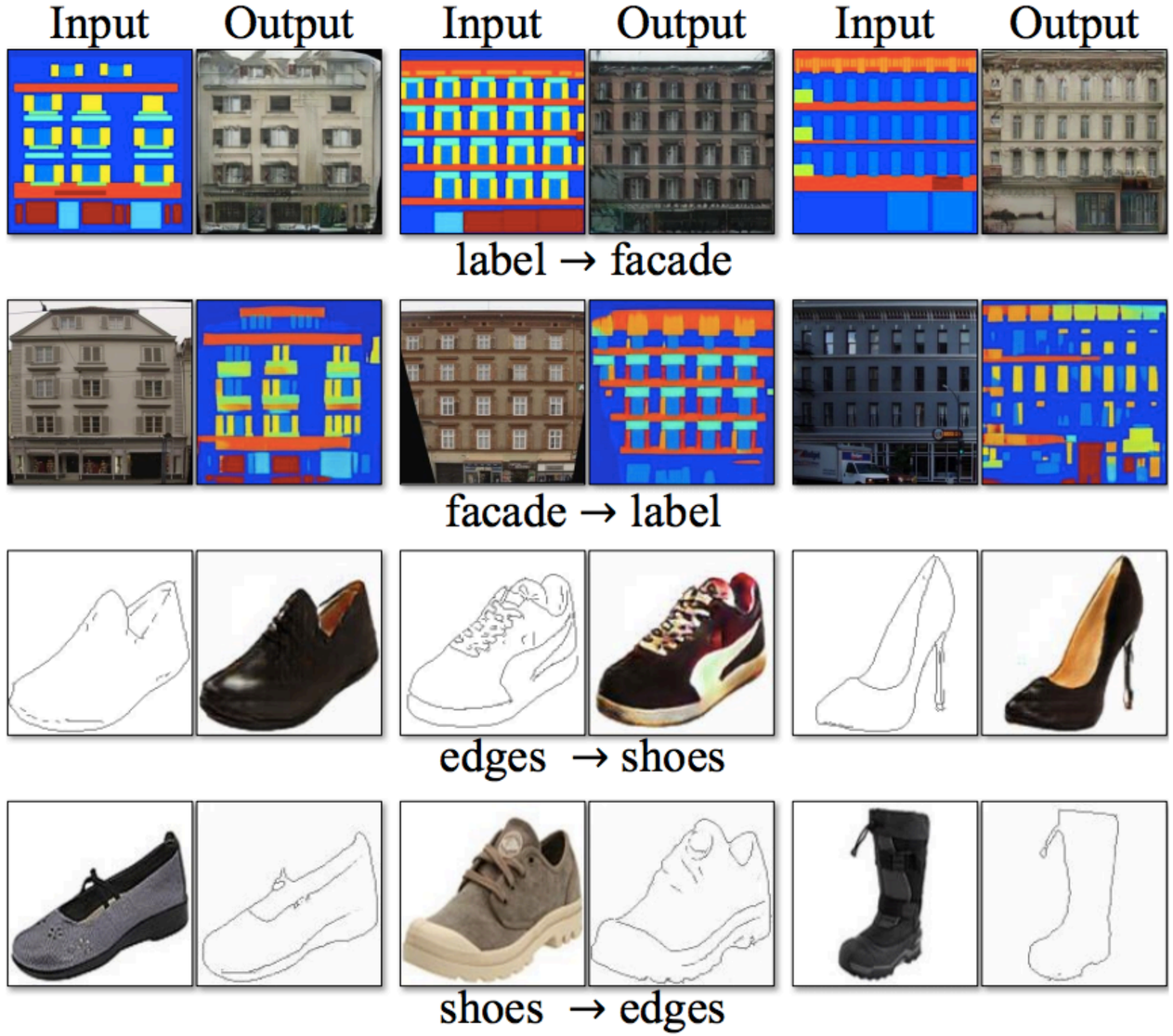
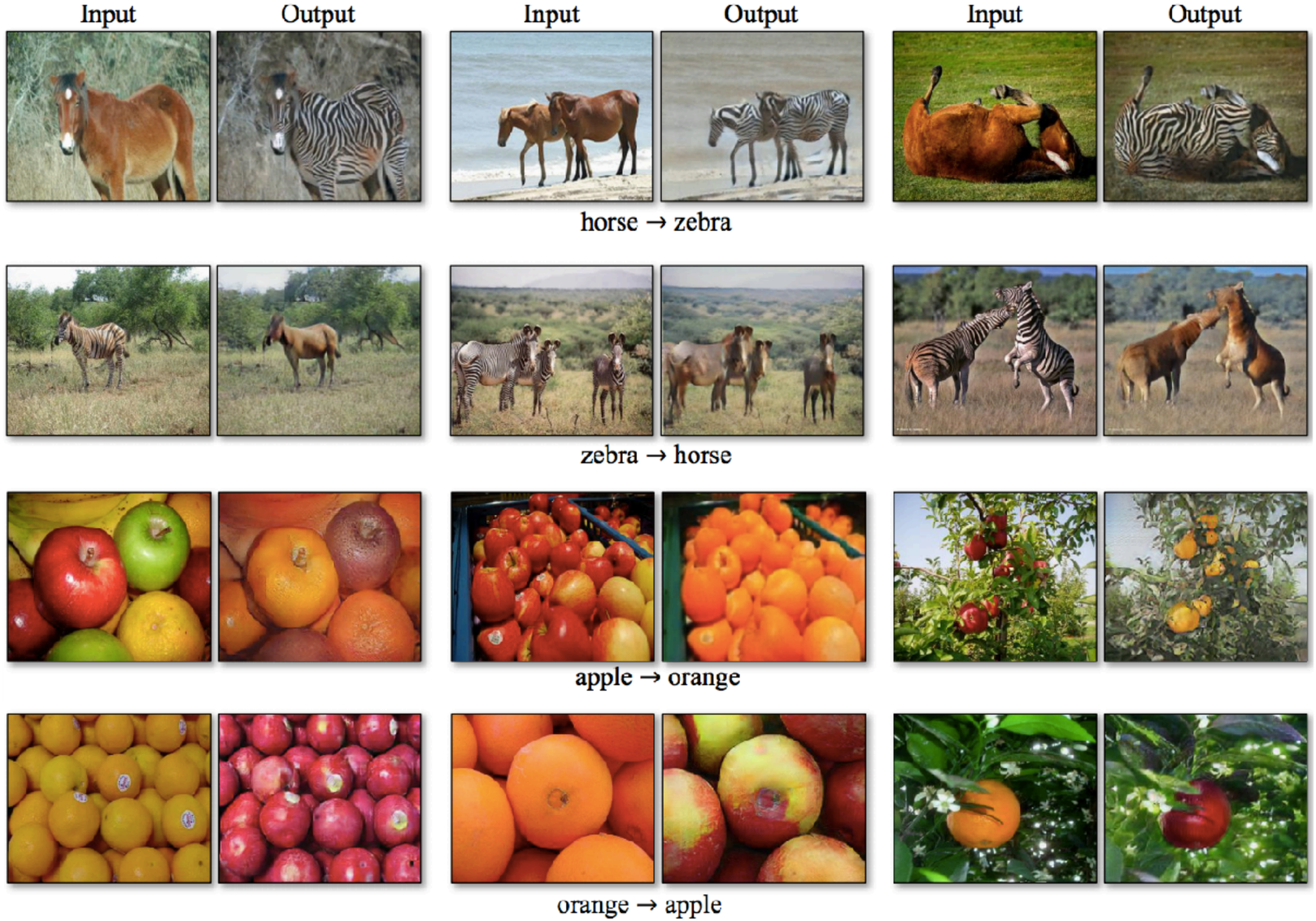
CycleGAN: Results

