

CS-503 Visual Intelligence: Machines and Minds

Roman Bachmann

15.04.2025



Generative modeling

Overview

Post-training, evaluations & reasoning



Task: Tune the base models to follow instructions, be more aligned, do reasoning, etc...

Result: More useful, aligned, and performant models during test-time

\$

Pre-training & scaling

Task: Model full data distribution $p(x)$, e.g. through autoregressive modeling

Result: Distilled web-scale world knowledge into a base model

\$\$\$

Pre-training data & tokenization

Task: Collect large-scale datasets of broad world knowledge and tokenize

Result: Large-scale pre-training corpus

\$



Data & Tokenization

Overview

Goals

- Base for the kind of world knowledge we want models to have
- Should be easy to collect or generate in a scalable manner

How to collect?

- Often scraped from the entire internet: Web pages, books, papers, images, videos, etc... → TBs of text data, PBs of image and video data
 - Mixed quality and large diversity
- Synthetically generated using existing models
 - High quality, but less diversity

EPFL Pre-training data

Text-only datasets

- **Broad world knowledge:** Everything ever written on the internet, books, papers, ...
- **Large-scale:** Trillions of tokens, hundreds of TB (unfiltered)
- **Mixed quality:** Some are high quality (e.g. Wikipedia, Books, Github, ...), some very low quality (e.g. scraped websites)

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

[LLaMA: Open and Efficient Foundation Language Models, Meta 2023]

Source	Type	Tokens	Words	Bytes	Docs
Pretraining + OLMo 2 1124 Mix					
DCLM-Baseline	Web pages	3.71T	3.32T	21.32T	2.95B
StarCoder filtered version from OLMoE Mix	Code	83.0B	70.0B	459B	78.7M
peS2o from Dolma 1.7	Academic papers	58.6B	51.1B	413B	38.8M
arXiv	STEM papers	20.8B	19.3B	77.2B	3.95M
OpenWebMath	Math web pages	12.2B	11.1B	47.2B	2.89M
Algebraic Stack	Math proofs code	11.8B	10.8B	44.0B	2.83M
Wikipedia & Wikibooks from Dolma 1.7	Encyclopedic	3.7B	3.16B	16.2B	6.17M
Total		3.90T	3.48T	22.38T	3.08B

[2 OLMo 2 Furious, Team OLMo 2025]

Text-only datasets: Fineweb

Random samples

spotlight provides a convenient rechargeable LED light for work play and everyday life. choose from many vibrant colors to match your car, home, or personal style.

- high power 0.5 watt LED bulb (35+ lumens)
- colorful anodized aluminum body
- 180+ minutes of light per charge
- water resistant / submersible
- red glow "charging" indicator
- rechargeable Ni - MH battery
- shines 50 meter / 150 feet

San Francisco 49ers cornerback Shawntae Spencer will miss the rest of the season with a torn ligament in his left knee.

Spencer, a fifth-year pro, will be placed on injured reserve soon after undergoing surgery Wednesday to repair the ligament. He injured his knee late in the 49ers' road victory at Seattle on Sept. 14, and missed last week's victory over Detroit.

Tarell Brown and Donald Strickland will compete to replace Spencer with the 49ers, who kept 12 defensive backs on their 53-man roster to start the season. Brown, a second-year pro, got his first career interception last weekend while filling in for Strickland, who also sat out with a knee injury.

personalized baby Gifts I site map I personalized name trains I new affiliates I privacy I personalized children's music I personalized children's books I personalized children's clocks I personalized lovies I personalized baby's first christmas gifts I personalized first birthday gifts I Children's Valentine's Day Gifts I Children's Easter Gifts I Kids Easter Gifts I Easter Baskets I Easter Bunny I Baby Bibs I Comfy Cozy Baby Gund I Get Well Gifts for Kids I Sesame Street Characters I Elmo Dolls I Sesame Street Dolls I Sesame Street Gifts I Sesame Street Elmo I Sesame Street Big Bird Sesame Street Cookie Monster I Personalized Kids Music I Easter Baskets for Infants I Easter Baskets for Kids I Kids Christmas Gifts I Children's Christmas Gifts I Unique Baby Blankets I Personalized Children's Books I Baby Christmas Baskets I Christmas Baby Gifts I Boston Red Sox Baby Gifts I 2Blockheads.com Personalized Children's Gifts "Where Kids are Stars"

1786 St. Peters Road
Pottstown, PA 19465
Phone: 484 824-8500

Hours of Operation: Monday - Friday 8AM-5PM EST
Not an affiliated company of Gund, Inc. or Sesame Workshop. The representations made on this website are those of 2Blockheads Baby Store. Gund Images © Gund, Inc. Gund®, babyGund® and Gotta Getta Gund® are trademarks of Gund, Inc. Sesame Workshop, Sesame Street are owned and licensed by Sesame Workshop. Copyright Sesame Workshop. All Rights Reserved.

It's be kind of a rough week photographically. My TS-E 24mm f/3.5L ii broke on the outing where I made this image (the shift locking knob fell right off) and someone stole my crampons when I left the outside this cave.

The very next day, I went on a great hike across the Mt Juneau ridge carrying a camera body, three lenses, and a tripod. What I wasn't carrying was a memory card of any type (iphone photos only on that trip...despite the sore shoulders). Still, I laughed out loud when I saw the LCD on my camera read "No CF Card".

Car Wash For Clara!

Now is your chance to help! 2 year old Clara Woodward has Cancer! Clara can't say "Neuroblastoma" but she knows how it feels. You can help!!

A Car Wash will be held Saturday July 23, 11am-2pm at Java Jet on the corner of Edison & Canal Drive in Kennewick.

There is also an account set up in Clara's name. Her family lives in Pasco and is travelling to Spokane for treatment. For further information contact" Kelly Gammon at 509-380-2321

Text-only datasets: Fineweb-edu

Random samples

Discover the cosmos! Each day a different image or photograph of our fascinating universe is featured, along with a brief explanation written by a professional astronomer. 2010 August 12 Explanation:

Each August, as planet Earth swings through dust trailing along the orbit of periodic comet Swift-Tuttle, skygazers can enjoy the Perseid Meteor Shower. The shower should build to its peak now, best seen from later tonight after moonset, until dawn tomorrow morning when Earth moves through the denser part of the wide dust trail. But shower meteors have been spotted for many days, like this bright Perseid streaking through skies near Lake Balaton, Hungary on...

Coyotes spend a good deal of their day sleeping. Members of a pack or family may sleep within close proximity of each other, or they may sleep much further apart, but probably within the same couple of acres of each other. They have amazing built-in time clocks, but they also are influenced by circumstances of the moment. My own dog could tell the time and knew what was to be done at that time. For example, I always set off, with my dog, at exactly 2:40 to pick up one of my kids at school. But one day I fell asleep — I would not have made it on time except that my dog began poking me with her muzzle at exactly 2:40. Needless to say, I was amazed. The same is true for coyotes — they seem to know when it is time to meet up, but if people or dogs are around, they will delay.

Mexican America - Introduction

"Mexican America" is a sampling of objects from the collections of the National Museum of American History. The stories behind these objects reflect the history of the Mexican presence in the United States. They illustrate a fundamentally American story about the centuries-old encounter between distinct (yet sometimes overlapping) communities that have coexisted but also clashed over land, culture, and livelihood.

Who, where, and what is Mexico? Over time, the definitions and boundaries of Mexico have changed. The Aztec Empire and the area where Náhuatl was spoken—today the region surrounding modern Mexico City—was known as Mexico. For 300 years, the Spanish colonizers renamed it New Spain.

When Mexico was reborn in 1821 as a sovereign nation, its borders stretched from California to Guatemala. It was a huge and ancient land of ethnically, linguistically, and economically diverse regions that struggled for national unity. Texas, (then part of the Mexican state of Coahuila y Tejas) was a frontier region far from the dense cities and fertile valleys of central Mexico, a place where immigrants were recruited from the United States. The immigrants in turn declared the Mexican territory an independent republic in 1836 (later a U.S. state), making the state the first cauldron of Mexican American culture. By 1853, the government of Mexico, the weaker neighbor of an expansionist United States, had lost what are today the states of California, Nevada, Utah, Arizona, New Mexico, Texas, and parts of Colorado and Wyoming. In spite of the imposition of a new border, the historical and living presence of Spaniards, Mexicans, indigenous peoples, and their mixed descendants remained a defining force in the creation of the American West.

"La América Mexicana" es una muestra conformada por objetos provenientes de las distintas colecciones del Museo Nacional de Historia Americana. Estos objetos reflejan la historia de la presencia mexicana en los Estados Unidos e ilustran una crónica fundamentalmente americana acerca del encuentro centenario entre comunidades diferentes que han coexistido, pero que también se han enfrentado, en la pugna por la tierra, la cultura y el sustento.

¿Quién, dónde y qué es México? Con el transcurso del tiempo, las definiciones y los límites de México han ido cambiando. Se conocía como México al Imperio Azteca y toda el área donde se hablaba náhuatl —actualmente la región circundante a la ciudad de México. Durante 300 años los colonizadores españoles se refirieron a ella como Nueva España. Cuando en 1821 México resurgió como una nación soberana, sus fronteras se extendían desde California a Guatemala. En ese entonces era un antiguo e inmenso territorio conformado por regiones étnica, lingüística y económicamente diversas que luchaban por adquirir unidad nacional. Texas (en ese entonces parte de los estados mexicanos de Coahuila y Tejas) era una región fronteriza lejos de las densas urbes y de los fértiles valles de México central, donde se reclutaban inmigrantes de los Estados Unidos. ...

Text-only datasets: Fineweb → Fineweb-edu

"Junk" data significantly harm LLM's knowledge capacity on good data (sometimes by 20x times!)

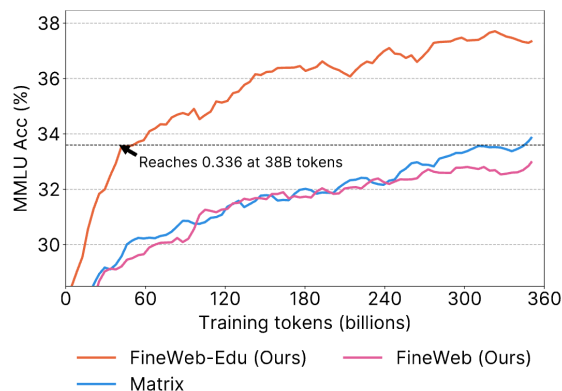


Figure 11: Performance Comparison on MMLU. FineWeb-Edu achieves a 33.6% accuracy on the MMLU benchmark at only 38 billion tokens, significantly outperforming Matrix (second best on the metric), which reaches similar accuracy at 300 billion tokens.

