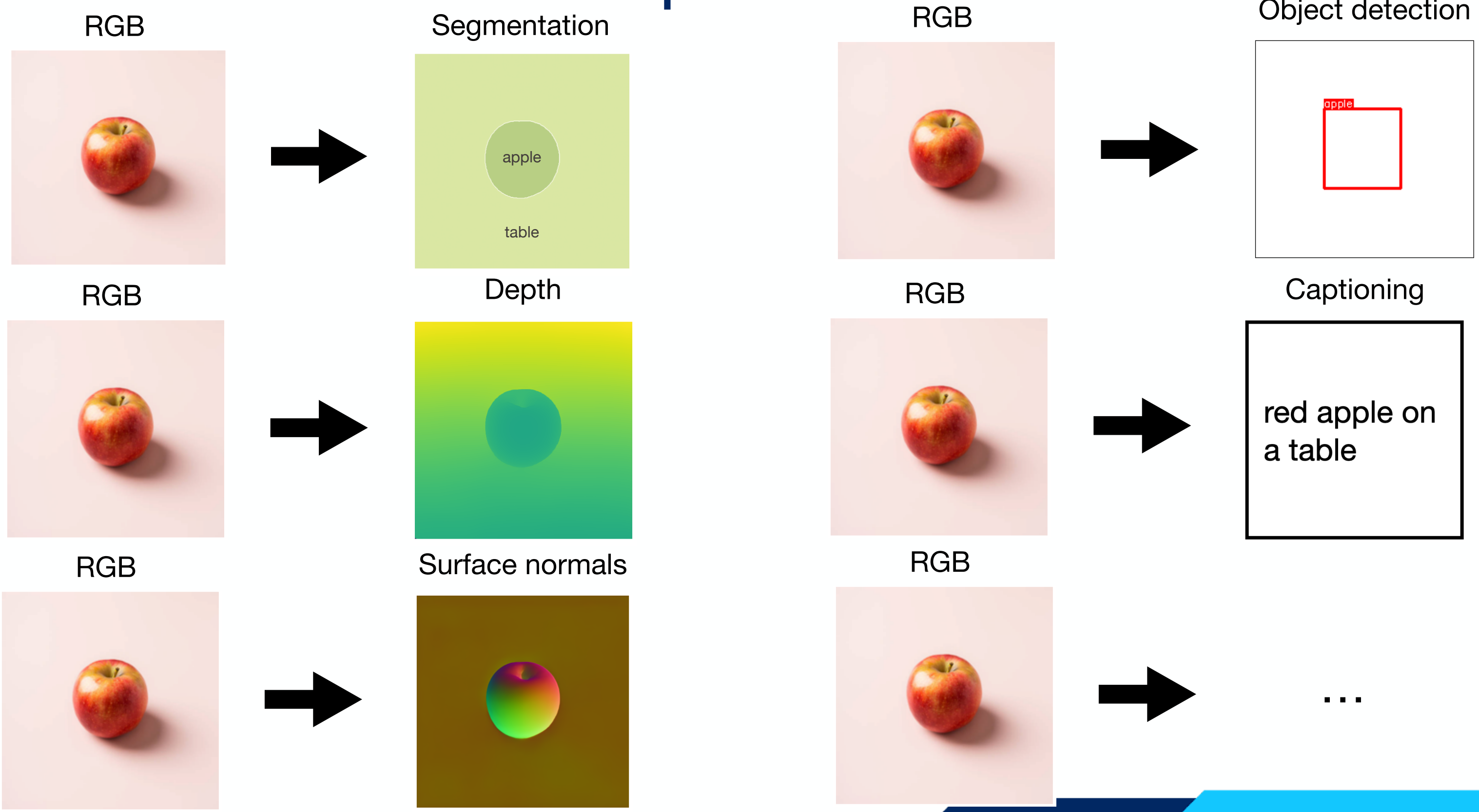


Figure 1: We demonstrate training a single model on tens of highly diverse modalities *without a loss in performance* compared to specialized single/few task models. The modalities are mapped to discrete tokens using modality-specific tokenizers. The model can generate *any* of the modalities from *any subset* of them.

Abstract

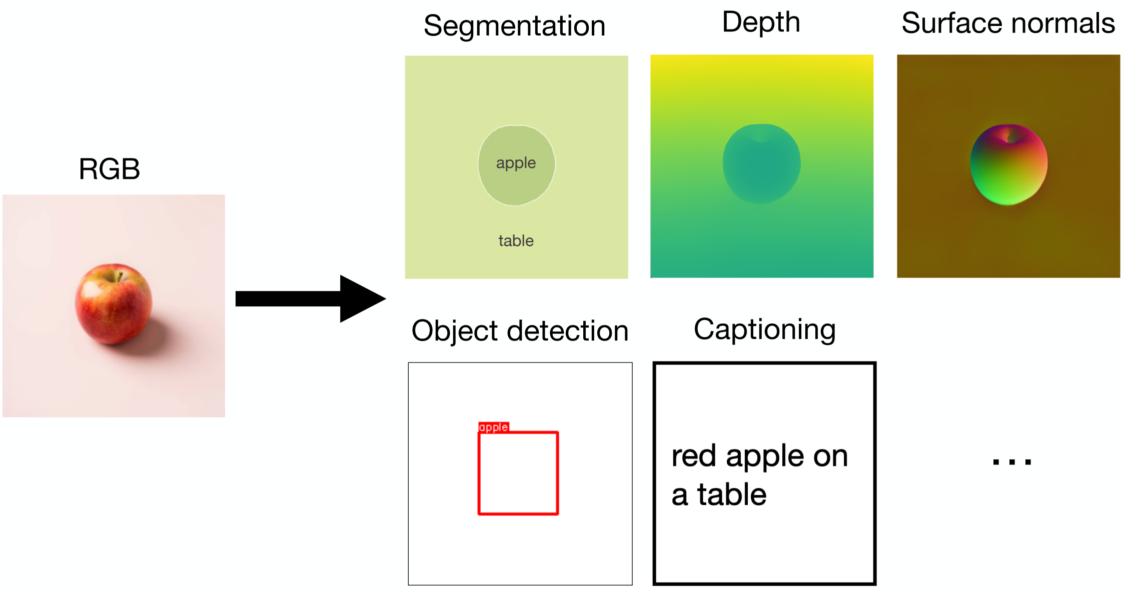
Current multimodal and multitask foundation models, like 4M [62] or UnifiedIO [59, 58], show promising results. However, their out-of-the-box abilities to accept diverse inputs and perform diverse tasks are limited by the (usually small) number of modalities and tasks they are trained on. In this paper, we develop a single any-to-any model trained on tens of highly diverse modalities and by

We want to solve multiple tasks



Slide credit: https://4m.epfl.ch/

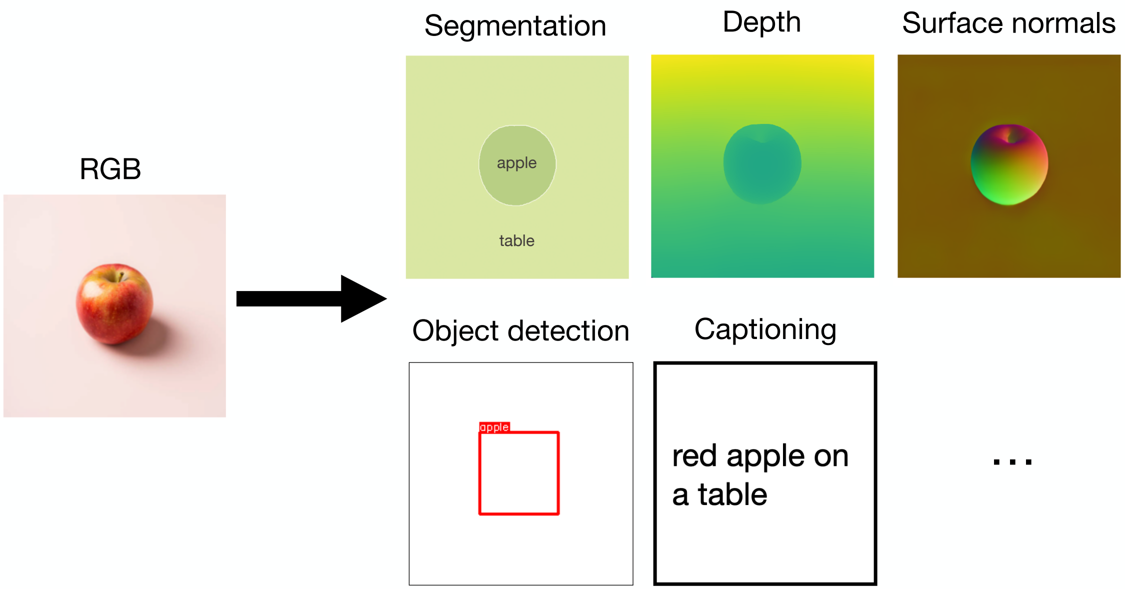
We want to solve multiple tasks



We want to solve multiple tasks

Efficiency

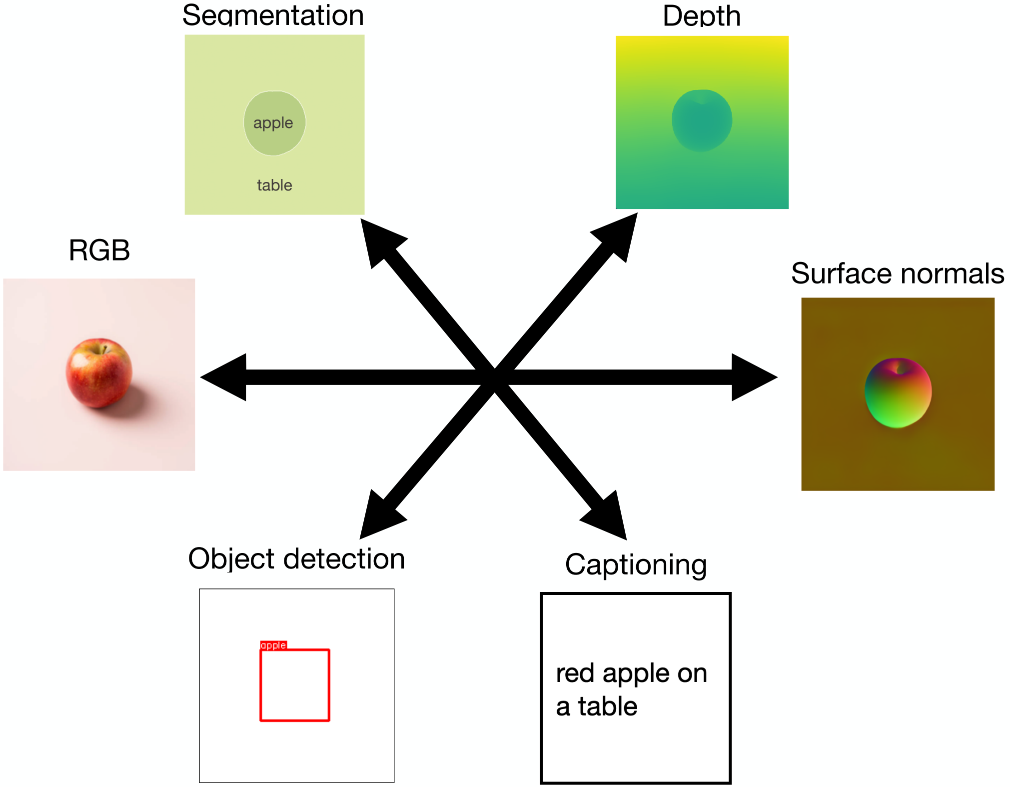
- Avoid training one model for each task



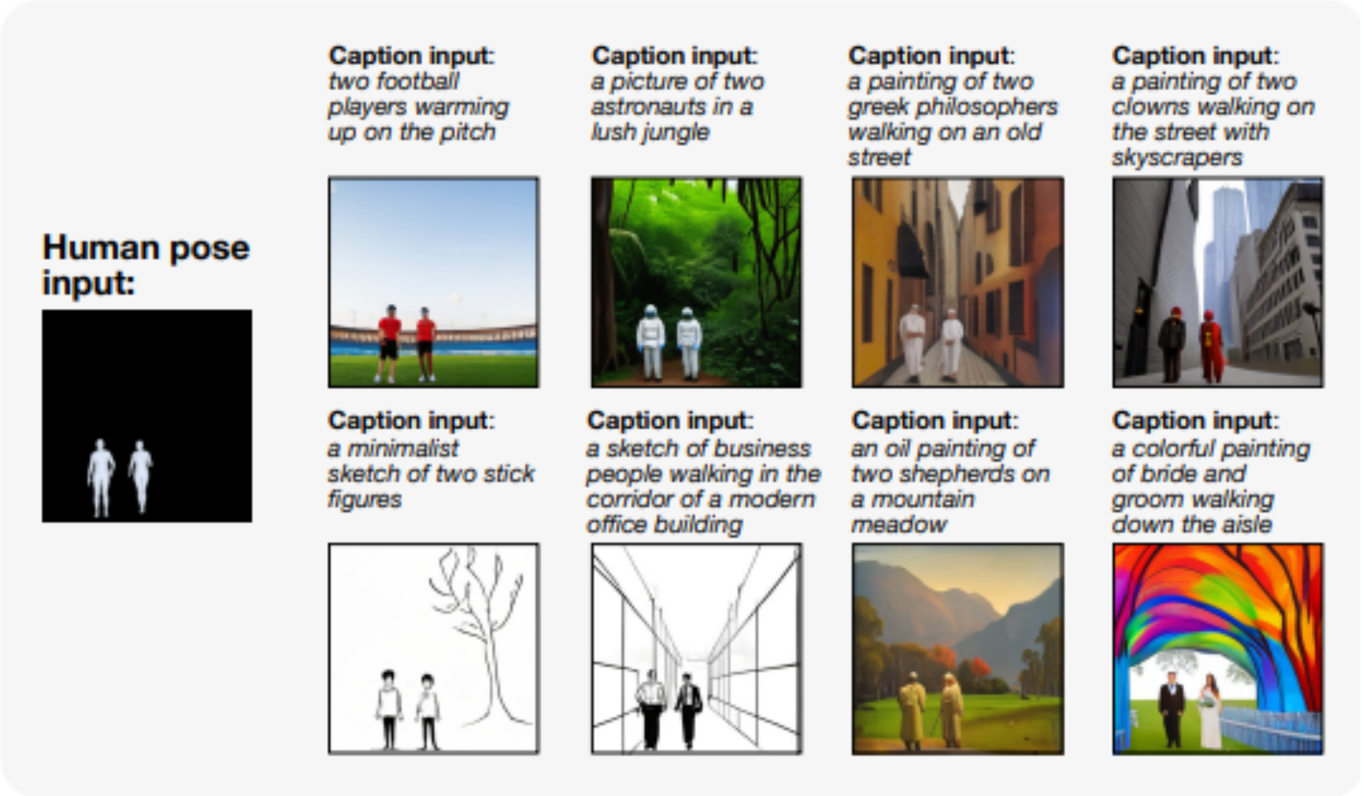
Solve multiple tasks & understand multiple modalities