- Other pix2pix tasks



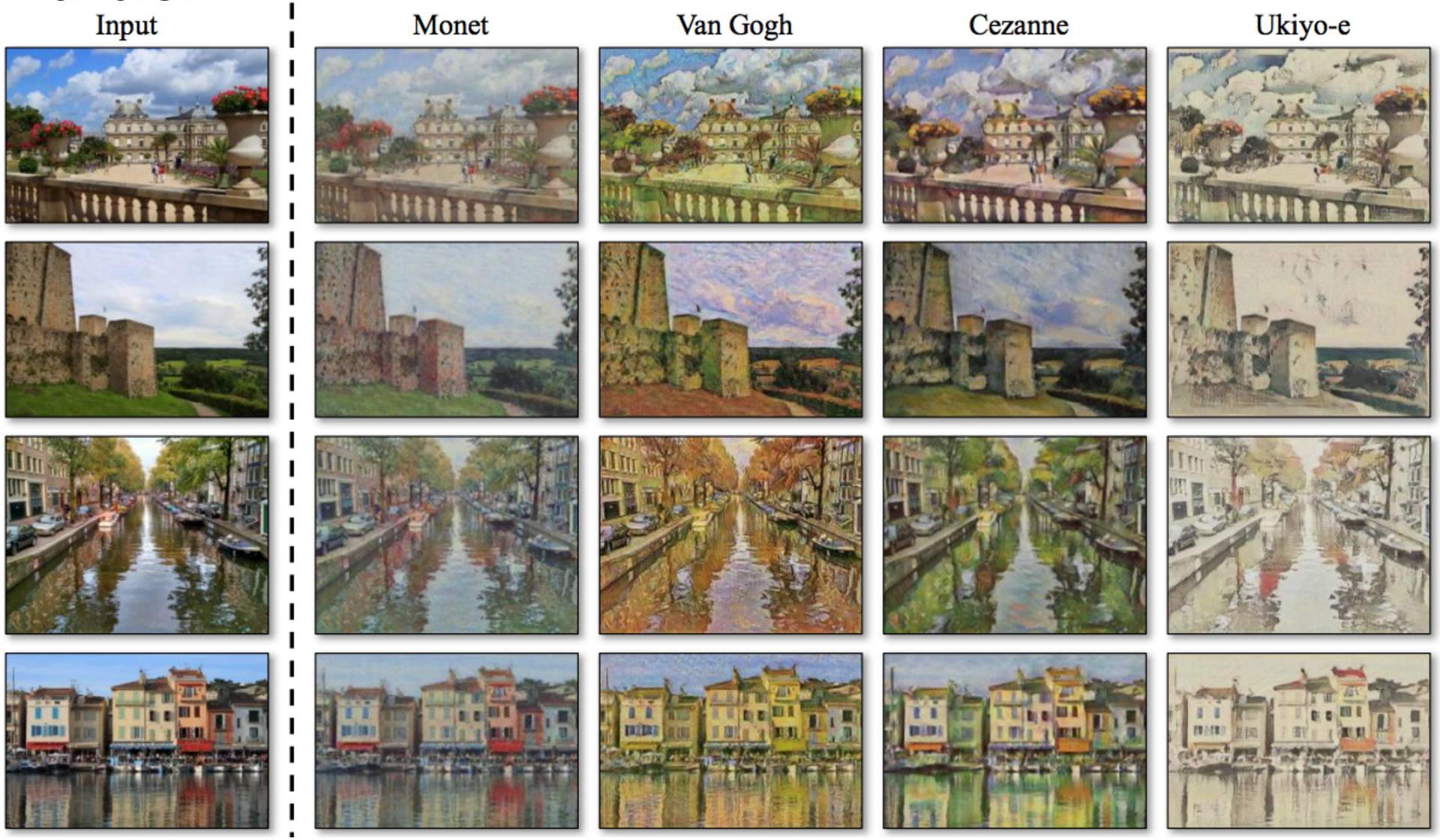
CycleGAN: Results

- Tasks for which paired data is unavailable

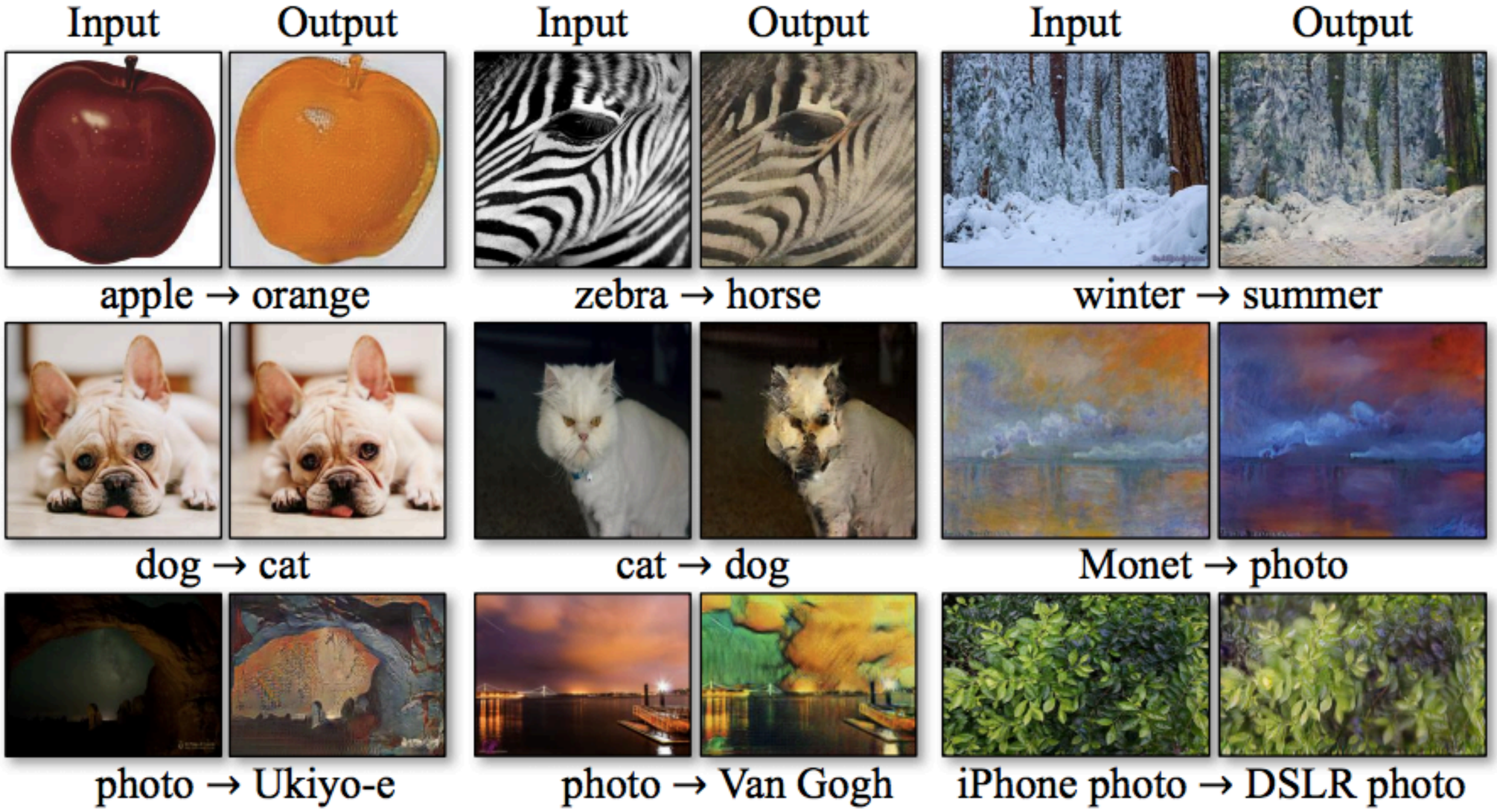


CycleGAN: Results

- Style transfer



CycleGAN: Failure cases



CycleGAN: Failure cases

Input

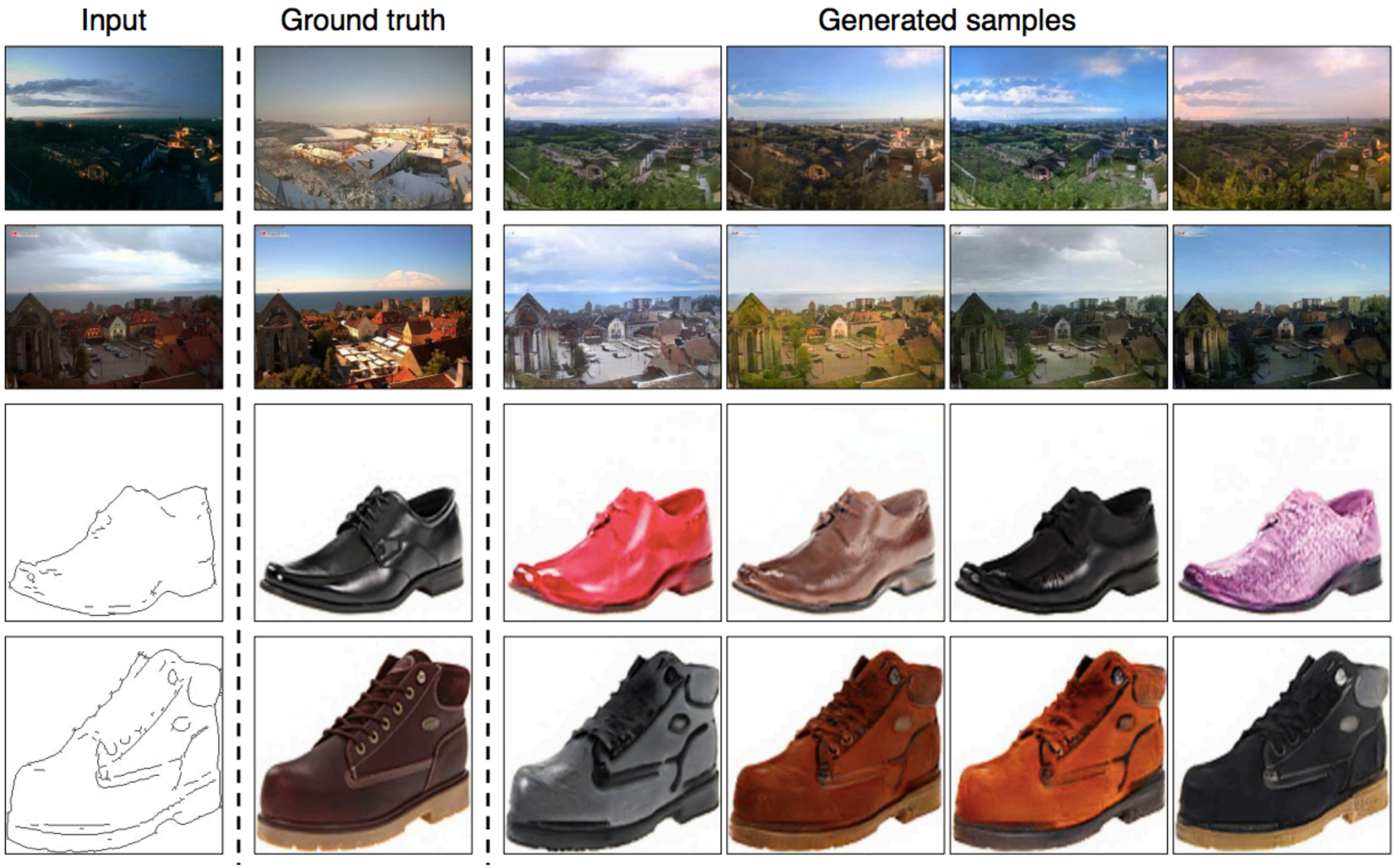
Output

horse → zebra

CycleGAN: Limitations

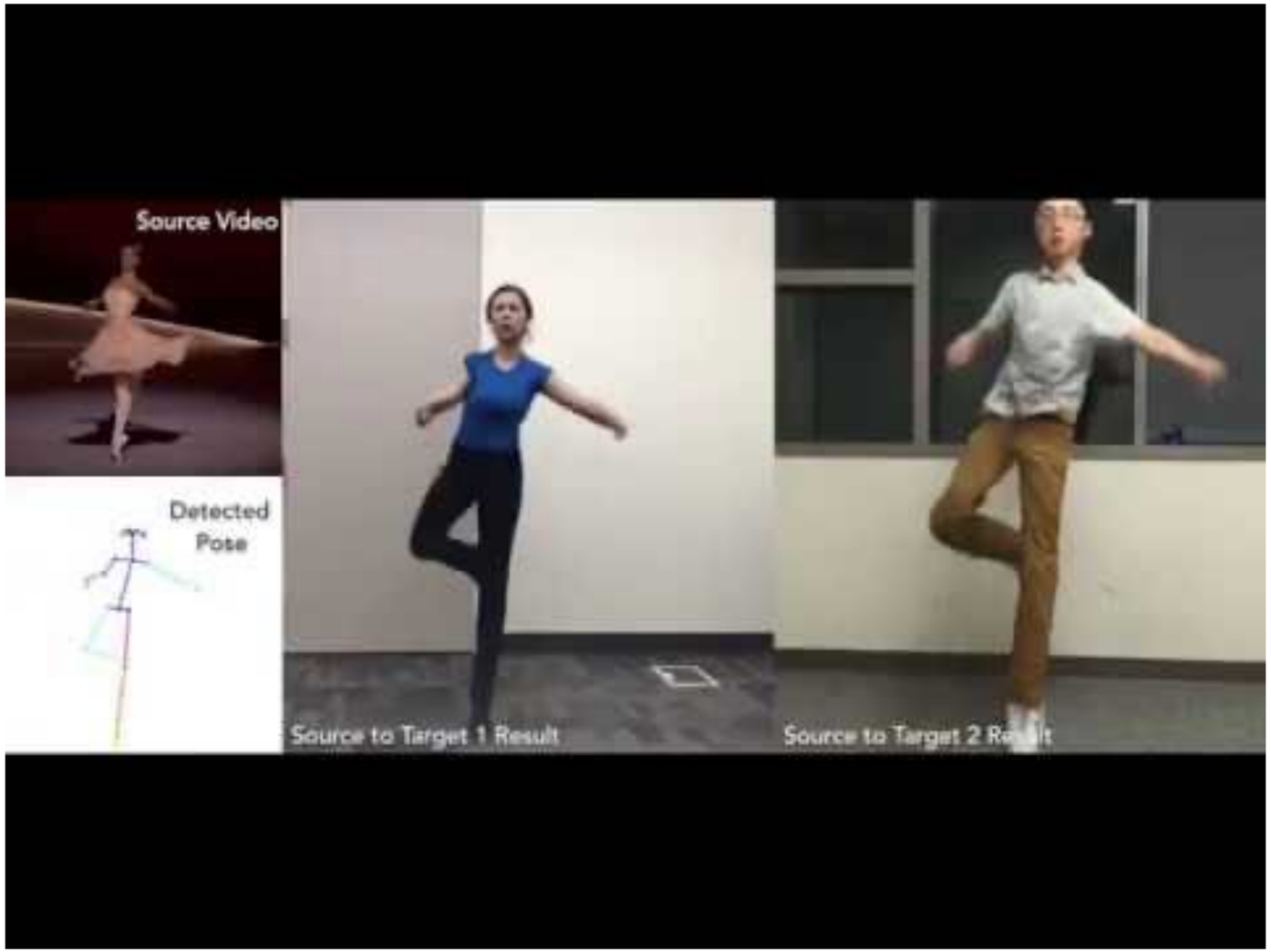
- Cannot handle shape changes (e.g., dog to cat)
- Can get confused on images outside of the training domains (e.g., horse with rider)
- Cannot close the gap with paired translation methods
- Does not account for the fact that one transformation direction may be more challenging than the other

Multimodal image-to-image translation



J.Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman,
Toward Multimodal Image-to-Image Translation, NIPS 2017

Human generation conditioned on pose



https://www.youtube.com/watch?v=PCBTZh41Ris

C. Chan, S. Ginosar, T. Zhou, A. Efros. Everybody Dance Now. ICCV 2019 **https://carolineec.github.io/everybody_dance_now/**

Agenda

1. Generative neural networks

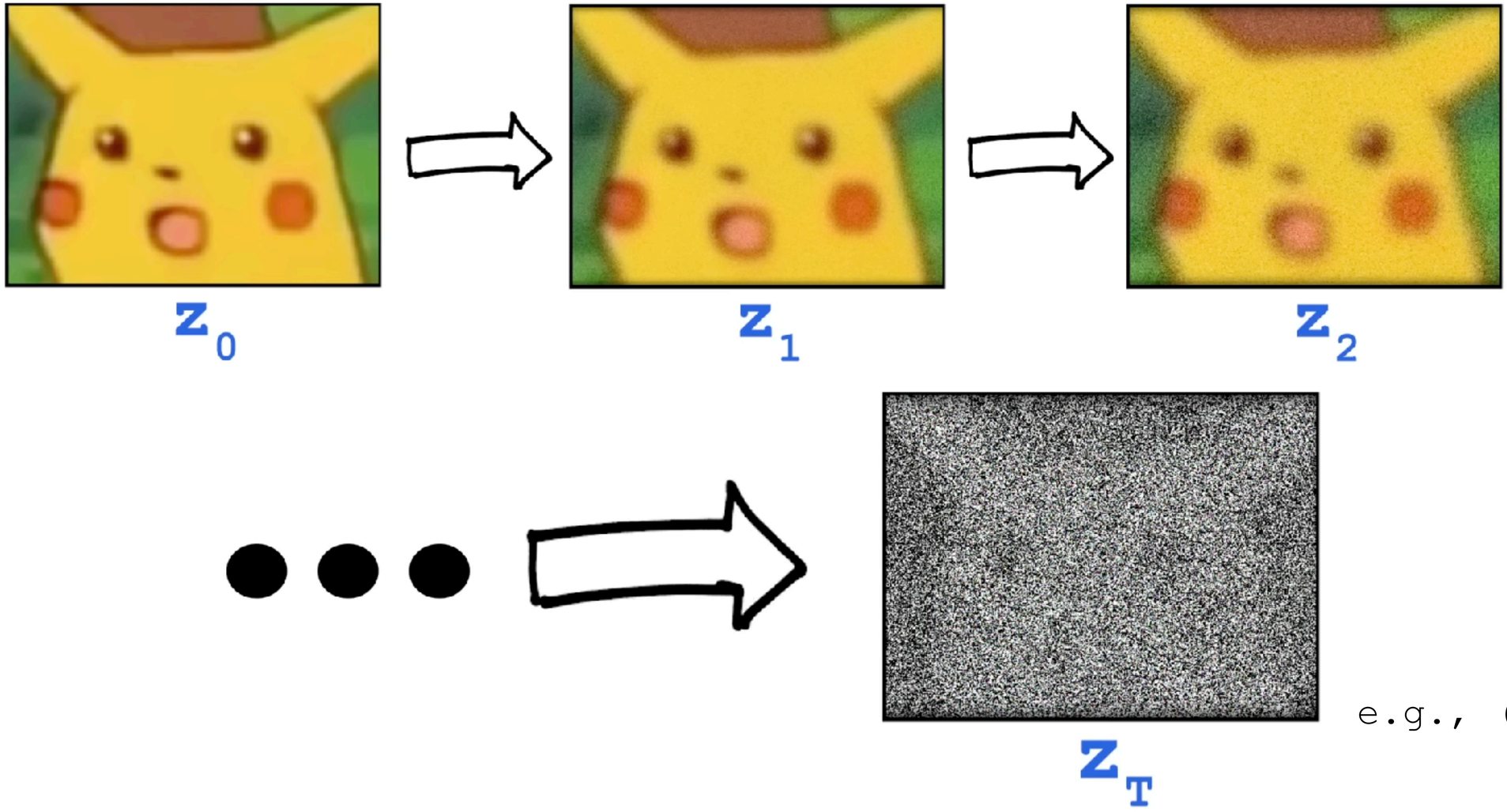
- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

Diffusion models: Forward (diffusion) process

"easy to convert structured data into noise"*



e.g., (T ~ 1000)

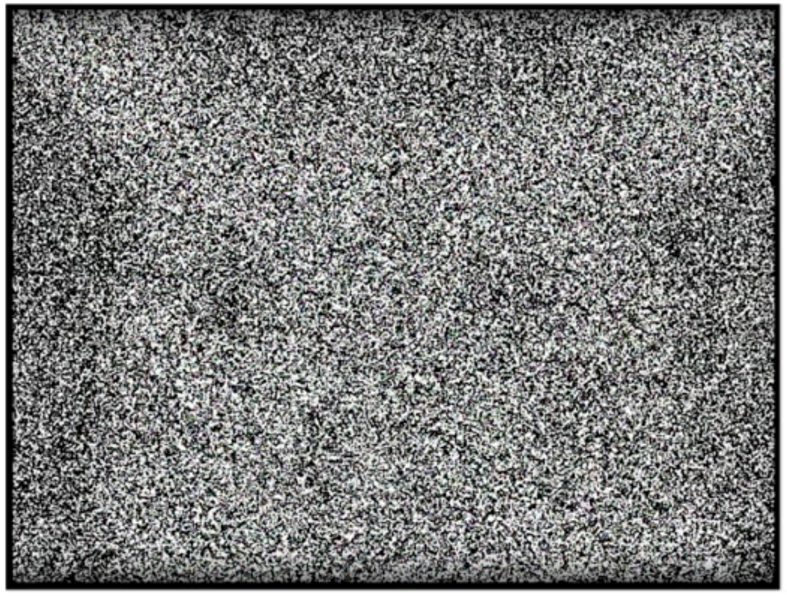
After enough steps, we have $z_T \sim N(0,1)$

*[Murphy 2023]

Edan Meyer: Stable Diffusion - What, Why, How?

Diffusion models

"hard to convert noise into structured data"*



z_T

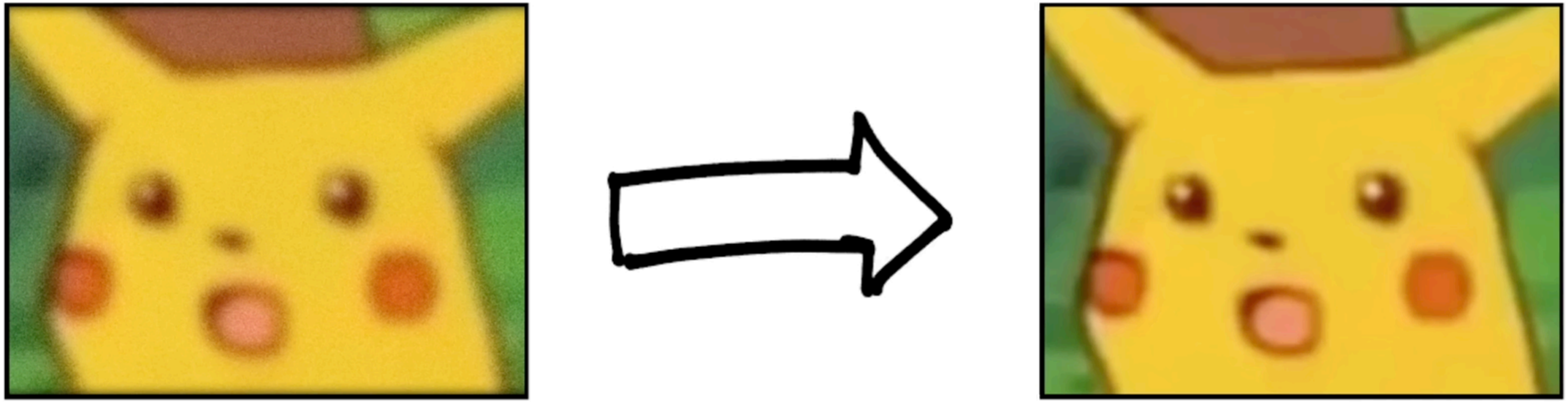
z_0

*[Murphy 2023]

Edan Meyer: [Stable Diffusion - What, Why, How?]

Diffusion models

Reverse (denoising) process



Edan Meyer: [Stable Diffusion - What, Why, How?]