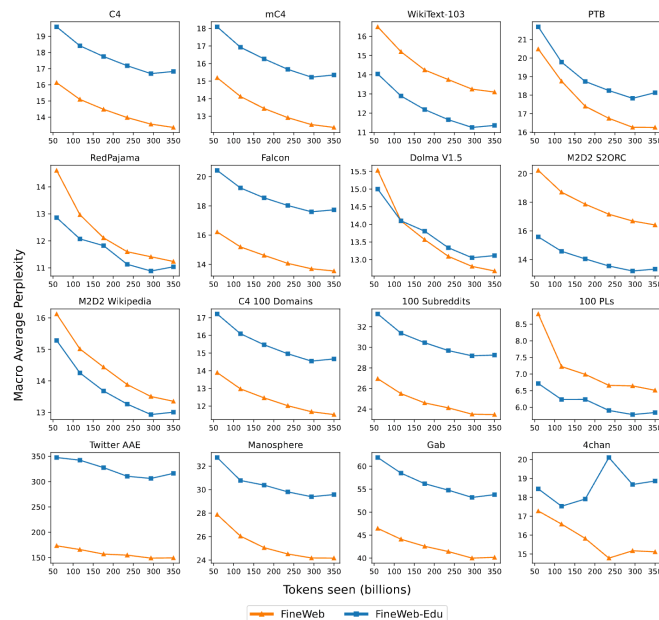


Figure 12: FineWeb and FineWeb-Edu fit to Paloma domains. FineWeb has lower perplexity on broad web sources while FineWeb-Edu has better coverage of Wikipedia and programming content.

Text-only datasets: Filtering

300B documents

370TB

240T tokens

4T tokens

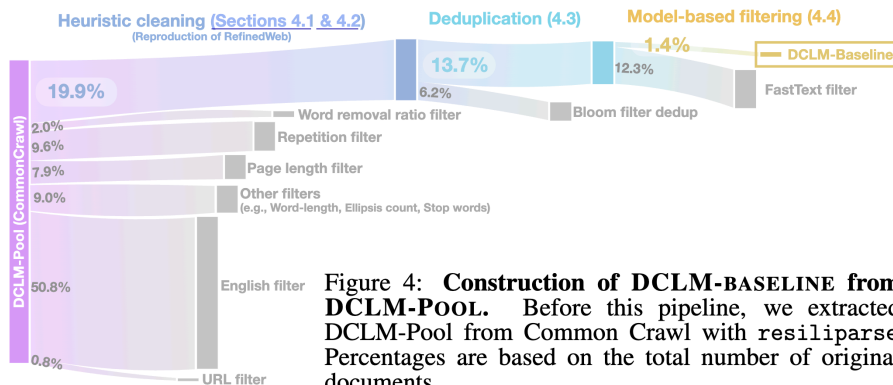


Figure 4: Construction of DCLM-BASELINE from DCLM-POOL. Before this pipeline, we extracted DCLM-Pool from Common Crawl with resiliparse. Percentages are based on the total number of original documents.

EPFL Pre-training data

Text-only datasets: Filtering

300B documents

370TB

240T tokens

4T tokens

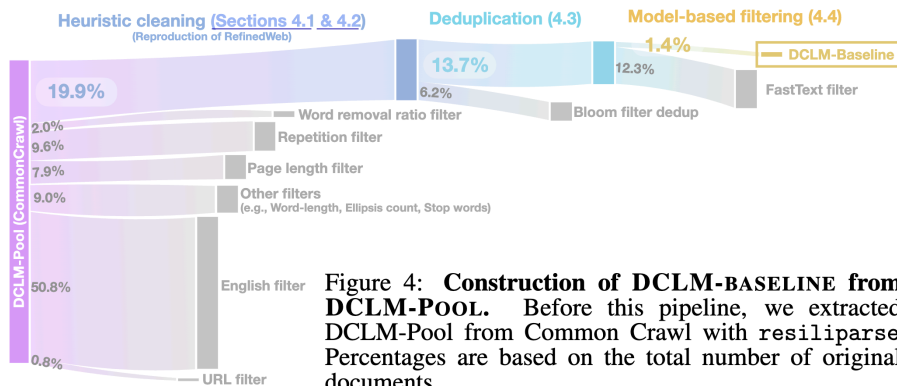


Figure 4: **Construction of DCLM-BASELINE from DCLM-POOL.** Before this pipeline, we extracted DCLM-Pool from Common Crawl with resiliiparse. Percentages are based on the total number of original documents.

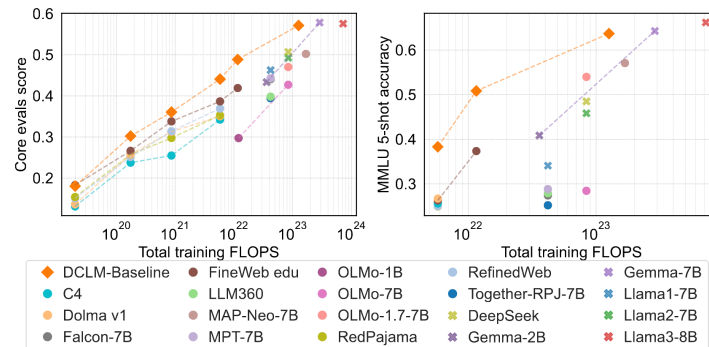
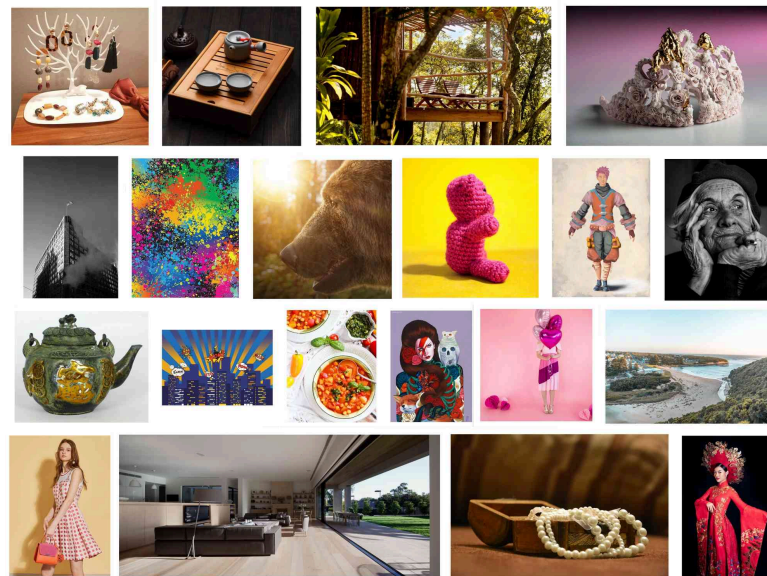


Figure 1: **Improving training sets leads to better models that are cheaper to train.** Using DataComp-LM, we develop a high-quality dataset, DCLM-BASELINE, which we use to train models with state-of-the-art trade-off between compute and performance. We compare on both (left) a CORE set of tasks and on (right) MMLU 5-shot. Specifically DCLM-BASELINE (orange) shows favorable performance relative to both close-source models (crosses) and other open-source datasets and models (circles). Models in this figure are from [\[4, 10, 22, 43, 68, 97, 100, 121, 130, 150, 154, 156, 160–162, 189\]](#).

Image-caption datasets

- **Broad world knowledge:**
Web-scale scraped images + alt or nearby text
- **Large-scale:** 10B+ images, several PB
- **Mixed quality:** Some are high quality (e.g. Stock images, art, ...), some very low quality (e.g. ads, irrelevant alt text, watermarks, ...)
- **Poor alignment:** Often alt text captures context instead of describing image content
- **Problematic content:** Consent & copy-right, PII, explicit content, ...



[<https://laion.ai/blog/laion-pop/>]

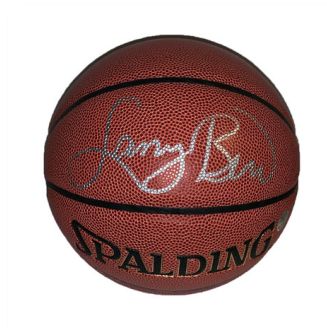
Image-caption datasets



*Rhododendron 'Princess Anne';
Dwarf rhododendron (Oslash; 17cm pot)*



*Michelle & Karl at The Granary Barns,
Newmarket 66*



*Larry Bird // Boston Celtics // Signed
Basketball*

Image-caption datasets



*Rhododendron 'Princess Anne';
Dwerghododendron (Ø 17cm pot)*



Re-captioning

*A compact shrub with clusters of soft yellow, trumpet-shaped flowers and glossy dark green leaves, blooming in abundance—
Rhododendron 'Princess Anne'.*

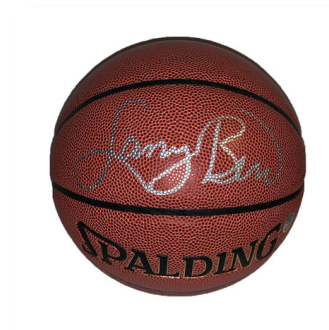


*Michelle & Karl at The Granary Barns,
Newmarket 66*



Re-captioning

A group of people gathers in a beautifully maintained courtyard between rustic stone and brick buildings, one with a red-tiled roof and large windows. The scene is bright and airy, suggesting a casual outdoor event or celebration. Children play on the grass while adults chat near the building entrance, surrounded by manicured gardens and shrubs. The overall atmosphere is relaxed and cheerful, set against the charm of a countryside venue.



*Larry Bird // Boston Celtics // Signed
Basketball*



Re-captioning

A Spalding basketball featuring a prominent silver autograph of former NBA player Larry Bird. The signature is clearly visible on the textured surface of the ball, positioned between the black seams.

Examples of MINT Multimodal Documents

[illegible]

Variazione degli stock di debito (1999-2007, punti di PIL)


Paese	debito pubblico (punti di PIL)	debito privato (punti di PIL)
France	0	25
Germany	-5	-10
Greece	-5	55
Ireland	-15	100
Italy	-10	35
Portugal	20	55
Spain	-15	100

A diagram showing a square domain with a central circular hole. The domain is divided into two regions: the outer region is labeled ϵ_1 and the inner region (the hole) is labeled ϵ_2 .


CS-503 – Visual Intelligence: Machines and Minds

Multimodal datasets: interleaved text & image

document text



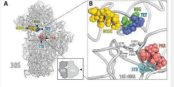
It appears that, while I was otherwise engaged watching Forest bore Burnley into a late submission last night, it seems there was an altogether more entertaining game cracking off at Old Trafford. Thanks to goals from Javier Hernandez and Ji-Sung Park, Manchester United emerged from their Champions League quarter-final fixture against Chelsea with a 2-1 victory on the night (and a 3-1 aggregate victory overall, with Ryan Giggs directly assisting all three of United's goals over both legs) and never once, by all accounts, looked like being caught. It also seems that Chelsea manager Carlo Ancelotti's decision to gamble on handing Fernando Torres a start ahead of Didier Drogba backfired in a rather mighty way, with the Spaniard forcibly yanked from the stage at half-time after he endured another 45 minutes of something close to nothing. When asked in a post-match presser whether he had made a mistake by plumping to start with Torres, Ancelotti said: "Maybe. Could be. I told you a lot of times this season I wanted to start with Fernando for this kind of game, these type of tactics. I wanted to put more pressure up front because we needed to score. Didier was fresh and he could use his power up front. This was the reason I took out Fernando." And, just maybe, because he was a complete non-entity... again – though, to be fair, it's not like Nicolas Anelka, Salomon Kalou, Frank Lampard et al had sterling games either. So that's the long and short of it peeps, United are through to the last four of the Champions League with their treble dream still firmly intact, while Chelsea are left to mope home with their season rendered almost completely null and void – Mr Abramovich will be pleased.