Efficiency

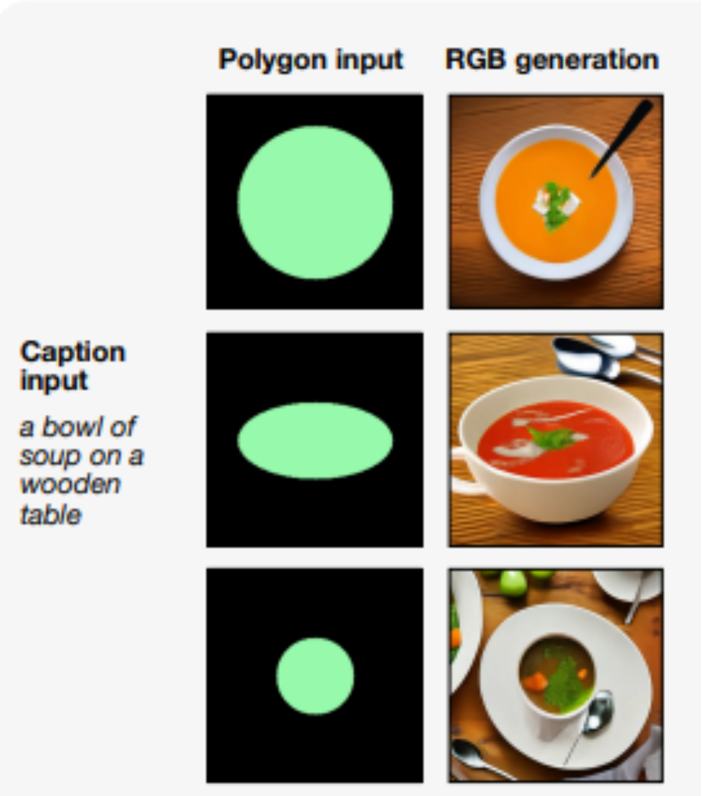
- Avoid training one model for each task
- Operate on a wide range of modalities and solves many tasks
- Anything in, anything out (any-to-any)
- Scale to large model sizes
- Benefit from large datasets



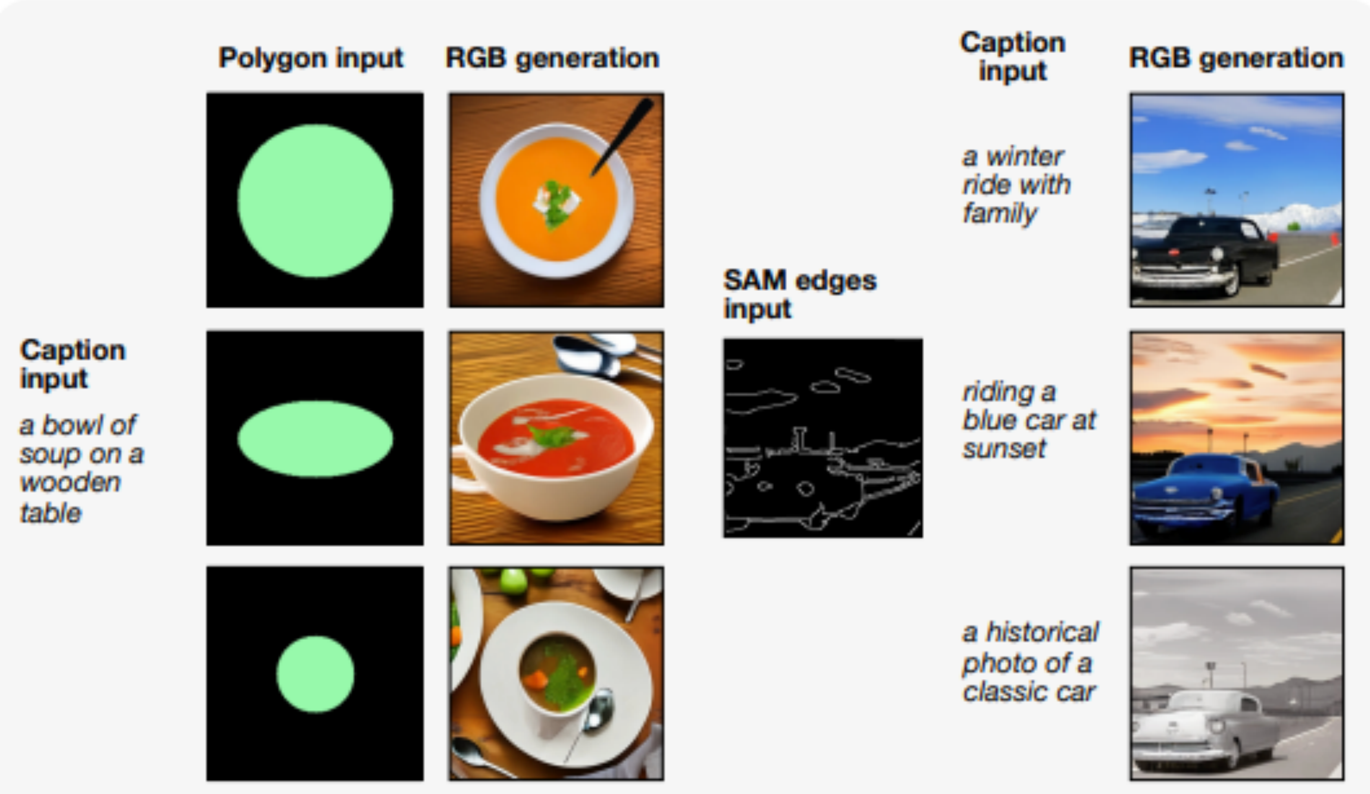
Fine-grained multimodal conditions control



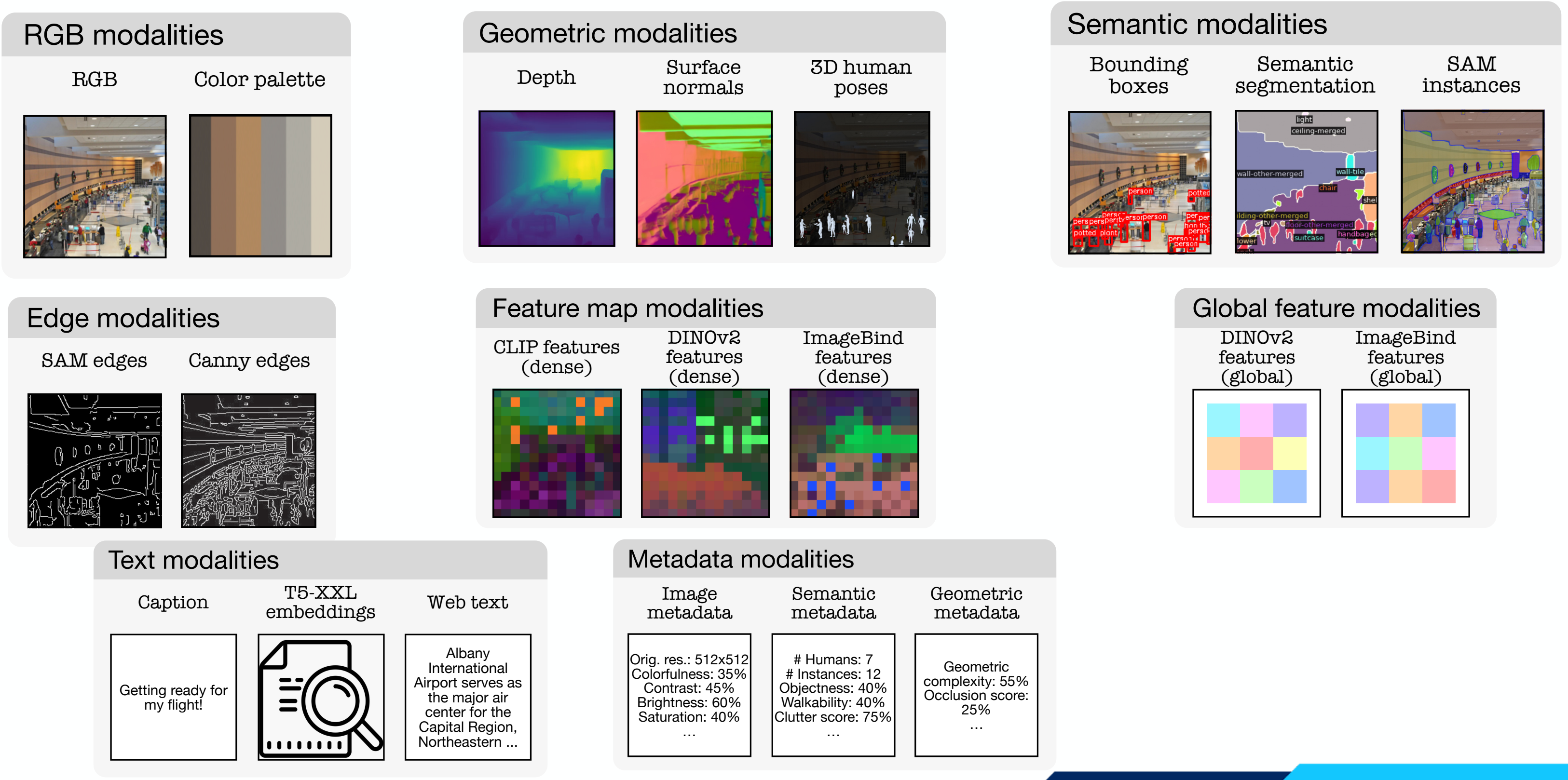
Probing with grounded generation



Probing with grounded generation



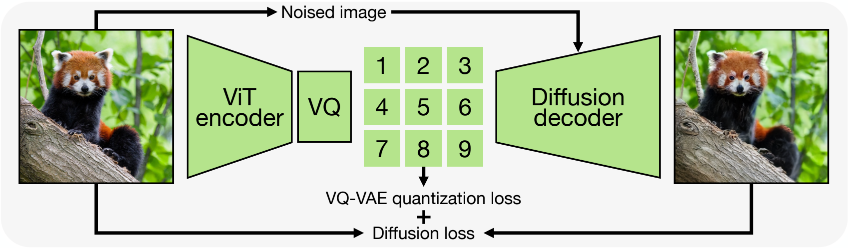
21 “types” of modalities



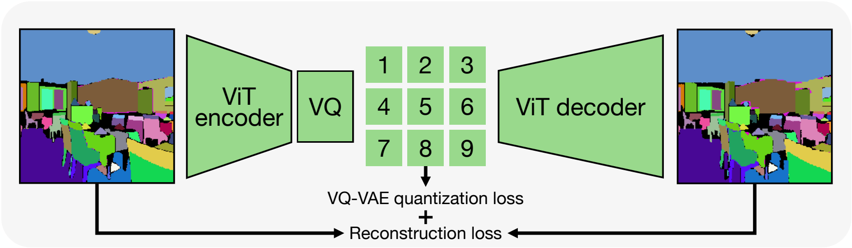
Modality specific tokenizer..... how to deal with them ?

Multimodal tokenization

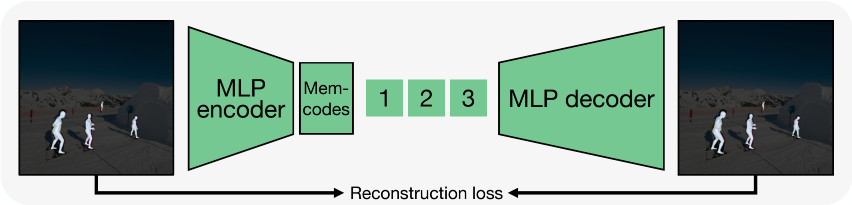
Spatial discrete VAE with diffusion decoder: RGB, normal, depth, edges



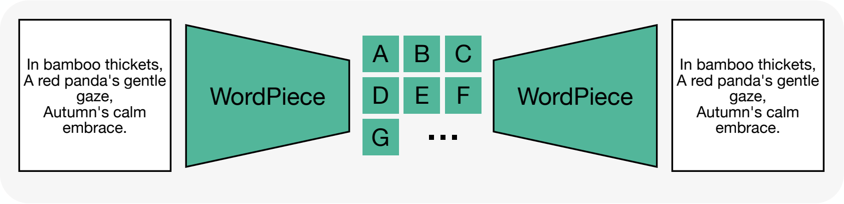
Spatial discrete VAE: Segmentation, CLIP, DINOv2, ImageBind, SAM inst.



MLP discrete VAE: Human poses, DINOv2 & ImageBind global tokens



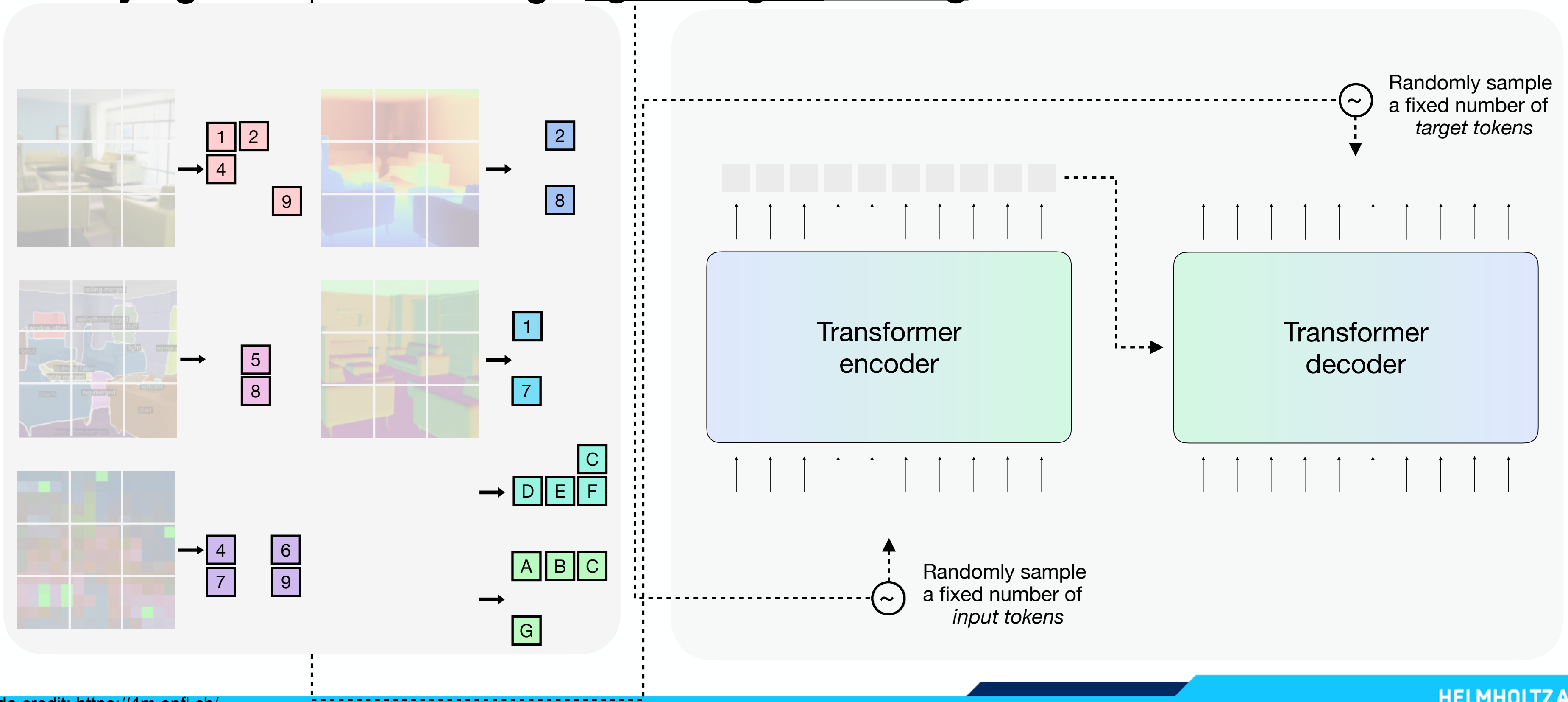
Sequence tokenizer: Text, bounding boxes, metadata, color palette