Diffusion models: Learning to denoise/reverse

Inputs

Step 1

z_1

Denoiser Network

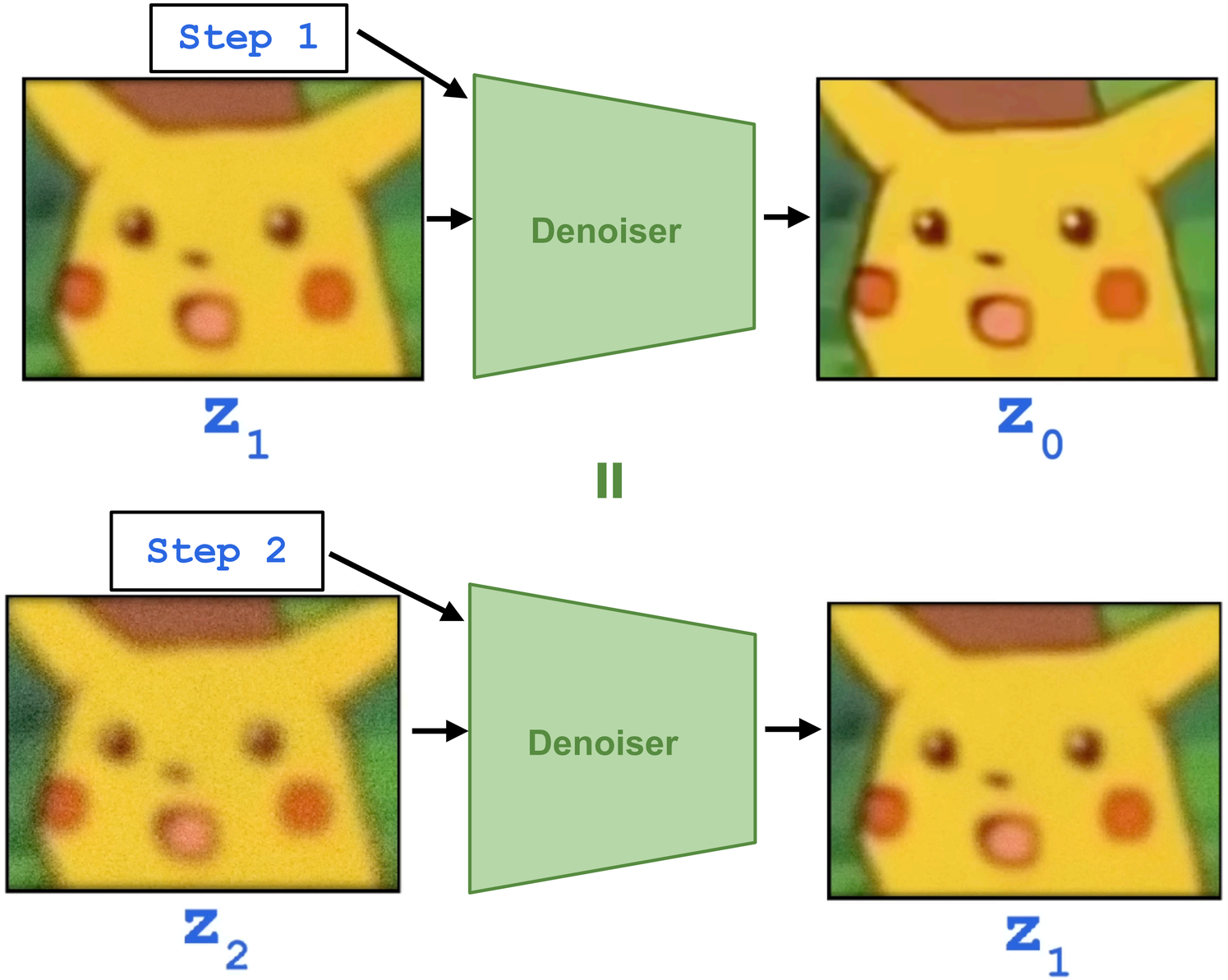
Label

Estimate either:
the denoised image, or
the noise itself

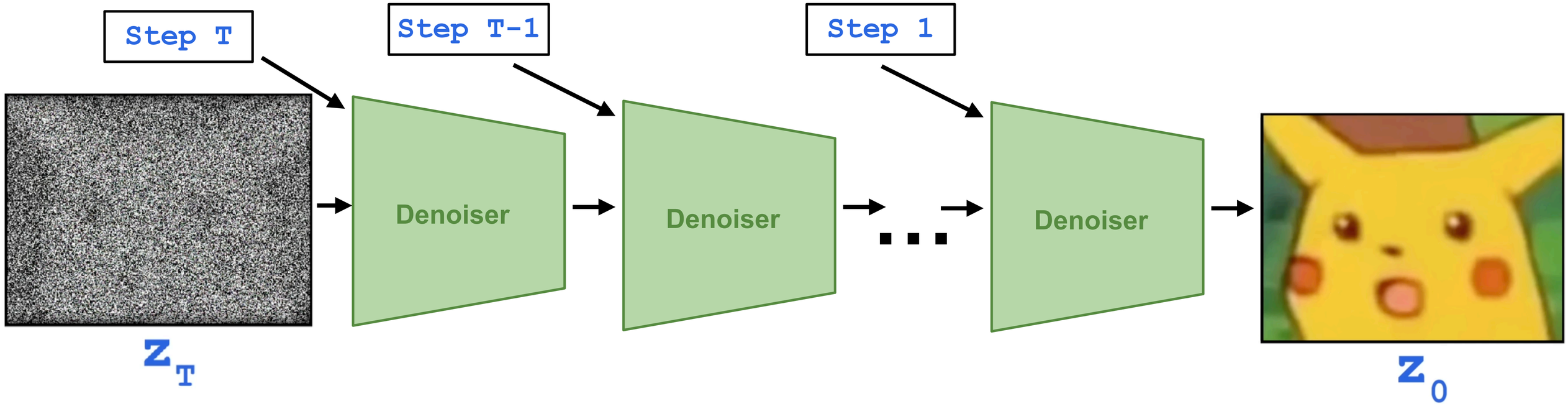
z_0

Can add: Condition

Diffusion models: Learning to denoise/reverse



Diffusion models: Test time



Diffusion models

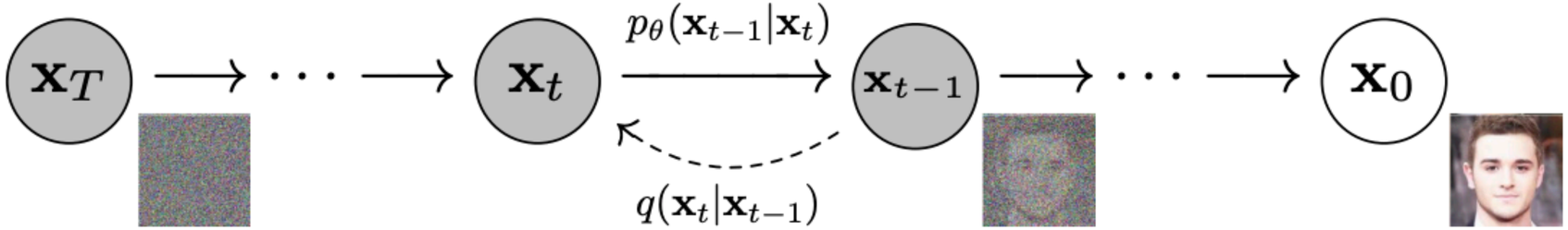


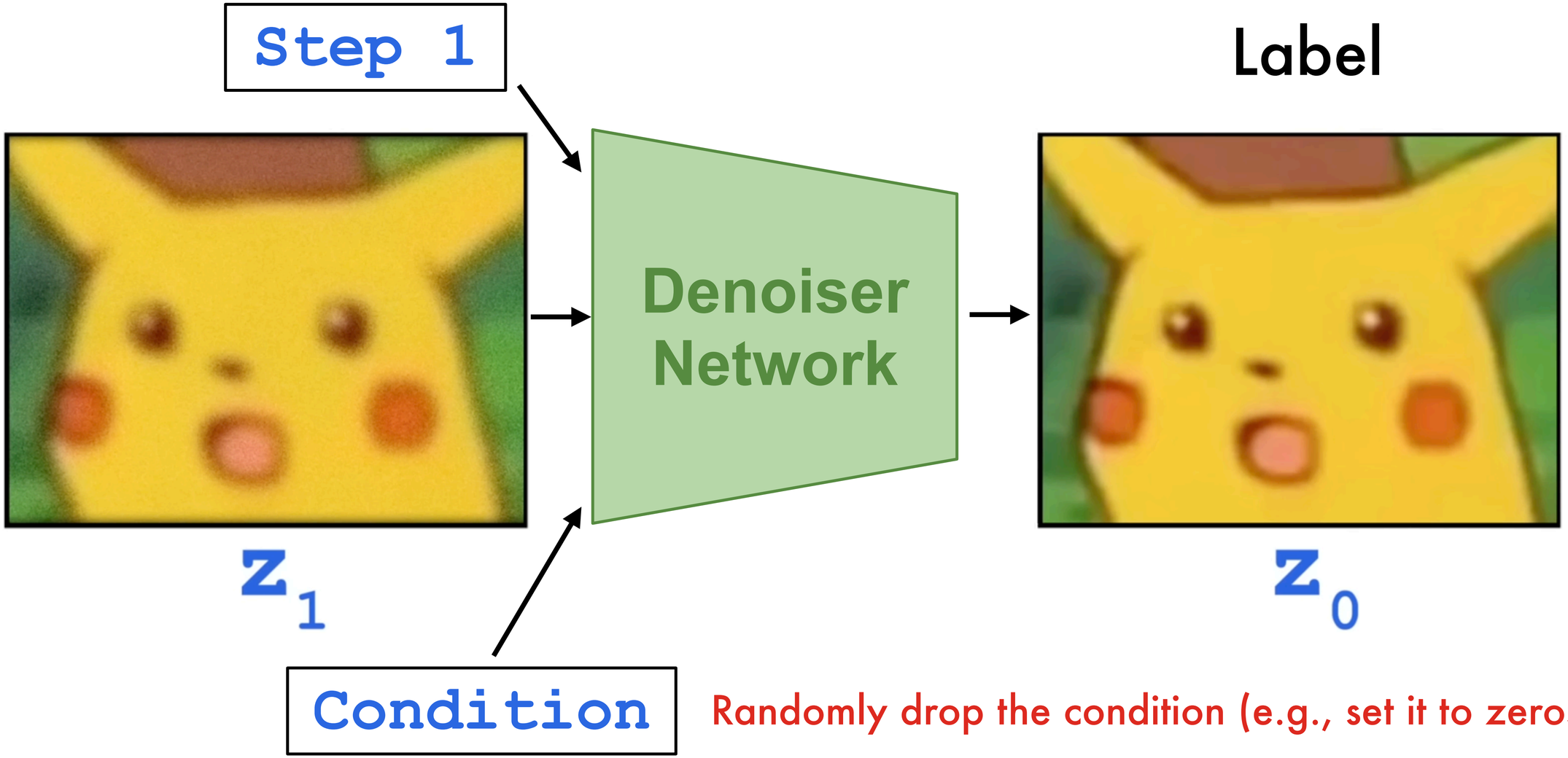
Figure 2: The directed graphical model considered in this work.

Unconditional CIFAR10 sample generation

"Noise schedule"?: linear, cosine etc

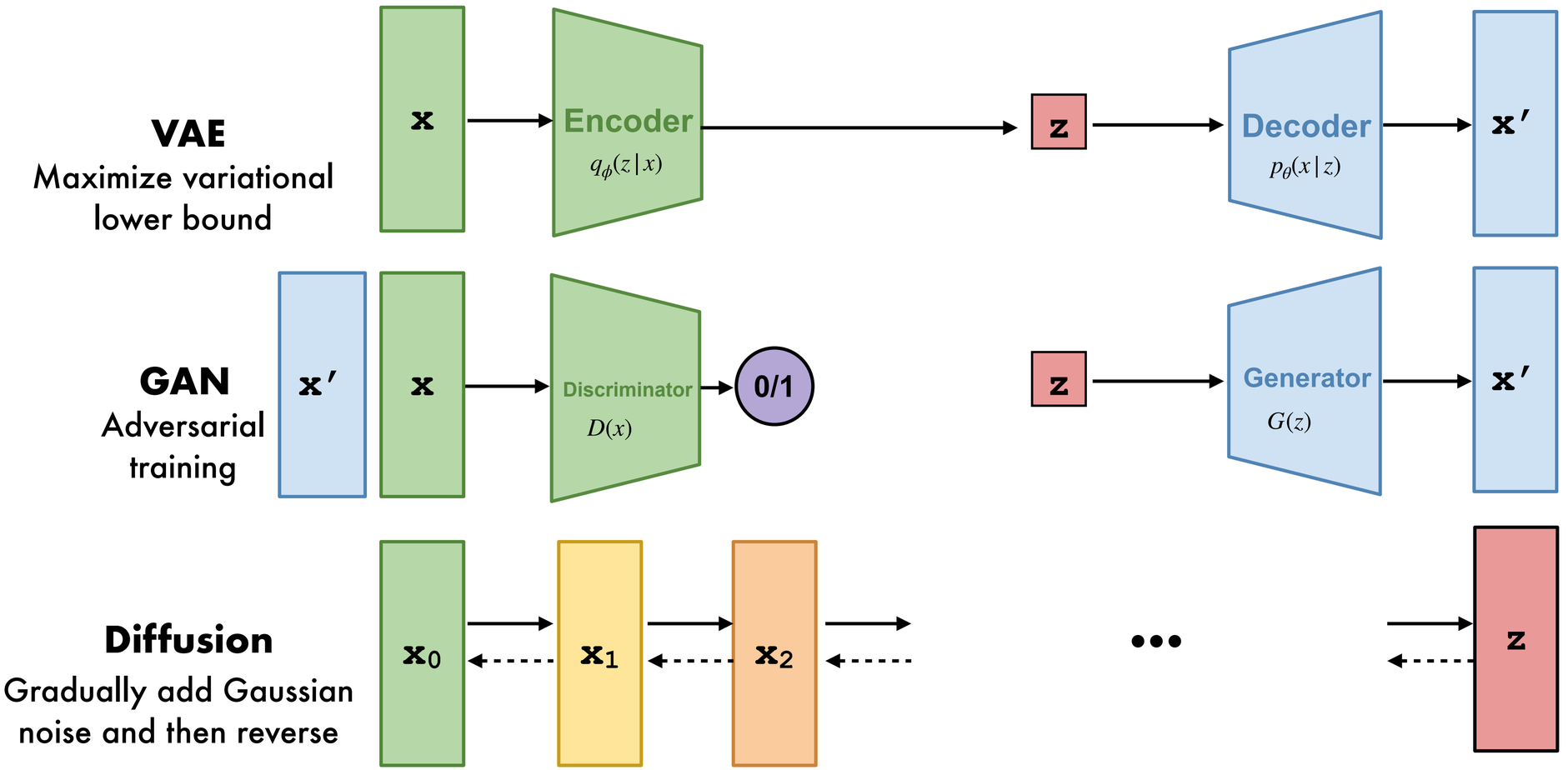
J. Ho et al. Denoising diffusion probabilistic models. NeurIPS 2020
Blog introduction: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Diffusion models: Conditioning



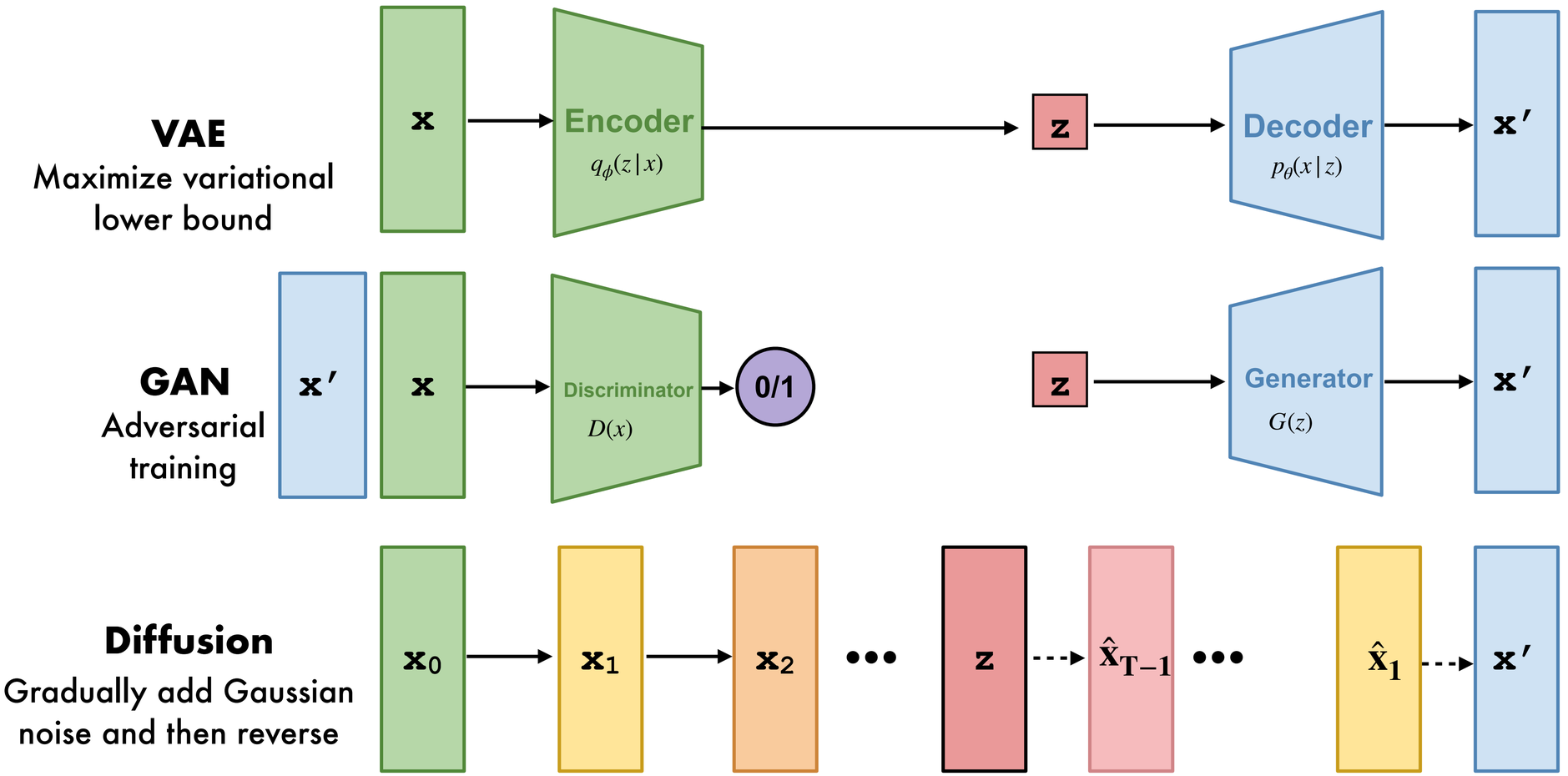
J. Ho and T. Salimans. **Classifier-free diffusion guidance**. In NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021.

Diffusion models vs GANs / VAEs

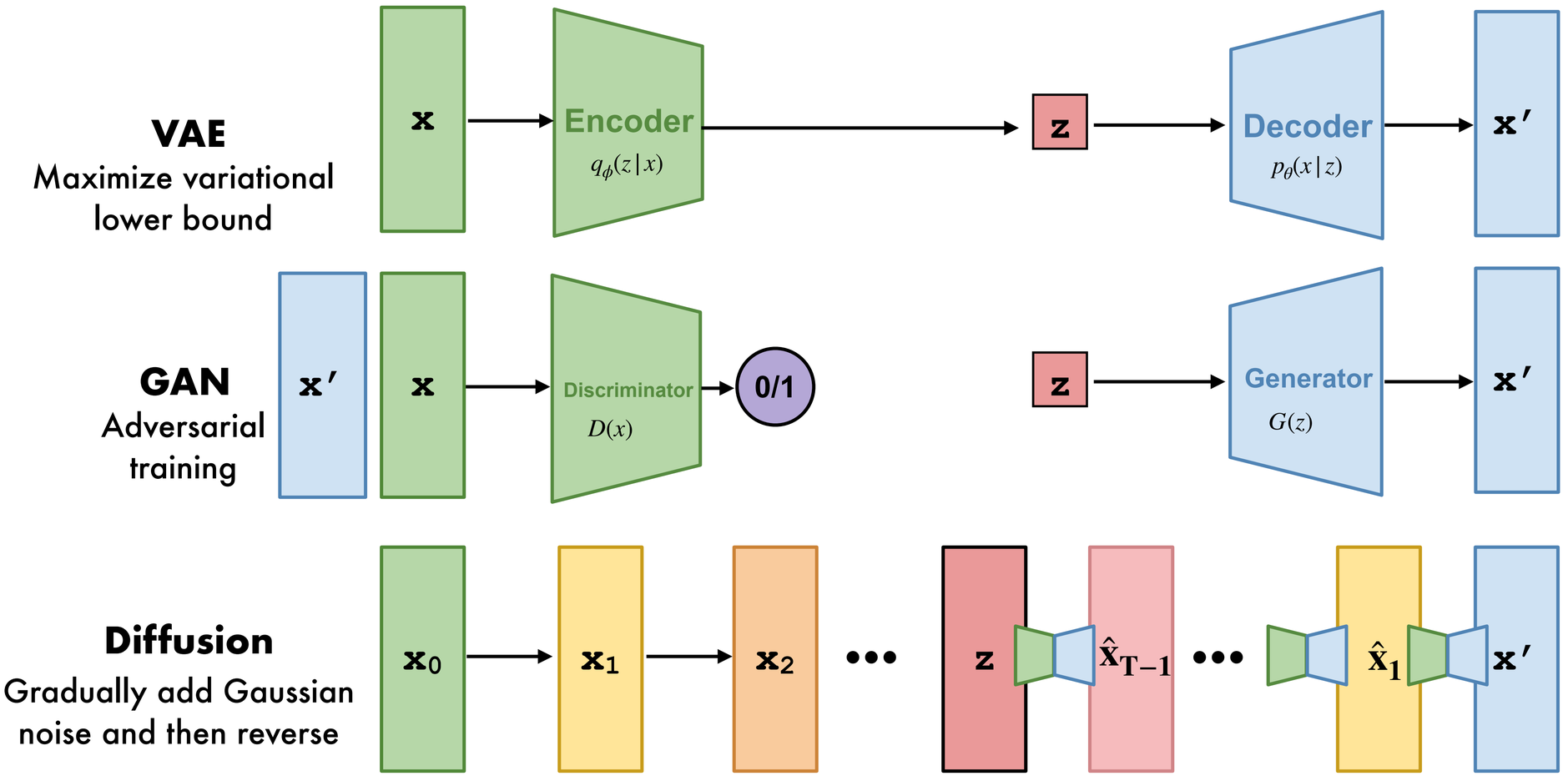


Source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

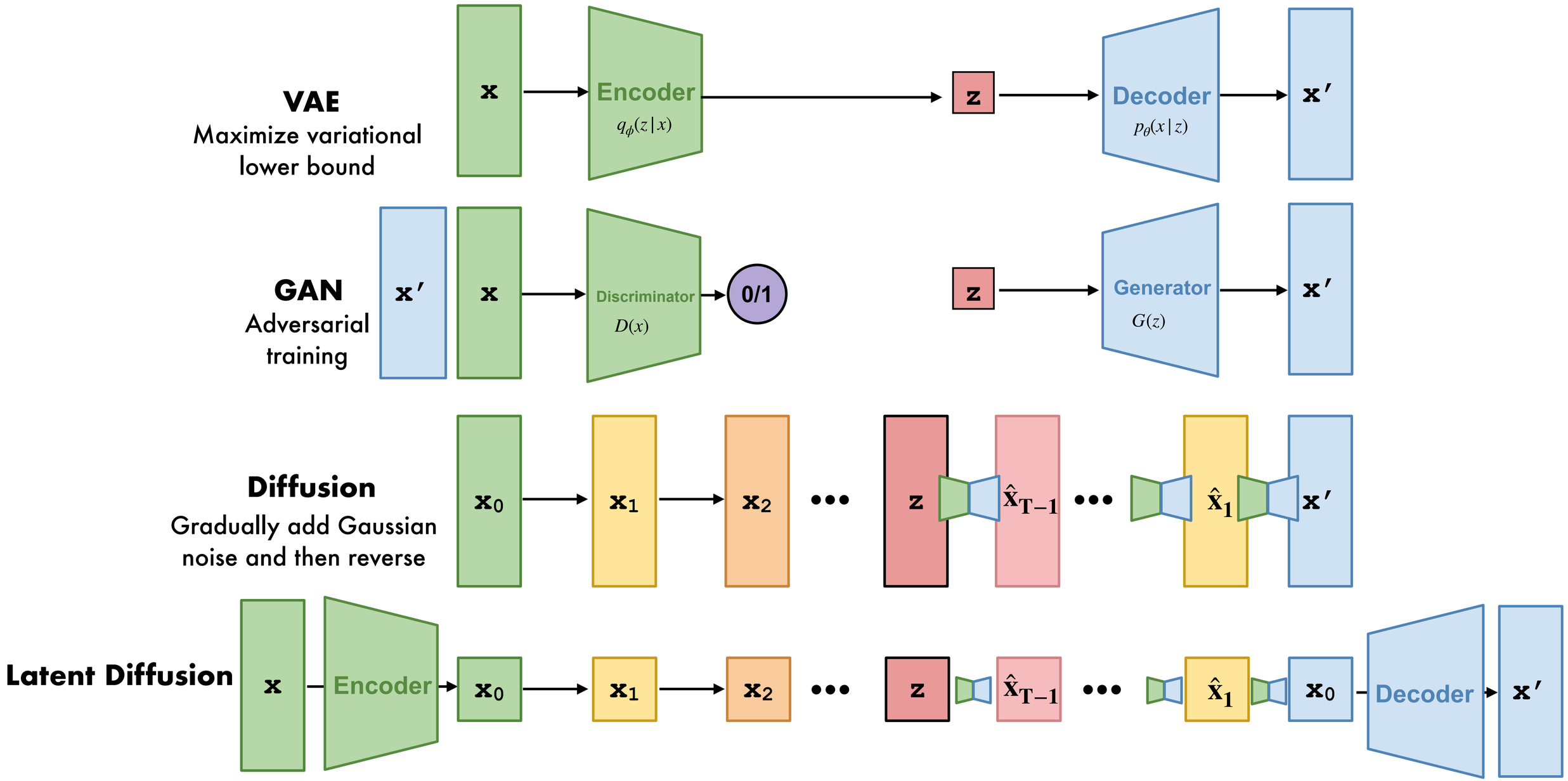
Diffusion models vs GANs / VAEs



Diffusion models vs GANs / VAEs



Diffusion models: Latent diffusion



Trends (!)