Fernando Torres and Wayne Rooney shake hands before the game

document text


By Jacqueline Carey, University of Illinois at Chicago April 6, 2018




ODLs (yellow) bind to a site on the ribosome not used by other antibiotics. Location of this site is shown here relative to the sites of other known antibiotics, such as neomycin (green), tetracycline (dark blue), aminoglycoside antibiotic paromomycin (red) and streptomycin (light blue). UC/Yury Polikanov, et al Researchers from the University of Illinois at Chicago and Nosopharm, a biotechnology company based in Lyon, France, are part of an international team reporting on the discovery of a new class of antibiotics. The antibiotic, first identified by Nosopharm, is unique and promising on two fronts: its unconventional source and its distinct way of killing bacteria, both of which suggest the compound may be effective at treating drug-resistant or hard-to-treat bacterial infections. Called obelisks (ODLs), the antibiotics are produced by symbiotic bacteria found in soil-dwelling nematode worms that colonize insects for food. The bacteria help to kill the insect and, importantly, secrete the antibiotic to keep competing bacteria away. Until now, these nematode-associated bacteria and the antibiotics they make have been largely understudied. To identify the antibiotic, the Nosopharm research team screened 80 cultured strains of the bacteria for antimicrobial activity. They then isolated the active compounds, studied their chemical structures and engineered more potent derivatives. The study, published in Molecular Cell, describes the new antibiotic and, for the first time, how it works. UIC's Alexander Mankin and Yury Polikanov are corresponding authors on the study and led the research on the antibiotic's mechanism of action. They found that ODLs act on the ribosome — the molecular machine of individual cells that makes the proteins it needs to function — of bacterial cells. "Like many clinically useful antibiotics, ODLs work by targeting the ribosome," said Polikanov, assistant professor of biological sciences in the UIC College of Liberal Arts and Sciences, "but ODLs are unique because they bind to a place on the ribosome that has never been used by other known antibiotics." The UIC researchers, including graduate student Tanja Florin and postdoctoral research associate Magdalena Dobosz-Bartoszek, also found that when bound to the ribosome, the antibiotic disrupts its ability to interpret and translate genetic code. "When ODLs are introduced to the bacterial cells, they impact the reading ability of the ribosome and cause the ribosome to make mistakes when it creates new proteins," said Mankin, director of the Center for Biomolecular Sciences in the UIC College of

document text

Hoshigaki, or Japanese dried persimmons as we know them in English, are a popular seasonal Japanese snack food made primarily during the autumn months when you can spot persimmon trees laden with their brightly colored orange fruits throughout the country and in supermarkets. Originally from China, persimmons, known as kaki in Japanese, are commonly found in Japan and Korea and make for a sweet, natural snack in their dried form. In Japan, hoshigaki are said to bring good luck and are often eaten on New Year's Day in a custom known as Hagatame in Japanese. How to make hoshigaki To make hoshigaki, the dehydrated persimmons are massaged by hand in a labor intense and lengthy process that takes between 4-6 weeks before they are ready to eat as a snack. Although there are hundreds of varieties of persimmons, there are two main categories, astringent and non-astringent, both of which can have quite different tastes.




For making hoshigaki, the astringent type of persimmon or hachisu are used. First the tops are cut off the persimmons while still leaving the stem in place, before the skin is peeled. The persimmons are then laid together by the stem with wings and hung in sunny windows. In Japan you can often see them hanging outdoors from bamboo poles during the winter months, particularly in rural areas.




The persimmons must be left undisturbed for a week before the daily massaging process begins. This massaging is a delicate process that begins as a light touch before progressing to more of a squeeze. The hoshigaki is ready to eat after a few weeks once a layer of powdery white sugar has formed on the surface and the persimmon becomes a burnt orange color. Once ready to eat, the hoshigaki can be stored for approximately several months in an airtight container or frozen. Where does Japanese persimmon fruit grow in Japan? Although persimmon trees are found throughout Japan, there are several regions that specialize in persimmon or kaki production. For example, prefectures to include Yamagata, Niigata, and Nara are all major persimmon growing regions. Other popular Japanese persimmon varieties are

document text



The templates additionally contain charts and graphs to make your initiation use useful and fascinating. Arch of the normal templates effectively accessible online are cardiology How To Change PowerPoint Template, dental templates, helps PowerPoint foundations, illness PowerPoint PPT, cerebrum PowerPoint launch and therefore upon you can even have the exchange of getting a template redid by the creators, attempt not to make more noticeable upon the off inadvertent that you are not finding the template, helpfully email them and acquire the ideal template. There are enlivened medicinal templates which looks remarkable. These templates contain enlivened impacts which stir the inauguration and make it worth viewing. Crowd are forever keen upon searching for the things which draws in initially. Energized templates are long-suffering gone you create an commencement which is displayed in the stomach of the outfit of spectators. The ideas become all the more effectively reasonable as soon as the alluring PowerPoint slides. Likewise it turns out to be anything but hard to give access and flavor considerations to the action of spectators. Search on the web and you will acquire vary rumored and prestigious sites which has an incredible assortment of medicinal How To Change PowerPoint Template.

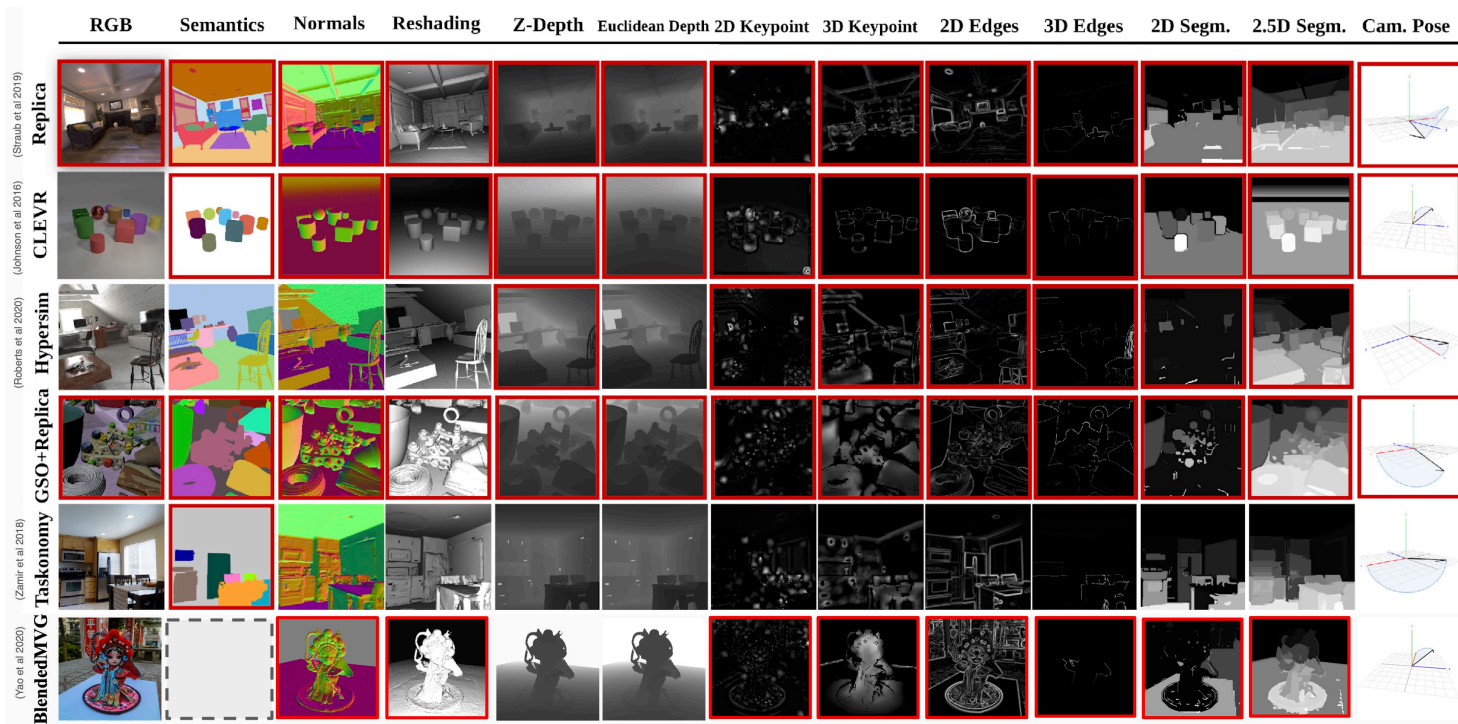


The best fragment of the How To Change PowerPoint Template is this that they can be altered. You can without much of a stretch create changes in the shading, text styles and pictures. Indeed you can be credited with various pictures as indicated by the topic in the slides you need. A ration of the templates are even planned particularly for the meetings and classes. They are alluring and eye appealing. PowerPoint is a unique creation instrument from Microsoft, it is the most generally utilized creation inauguration programming friendly today. It is utilized by experts, scholastics, understudies and others to feature thoughts and do its stuff data in a unique organization. A decent atmosphere How To Change PowerPoint Template is the start for each introduction, consequently it's valuable to hit the nail on the head. It might in some cases be standard to utilize one of the innumerable PowerPoint subjects or a free template, however upon the off inadvertent that you obsession to swell the effect of your Presentation and observation

document text

2014 Topps Archives Baseball just got a little bit wilder. Charlie Sheen has been confirmed as a Major League Autographs signer. Sheen played Ricky "Wild Thing" Vaughn in the classic baseball comedy. The movie turns 25 this year and Topps is including both signed and unsigned cards from the film in the retro-themed product. Based on the 1989 Topps Baseball design, the card includes a shot from the film as well as the iconic mohawk-wearing baseball logo. This doesn't appear to be all for Sheen and Major League as far as 2014 Topps products go. They also plan to have swatches from a Sheen-worn Indians jersey (complete with Vaughn nameplate) in 2014 Topps Triple Threads; 2014 Topps Five Star and 2014 Topps Dynasty. Sheen, a noted baseball fan and memorabilia collector, has had autographs in a handful of products prior to this. 2011 Leaf Legends of Sport includes an autograph of the actor with artwork based on Major League, but it's not licensed by Paramount, the company that made the movie. 2014 Topps Archives Baseball is slated for release in May.