4M pre-training objective

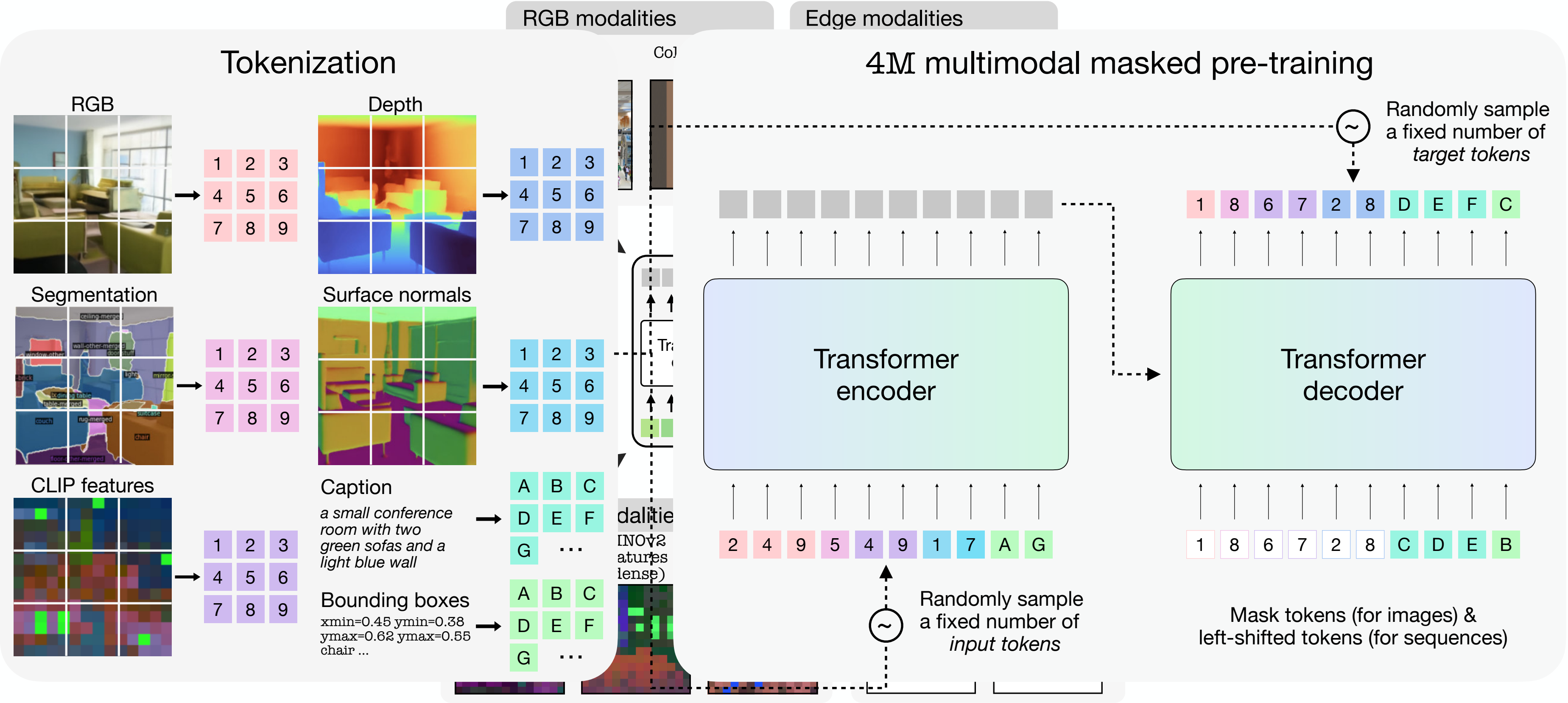
Unifying modalities through tokenization

Scalable cross-modal learning through masking



Slide credit: https://4m.epfl.ch/

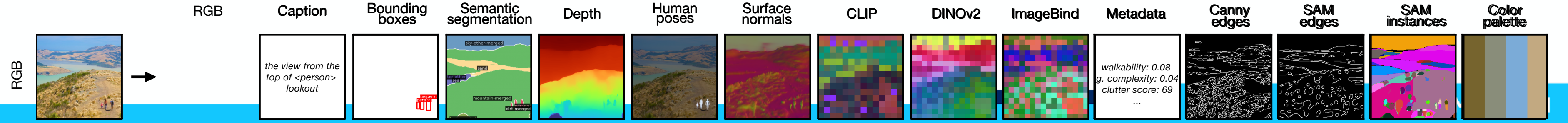
Multimodal masked modeling



Generate any modality conditioned on any other

RGB

Generate any modality conditioned on any other



Self-consistent prediction through chained multimodal generation

Tokenization

Bounding boxes

xmin=0.30 ymin=0.51
xmax=0.68 ymax=0.99
horse → A B C D E

4M chained multimodal generation

Iteration 1 2 3 4 5 6 7 8 9 10

Self-consistent prediction through chained multimodal generation

Tokenization

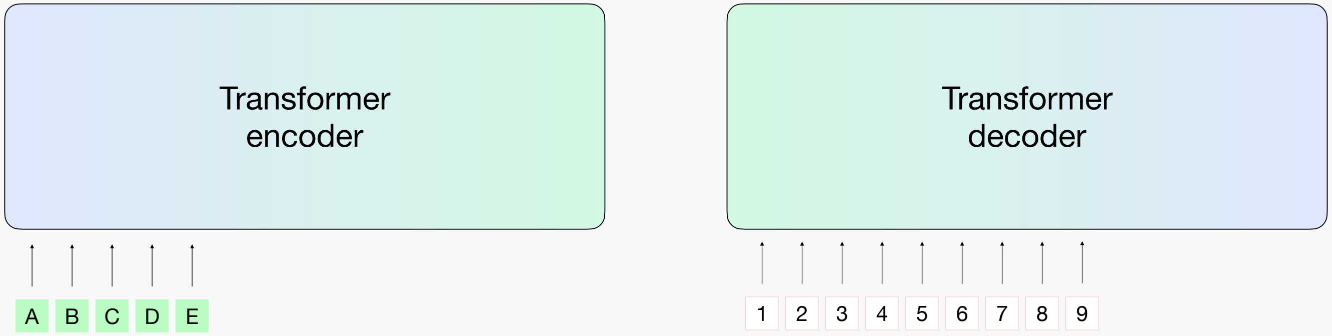
Bounding boxes

xmin=0.30 ymin=0.51
xmax=0.68 ymax=0.99
horse

→ A B C D E

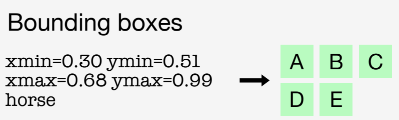
4M chained multimodal generation

Iteration 1 2 3 4 5 6 7
(Generate RGB with MaskGIT)

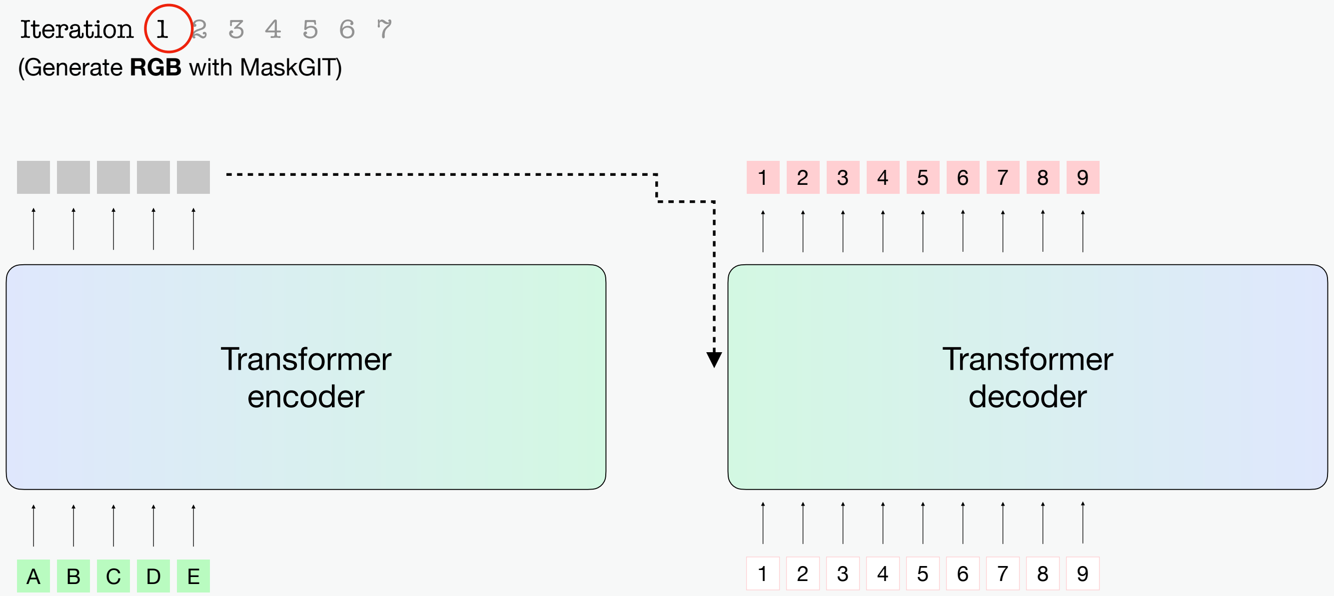


Self-consistent prediction through chained multimodal generation

Tokenization

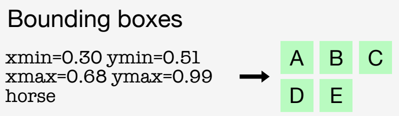


4M chained multimodal generation

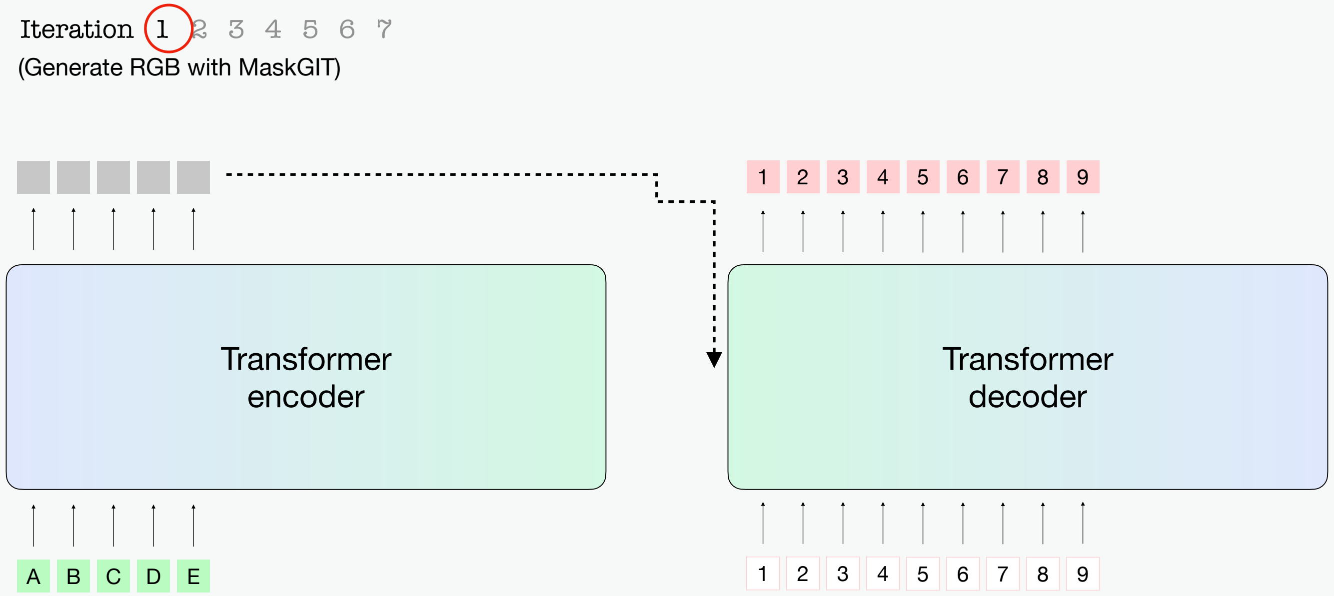


Self-consistent prediction through chained multimodal generation

Tokenization



4M chained multimodal generation



Self-consistent prediction through chained multimodal generation

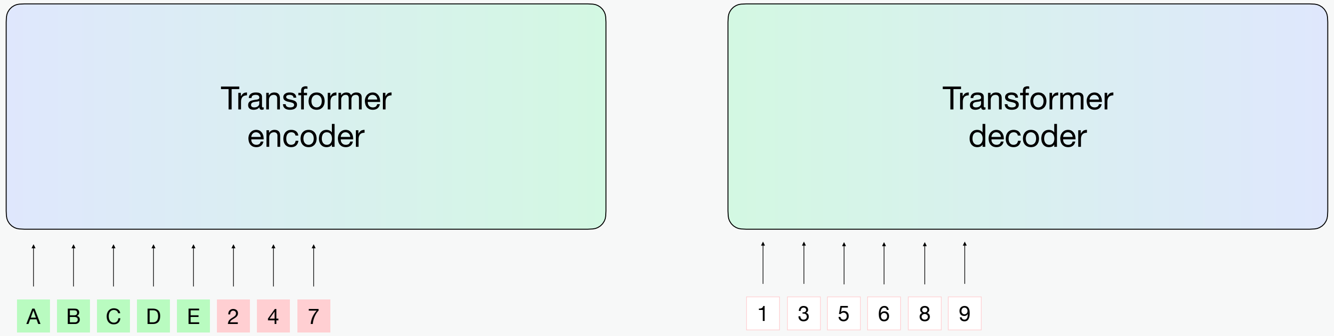
Tokenization

Bounding boxes

xmin=0.30 ymin=0.51
xmax=0.68 ymax=0.99
horse → A B C D E

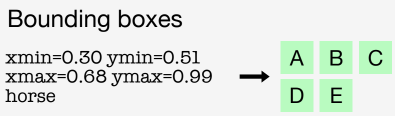
4M chained multimodal generation

Iteration 1 2 3 4 5 6 7
(Generate RGB with MaskGIT)