23 May 2022

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Chitwan Saharia*, William Chan*, Saurabh Saxena†, Lala Li†, Jay Whang†, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho†, David J Fleet†, Mohammad Norouzi*

{sahariac,williamchan,mnorouzi}@google.com
{srbs,lala,jwhang,jonathanho,davidfleet}@google.com

Google Research, Brain Team
Toronto, Ontario, Canada

Imagen (Google)

13 Apr 2022

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh*
OpenAI
aramesh@openai.com

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

Alex Nichol*
OpenAI
alex@openai.com

Casey Chu*
OpenAI
casey@openai.com

Mark Chen
OpenAI
mark@openai.com

Dall-E 2 (OpenAI)

Paid access with API, Sep. 2022

Dall-E 3 (OpenAI)

Paid access with ChatGPT+, Oct. 2023

20 Dec 2021

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach[1] * Andreas Blattmann[1] * Dominik Lorenz[1] Patrick Esser[R] Björn Ommer[1]
[1]Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany [R]Runway ML
https://github.com/CompVis/latent-diffusion

Abstract

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulat… lows for a guiding mechanism to control the ima… eration process without retraining. However, sinc… models typically operate directly in pixel space, op… tion of powerful DMs often consumes hundreds of…

Latent Diffusion => Stable Diffusion (StabilityAI)

Open sourced, Aug. 2022

Diffusion models

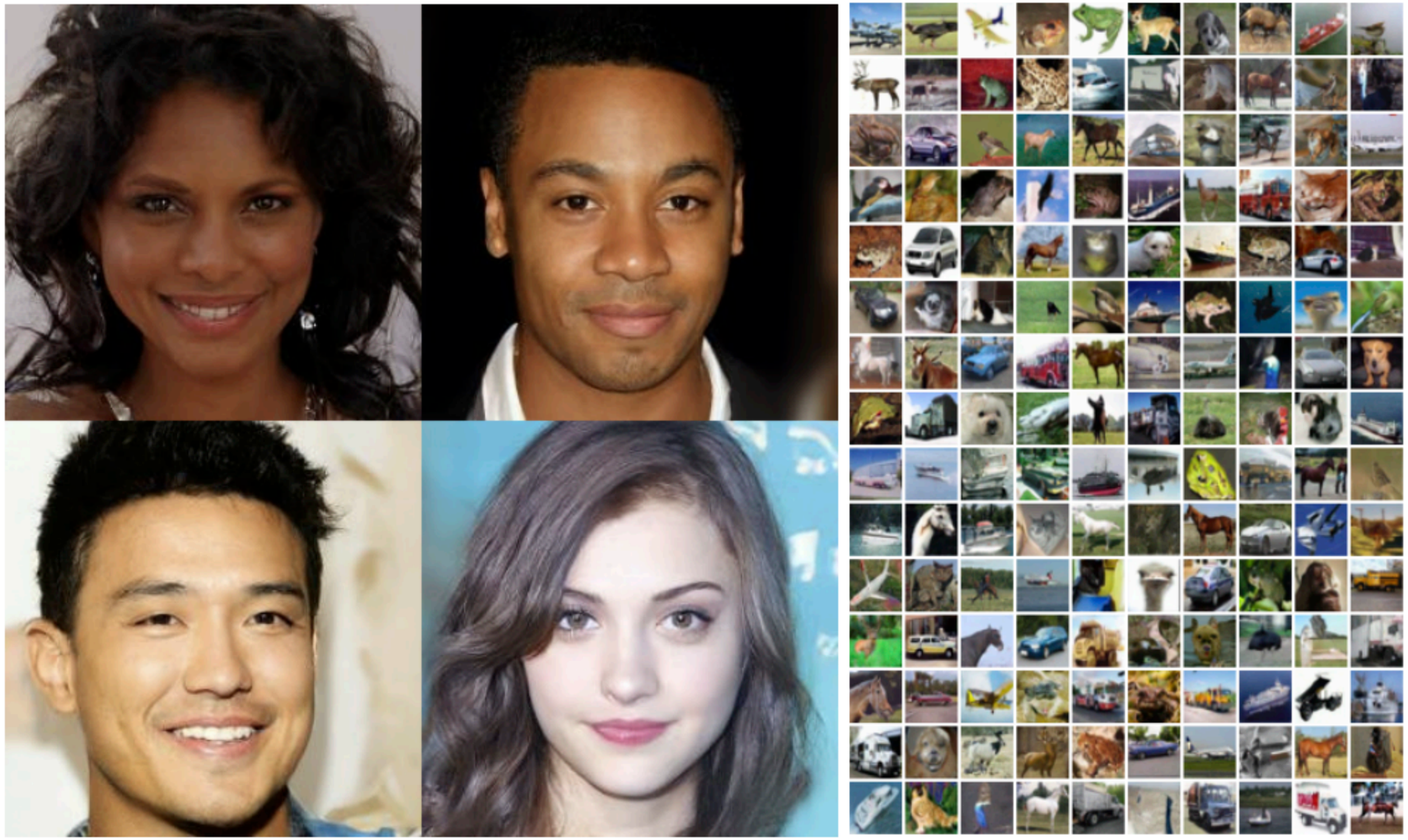


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

J. Ho et al. Denoising diffusion probabilistic models. NeurIPS 2020

Diffusion models

- “We can sample with as few as 25 forward passes while maintaining FIDs comparable to BigGAN”

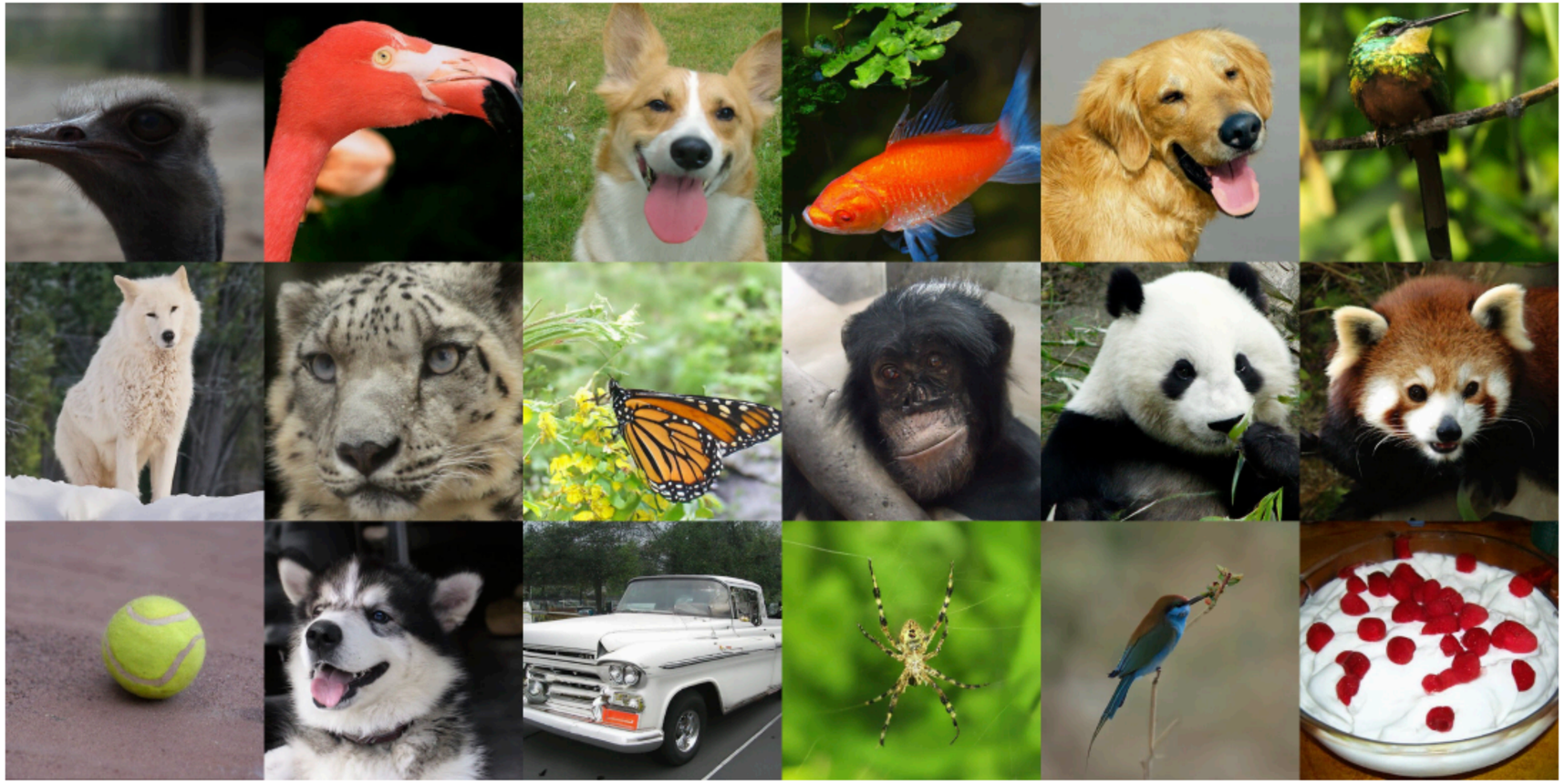
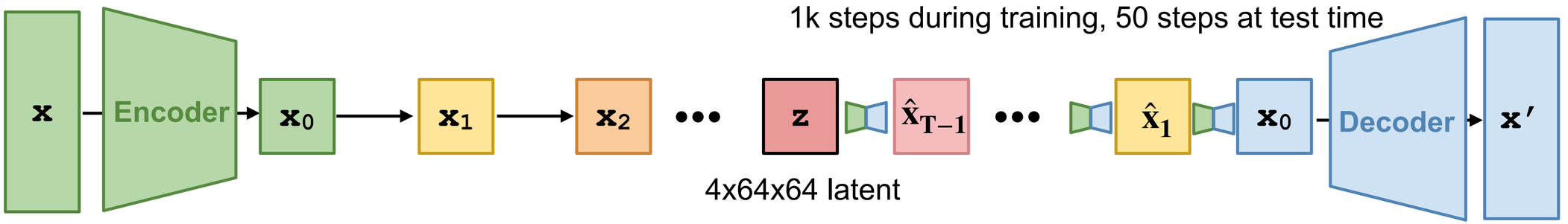


Figure 1: Selected samples from our best ImageNet 512×512 model (FID 3.85)

P. Dhariwal and A. Nichol. Diffusion Models Beat GANs on Image Synthesis. NeurIPS 2021

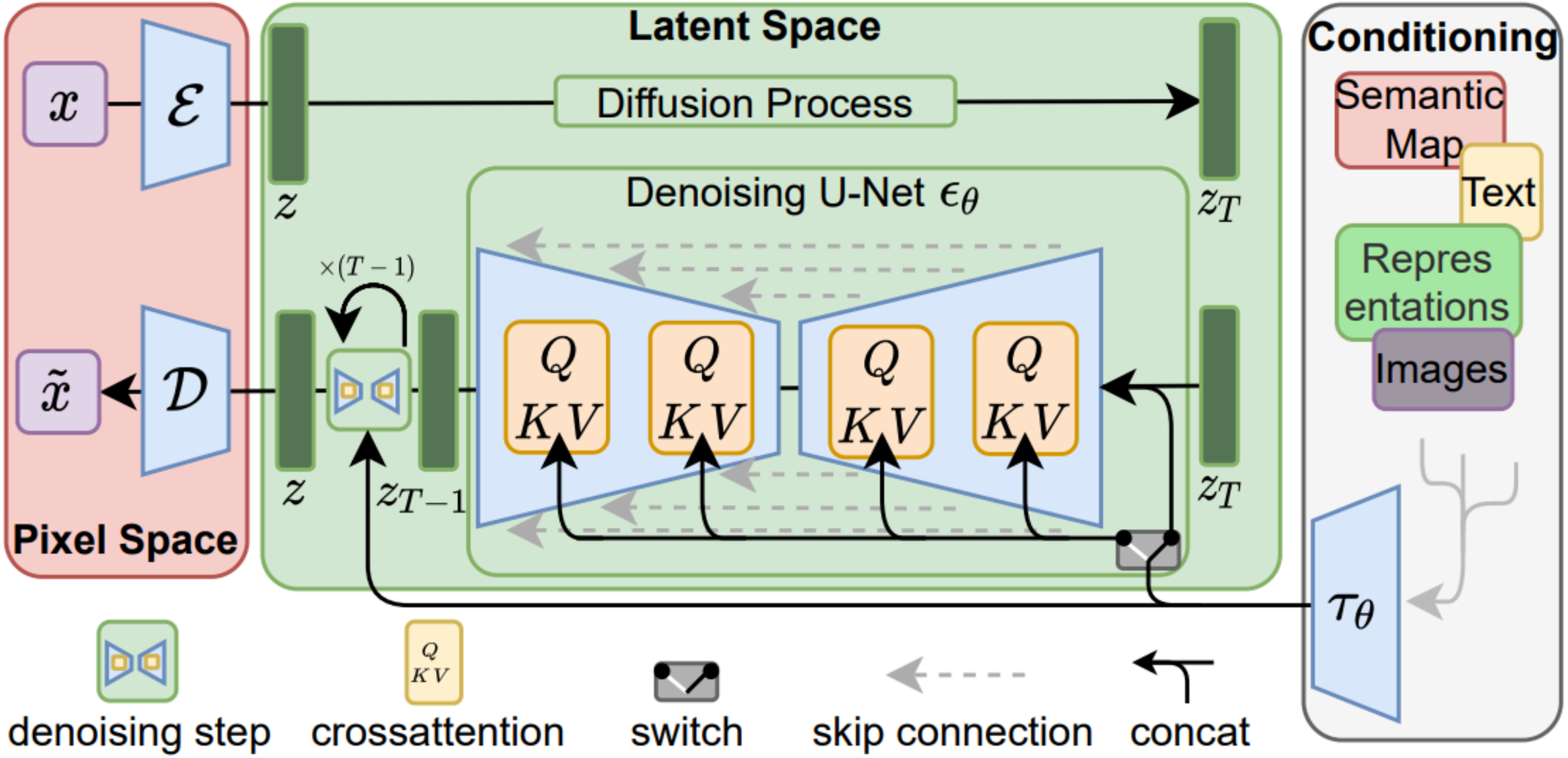
Latent diffusion models (aka Stable Diffusion)

- Trained on a 2B subset of LAION5B dataset (crawl of the internet)
- Unconditional image synthesis, inpainting, stochastic super-resolution. General-purpose conditioning: class-conditional, text-to-image, layout-to-image…



R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

Latent diffusion models (aka Stable Diffusion)



R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

Further reading

https://arxiv.org/pdf/2208.11970.pdf

Understanding Diffusion Models: A Unified Perspective

Calvin Luo
Google Research, Brain Team
calvinluo@google.com

August 26, 2022

Contents

Other Generative Models

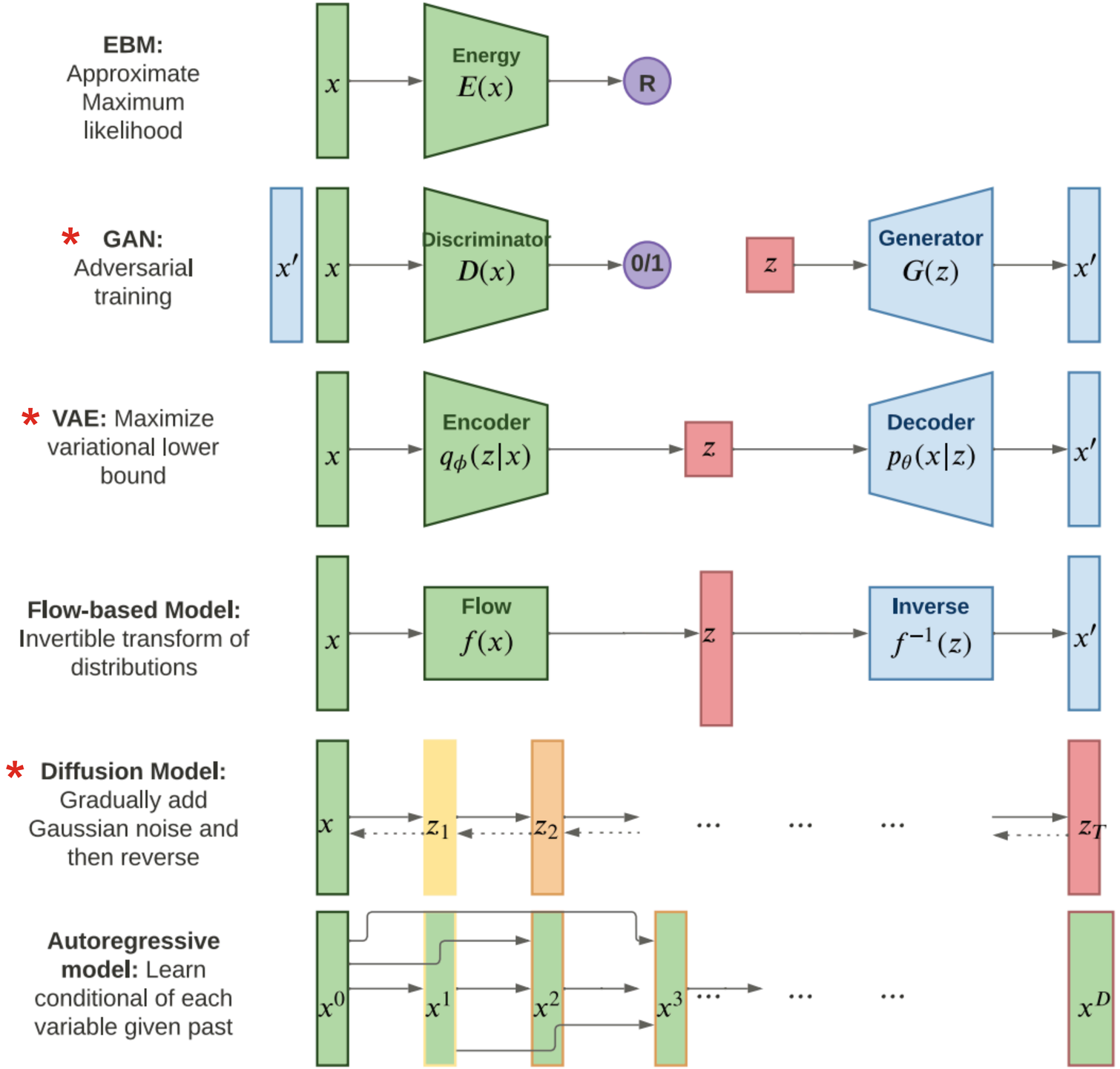


Fig. from Murphy 2023, adapted from

https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Agenda

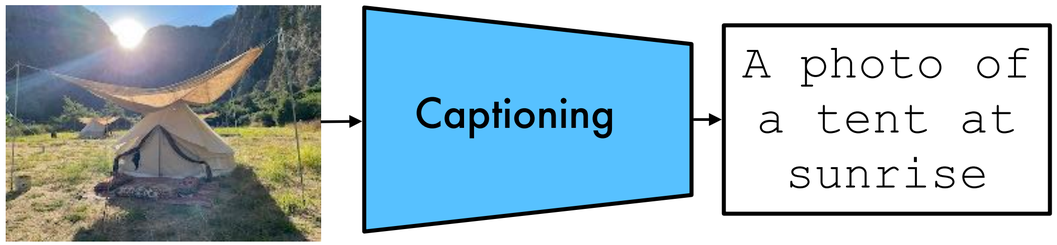
1. Generative neural networks

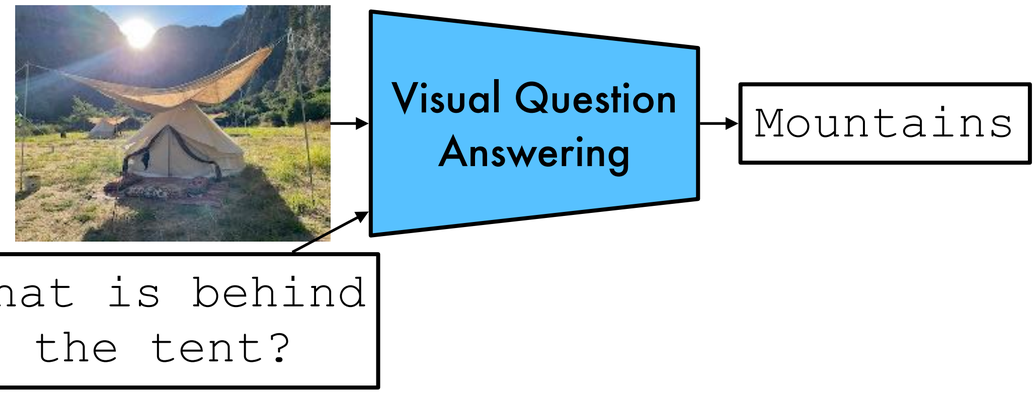
- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