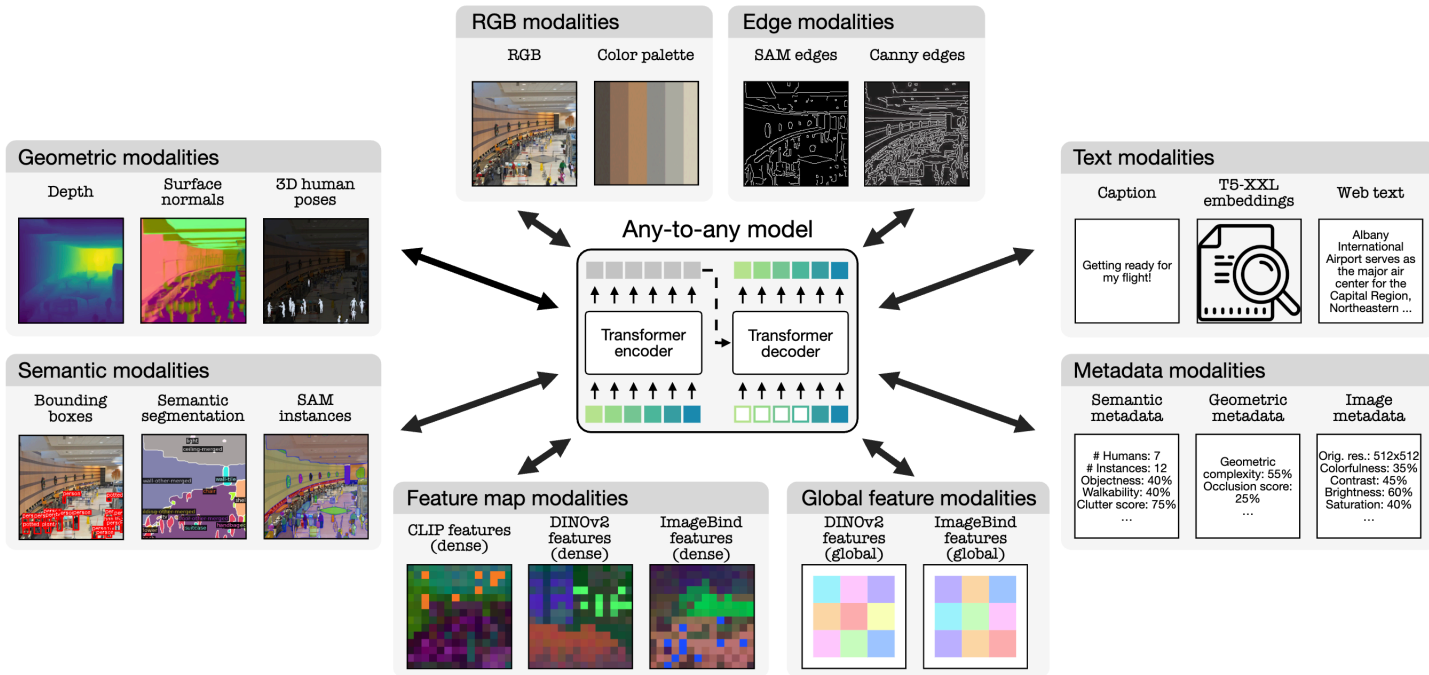
Multimodal datasets: Massively multimodal datasets



[OmniData: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets from 3D Scans, Eftekhar et al. 2021]

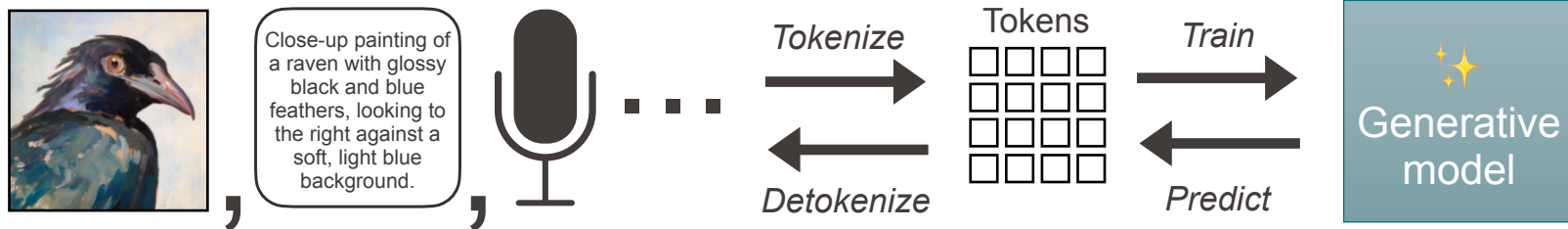
[Taskonomy: Disentangling Task Transfer Learning, Zamir et al. 2018]

Multimodal datasets: Massively multimodal datasets



Overview: Goals of tokenization

- Semantic compression (e.g. text is easier to model than audio)
- Reduced sequence length (e.g. millions of pixels → thousands of tokens)
- Regularized latent space (e.g. through discrete or soft regularizers)
- Unification of different modalities (images, text, audio, ... → tokens)



Overview

- Maps text to discrete tokens (indices)
- Typical vocabulary sizes: 10k - 200k+

"Tokenization is the process of breaking down text into smaller units called tokens, which can be words, phrases, or even individual characters. This is a fundamental step in natural language processing (NLP) and text analysis, as it helps computers understand and analyze human language by simplifying the text structure."



Tokenization is the process of breaking down text into smaller units called tokens, which can be words, phrases, or even individual characters. This is a fundamental step in natural language processing (NLP) and text analysis, as it helps computers understand and analyze human language by simplifying the text structure.

=

[30642, 1634, 318, 262, 1429, 286, 7163, 866, 2420, 656, 4833, 4991, 1444, 16326, 11, 543, 460, 307, 2456, 11, 20144, 11, 393, 772, 1981, 3435, 13, 770, 318, 257, 7531, 2239, 287, 3288, 3303, 7587, 357, 45, 19930, 8, 290, 2420, 3781, 11, 355, 340, 5419, 9061, 1833, 290, 16602, 1692, 3303, 416, 7106, 4035, 262, 2420, 4645, 13]

Tokenization schemes



- **Character- and byte-level tokenization:**

- 1 character/byte = 1 token
- **+** : Simple and language agnostic, handles rare words and typos
- **-** : Long sequence lengths, less semantic

- **Word-level tokenization:**

- 1 word = 1 token
- **+** : Short sequences, interpretable/semantic
- **-** : Large vocabulary, fails on OOV words/typos, language-specific

- **Subword tokenization** (e.g. BPE, WordPiece, SentencePiece):

- Text split into subwords, i.e. 1 word can be 1 or more tokens
- **+** : Balanced approach, handles common parts of words (e.g. prefixes)
- **-** : Fragmented splits (e.g. "9.11 and 9.8" gets encoded into  and , more complex encoding and decoding process

EPFL Text tokenization

Byte Pair Encoding (BPE)

1. Take large corpus of text
2. Start with one token per character
3. Merge common pairs of tokens into a token
4. Repeat until desired vocabulary size or all merged

tokenizer: text to token index



tokenizer: text to token index



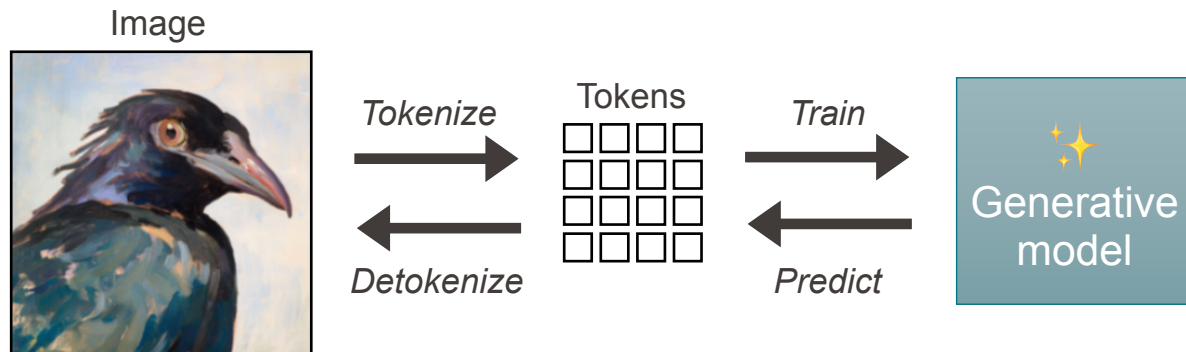
tokenizer: text to token index



tokenizer: text to token index

Image Tokenization

Goal: Project images into a sequence of tokens to model with a generative model



Goal: Project images into a sequence of tokens to model with a generative model

- **Why tokenize:**
 - Reduce sequence length
 - Abstract away imperceptible details (lossy compression)
- **Modeling-dependent properties of tokens:**
 - Regularized latent space
 - For autoregressive models: Provide a prediction target that can be sampled from
 - Semantic latent space
 - Ordering
 - ...