Self-consistent prediction through chained multimodal generation

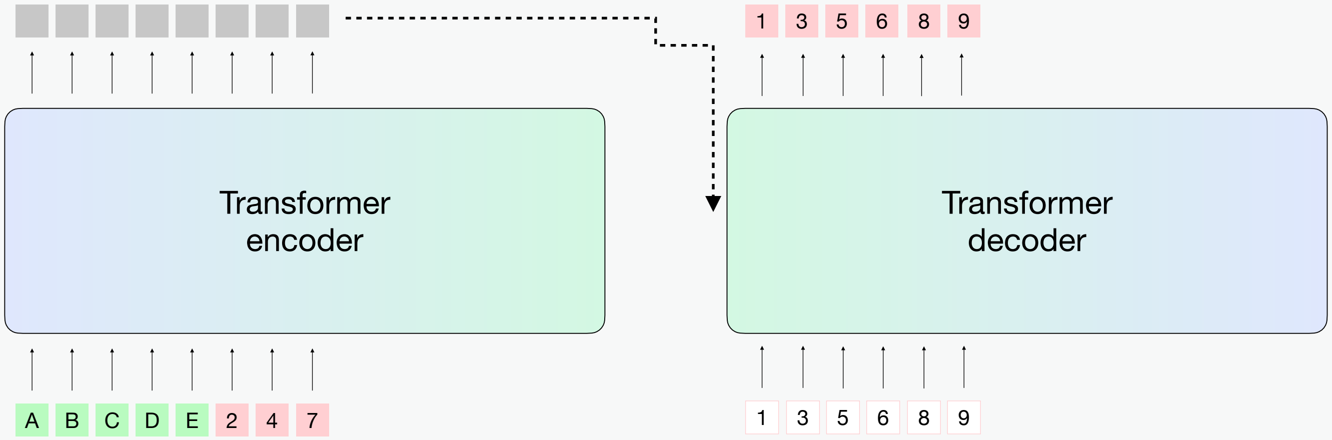
Tokenization



4M chained multimodal generation

Iteration 1 2 3 4 5 6 7

(Generate RGB with MaskGIT)

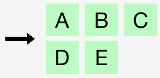


Self-consistent prediction through chained multimodal generation

Tokenization

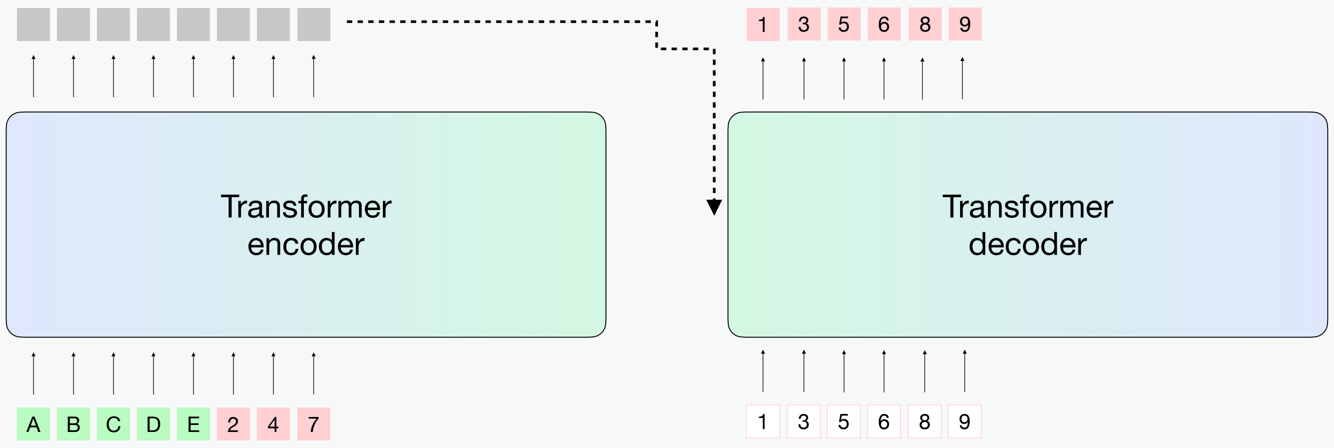
Bounding boxes

xmin=0.30 ymin=0.51
xmax=0.68 ymax=0.99
horse



4M chained multimodal generation

Iteration 1 2 3 4 5 6 7
(Generate RGB with MaskGIT)



Self-consistent prediction through chained multimodal generation

Tokenization

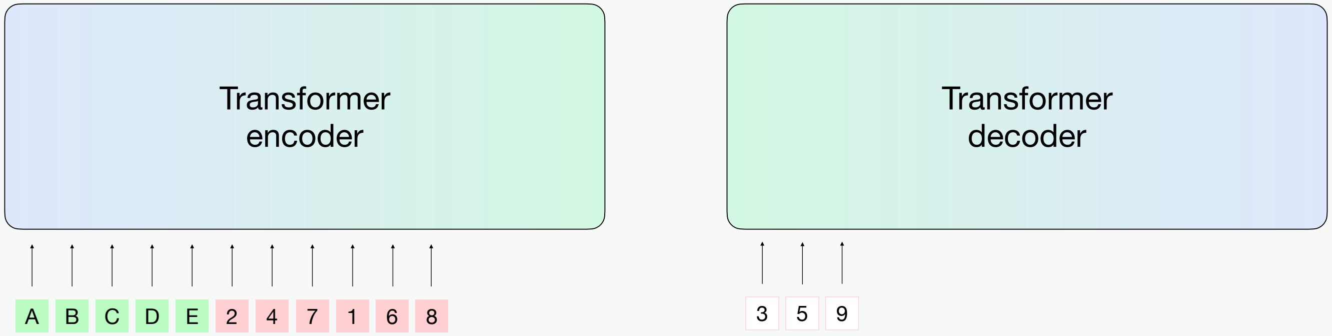
Bounding boxes

xmin=0.30 ymin=0.51
xmax=0.68 ymax=0.99
horse

→ A B C D E

4M chained multimodal generation

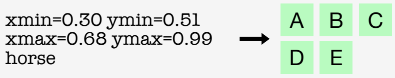
Iteration 1 2 3 4 5 6 7
(Generate RGB with MaskGIT)



Self-consistent prediction through chained multimodal generation

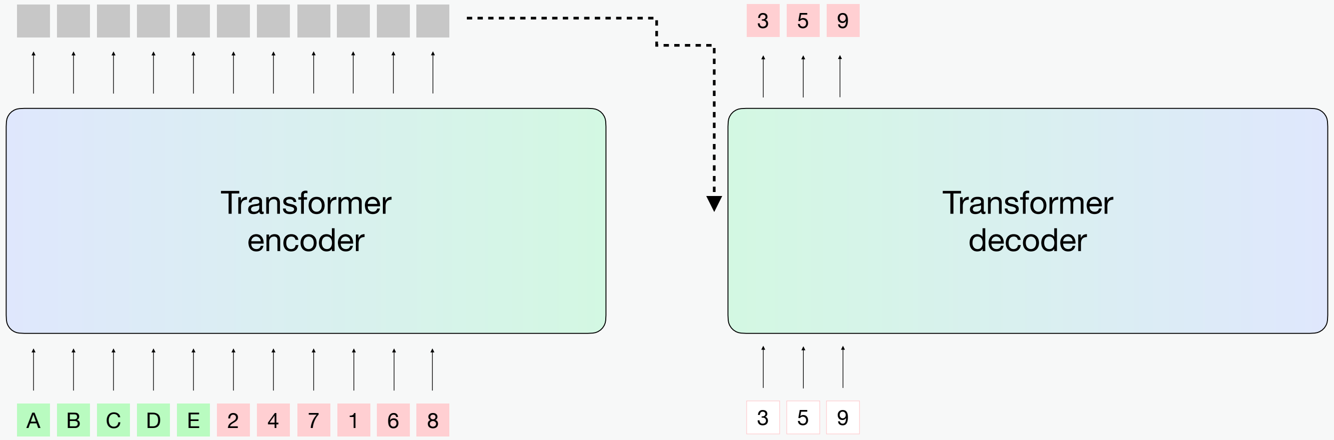
Tokenization

Bounding boxes



4M chained multimodal generation

Iteration 1 2 3 4 5 6 7
(Generate RGB with MaskGIT)



Self-consistent prediction through chained multimodal generation

Tokenization

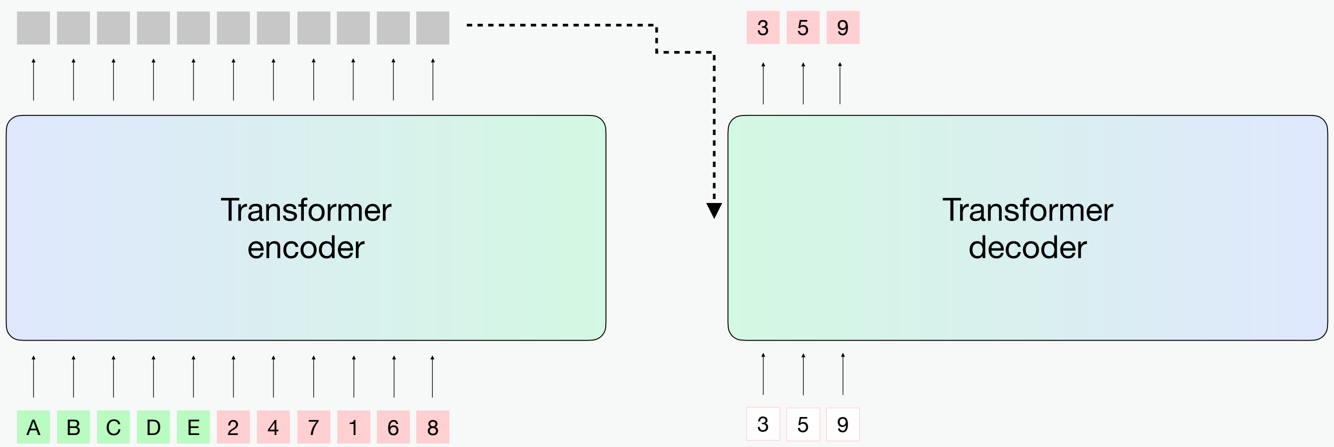
Bounding boxes

xmin=0.30 ymin=0.51
xmax=0.68 ymax=0.99
horse

→ A B C D E

4M chained multimodal generation

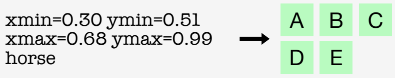
Iteration 1 2 3 4 5 6 7
(Generate RGB with MaskGIT)



Self-consistent prediction through chained multimodal generation

Tokenization

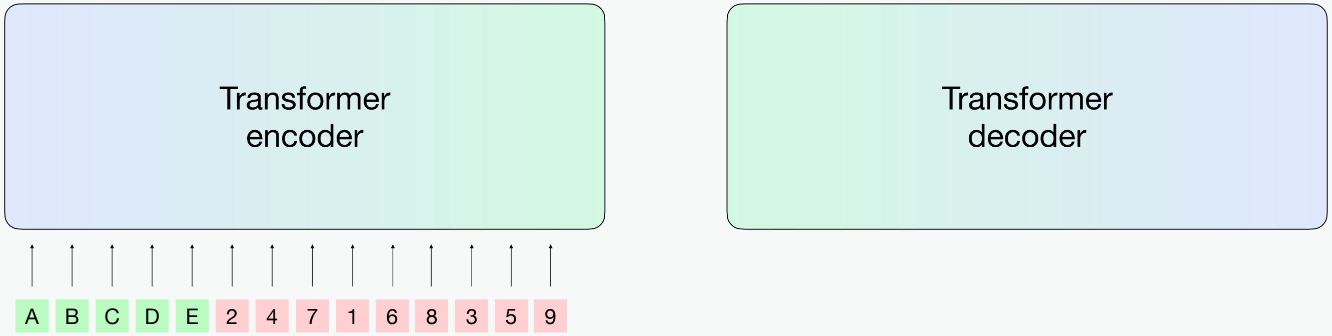
Bounding boxes



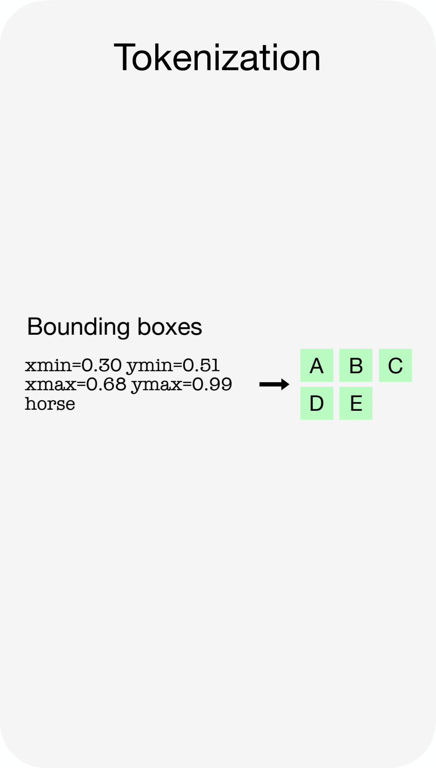
4M chained multimodal generation

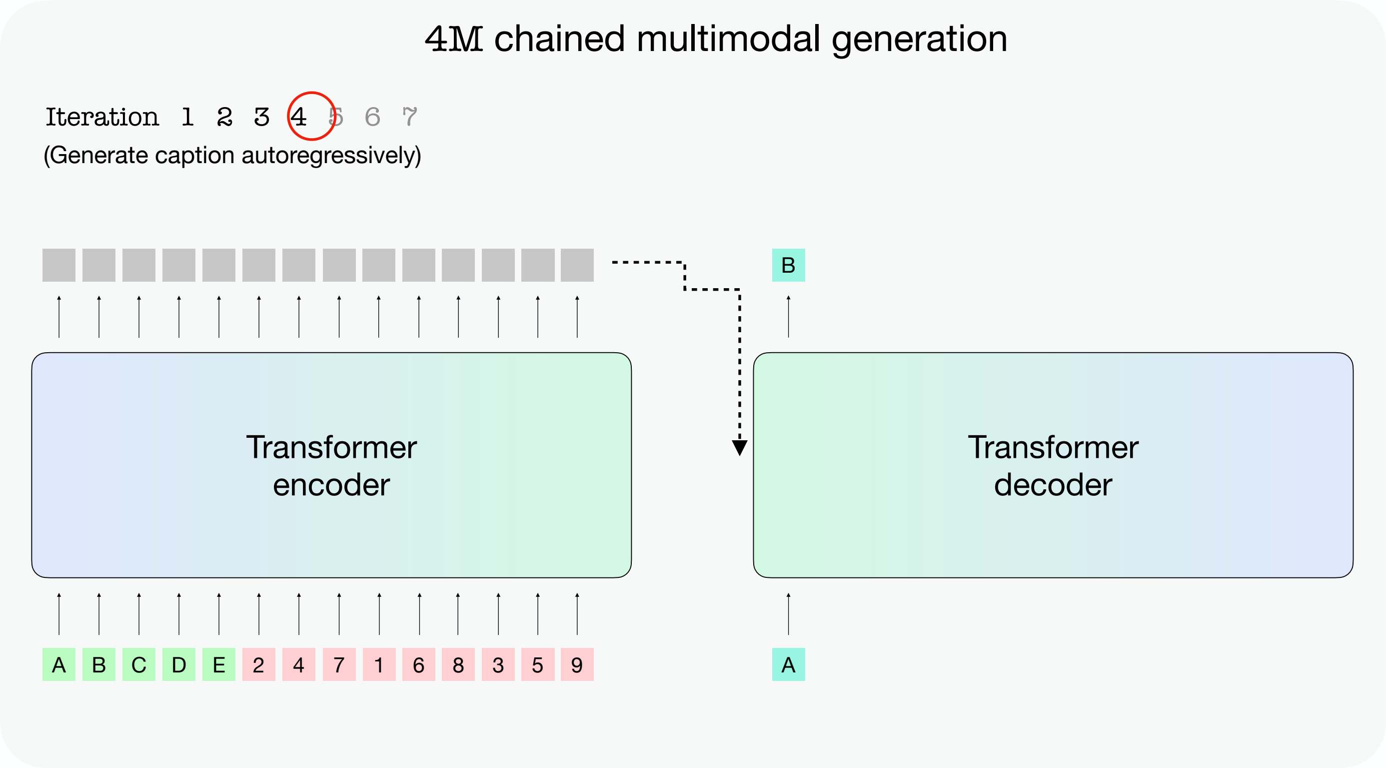
Iteration 1 2 3 4 5 6 7

(Generate caption autoregressively)