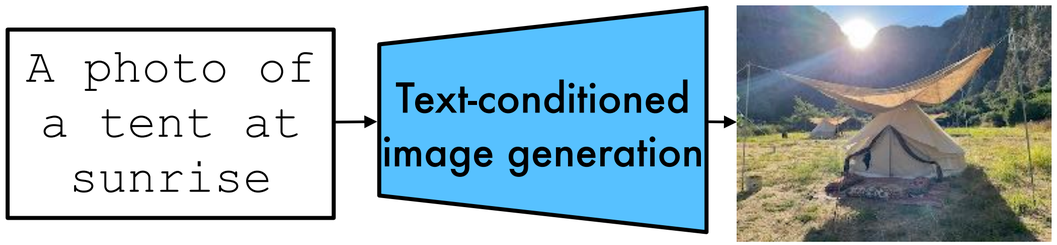
2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

Vision & Language: Tasks







Vision & Language: Tasks

Image in, text out

Captioning

A photo of a tent at sunrise

Image & text in, text out

Visual Question Answering

Mountains

What is behind the tent?

Text in, Image out

A photo of a tent at sunrise

Text-conditioned image generation

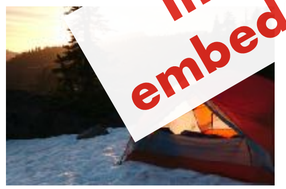
Text in, embedding out

A photo of a tent at sunrise

Text-to-image retrieval

Rank/search in an image gallery

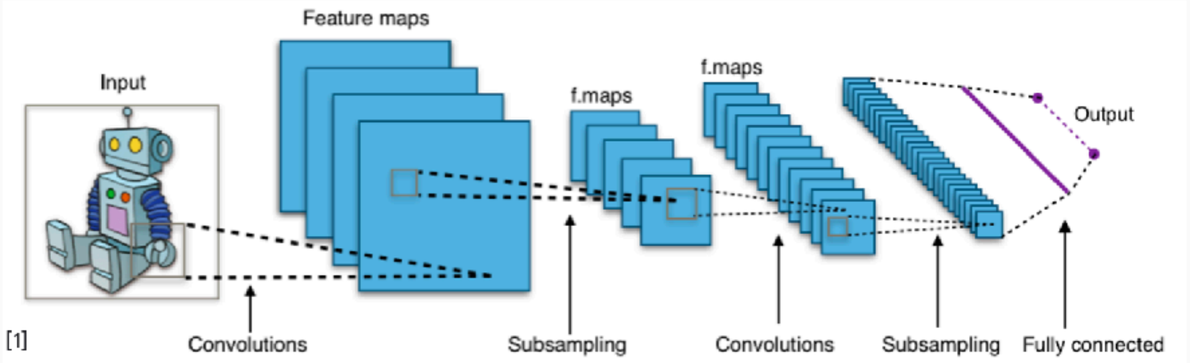
Image in, embedding out



Before:
One architecture per field

Computer Vision

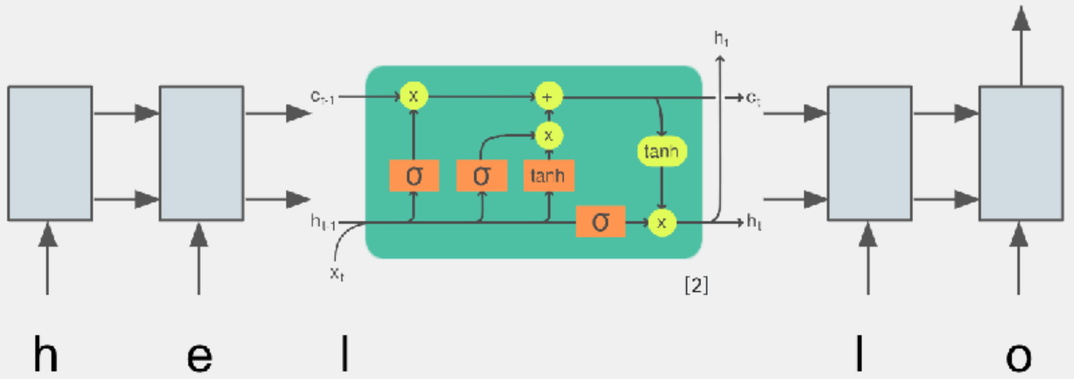
Convolutional NNs (+ResNets)



[1]

Natural Lang. Proc.

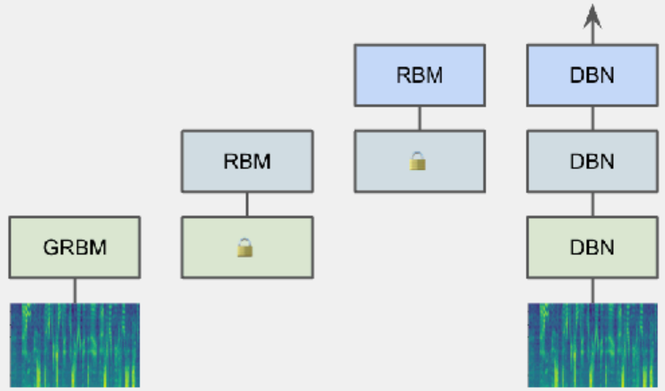
Recurrent NNs (+LSTMs)



h e l l o

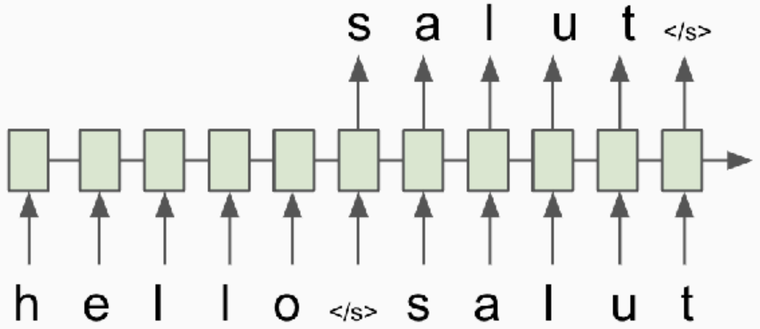
Speech

Deep Belief Nets (+non-DL)



Translation

Seq2Seq



RL

BC/GAIL

Algorithm 1 Generative adversarial imitation learning

1: **Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0
2: **for** $i = 0, 1, 2, \ldots$ **do**
3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4: Update the discriminator parameters from w_i to w_{i+1} with the gradient

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))] \quad (17)$$

5: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$. Specifically, take a KL-constrained natural gradient step with

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)] - \lambda \nabla_\theta H(\pi_\theta),$$
$$\text{where } Q(\bar{s},\bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \mid s_0 = \bar{s}, a_0 = \bar{a}] \quad (18)$$

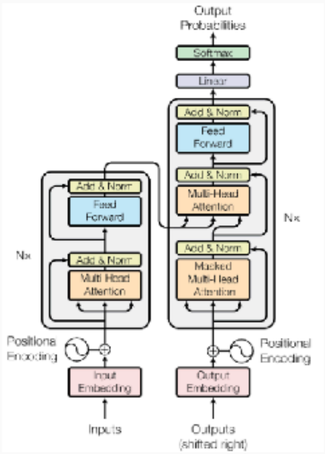
6: **end for**

[1] CNN image CC-BY-SA by Aphex34 for Wikipedia https://commons.wikimedia.org/wiki/File:Typical_cnn.png
[2] RNN image CC-BY-SA by GChe for Wikipedia https://commons.wikimedia.org/wiki/File:The_LSTM_Cell.svg

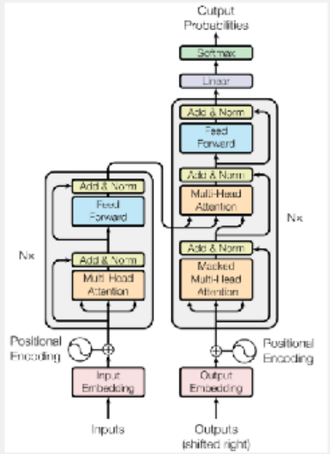
Credit: Lucas Beyer

After:
Unified architecture for all input types (Transformers)

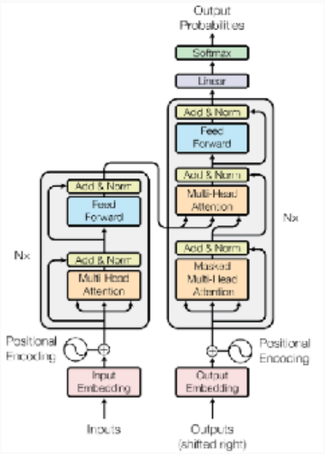
Computer Vision



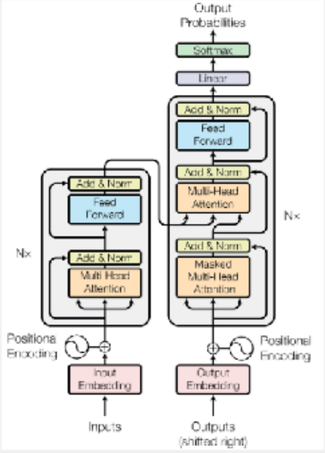
Natural Lang. Proc.



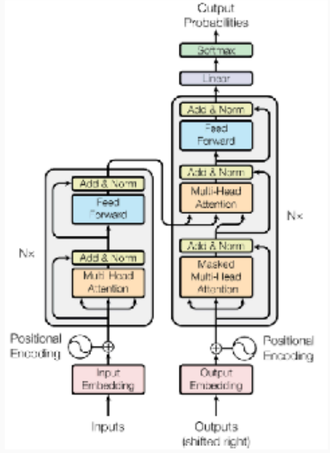
Reinf. Learning



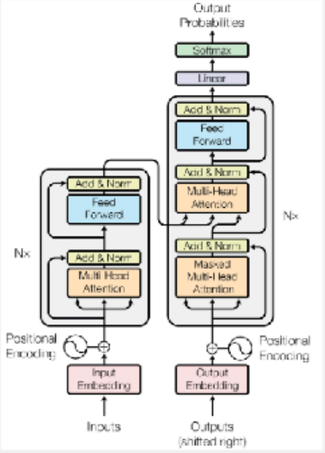
Speech



Translation



Graphs/Science



Transformer image source: "Attention Is All You Need" paper

Credit: Lucas Beyer

Agenda

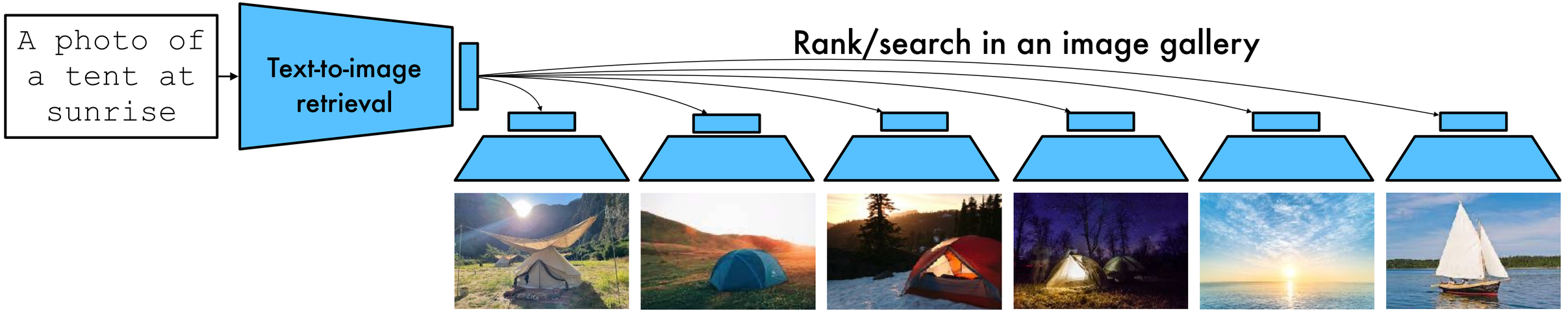
1. Generative neural networks

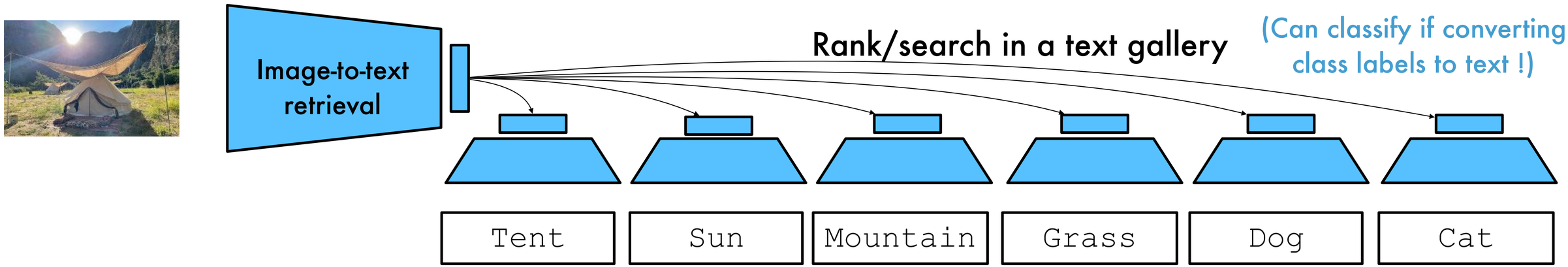
- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