Sequence length reduction

- **Before tokenization:**

512 x 512 image: $512 \times 512 = 262'144$ tokens

- **After tokenization:**

Downsample with patch size 16 x 16: $(512/16) \times (512/16) = 1'024$ tokens

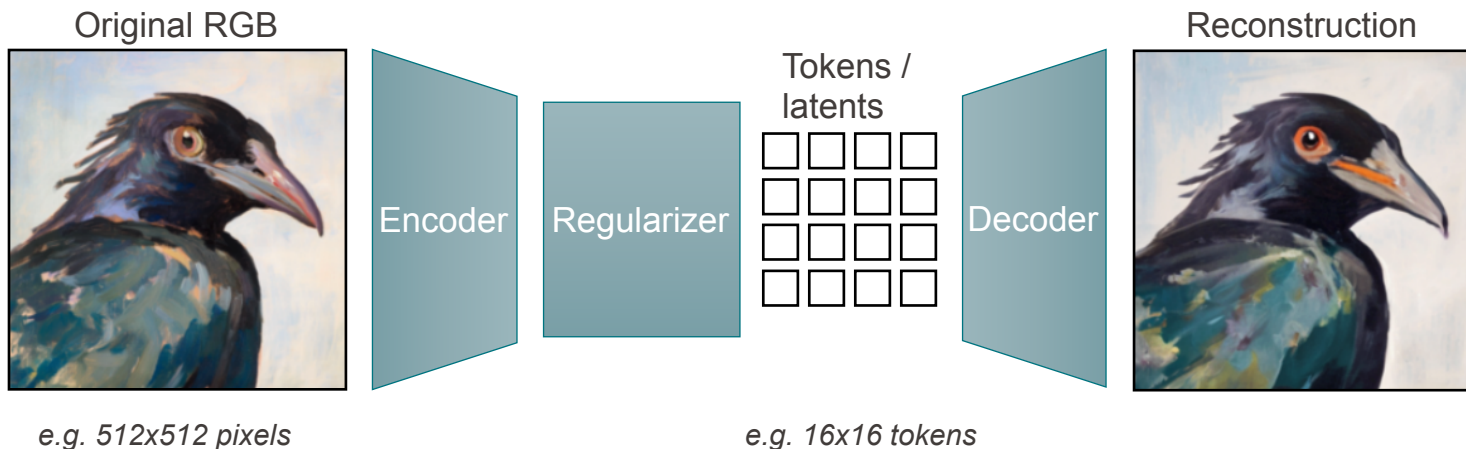
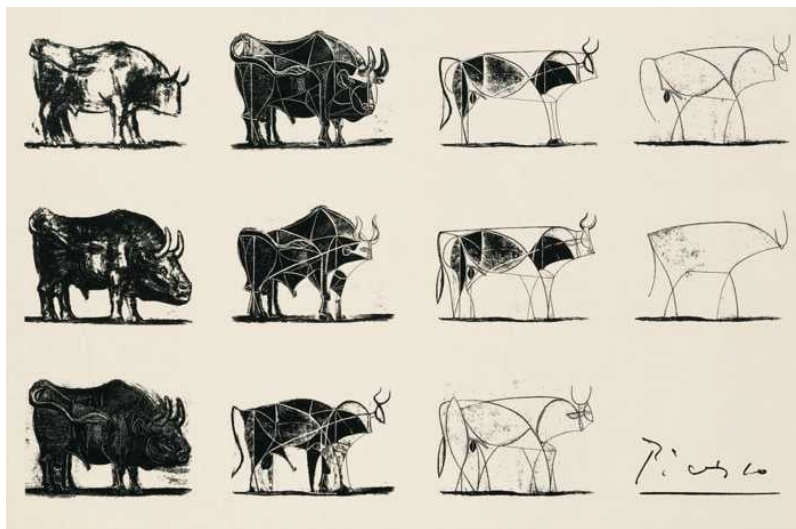


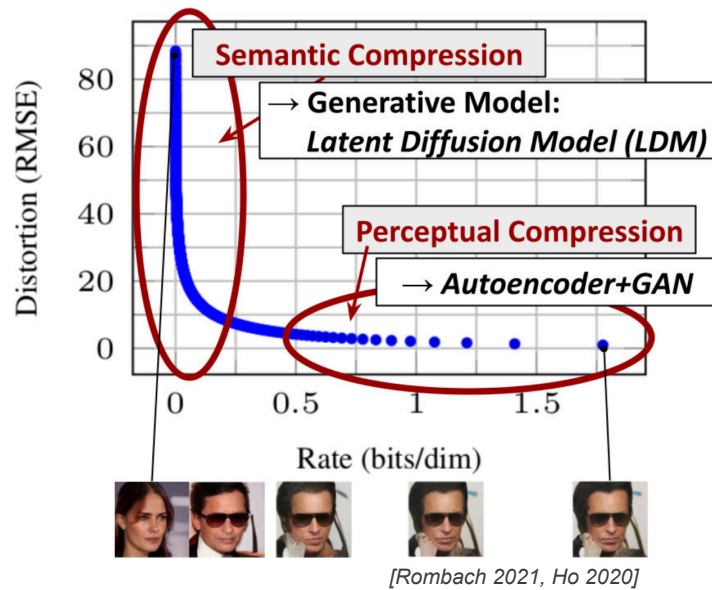
Image Tokenization

Abstract away fine-grained (imperceptible) details

- We want to spend model capacity to predict aspects that matter



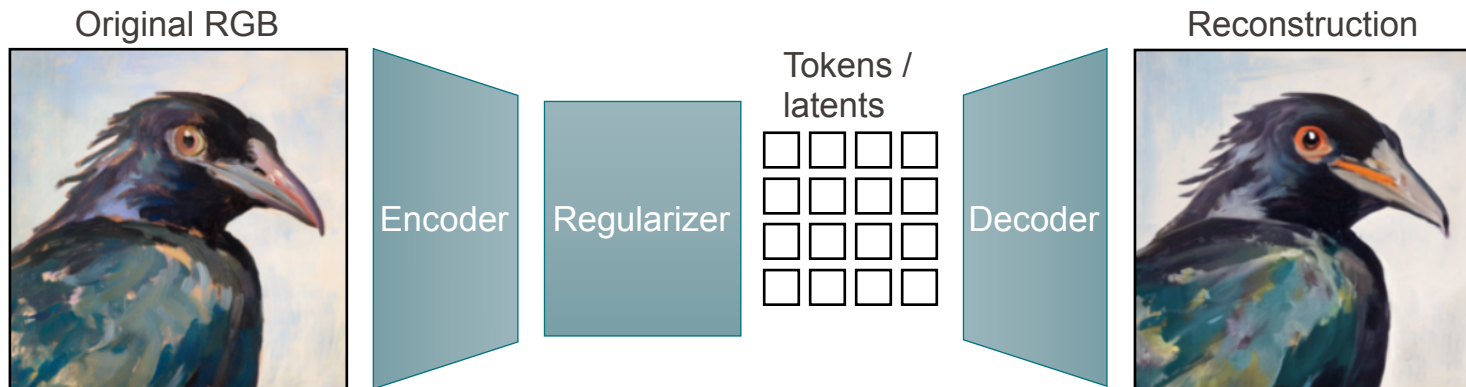
[The Bull, Pablo Picasso]



How to train an image tokenizer?

Overview

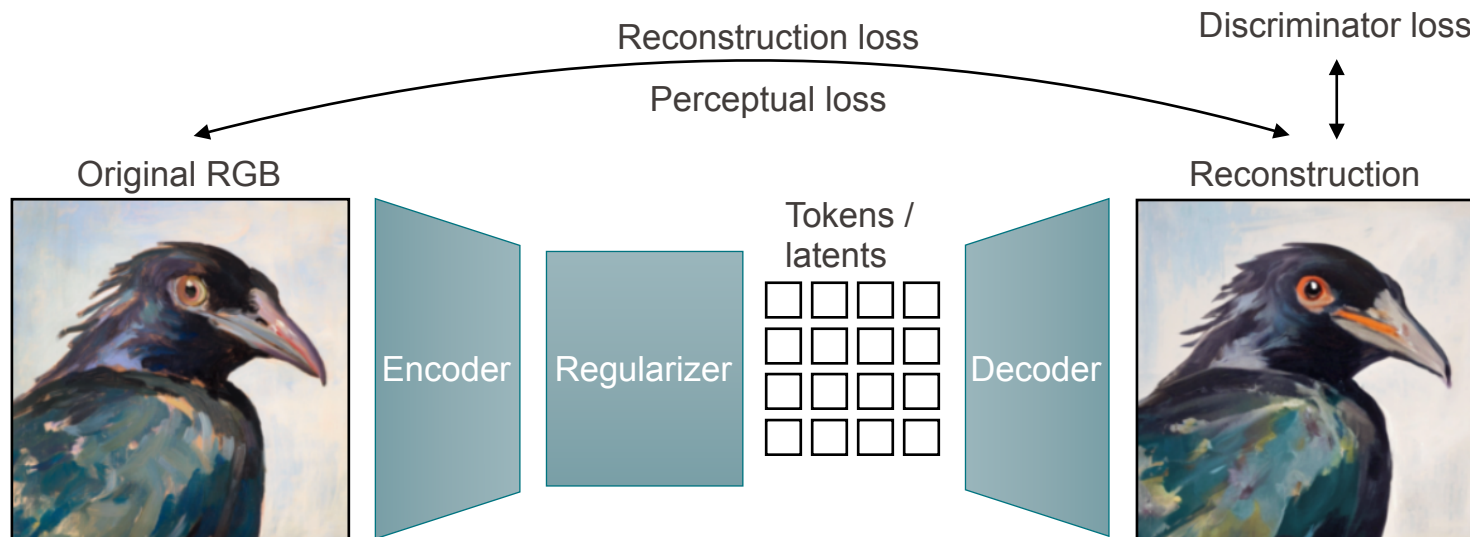
- **Architecture:** Bottleneck autoencoder
- **Bottleneck:** Discrete or continuous regularization
- **Objective:** Mostly autoencoding (reconstruction)



How to train an image tokenizer?

Objective

- **Main objective:** Autoencoding (i.e. reconstruction loss)
- **Auxiliary objectives:** Perceptual loss, Discriminator loss, etc...



EPFL How to train an image tokenizer?

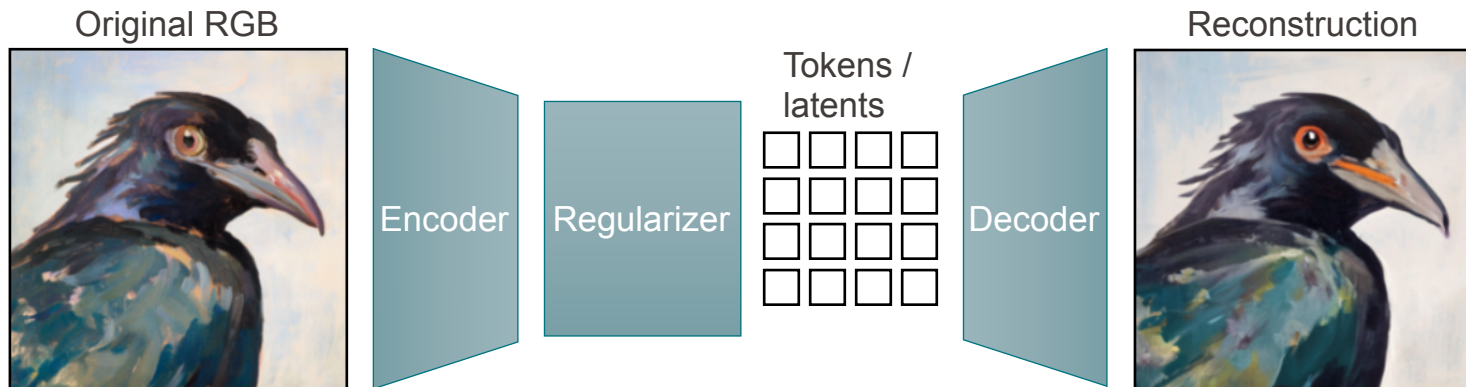
Regularizer / Bottleneck

■ Discrete:

- Each token can be one of K classes. Commonly $K = 4k, 16k, 64k, \dots$
- Train with a discrete bottleneck (e.g. FSQ, vector quantization, ...)