Self-consistent prediction through chained multimodal generation

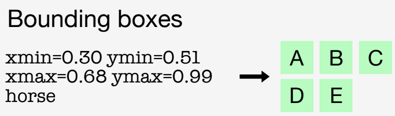




HELMHOLTZ AI Artificial Intelligence Cooperation Unit

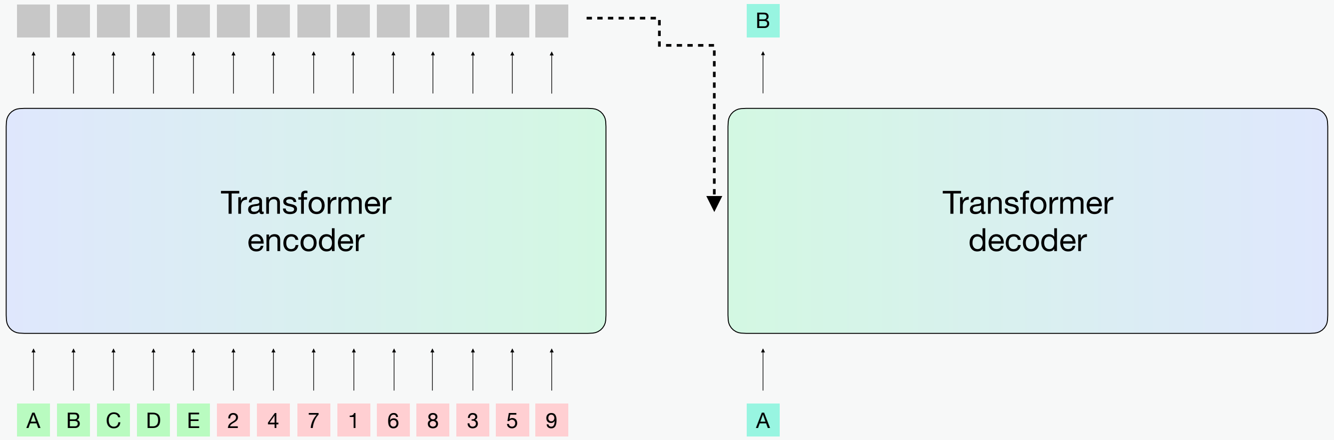
Self-consistent prediction through chained multimodal generation

Tokenization

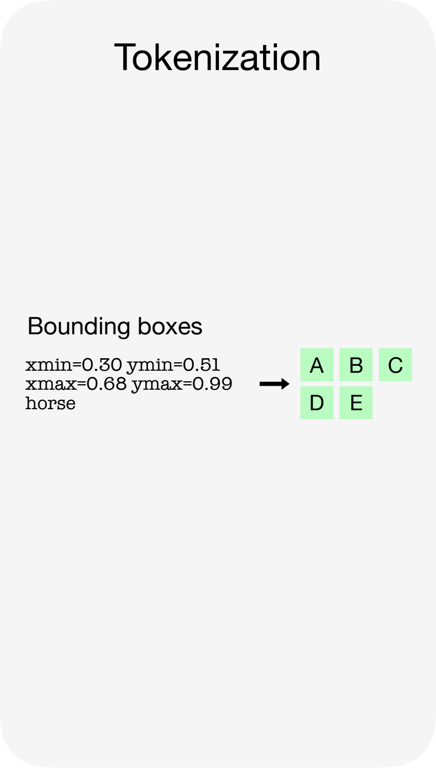


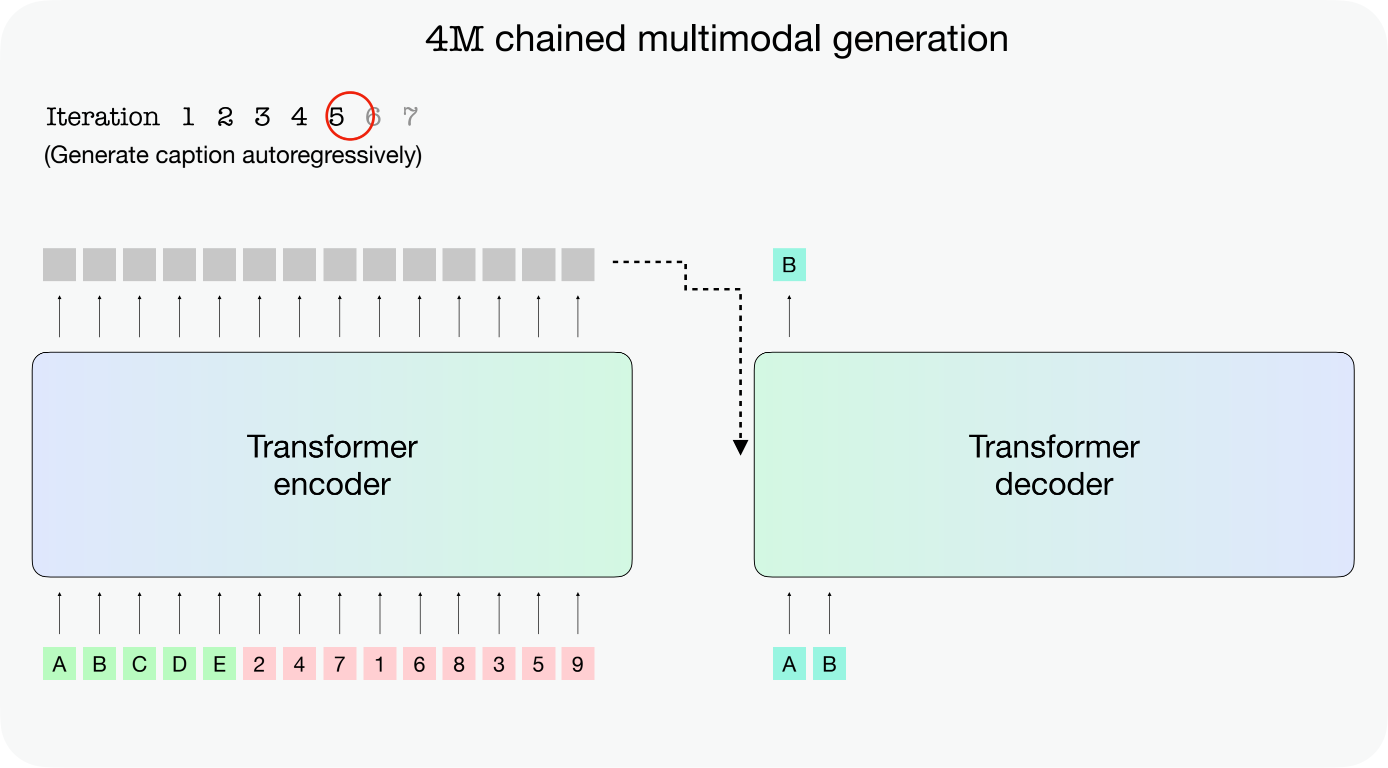
4M chained multimodal generation

Iteration 1 2 3 4 5 6 7
(Generate caption autoregressively)



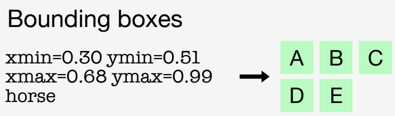
Self-consistent prediction through chained multimodal generation



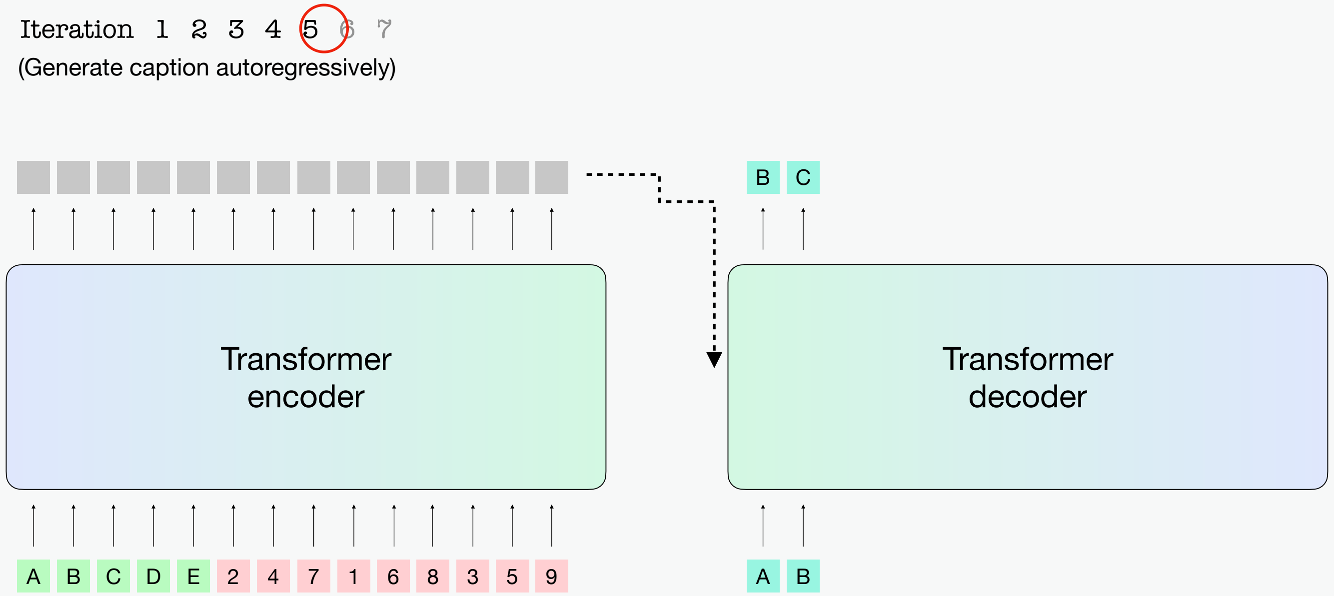


Self-consistent prediction through chained multimodal generation

Tokenization

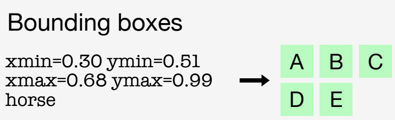


4M chained multimodal generation

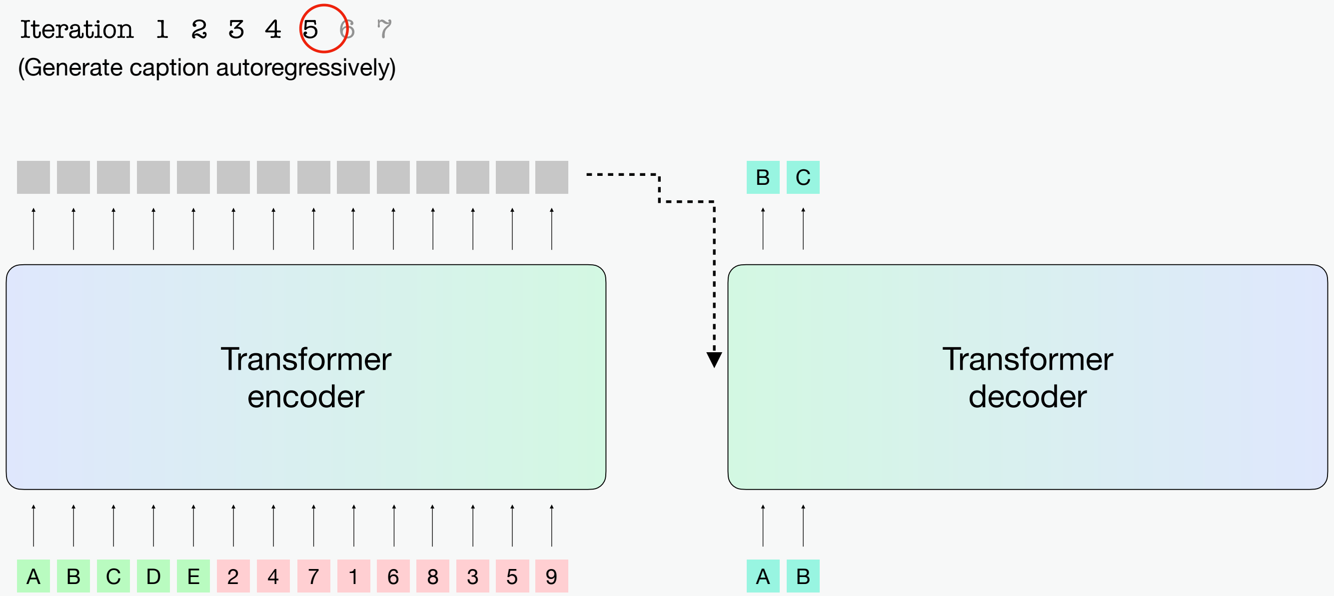


Self-consistent prediction through chained multimodal generation

Tokenization

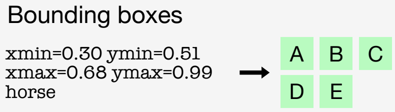


4M chained multimodal generation



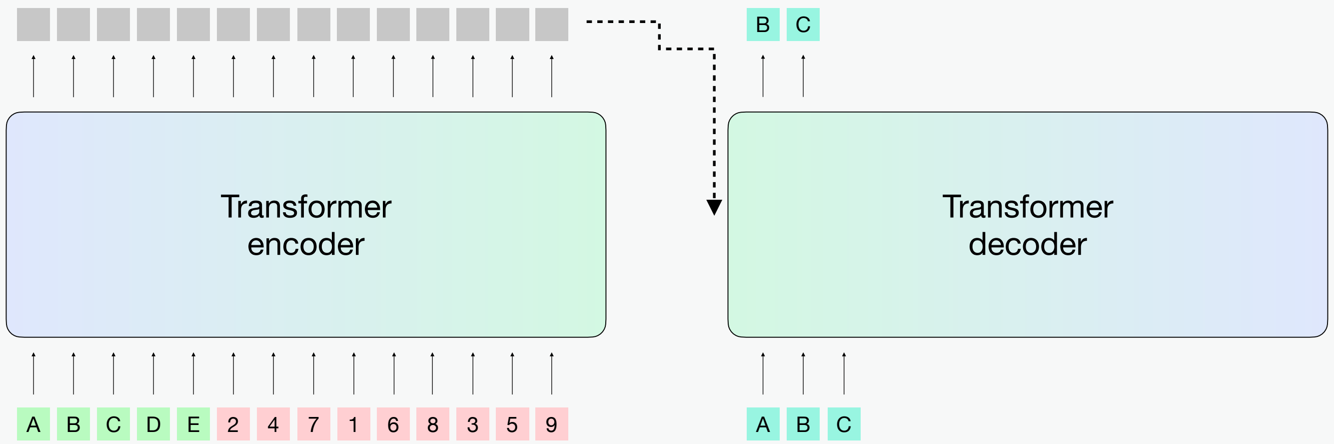
Self-consistent prediction through chained multimodal generation