Text-image retrieval

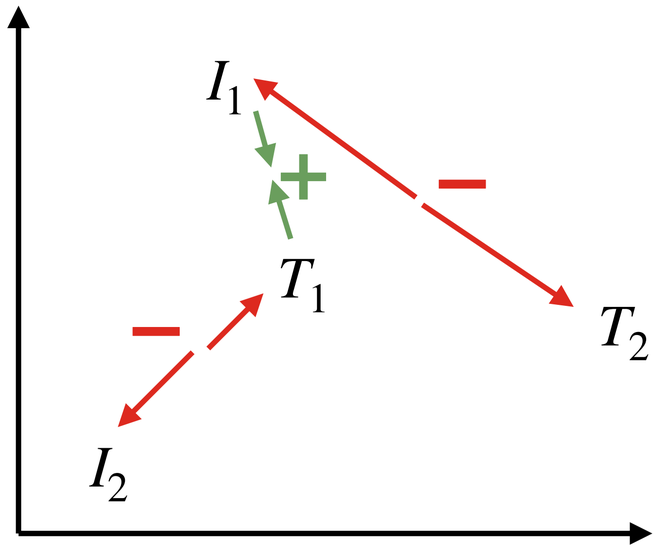


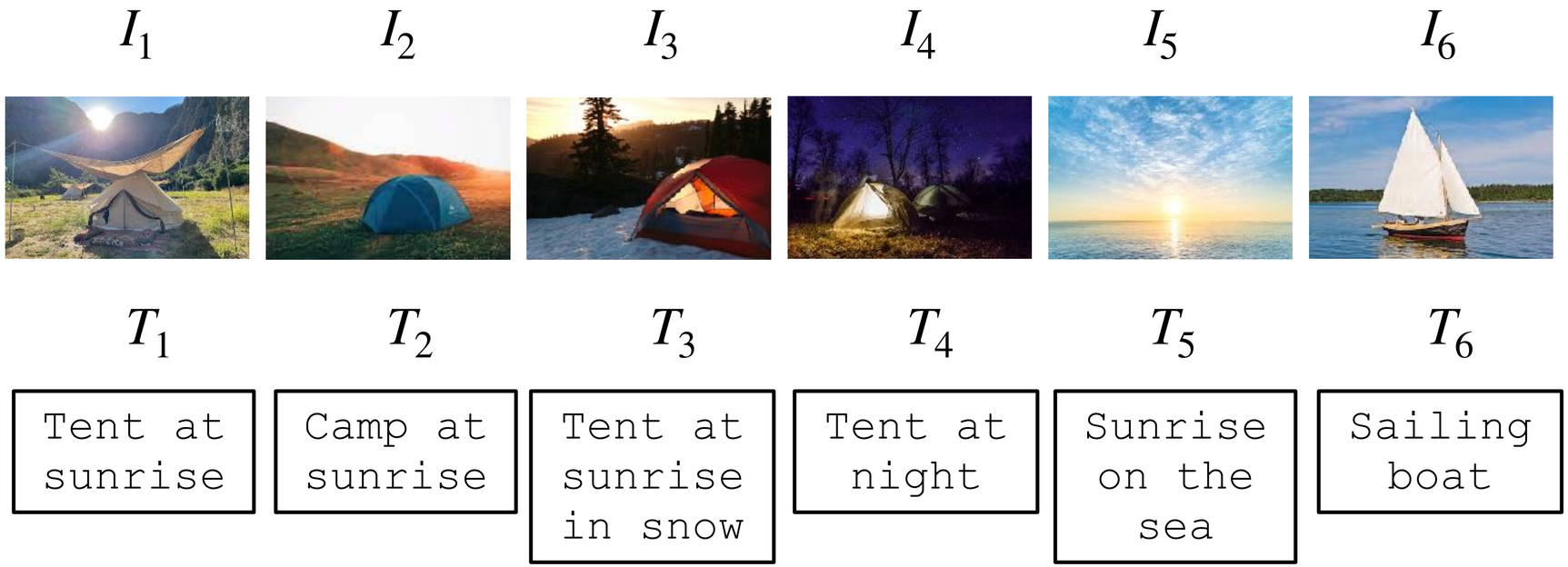


Text-image retrieval: training

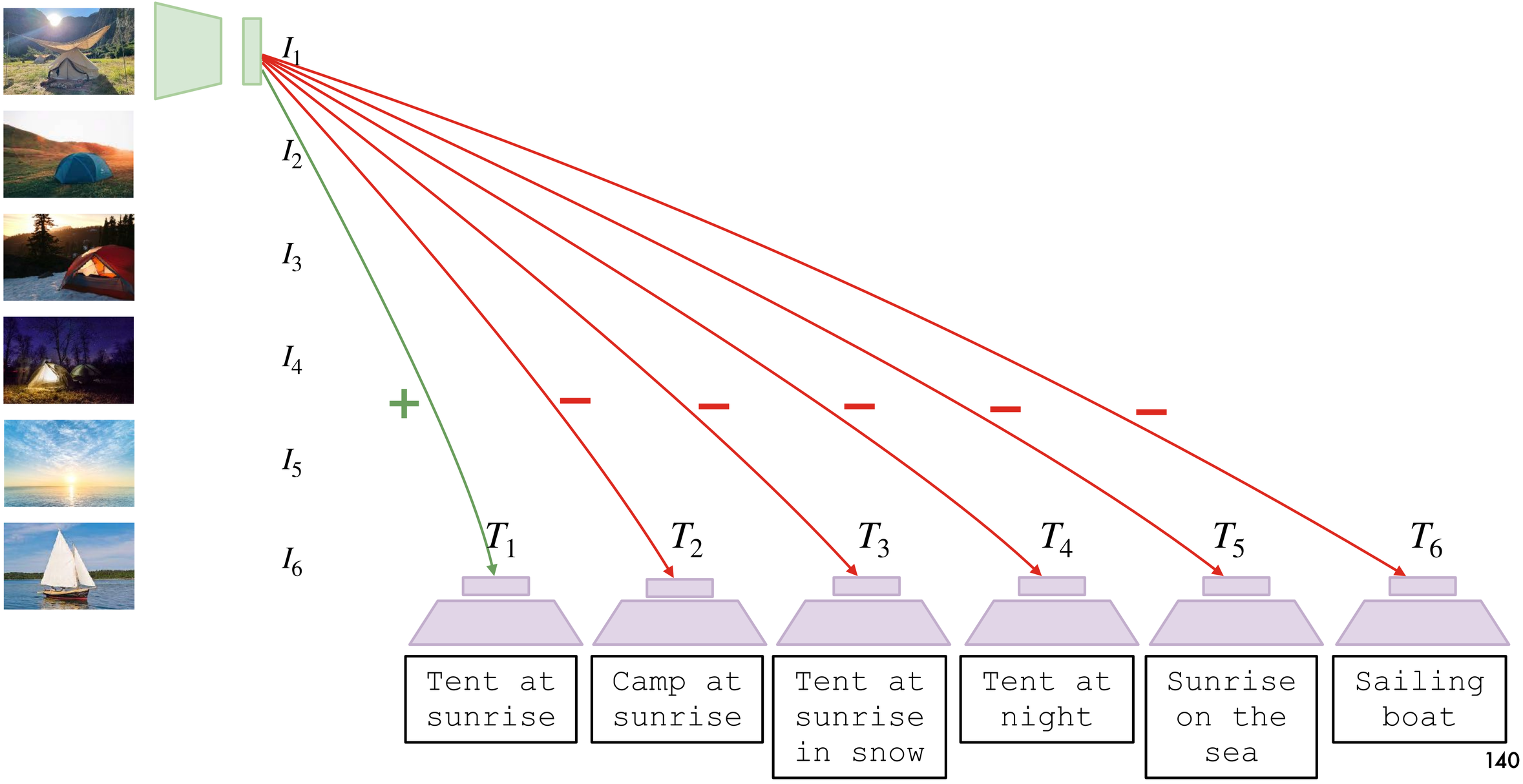
Training data: Text-image pairs (T_i, I_i)

Goal: Learn a joint embedding space

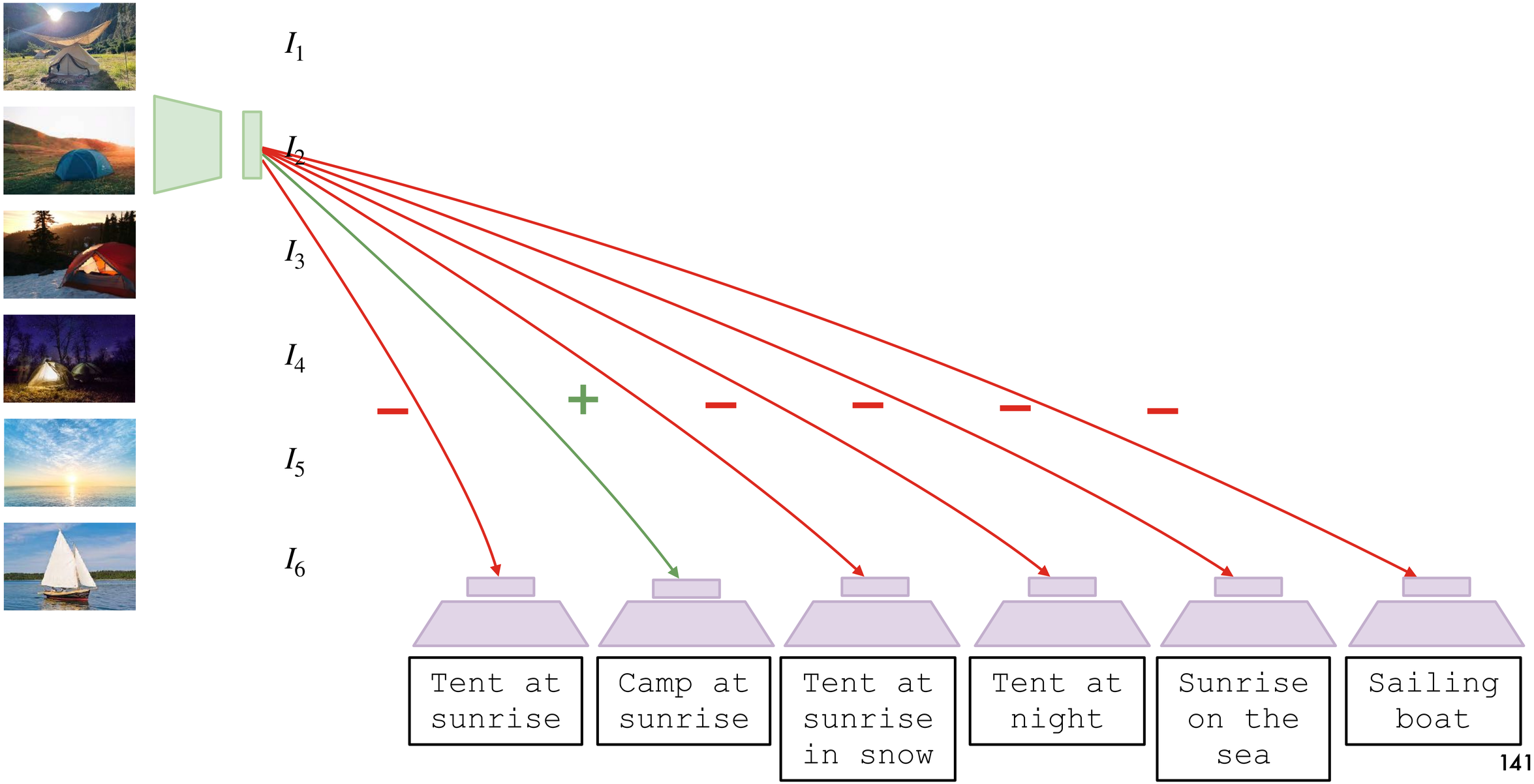




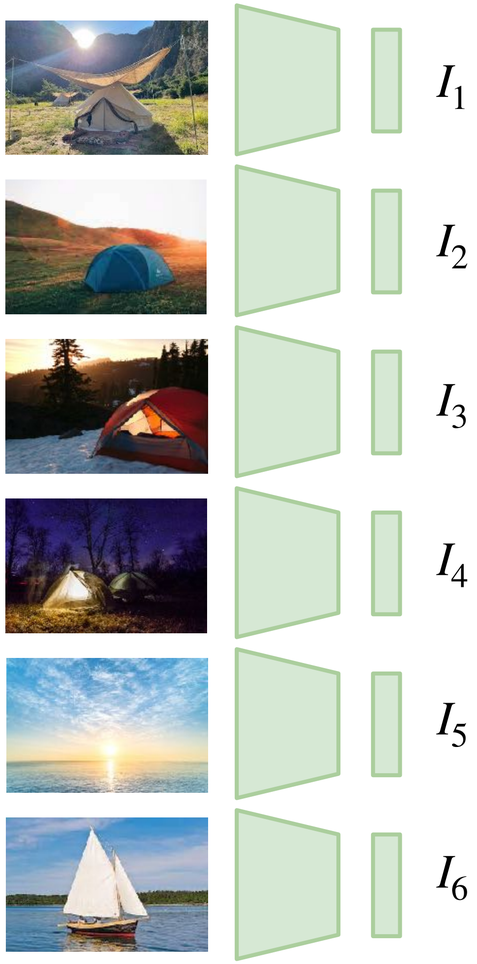
Text-image retrieval: training

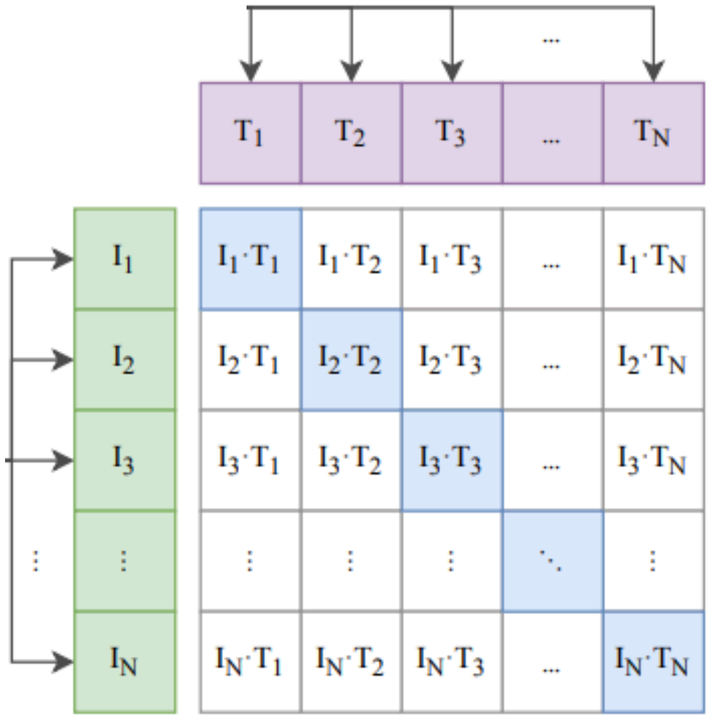


Text-image retrieval: training



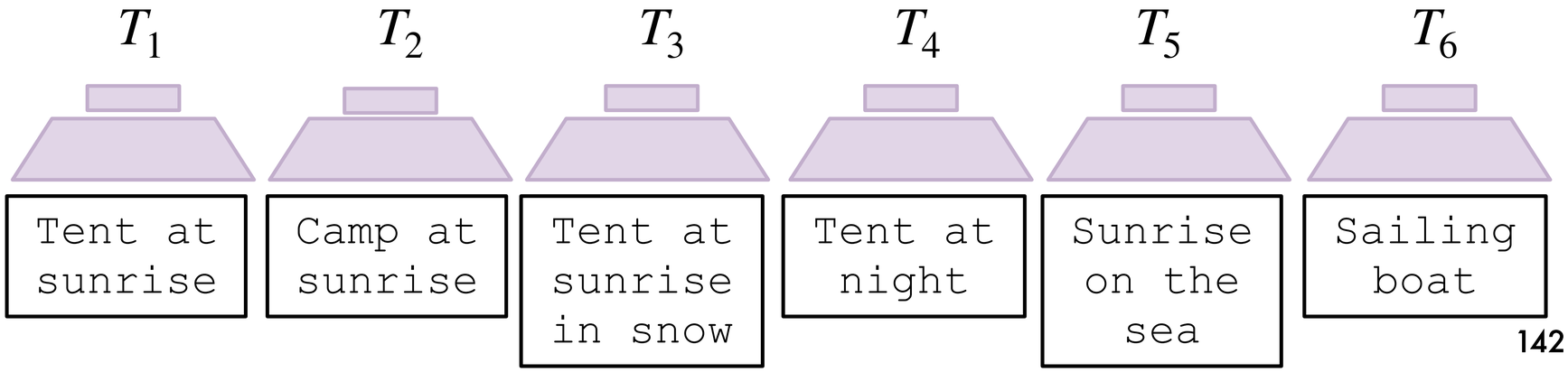
Text-image retrieval: training





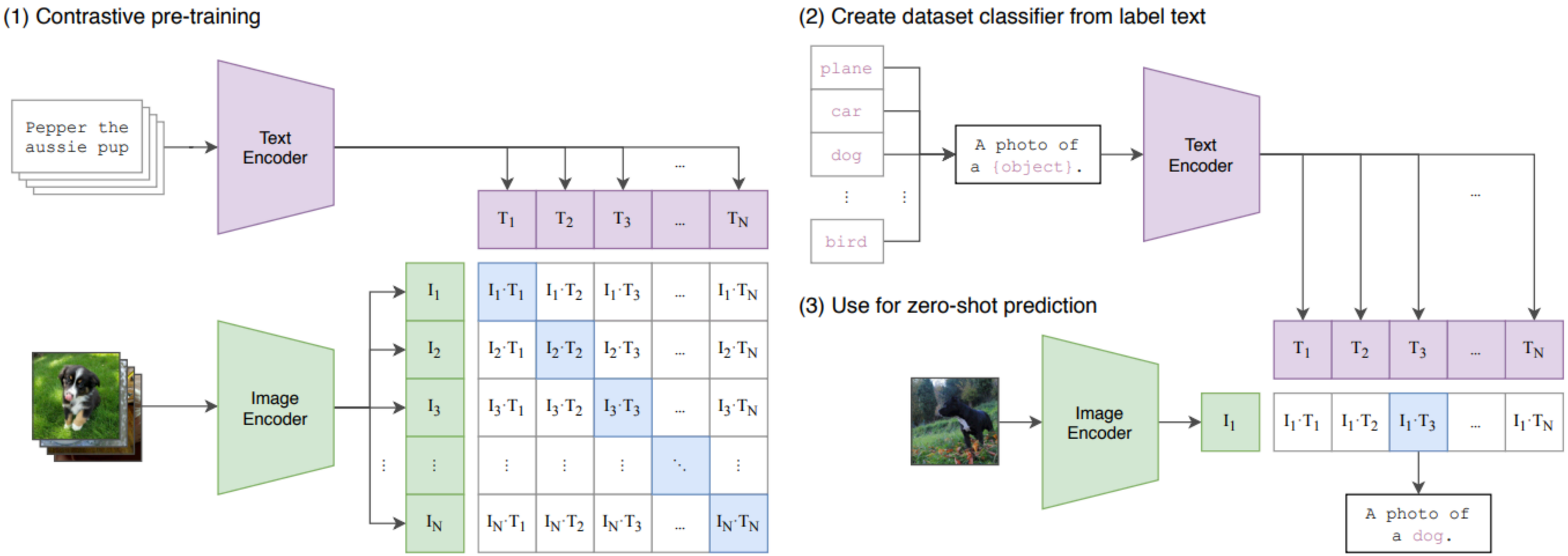
Contrastive objective:
in a batch of N image-text pairs, classify each text to the correct image and vice versa

(aka InfoNCE loss)



[Radford et al., CLIP, ICML 2021]

Contrastive Language-Image Pretraining (CLIP)



Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021 https://openai.com/blog/clip/

CLIP: Details

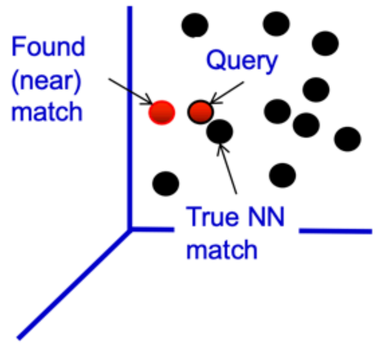
- Image encoders
  - ResNet-50 with self-attention layer on top of global average pooling
  - Vision transformer (ViT)
- Text encoder: GPT-style transformer with 63M parameters
- Dataset: 400M image-text pairs from the Web

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021 https://openai.com/blog/clip/

Remember: Efficient search

Finding *approximate* nearest neighbour vectors

- Approximate method is not guaranteed to find the nearest neighbour.
- Can be much faster, but at the cost of missing some nearest matches



- Approximate nearest neighbor search if the gallery size is millions.