■ Continuous:

- Each token is a d -dimensional continuous latent. Commonly $d = 4, 8, 16, \dots$
- Train with KL-regularizer to keep latent space well-behaved





Pre-training

Overview

■ Goals

- Approximate the data distribution $p(x)$
- Extract broad world knowledge from a corpus and distill it into a base model

■ Objectives

- Predictive: "Corrupt the data and predict the original"
- **Next-token prediction:** Mask the next token and predict it
- **Masked modeling:** Mask a random set of tokens and predict them
- **Diffusion:** Noise the data and predict the noise / clean data / flow

Autoregressive / next-token prediction

- **Goal:** Model the joint distribution of the data
- **How:** Factorize using chain-rule and model through next-token prediction

$$\begin{aligned} p(x) &= p(x_1, x_2, \dots, x_L) \\ &= p(x_1)p(x_2|x_1)\dots p(x_L|x_1, \dots, x_{L-1}) \\ &= \prod_{i=1}^L p(x_i|x_{<i}) \end{aligned}$$

- **Pros:** Powerful objective and efficient through teacher forcing
- **Cons:** Inference is fixed-order and slow

Pre-training objectives

Autoregressive / next-token prediction

We model $\prod_{i=1}^L p(x_i | x_{<i})$ with a single autoregressive Transformer.

""	→	$p(x_1 "")$
"I"	→	$p(x_2 "I")$
"I love"	→	$p(x_3 "I love")$
"I love drinking"	→	$p(x_4 "I love drinking")$
"I love drinking iced"	→	$p(x_5 "I love drinking iced")$
	⋮	

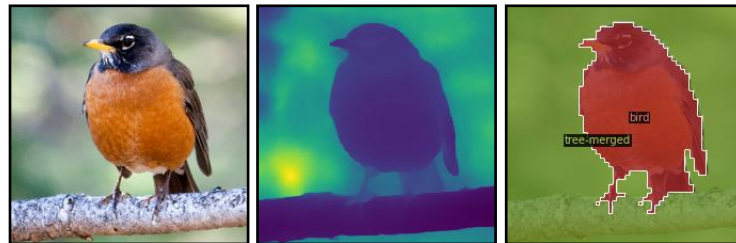
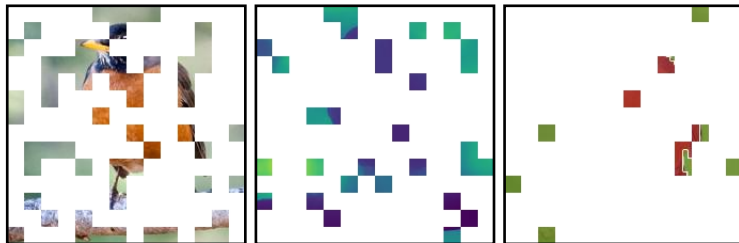
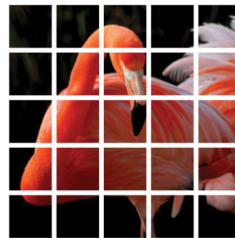
Pre-training objectives

Masked modeling ("discrete diffusion")

"I love drinking [MASK] tea
when [MASK] hot outside."



"iced" "it's"

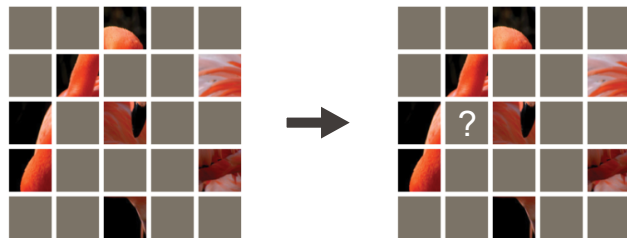


Masked modeling ("discrete diffusion")

- **Goal:** As in AR modeling, model the joint distribution of the data
- **How:** Factorize using chain-rule and model through **random** next-token prediction

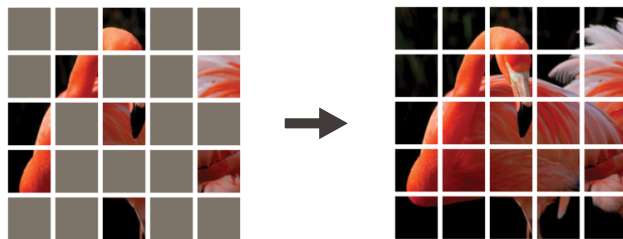
Let's look at this as AR modeling with a random order. I.e. define a permutation $\pi : \{1, 2, \dots, L\} \rightarrow \{1, 2, \dots, L\}$ of indices and factorize the probability:

$$\begin{aligned} p(x) &= p(x_1, x_2, \dots, x_L) \\ &= \prod_{i=1}^L p(x_{\pi(i)} | x_{<\pi(i)}) \end{aligned}$$



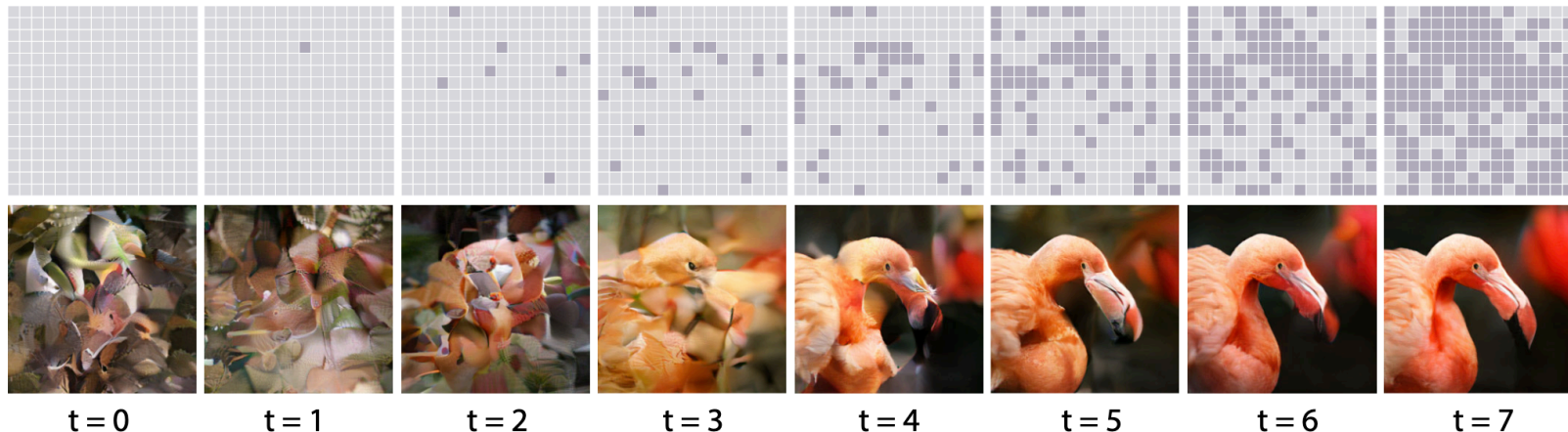
Masked modeling ("discrete diffusion"): Parallel decoding

- Tokens are conditionally dependent on each other, i.e. $p(x_i, x_j | c) = p(x_i | c)p(x_j | c, x_i)$
- **In practice:** Some token pairs are nearly conditionally independent (e.g. if they are far away), i.e. $p(x_i, x_j | c) \approx p(x_i | c)p(x_j | c)$
- **Consequence:** Some tokens can be predicted in parallel \rightarrow fewer decoding steps needed



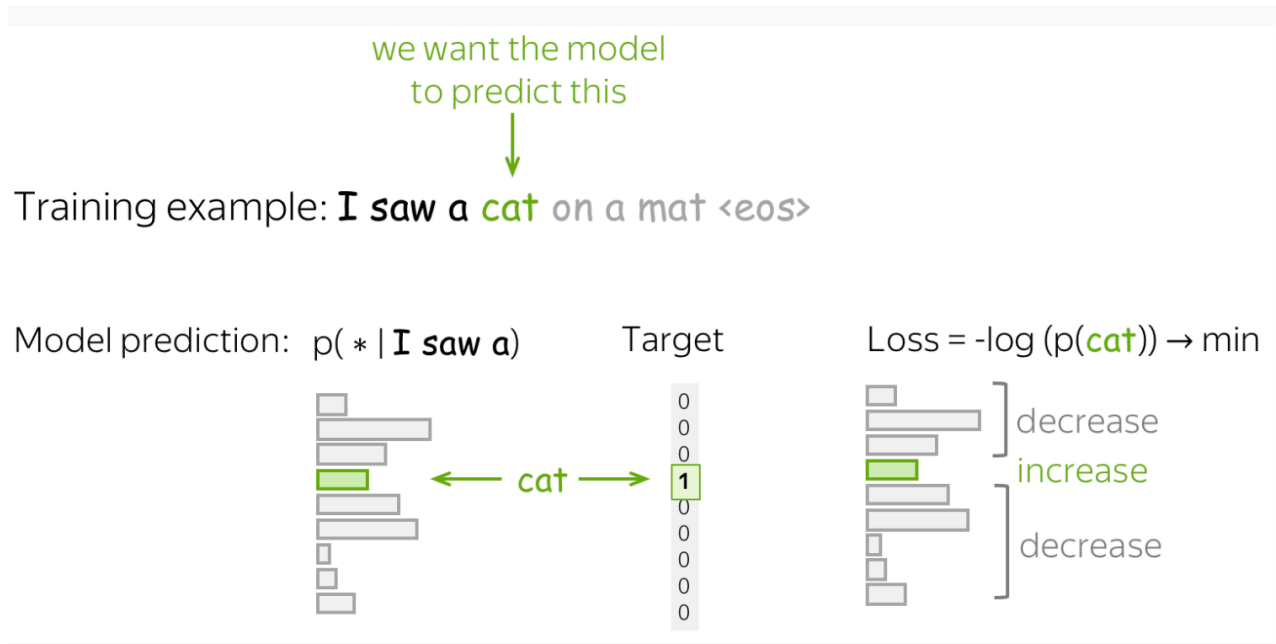
Masked modeling ("discrete diffusion"): Parallel decoding

- Tokens are conditionally dependent on each other, i.e. $p(x_i, x_j|c) = p(x_i|c)p(x_j|c, x_i)$
- **In practice:** Some token pairs are nearly conditionally independent (e.g. if they are far away), i.e. $p(x_i, x_j|c) \approx p(x_i|c)p(x_j|c)$
- **Consequence:** Some tokens can be predicted in parallel → fewer decoding steps needed



Training loss

Minimize cross entropy loss (= "next token classification")



Massive multitask learning

- Objective is simple, but data is rich
- Predicting masked data / next token = implicitly massively multitask learning

<i>Task</i>	<i>Example sentence in pre-training that would teach that task</i>
<i>Grammar</i>	In my free time, I like to {code, banana}
<i>Lexical semantics</i>	I went to the store to buy papaya, dragon fruit, and {durian, squirrel}
<i>World knowledge</i>	The capital of Azerbaijan is {Baku, London}
<i>Sentiment analysis</i>	Movie review: I was engaged and on the edge of my seat the whole time. The movie was {good, bad}
<i>Translation</i>	The word for “pretty” in Spanish is {bonita, hola}
<i>Spatial reasoning</i>	Iroh went into the kitchen to make tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the {kitchen, store}
<i>Math question</i>	Arithmetic exam answer key: $3 + 8 + 4 = \{15, 11\}$
[millions more] Extreme multi-task learning!	