Tokenization



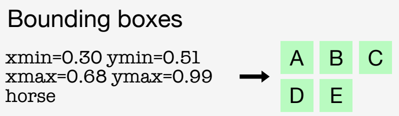
4M chained multimodal generation

Iteration 1 2 3 4 5 6 7
(Generate caption autoregressively)

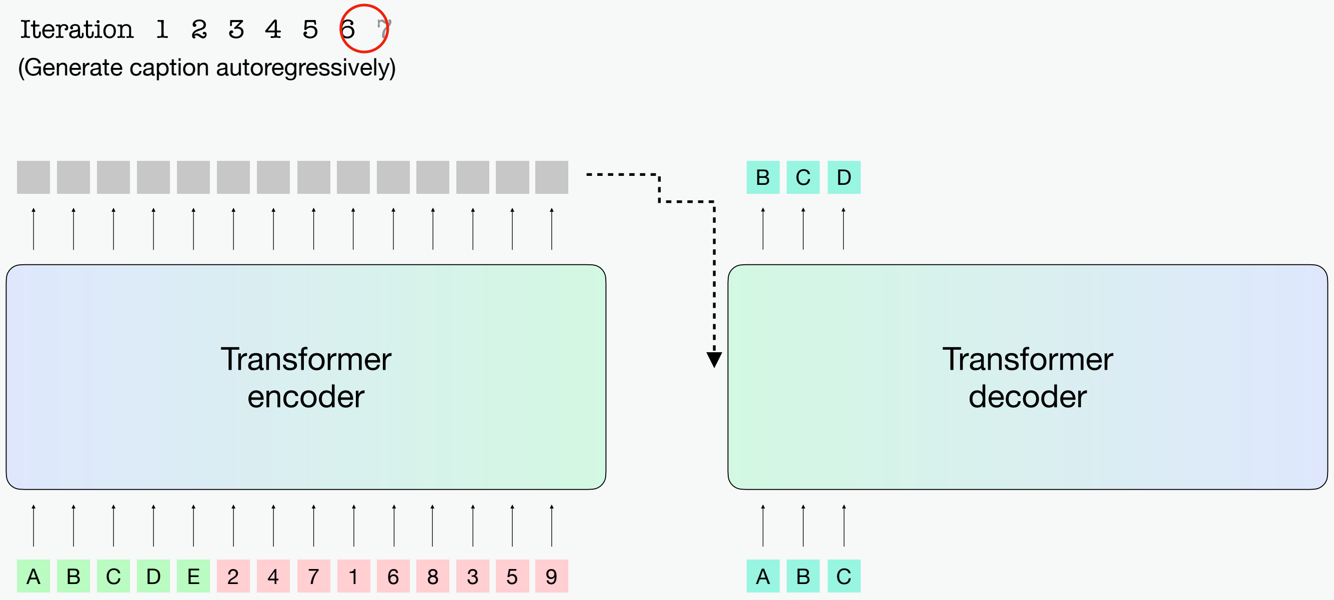


Self-consistent prediction through chained multimodal generation

Tokenization

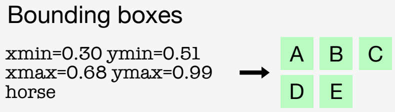


4M chained multimodal generation



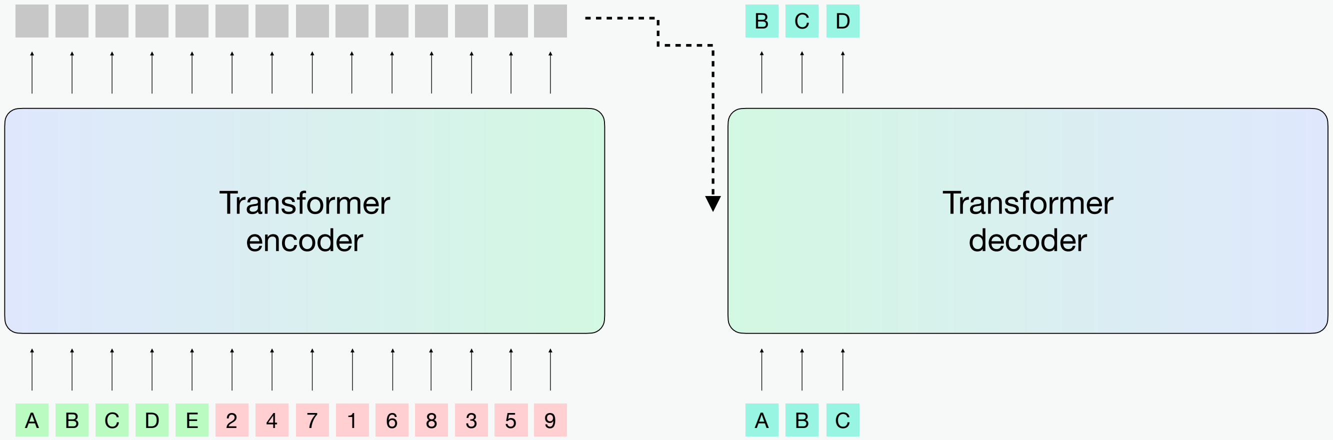
Self-consistent prediction through chained multimodal generation

Tokenization



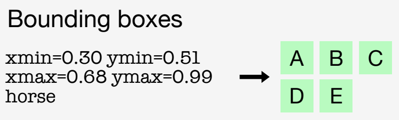
4M chained multimodal generation

Iteration 1 2 3 4 5 6 7
(Generate caption autoregressively)

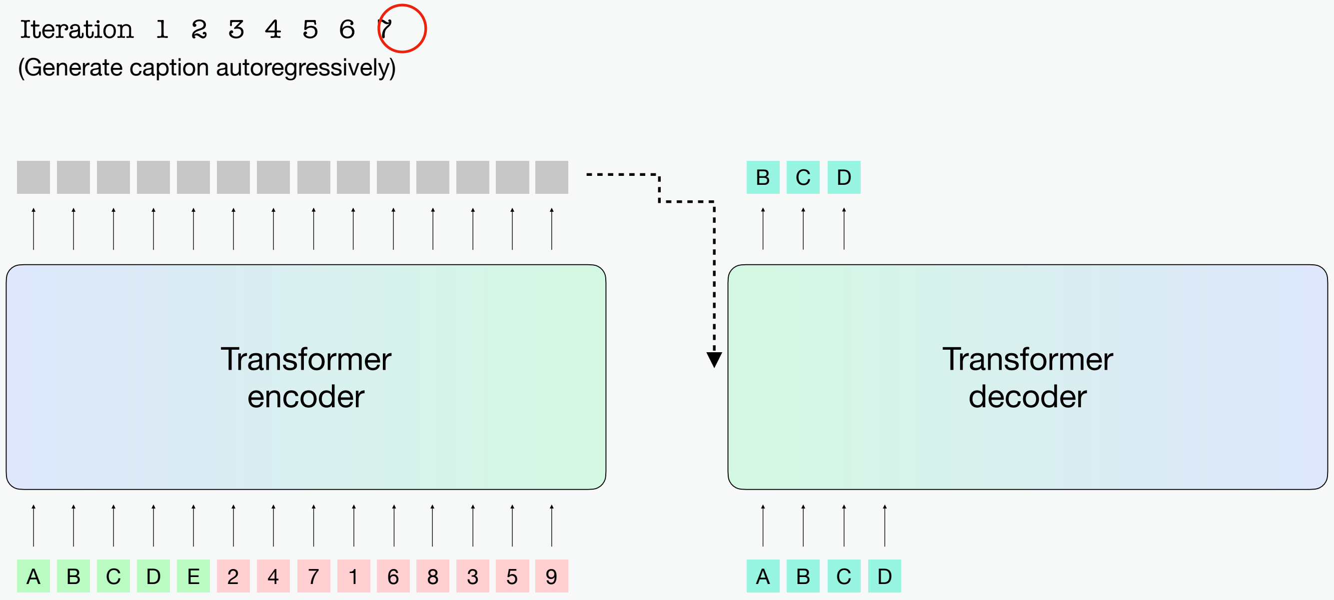


Self-consistent prediction through chained multimodal generation

Tokenization



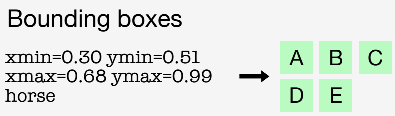
4M chained multimodal generation



HELMHOLTZ AI Artificial Intelligence Cooperation Unit

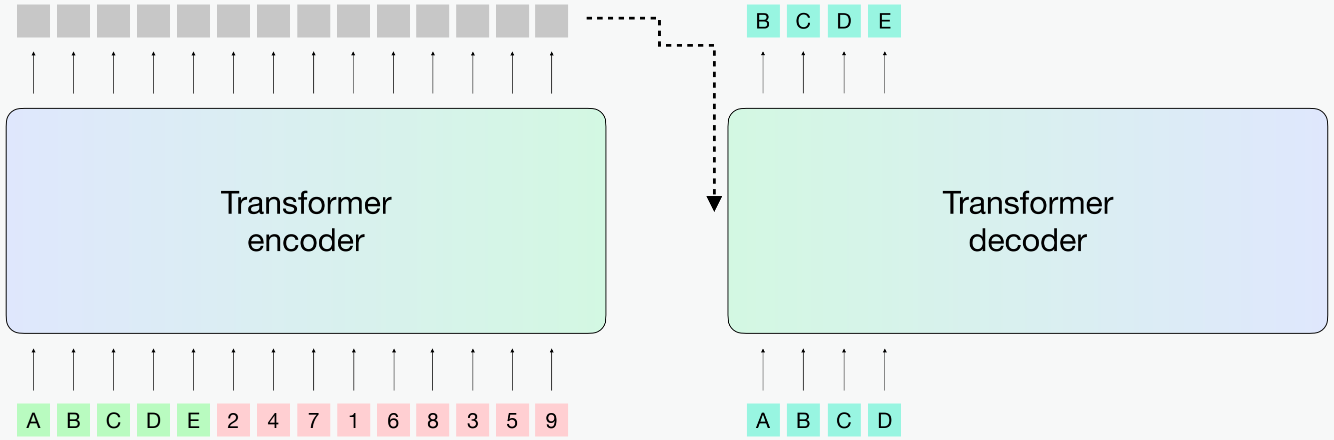
Self-consistent prediction through chained multimodal generation

Tokenization



4M chained multimodal generation

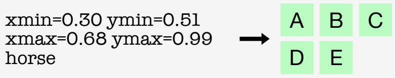
Iteration 1 2 3 4 5 6 7
(Generate caption autoregressively)



Self-consistent prediction through chained multimodal generation

Tokenization

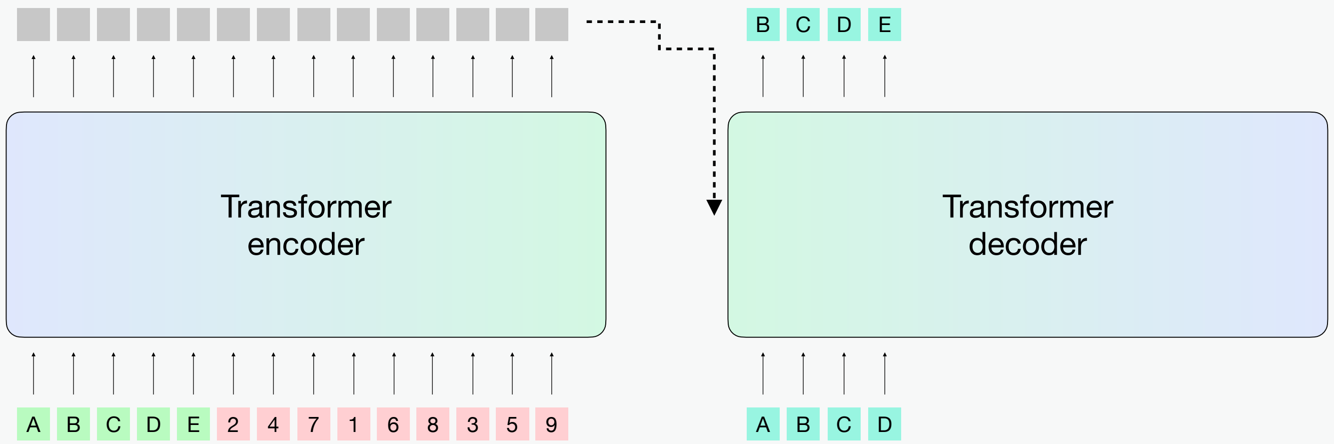
Bounding boxes



4M chained multimodal generation

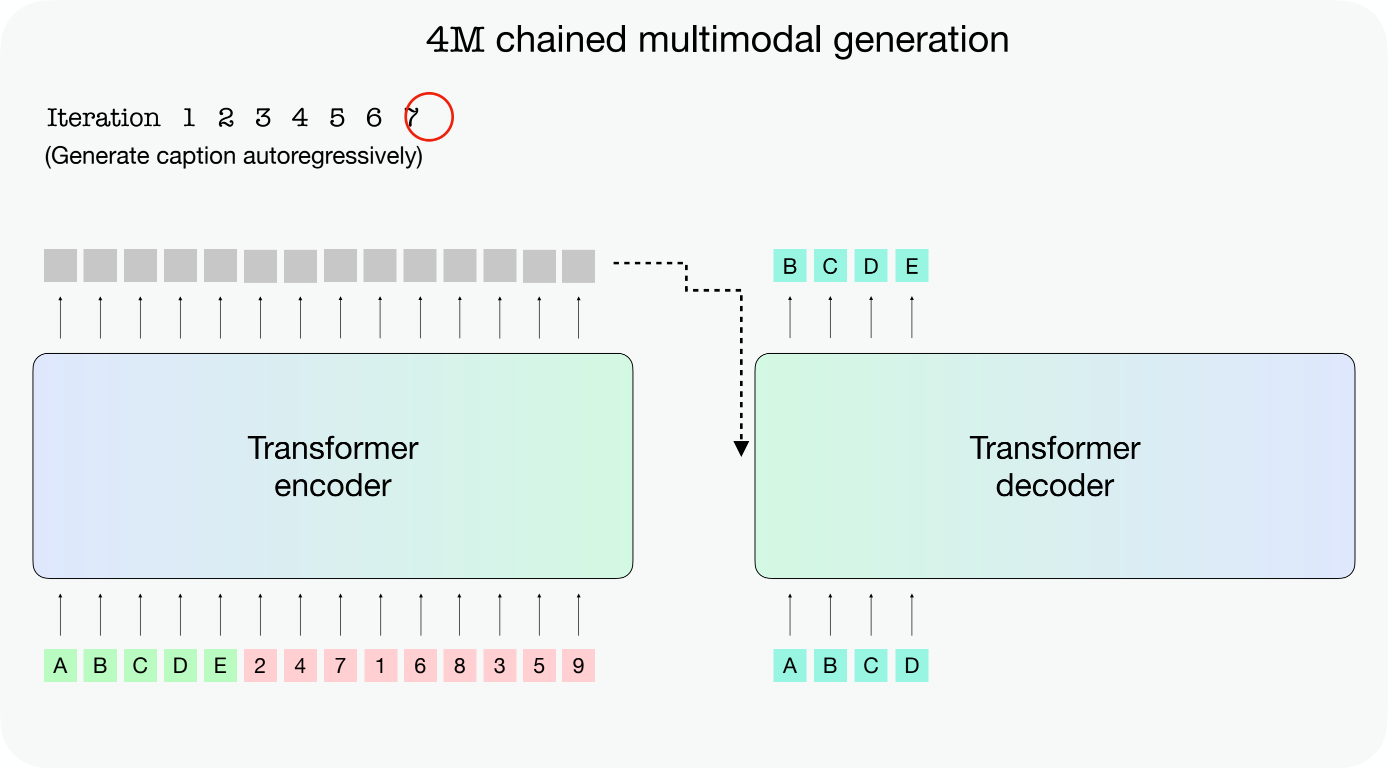
Iteration 1 2 3 4 5 6 7

(Generate caption autoregressively)