FAISS Library: https://github.com/facebookresearch/faiss

Agenda

1. Generative neural networks

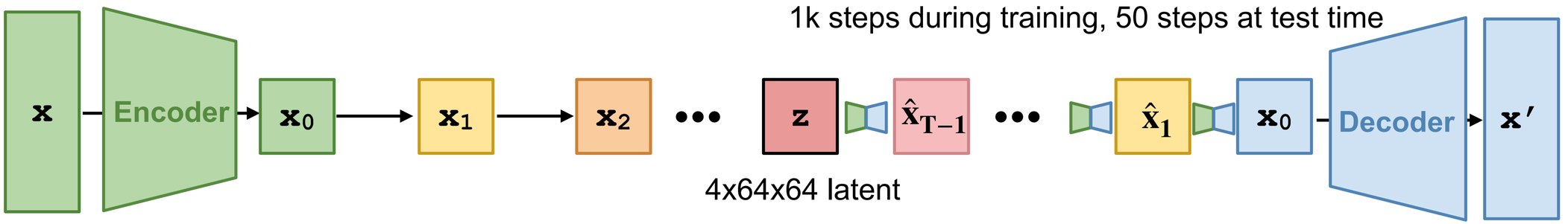
- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

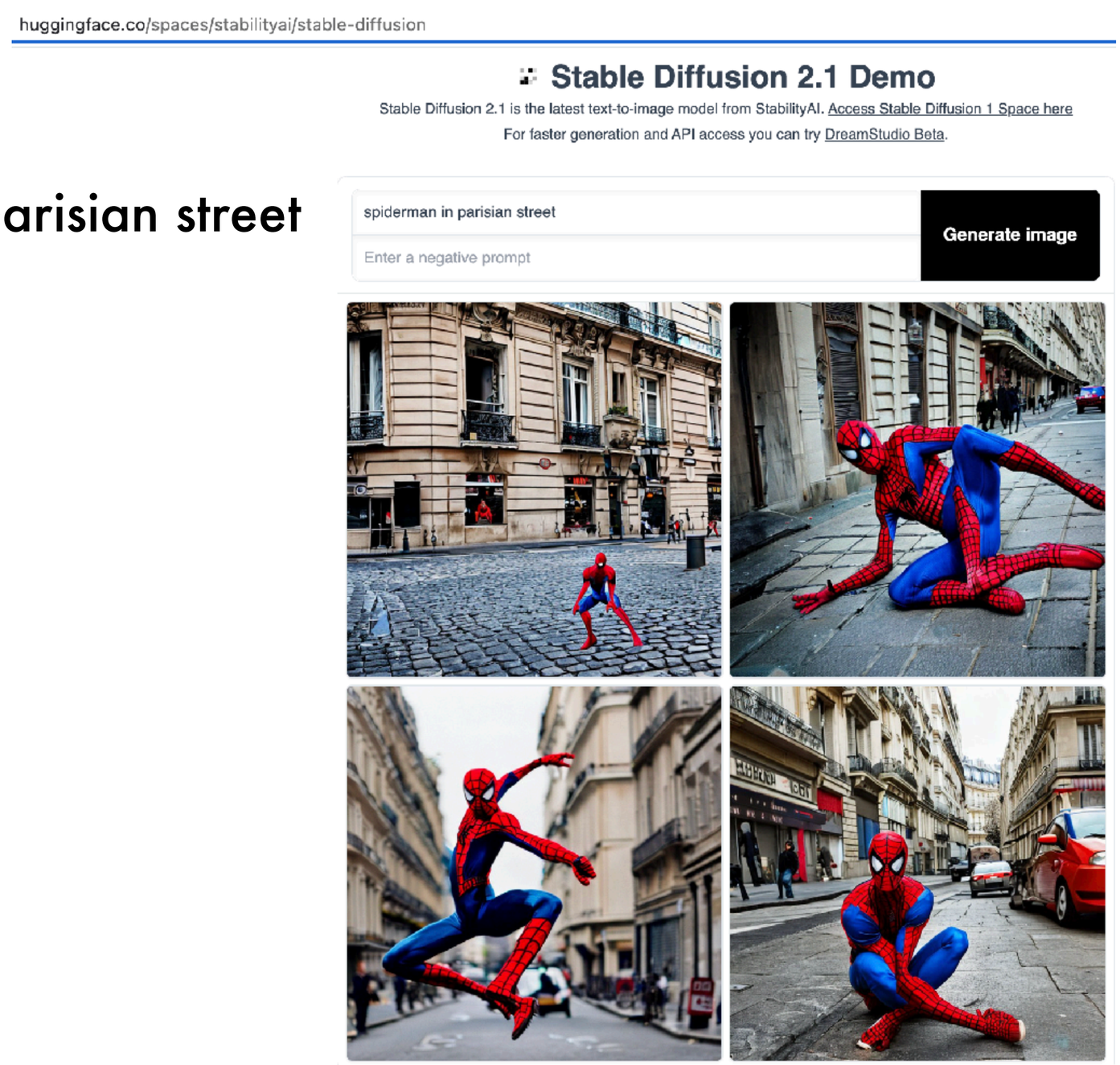
Latent diffusion models (aka Stable Diffusion)

- Demo: https://huggingface.co/spaces/stabilityai/stable-diffusion



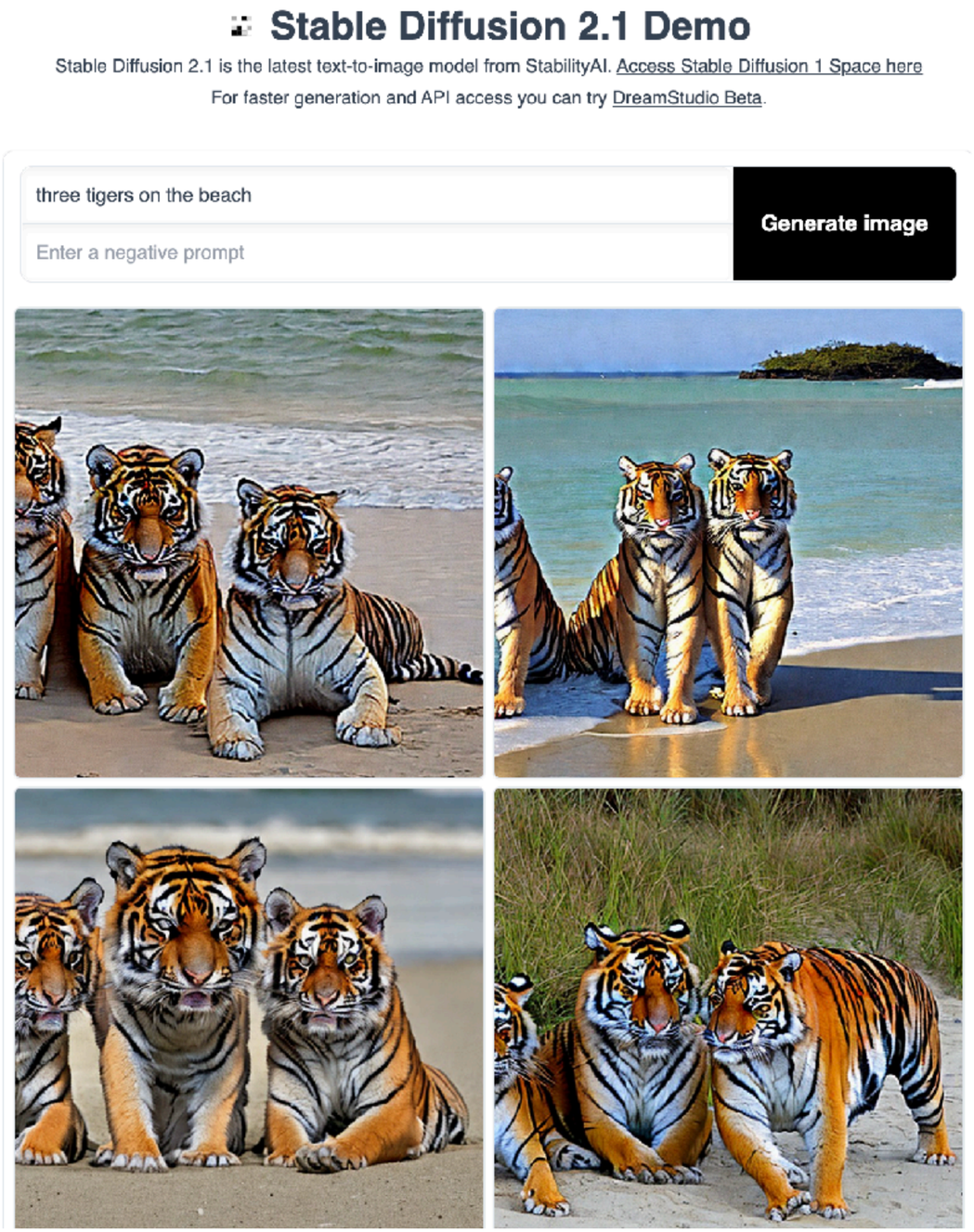
R. Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022

spiderman in parisian street



R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

three tigers on the beach



R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

huggingface.co/spaces/stabilityai/stable-diffusion

Spaces: stabilityai / stable-diffusion like 5.14k Running on CUSTOM ENV

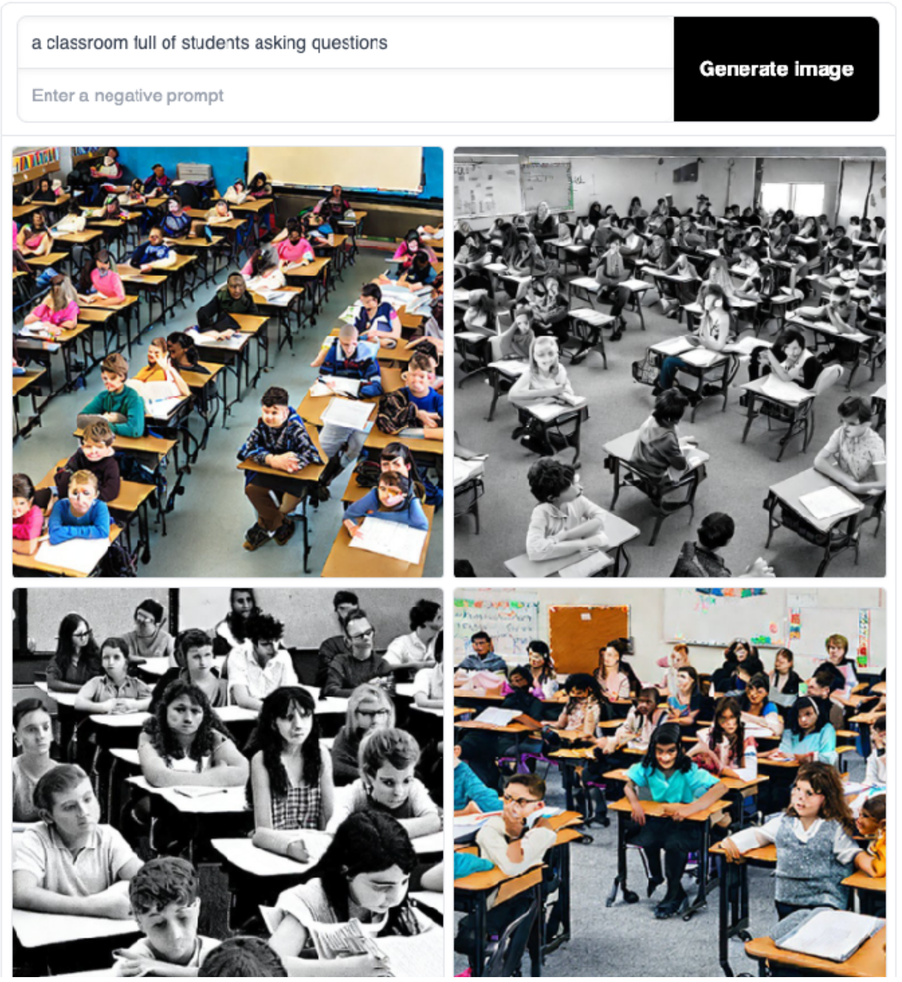
App Files and versions Community 11914

Stable Diffusion 2.1 Demo

Stable Diffusion 2.1 is the latest text-to-image model from StabilityAI. Access Stable Diffusion 1 Space here

For faster generation and API access you can try DreamStudio Beta.

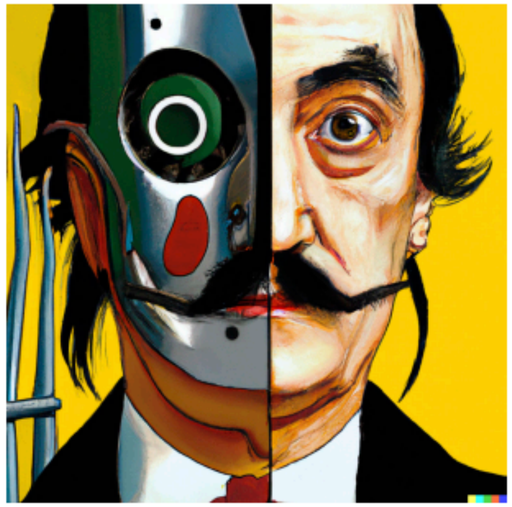
a classroom full of students asking questions



a classroom full of students asking questions

R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

DALL-E-2



vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

A. Ramesh et al. Hierarchical text-conditional image generation with CLIP latents. 2022

DALL-E-2

“A closeup of a handpalm with leaves growing from it.”

Figure 19: Random samples from unCLIP for prompt “A close up of a handpalm with leaves growing from it.”

DALL-E-2

“Vibrant portrait painting of Salvador Dali with a robotic half face”

Figure 18: Random samples from unCLIP for prompt “Vibrant portrait painting of Salvador Dali with a robotic half face”

Imagen

“We discover that **large frozen language models** trained only on text data are surprisingly very effective text encoders for text-to-image generation, and that **scaling the size of frozen text encoder** improves sample quality significantly more than scaling the size of image diffusion model”

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

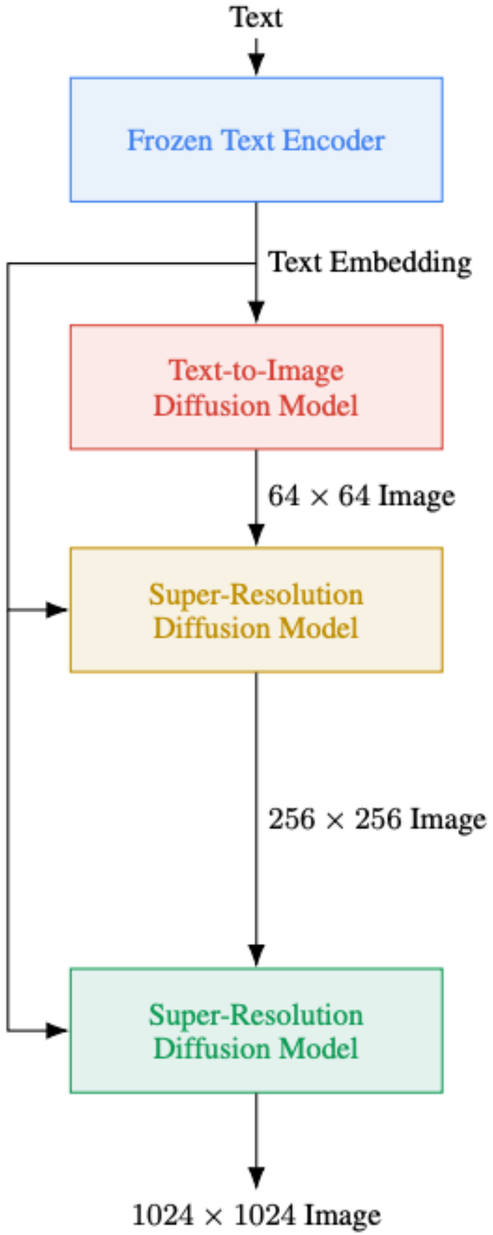
A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

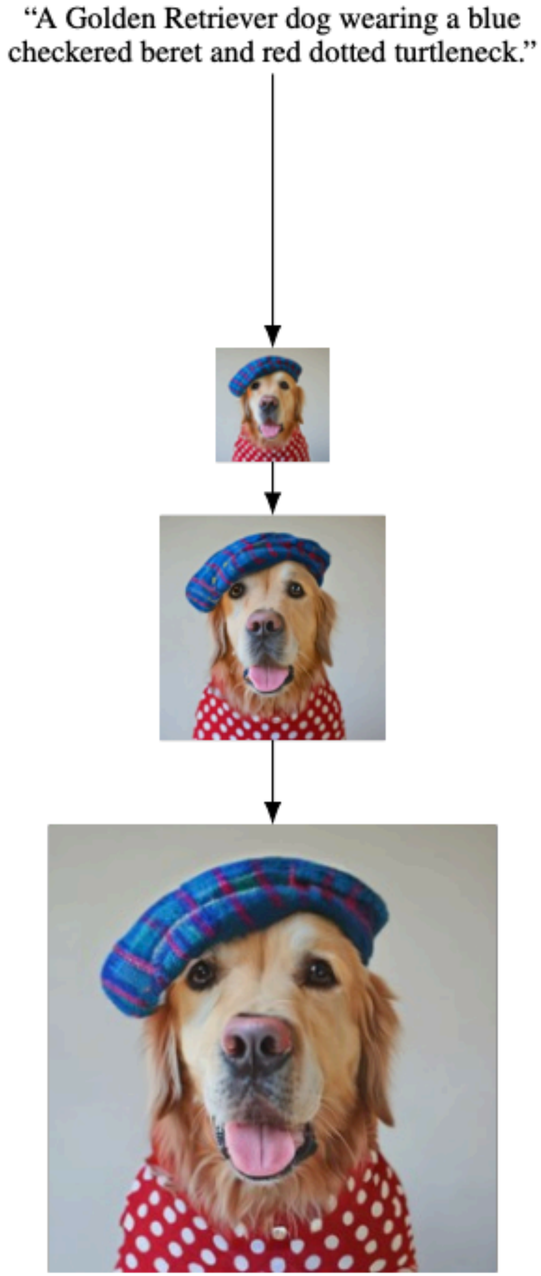
A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

C. Sharia et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. NeurIPS 2022

Imagen

Cascade of conditional diffusion models

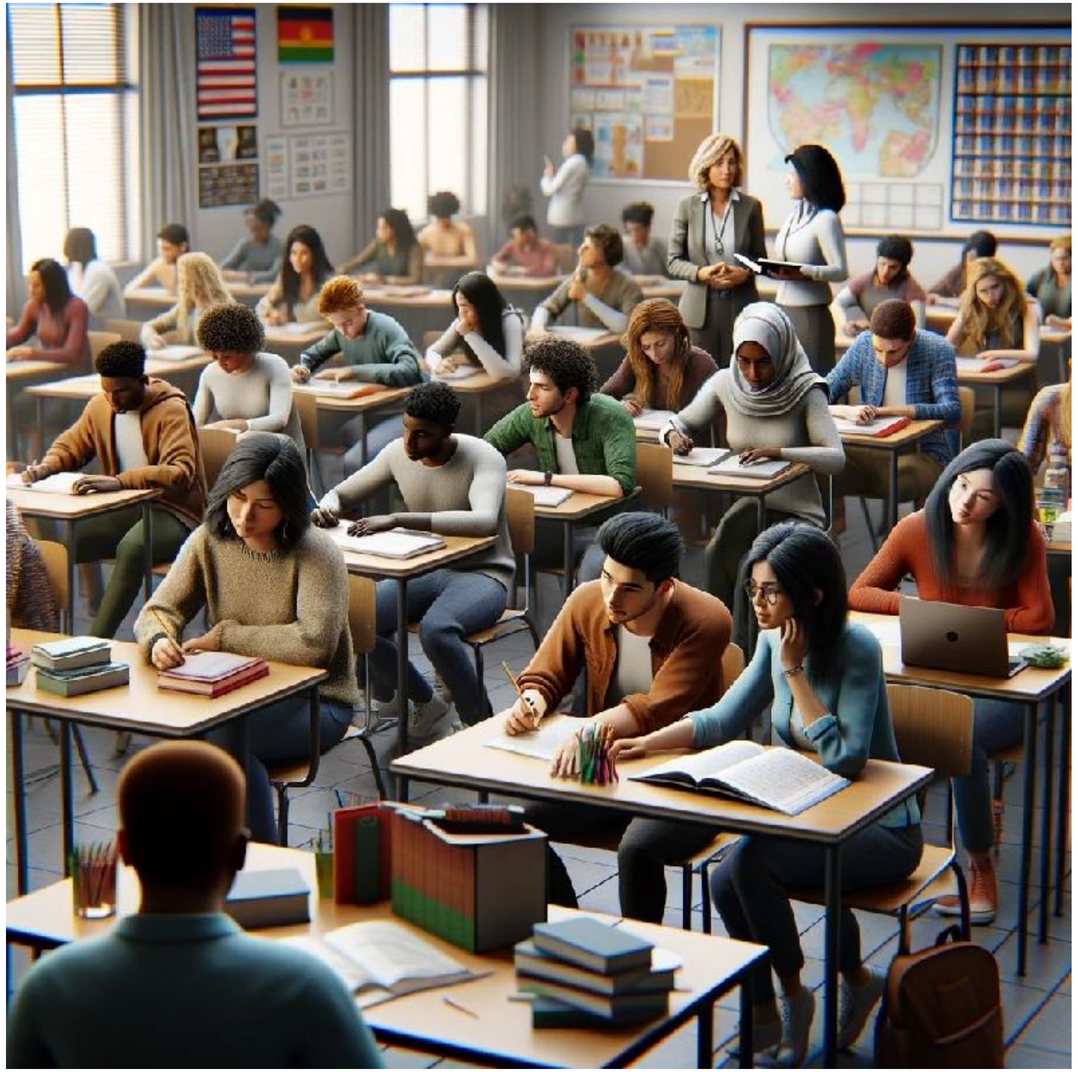




C. Sharia et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. NeurIPS 2022

DALL-E-3

"Classroom full of students"



Betker et al. Improving Image Generation with Better Captions. 2023

DALL-E-3

"People dancing"

"For DALL-E 3, we trained our own diffusion decoder on top of the latent space learned by the VAE trained by Rombach et al. (2022). We found that using a diffusion decoder here provided marked improvements to fine image details, for example text or human faces."