$$\begin{aligned}
 \mathcal{L}_{\text{overall}} = & 10^{-3} \mathcal{L}_{\text{grammar}} \\
 & + 10^{-6} \mathcal{L}_{\text{sentiment}} \\
 & + 10^{-3} \mathcal{L}_{\text{knowledge}} \\
 & \dots \\
 & + 10^{-4} \mathcal{L}_{\text{math}}
 \end{aligned}$$

EPFL Examples: AR

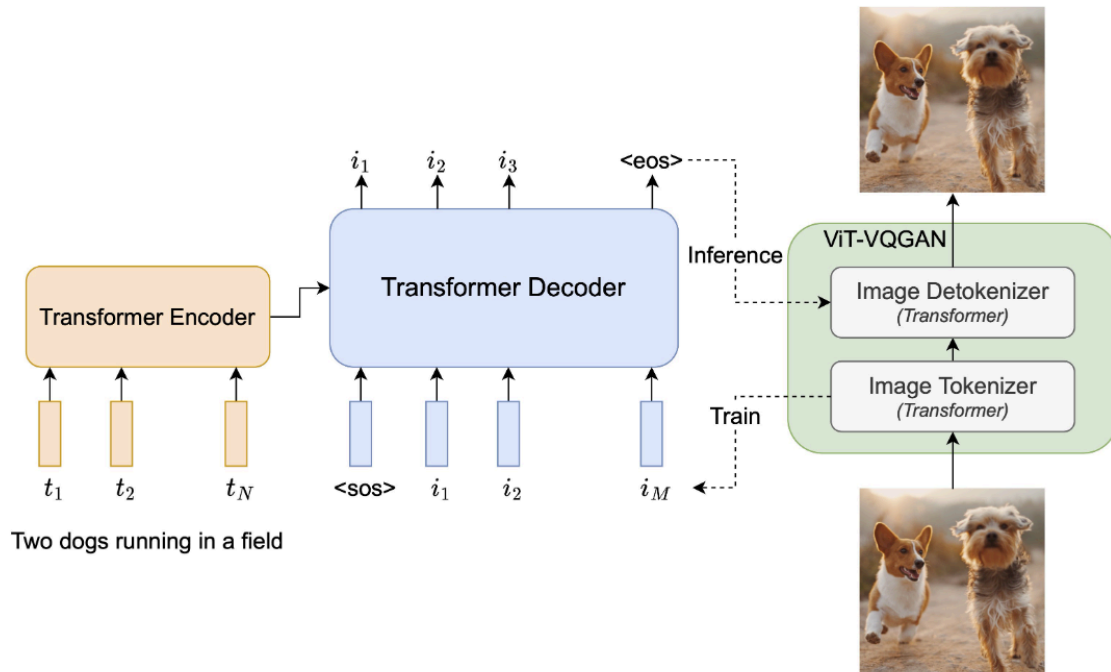
Example: Llama 3

- 3 model sizes: 8B, 70B, 405B
- Trained on 15T multilingual tokens (Llama 2 was 1.8T)
 - Chinchilla-optimal (~20 tokens per parameter) would be 160B, 1.4T, 8.1T tokens
- 405B model stats:
 - 3.8×10^{25} FLOPs (50x Llama 2)
 - Up to 16k H100s (700W TDP)
 - 38-43% BF16 model FLOPs utilization (MFU)

	8B	70B	405B
Layers	32	80	126
Model Dimension	4,096	8192	16,384
FFN Dimension	14,336	28,672	53,248
Attention Heads	32	64	128
Key/Value Heads	8	8	8
Peak Learning Rate	3×10^{-4}	1.5×10^{-4}	8×10^{-5}
Activation Function	SwiGLU		
Vocabulary Size	128,000		
Positional Embeddings	RoPE ($\theta = 500,000$)		

Example: Parti

Autoregressive modeling on image tokens



Example: Parti

Autoregressive modeling on image tokens

Parti-350M



Parti-750M



Parti-3B



Parti-20B



A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!



A green sign that says "Very Deep Learning" and is at the edge of the Grand Canyon. Puffy white clouds are in the sky.

Examples: AR

Example: LlamaGen

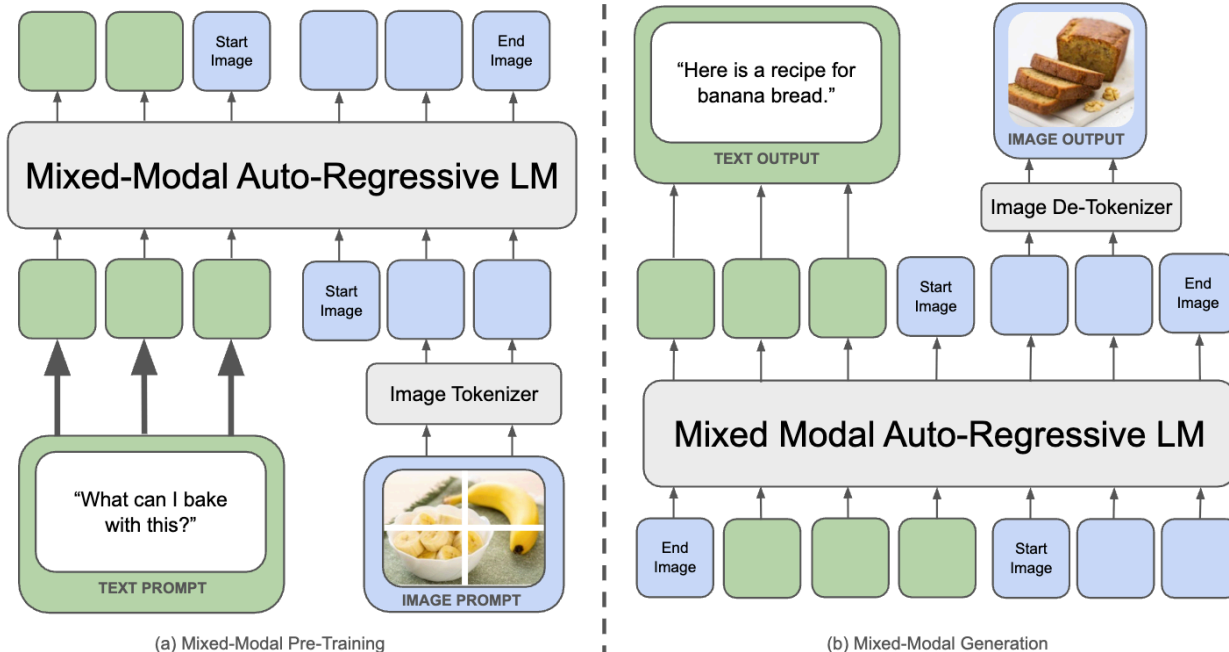
Autoregressive modeling on image tokens with GPT (Llama) architecture



Examples: AR

Example: Chameleon

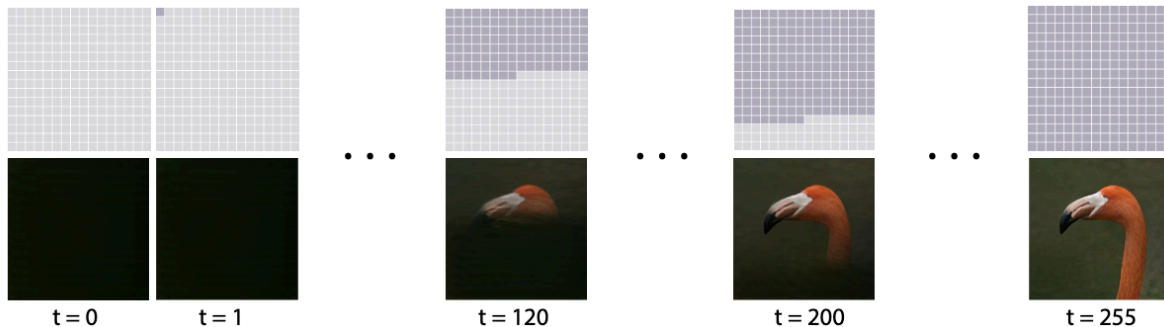
Autoregressive modeling on text + image tokens



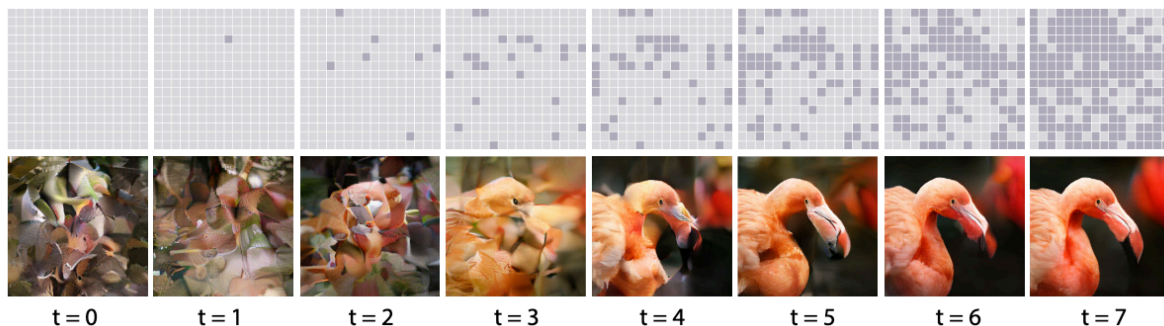
Example: MaskGIT

Masked modeling on image tokens

Sequential
Decoding
with Autoregressive
Transformers



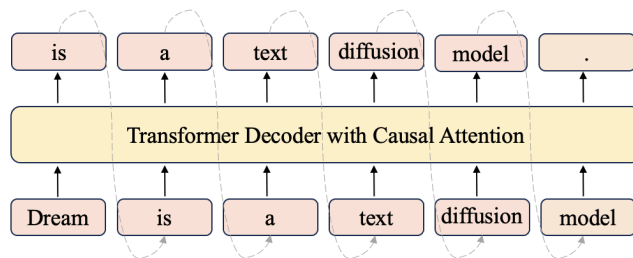
Scheduled
Parallel
Decoding
with MaskGIT



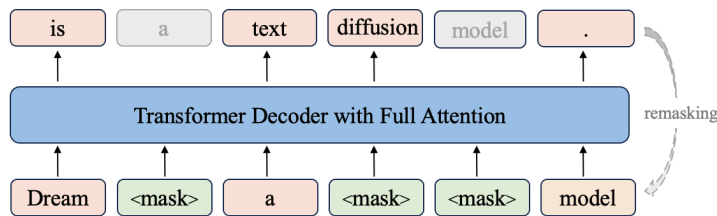
Examples: Masked modeling

Example: Dream

Masked modeling on text tokens



(a) Autoregressive Modeling



(b) Diffusion Modeling in Dream

Examples: Masked modeling

Example: Dream

Masked modeling on text tokens



Examples: Masked modeling

Example: Dream

Masked modeling on text tokens

Please write a Python class that implements a PyTorch trainer capable of training a model on a toy dataset.