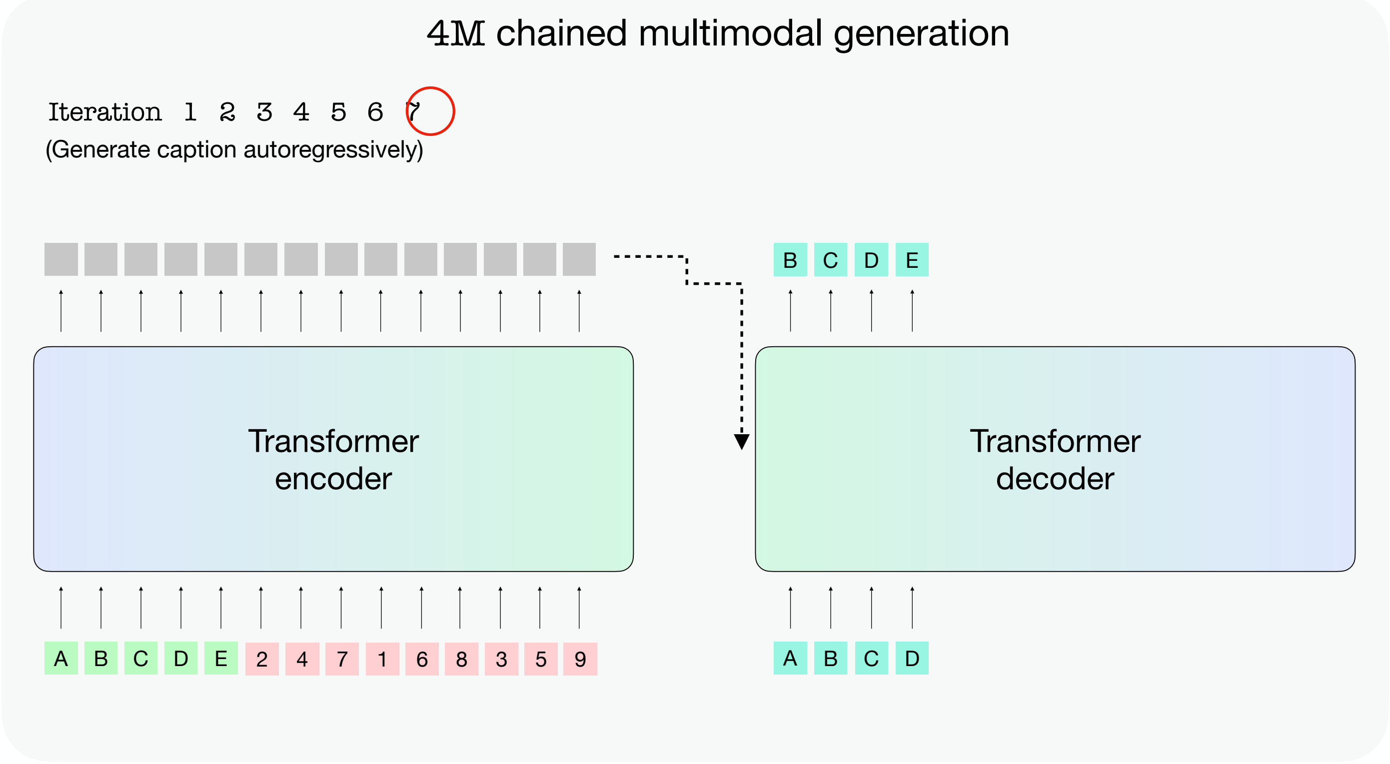


HELMHOLTZ AI Artificial Intelligence Cooperation Unit

Self-consistent prediction through chained multimodal generation

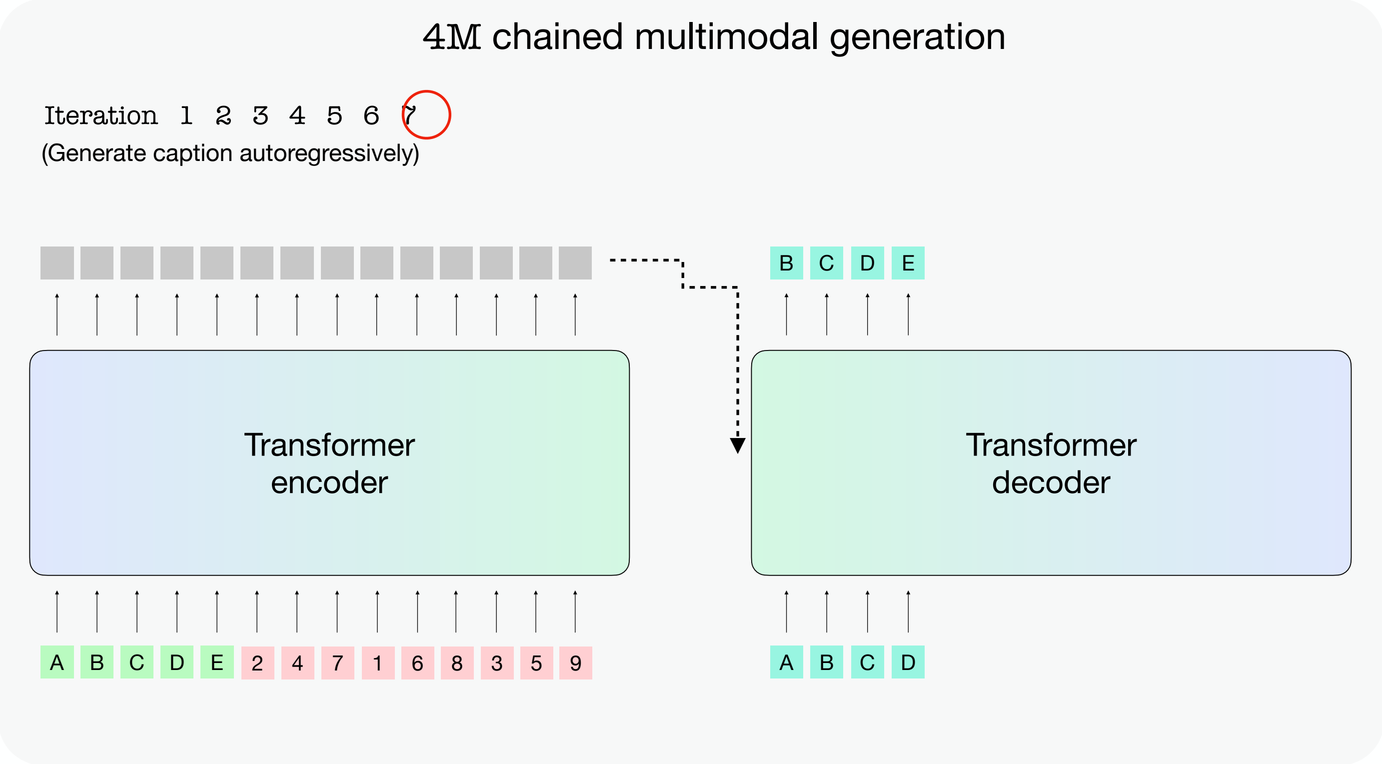


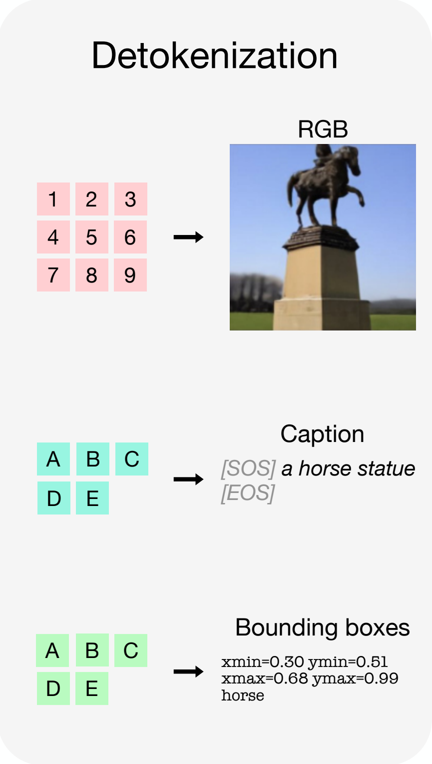
Self-consistent prediction through chained multimodal generation



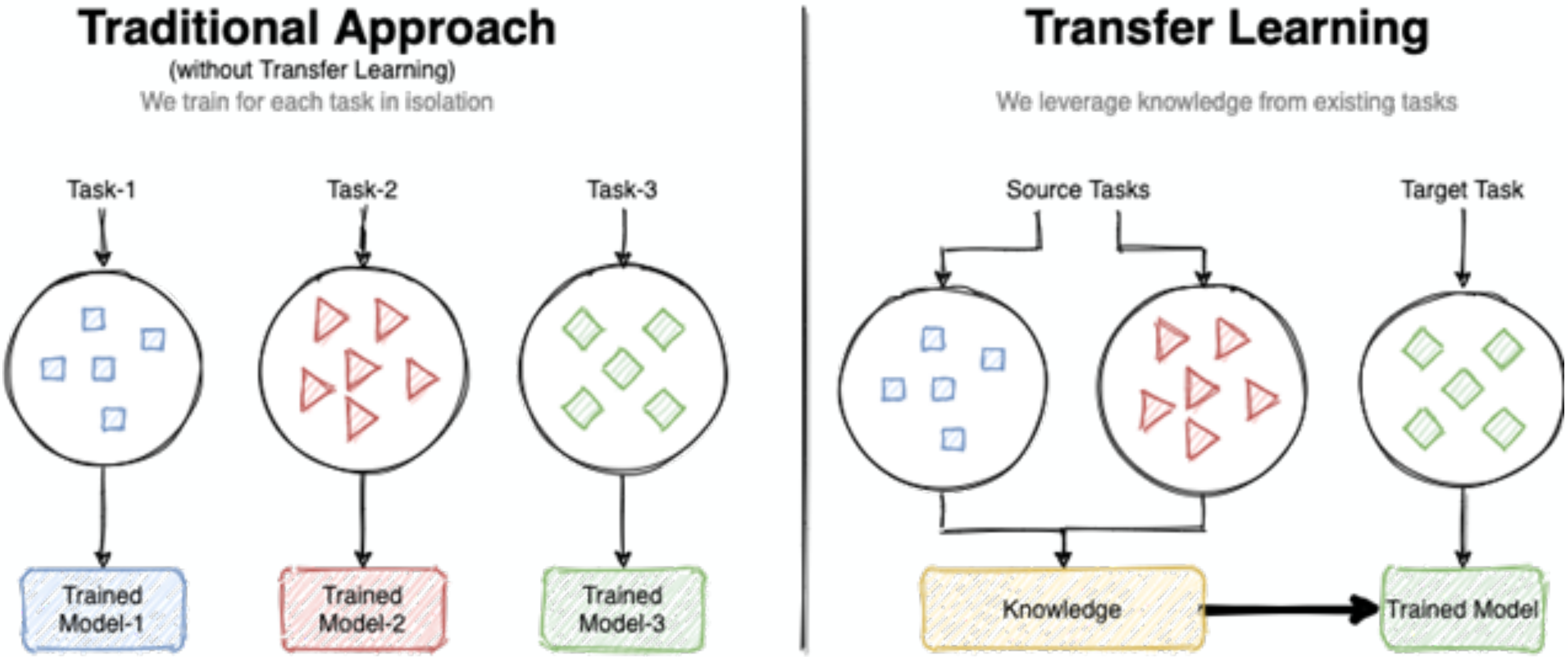
Detokenization

Self-consistent prediction through chained multimodal generation





Transfer learning



https://medium.com/@aimathavan14/transfer-learning-74cec7927b9f

HELMHOLTZ AI Artificial Intelligence Cooperation Unit

apple / ml-4m

Type / to search

Code | Issues 15 | Pull requests 2 | Security | Insights

ml-4m Public

Watch 33 | Fork 107 | Star 1.7k

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

roman-bachmann Added missing MAX_LEN_YAML_PARSE cda590f · last month 26 Commits

assets	Added arXiv link and demo sample output	last year
cfgs/default	4M-21 release	last year
fourm	Added missing MAX_LEN_YAML_PARSE	last month
notebooks	4M-21 retrieval demo notebook	last year
.gitattributes	first commit	last year
.gitignore	first commit	last year
ACKNOWLEDGEMENTS.md	4M-21 release	last year
CODE_OF_CONDUCT.md	first commit	last year
CONTRIBUTING.md	first commit	last year
LICENSE	first commit	last year
LICENSE_WEIGHTS	first commit	last year
README.md	Remove call to ast.literal_eval	2 months ago
README_DATA.md	4M-21 release	last year
README_GENERATION.md	4M-21 release	last year
README_TOKENIZATION.md	docs: update README_TOKENIZATION.md	last year
README_TRAINING.md	4M-21 release	last year

About

4M: Massively Multimodal Masked Modeling

4m.epfl.ch

- Readme
- Apache-2.0 license
- Code of conduct
- Activity
- Custom properties
- 1.7k stars
- 33 watching
- 107 forks

Report repository

Releases

No releases published

Packages

No packages published

Contributors 7

Languages

https://github.com/apple/ml-4m

4M models

Model	# Mod.	Datasets	# Params	Config	Weights
4M-B	7	CC12M	198M	Config	Checkpoint / HF Hub
4M-B	7	COYO700M	198M	Config	Checkpoint / HF Hub
4M-B	21	CC12M+COYO700M+C4	198M	Config	Checkpoint / HF Hub
4M-L	7	CC12M	705M	Config	Checkpoint / HF Hub
4M-L	7	COYO700M	705M	Config	Checkpoint / HF Hub
4M-L	21	CC12M+COYO700M+C4	705M	Config	Checkpoint / HF Hub
4M-XL	7	CC12M	2.8B	Config	Checkpoint / HF Hub
4M-XL	7	COYO700M	2.8B	Config	Checkpoint / HF Hub
4M-XL	21	CC12M+COYO700M+C4	2.8B	Config	Checkpoint / HF Hub

To load models from Hugging Face Hub:

```
from fourm.models.fm import FM

fm7b_cc12m  = FM.from_pretrained('EPFL-VILAB/4M-7_B_CC12M')
fm7b_coyo   = FM.from_pretrained('EPFL-VILAB/4M-7_B_COYO700M')
fm21b       = FM.from_pretrained('EPFL-VILAB/4M-21_B')

fm7l_cc12m  = FM.from_pretrained('EPFL-VILAB/4M-7_L_CC12M')
fm7l_coyo   = FM.from_pretrained('EPFL-VILAB/4M-7_L_COYO700M')
fm21l       = FM.from_pretrained('EPFL-VILAB/4M-21_L')

fm7xl_cc12m = FM.from_pretrained('EPFL-VILAB/4M-7_XL_CC12M')
fm7xl_coyo  = FM.from_pretrained('EPFL-VILAB/4M-7_XL_COYO700M')
fm21xl      = FM.from_pretrained('EPFL-VILAB/4M-21_XL')
```

Tokenizers

Modality	Resolution	Number of tokens	Codebook size	Diffusion decoder	Weights
RGB	224-448	196-784	16k	✓	Checkpoint / HF Hub
Depth	224-448	196-784	8k	✓	Checkpoint / HF Hub
Normals	224-448	196-784	8k	✓	Checkpoint / HF Hub
Edges (Canny, SAM)	224-512	196-1024	8k	✓	Checkpoint / HF Hub
COCO semantic segmentation	224-448	196-784	4k	✕	Checkpoint / HF Hub
CLIP-B/16	224-448	196-784	8k	✕	Checkpoint / HF Hub
DINOv2-B/14	224-448	256-1024	8k	✕	Checkpoint / HF Hub
DINOv2-B/14 (global)	224	16	8k	✕	Checkpoint / HF Hub
ImageBind-H/14	224-448	256-1024	8k	✕	Checkpoint / HF Hub
ImageBind-H/14 (global)	224	16	8k	✕	Checkpoint / HF Hub
SAM instances	-	64	1k	✕	Checkpoint / HF Hub
3D Human poses	-	8	1k	✕	Checkpoint / HF Hub

Out-of-the-box capabilities evaluation

	Method	Normals ↓	Depth ↓	Sem. seg. ↑	Inst. seg. ↑	IN1K kNN ↑	3D human KP ↓
Pseudo labelers	Omnidata [37]	22.5	**0.68**	✗	✗	✗	✗
	M2F-B [18]	✗	✗	45.7	✗	✗	✗
	SAM [39]	✗	✗	✗	**32.9**	✗	✗
	DINOv2-B14 [53]	✗	✗	✗	✗	**82.1** / 93.9	✗
	ImageBind-H14 [28]	✗	✗	✗	✗	81.1 / **94.4**	✗
	4D-Humans [30]	✗	✗	✗	✗	✗	**81.3**
	OASIS [17]	34.3	✗	✗	✗	✗	✗
	MiDaS DPT [58]	✗	0.73	✗	✗	✗	✗
	M2F-S [18]	✗	✗	44.6	✗	✗	✗
	M2F-L [18]	✗	✗	48.0	✗	✗	✗
	HMR [36]	✗	✗	✗	✗	✗	130.0
	UnifiedIO-B [47]	35.7	1.00	32.9	✗	✗	✗
	UnifiedIO-L [47]	33.9	0.87	41.6	✗	✗	✗
	UnifiedIO-XL [47]	31.0	0.82	44.3	✗	✗	✗
	4M B [50]	21.9	0.71	43.3	✗	✗	✗
	Ours B	21.7	0.71	42.5	15.9	73.1 / 89.7	108.3
	4M L [50]	21.5	0.69	47.2	✗	✗	✗
	Ours L	21.1	0.69	46.4	31.2	77.0 / 91.9	97.4
	4M XL [50]	**20.6**	0.69	**48.1**	✗	✗	✗
	Ours XL	20.8	**0.68**	**48.1**	32.0	78.3 / 92.4	92.0
	Tokenizer bound*	4.0	0.06	90.5	91.2	80.2 / 93.0	17.5

- Outperform or approach performance of **strong specialist models** (incl. pseudo labelers)

✗ = given model does not have capability out-of-the-box

Out-of-the-box capabilities evaluation

	Method	Normals ↓	Depth ↓	Sem. seg. ↑	Inst. seg. ↑	IN1K kNN ↑	3D human KP ↓
Pseudo labelers	Omnidata [37]	22.5	**0.68**	✗	✗	✗	✗
	M2F-B [18]	✗	✗	45.7	✗	✗	✗
	SAM [39]	✗	✗	✗	**32.9**	✗	✗
	DINOv2-B14 [53]	✗	✗	✗	✗	**82.1** / 93.9	✗
	ImageBind-H14 [28]	✗	✗	✗	✗	81.1 / **94.4**	✗
	4D-Humans [30]	✗	✗	✗	✗	✗	**81.3**
	OASIS [17]	34.3	✗	✗	✗	✗	✗
	MiDaS DPT [58]	✗	0.73	✗	✗	✗	✗
	M2F-S [18]	✗	✗	44.6	✗	✗	✗
	M2F-L [18]	✗	✗	48.0	✗	✗	✗
	HMR [36]	✗	✗	✗	✗	✗	130.0
	UnifiedIO-B [47]	35.7	1.00	32.9	✗	✗	✗
	UnifiedIO-L [47]	33.9	0.87	41.6	✗	✗	✗
	UnifiedIO-XL [47]	31.0	0.82	44.3	✗	✗	✗
	4M B [50]	21.9	0.71	43.3	✗	✗	✗
	Ours B	21.7	0.71	42.5	15.9	73.1 / 89.7	108.3
	4M L [50]	21.5	0.69	47.2	✗	✗	✗
	Ours L	21.1	0.69	46.4	31.2	77.0 / 91.9	97.4
	4M XL [50]	**20.6**	0.69	**48.1**	✗	✗	✗
	Ours XL	20.8	**0.68**	**48.1**	32.0	78.3 / 92.4	92.0
	Tokenizer bound*	4.0	0.06	90.5	91.2	80.2 / 93.0	17.5

✗ = given model does not have capability out-of-the-box

- Outperform or approach performance of **strong specialist models** (incl. pseudo labelers)
- Outperform strong multimodal/multitask models

Out-of-the-box capabilities evaluation

	Method	Normals ↓	Depth ↓	Sem. seg. ↑	Inst. seg. ↑	IN1K kNN ↑	3D human KP ↓
Pseudo labelers	Omnidata [37]	22.5	**0.68**	✗	✗	✗	✗
	M2F-B [18]	✗	✗	45.7	✗	✗	✗
	SAM [39]	✗	✗	✗	**32.9**	✗	✗
	DINOv2-B14 [53]	✗	✗	✗	✗	**82.1** / 93.9	✗
	ImageBind-H14 [28]	✗	✗	✗	✗	81.1 / **94.4**	✗
	4D-Humans [30]	✗	✗	✗	✗	✗	**81.3**
	OASIS [17]	34.3	✗	✗	✗	✗	✗
	MiDaS DPT [58]	✗	0.73	✗	✗	✗	✗
	M2F-S [18]	✗	✗	44.6	✗	✗	✗
	M2F-L [18]	✗	✗	48.0	✗	✗	✗
	HMR [36]	✗	✗	✗	✗	✗	130.0
	UnifiedIO-B [47]	35.7	1.00	32.9	✗	✗	✗
	UnifiedIO-L [47]	33.9	0.87	41.6	✗	✗	✗
	UnifiedIO-XL [47]	31.0	0.82	44.3	✗	✗	✗
	4M B [50]	21.9	0.71	43.3	✗	✗	✗
	Ours B	21.7	0.71	42.5	15.9	73.1 / 89.7	108.3
	4M L [50]	21.5	0.69	47.2	✗	✗	✗
	Ours L	21.1	0.69	46.4	31.2	77.0 / 91.9	97.4
	4M XL [50]	**20.6**	0.69	**48.1**	✗	✗	✗
	Ours XL	20.8	**0.68**	**48.1**	32.0	78.3 / 92.4	92.0
	Tokenizer bound*	4.0	0.06	90.5	91.2	80.2 / 93.0	17.5

✗ = given model does not have capability out-of-the-box

- Outperform or approach performance of **strong specialist models** (incl. pseudo labelers)
- Outperform strong multimodal/multitask models
- Match performance of 4M on common tasks...

Out-of-the-box capabilities evaluation

	Method	Normals ↓	Depth ↓	Sem. seg. ↑	Inst. seg. ↑	IN1K kNN ↑	3D human KP ↓
Pseudo labelers	Omnidata [37]	22.5	**0.68**	✗	✗	✗	✗
	M2F-B [18]	✗	✗	45.7	✗	✗	✗
	SAM [39]	✗	✗	✗	**32.9**	✗	✗
	DINOv2-B14 [53]	✗	✗	✗	✗	**82.1** / <u>93.9</u>	✗
	ImageBind-H14 [28]	✗	✗	✗	✗	<u>81.1</u> / **94.4**	✗
	4D-Humans [30]	✗	✗	✗	✗	✗	**81.3**
	OASIS [17]	34.3	✗	✗	✗	✗	✗
	MiDaS DPT [58]	✗	0.73	✗	✗	✗	✗
	M2F-S [18]	✗	✗	44.6	✗	✗	✗
	M2F-L [18]	✗	✗	<u>48.0</u>	✗	✗	✗
	HMR [36]	✗	✗	✗	✗	✗	130.0
	UnifiedIO-B [47]	35.7	1.00	32.9	✗	✗	✗
	UnifiedIO-L [47]	33.9	0.87	41.6	✗	✗	✗
	UnifiedIO-XL [47]	31.0	0.82	44.3	✗	✗	✗
	4M B [50]	21.9	0.71	43.3	✗	✗	✗
	Ours B	21.7	0.71	42.5	15.9	73.1 / 89.7	108.3
	4M L [50]	21.5	<u>0.69</u>	47.2	✗	✗	✗
	Ours L	21.1	<u>0.69</u>	46.4	31.2	77.0 / 91.9	97.4
	4M XL [50]	**20.6**	<u>0.69</u>	**48.1**	✗	✗	✗
	Ours XL	<u>20.8</u>	**0.68**	**48.1**	<u>32.0</u>	78.3 / 92.4	<u>92.0</u>
	Tokenizer bound*	4.0	0.06	90.5	91.2	80.2 / 93.0	17.5

✗ = given model does not have capability out-of-the-box

- Outperform or approach performance of **strong specialist models** (incl. pseudo labelers)
- Outperform strong multimodal/multitask models
- Match performance of 4M on common tasks…

… while being able to solve 3x more tasks/modalities

Out-of-the-box capabilities evaluation

	Method	Normals ↓	Depth ↓	Sem. seg. ↑	Inst. seg. ↑	IN1K kNN ↑	3D human KP ↓
Pseudo labelers	Omnidata [37]	22.5	**0.68**	✗	✗	✗	✗
	M2F-B [18]	✗	✗	45.7	✗	✗	✗
	SAM [39]	✗	✗	✗	**32.9**	✗	✗
	DINOv2-B14 [53]	✗	✗	✗	✗	**82.1** / <u>93.9</u>	✗
	ImageBind-H14 [28]	✗	✗	✗	✗	<u>81.1</u> / **94.4**	✗
	4D-Humans [30]	✗	✗	✗	✗	✗	**81.3**
	OASIS [17]	34.3	✗	✗	✗	✗	✗
	MiDaS DPT [58]	✗	0.73	✗	✗	✗	✗
	M2F-S [18]	✗	✗	44.6	✗	✗	✗
	M2F-L [18]	✗	✗	<u>48.0</u>	✗	✗	✗
	HMR [36]	✗	✗	✗	✗	✗	130.0
	UnifiedIO-B [47]	35.7	1.00	32.9	✗	✗	✗
	UnifiedIO-L [47]	33.9	0.87	41.6	✗	✗	✗
	UnifiedIO-XL [47]	31.0	0.82	44.3	✗	✗	✗
	4M B [50]	21.9	0.71	43.3	✗	✗	✗
	Ours B	21.7	0.71	42.5	15.9	73.1 / 89.7	108.3
	4M L [50]	21.5	<u>0.69</u>	47.2	✗	✗	✗
	Ours L	21.1	<u>0.69</u>	46.4	31.2	77.0 / 91.9	97.4
	4M XL [50]	**20.6**	<u>0.69</u>	**48.1**	✗	✗	✗
	Ours XL	<u>20.8</u>	**0.68**	**48.1**	<u>32.0</u>	78.3 / 92.4	<u>92.0</u>
	Tokenizer bound*	4.0	0.06	90.5	91.2	80.2 / 93.0	17.5

✗ = given model does not have capability out-of-the-box

- Outperform or approach performance of **strong specialist models** (incl. pseudo labelers)
- Outperform strong multimodal/multitask models
- Match performance of 4M on common tasks...

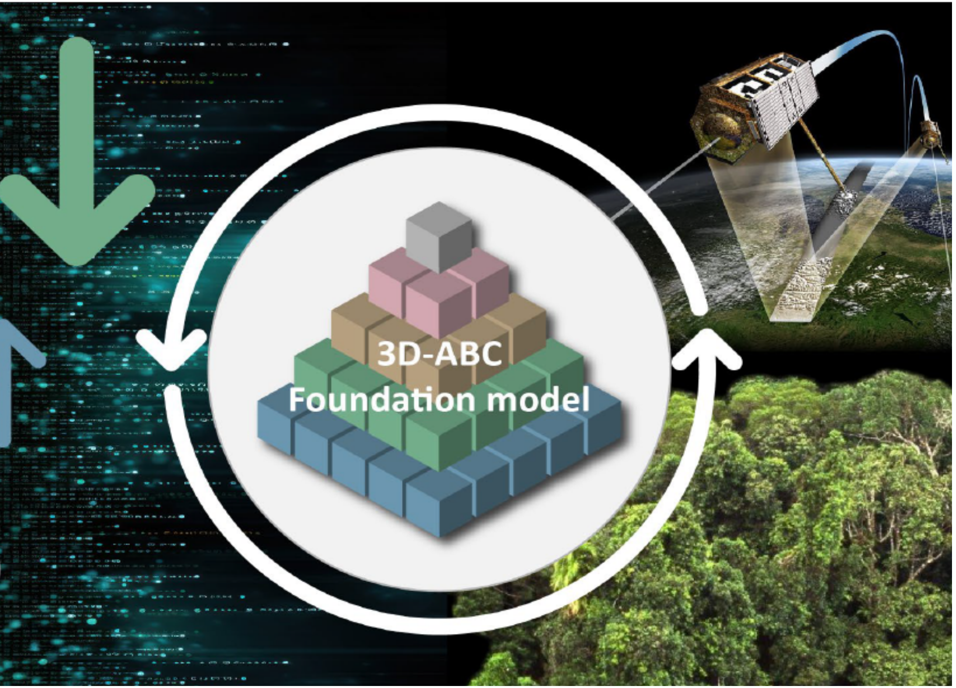
  ... while being able to solve 3x more tasks/modalities
- Tokenization does not create a performance bottleneck

Take home message

The future of foundation models is unification, not just scale. The 4M model shows that with the right architecture, we can train a single model that learns shared representations across vision, text, audio, and structured data. This paves the way for foundation models that are flexible and easily adaptable and not just bigger.

More resources : https://4m.epfl.ch/

What’s Next



“*Foundation Model Approach for Global Terrestrial Carbon Stock Mapping*” by Aldino Rizaldy

at Oncoray, Dresden on July 7th 2025

https://events.hifis.net/event/2680/