Please write a Python class that implements a PyTorch trainer capable of training a model on a toy dataset.

Please write a Python class that implements a PyTorch trainer capable of training a model on a toy dataset.

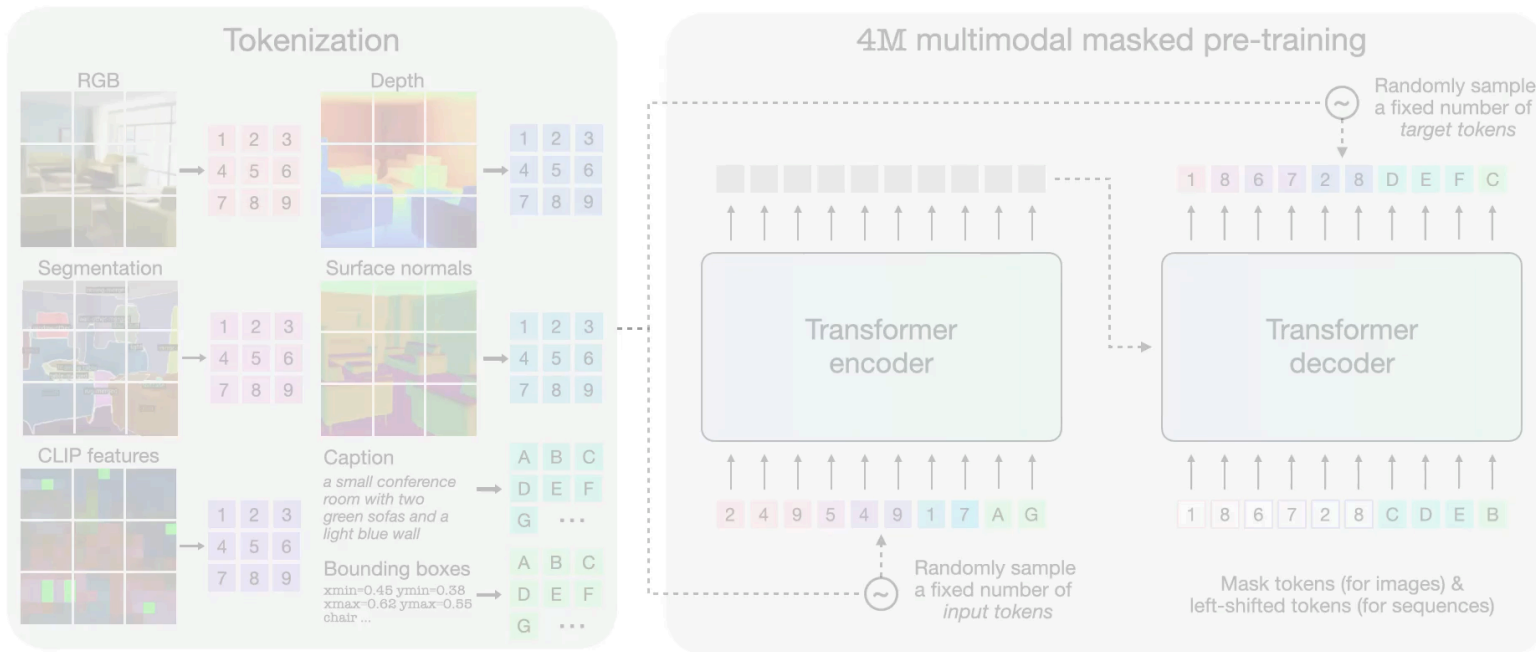
torch

the

Examples: Masked modeling

Example: 4M

Masked modeling on multimodal tokens



Examples: Masked modeling

Example: 4M

Masked modeling on multimodal tokens

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 8 9 10

Transformer
encoder

Transformer
decoder

In-context learning

- Perform a novel task from few demonstrations
- Instead of fine-tuning, provide task examples

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



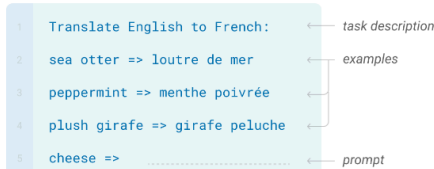
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



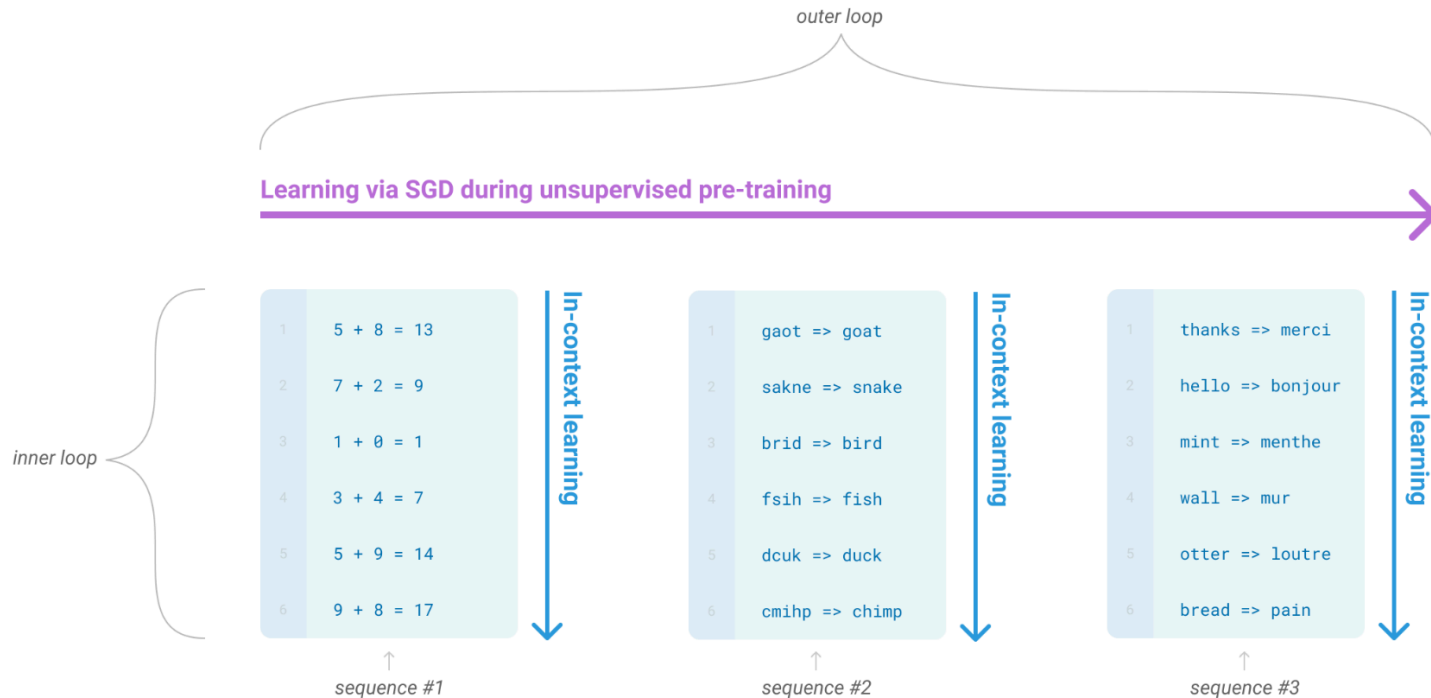
Traditional fine-tuning (not used for GPT-3)

Fine-tuning


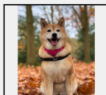
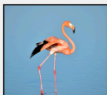






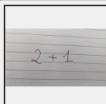
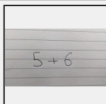
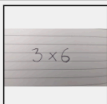

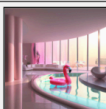

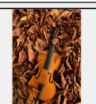
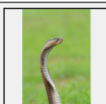
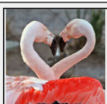
The model is trained via repeated gradient updates using a large corpus of example tasks.



In-context learning



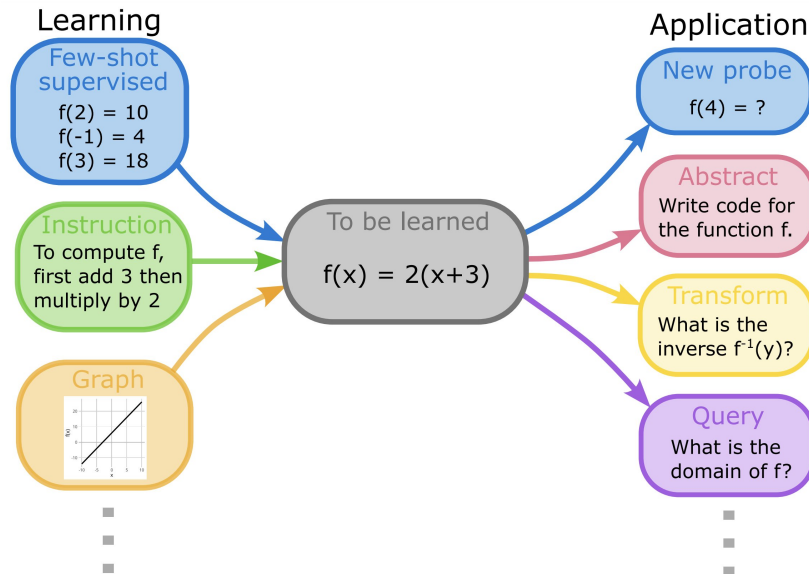
In-context learning

Input Prompt					Completion	
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.		This is	a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.		What is the name of the city where this was painted? Answer:	Arles.
	Output: "Underground"		Output: "Congress"		Output:	"Souloumi"
	2+1=3		5+6=11			3x6=18
	Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.		Output: A pink room with a flamingo pool float.		Output:	A portrait of Salvador Dali with a robot head.
	Les sanglots longs des violons de l'automne blessent mon cœur d'une langueur monotone.		Pour qui sont ces serpents qui sifflent sur vos têtes?			Je suis un cœur qui bat pour vous.

[Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac et al. 2022]

In-context learning

"Any distribution of sequences in which context non-trivially decreases loss on subsequent predictions can be interpreted as eliciting a kind of in-context learning."



Chain-of-thought

Prompt the model to provide intermediate "reasoning" steps, rather than answering directly.

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Chain-of-thought

Prompt the model to provide intermediate "reasoning" steps, rather than answering directly.

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Enjoy the Course!

Amir Zamir (amir.zamir@epfl.ch)

Rishubh Singh (rishubh.singh@epfl.ch head TA)

Zhitong Gao (zhitong.gao@epfl.ch)

Roman Bachmann (roman.bachmann@epfl.ch)

<https://vilab.epfl.ch/>