Betker et al. Improving Image Generation with Better Captions. 2023

DALL-E-3

"*Photo of Paris*"

Betker et al. Improving Image Generation with Better Captions. 2023

...

Stable Diffusion

Latent Diffusion

Dec 2021 / Aug 2022 (CVPR'22)

Imagen

Cascaded Diffusion

May 2022 (NeurIPS'22)

...

DALL-E

VQ-VAE + Transformers

Feb 2021 (ICML'21)

...

DALL-E-2

Diffusion + CLIP latents

Apr 2022

...

DALL-E-3

Latent Diffusion + Better captions + Other undocumented improvements

Oct 2023

[DALL-E] A. Ramesh et al., Zero-Shot Text-to-Image Generation, ICML 2021

[StableDiffusion] R. Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022

[Imagen] C. Sharia et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. NeurIPS 2022

[DALL-E-2] A. Ramesh et al. Hierarchical text-conditional image generation with CLIP latents. 2022

[DALL-E-3] Betker et al. Improving Image Generation with Better Captions. 2023

...

Agenda

1. Generative neural networks

- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning

Parenthesis:
(Language Models

(Large) Language models (LLMs)

[GPT] Radford, Narasimhan, Salimans, Sutskever, Improving Language Understanding by Generative Pre-Training, 2018

Before: RNNs, Supervised *GPT: Transformers, Unsupervised*

[BERT] Devlin, Chang, Lee, Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL 2019

Before: Autoregressive encoder-decoder generation *BERT: Non-autoregressive, encoder-only, masked modeling*

[GPT-2] Radford, Wu, Child, Luan, Amodei, Sutskever, Language Models are Unsupervised Multitask Learners, 2019

***1.5B parameter** Transformer + a new dataset of millions of webpages (WebText), SOTA zero-shot results on 7/8 datasets, still underfits WebText*

[T5] Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, JMLR 2020.

***11 billion parameters,** survey-like controlled study, CommonCrawl data*

[GPT-3] Brown, Mann, Ryder, Subbiah, … Radford, Sutskever, Amodei, Language Models are Few-Shot Learners, NeurIPS 2020

***175 billion parameters**, 10x more than any previous non-sparse language model, **trained on 400B tokens** from CommonCrawl data*

[GPT-4] [LlaMa] [LlaMa-2] …

(Large) Language models (LLMs)

[GPT] Radford, Narasimhan, Salimans, Sutskever, Improving Language Understanding by Generative Pre-Training, 2018

Before: RNNs, Supervised *GPT: Transformers, Unsupervised*

Transformers

[BERT] Devlin, Chang, Lee, Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL 2019

Before: Autoregressive encoder-decoder generation *BERT: Non-autoregressive, encoder-only, masked modeling*

[GPT-2] Radford, Wu, Child, Luan, ... pervised Multitask Learners, 2019

1.5B parameter *Transformer + a new ... -shot results on 7/8 datasets, still underfits WebText*

Mostly unsupervised, e.g., next word prediction

[T5] Raffel, Shazeer, Roberts, Lee, N... ts of Transfer Learning with a Unified Text-to-Text Transformer, JMLR 2020.

11 billion parameters, *survey-like controlled study, CommonCrawl data*

[GPT-3] Brown, Mann, Ryder, Subbiah, ... Radford, Sutskever, Amo... rners, NeurIPS 2020

175 billion parameters*, 10x more than any previous non-sparse language m... Crawl data*

Models & data getting bigger

[GPT-4] [LlaMa] [LlaMa-2] ...

(a) Datasets used to train GPT-3

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

"Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset [RSR+19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models ***without ever updating on the same sequence twice****."*

(b) Total Compute Used During Training

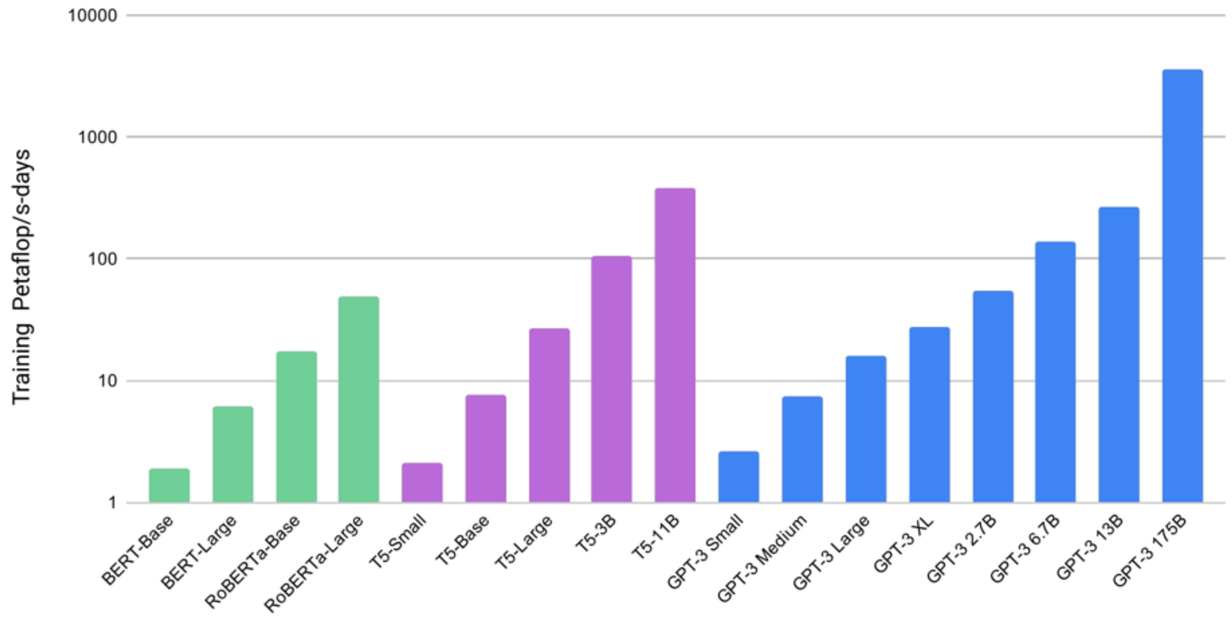


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

Model	Total train compute (PF-days)	Total train compute (flops)	Params (M)	Training tokens (billions)
T5-Small	2.08E+00	1.80E+20	60	1,000
T5-Base	7.64E+00	6.60E+20	220	1,000
T5-Large	2.67E+01	2.31E+21	770	1,000
T5-3B	1.04E+02	9.00E+21	3,000	1,000
T5-11B	3.82E+02	3.30E+22	11,000	1,000
BERT-Base	1.89E+00	1.64E+20	109	250
BERT-Large	6.16E+00	5.33E+20	355	250
RoBERTa-Base	1.74E+01	1.50E+21	125	2,000
RoBERTa-Large	4.93E+01	4.26E+21	355	2,000
GPT-3 Small	2.60E+00	2.25E+20	125	300
GPT-3 Medium	7.42E+00	6.41E+20	356	300
GPT-3 Large	1.58E+01	1.37E+21	760	300
GPT-3 XL	2.75E+01	2.38E+21	1,320	300
GPT-3 2.7B	5.52E+01	4.77E+21	2,650	300
GPT-3 6.7B	1.39E+02	1.20E+22	6,660	300
GPT-3 13B	2.68E+02	2.31E+22	12,850	300
GPT-3 175B	3.64E+03	3.14E+23	174,600	300

[GPT-3] Brown, Mann, Ryder, Subbiah, Kaplan, … Radford, Sutskever, Amodei, Language Models are Few-Shot Learners

Reviews of the GPT-3 paper

https://papers.nips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Review.html

Review 4

Summary and Contributions: The paper shows scaling up language models can achieve task-agnostic few-shot performances on various NLP tasks. Besides other promising results on various tasks and examples, this paper has a clear contribution to the community; industry-level, heavy engineering efforts, and their analyses on various aspects. I do appreciate such efforts and empirical findings described in the paper.

Strengths: The paper has the following strengths: (1) A comprehensive analysis has been made to evaluate the model in terms of task-agonistic behavior, memorization, and weakness based on the prior works/critiques made on the previous version of this work. (2) The empirical observations about the few-shot models' capabilities are made (Figure 1.1 and Section 3.4), showing the scaling effects, limitations, and upper-bound of few-shot settings. (3) Experiments on different tasks in various applications indicate how much existing datasets/tasks are getting benefit from these huge-size language models and what kinds of operations (e.g., bidirectional, reasoning, external knowledge) should be done in the future toward that direction. (4) I pretty much enjoyed reading the Broader Impact section, which tries to adopt the feedbac
fairness and bias, energy usage, and potential misuse of the model. (5) For a perspective of usefulness, I think this c

Weaknesses: Improvements from few shot LMs are not that surprising because it is mainly because the model uses
not fair to compare with relatively smaller models, the improvements themselves may not be the main contribution of
engi
coul
few s
artifi
rathe
Compared to GPT2, the only difference made in this work is scaling up the training in terms of data size and compari
one-shot, and zero-shot. What scientific values does this paper bring to our community except for empirical observa
on NaturalQS shows GPT3 mean that it does not include any external knowledge in Wikipedia and their appropriate r
Figure 1.1., let's say you use 175B x 1000 parameters, do you think the improvement from {zero,one,few}-shots still li
of contexts increases, the degree of improvements from the few-shot GPT3 seems to be not that steep. Does this in
around K=100 or something? Also, please show me the zero/one-short cases as well.

Improvements … not that surprising … mainly because … more training data/parameters/computing resources. … an empirical paper with huge engineering efforts

Correctness: Please see many comments above.

Clarity: Yes, the paper is written well and easy to follow.

Relation to Prior Work: Yes.

Reproducibility: Yes

Review 2

Summary and Contributions: In this paper, the authors empirically demonstrate that increasing the model size -- in term of depth and width, and thus number of parameters -- of language models (LM) result in a better task-agnostic learner, which can zero/one/few-shots multiple well-know NLP tasks. - The authors use the same transformer base architecture as GPT-2, except for the Sparse attention (Child, et.al. 2019), which improve the model efficiency. They trained 8 models from 125 M to 175 B parameters to study the effect of the model size in the zero/one/few-shots settings. - The authors train the LM using 300 billion tokens from 5 sources (i.e., Common Crawl, WebText2, Book Collection 1 and 2, Wikipedia). - The authors evaluate the models' performance in a zero/one/few-shot setting on a large variety of NLP tasks such as LM perplexity, QA, CQA, SuperGLUE, MT, etc. Importantly, the zero/one/few-shots is done without fine-tuning the model, but by providing as context --priming-- the task-description (i.e., for zero-shot) or pairs of examples (one/few-shots), and making the model auto-regressively generate the response. As also clearly stated by the authors, this approach is not novel, since also GPT-2 used the same mechanism, but in this paper, the author extended the evaluation to way more tasks and showed that by increasing the model size the few-shot ability of the model greatly increases. - The authors compare the performance of the model to the current state-of-the-art and they highlight the advantages and disadvantages of the proposed model. I really appreciated the openness of the authors when they described their results, avoiding st
audience and the broader impact of the paper is clea
not an expert in the bias/ethical ML, but in my perspe

although not novel per se (i.e. GPT-2), … can have a big impact

Strengths: - the zero/one/few-shots methodology is
includes a large variety of NLP tasks and SOTA baseli
found hard (52% accuracy) to recognise which article is written by humans or the GPT-3 model (175B)

Weaknesses: - the authors already discussed most of the limitation of the current model (e.g. missing of bi-directional attection etc.). I found that one limitation could be the length of the context when increasing the number of shots. To elaborate, in some tasks (e.g. QA, summurization) where the input are entire articles, going beyond the 25/30 shots would be very challenging. GPT-3 already double (2048 tokens) the context size compare to GPT-2, but scaling very long inputs remains challenging, both in term of memory consumption and models inference (although the authors already use Sparse Transformer).

Correctness: Yes.

Clarity: Yes.

Relation to Prior Work: yes, to the best they could do in 8 pages. I think the citation format is not the NeurIPS 2020 template, but this can be easily change in the camera ready.

Reproducibility: Yes

Additional Feedback:

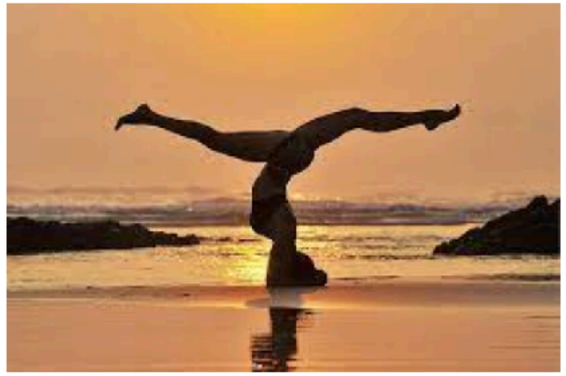
[GPT-3] Brown, Mann, Ryder, Subbiah, Kaplan, … Radford, Sutskever, Amodei, Language Models are Few-Shot Learners

Parenthesis Closed:
Language Models)

Image Captioning: Image in, Text out

A politician receives a gift from politician.

A collage of different colored ties on a white background.



Silhouette of a woman practicing yoga on the beach at sunset.

Aerial view of a road in autumn.

Mokady et al., ClipCap: CLIP Prefix for Image Captioning, arXiv 2021

ClipCap: CLIP Prefix for Image Captioning

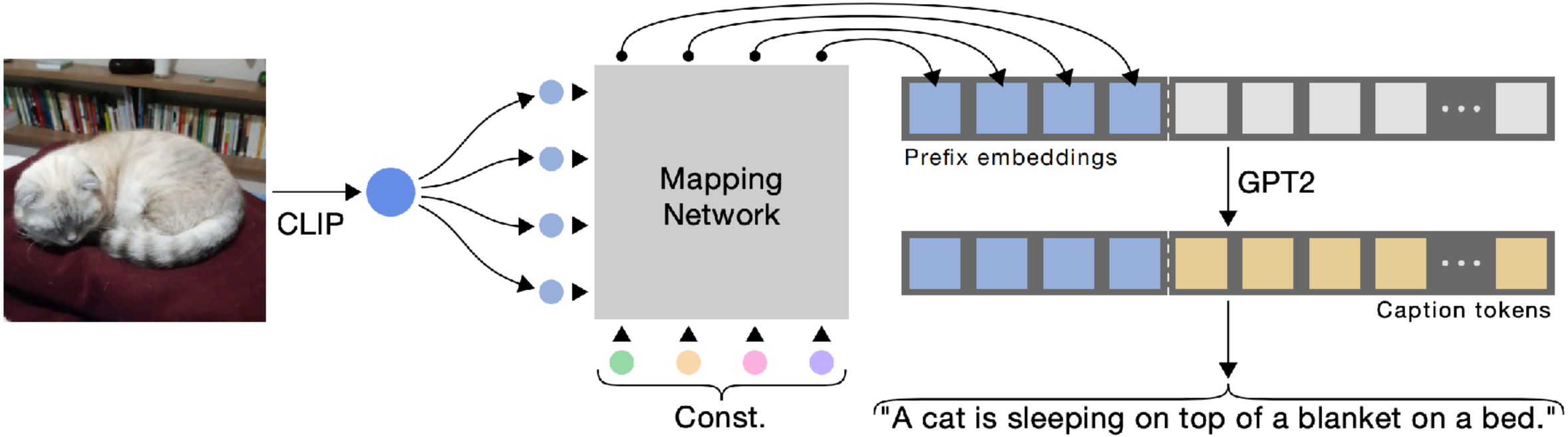
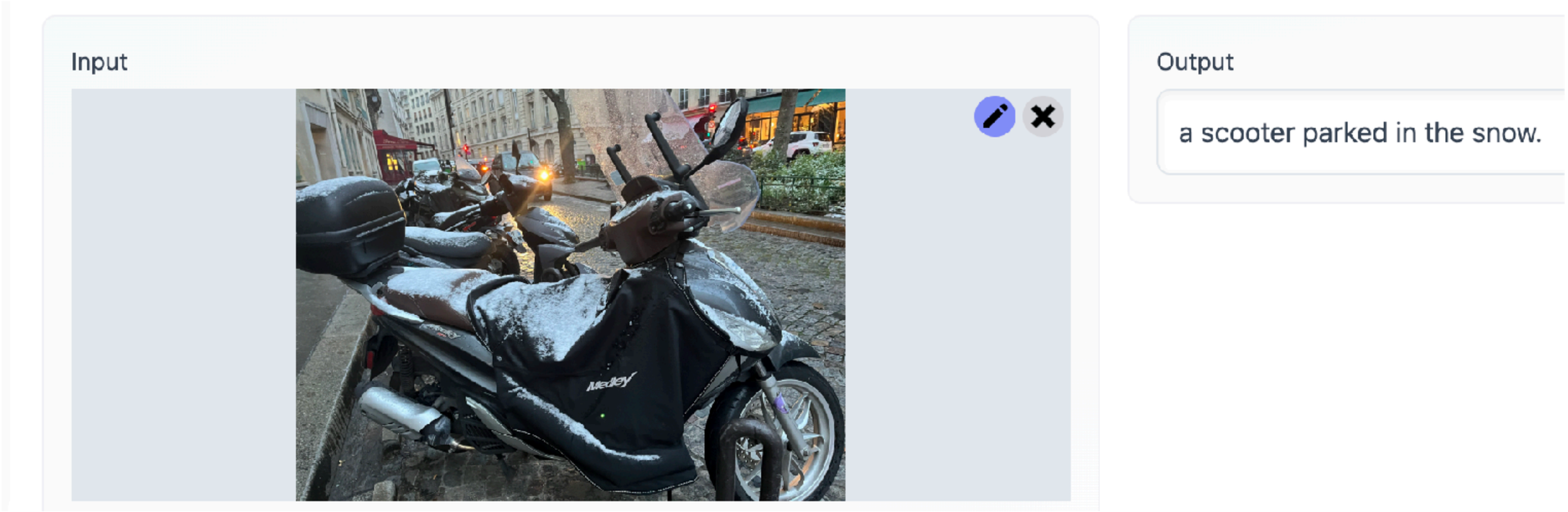


Figure 2. Overview of our transformer-based architecture, enabling the generation of meaningful captions while both CLIP and the language model, GPT-2, are frozen. To extract a fixed length prefix, we train a lightweight transformer-based mapping network from the CLIP embedding space and a learned constant to GPT-2. At inference, we employ GPT-2 to generate the caption given the prefix embeddings. We also suggest a MLP-based architecture, refer to Sec. 3 for more details.

Mokady et al., ClipCap: CLIP Prefix for Image Captioning, arXiv 2021

ClipCap: CLIP Prefix for Image Captioning

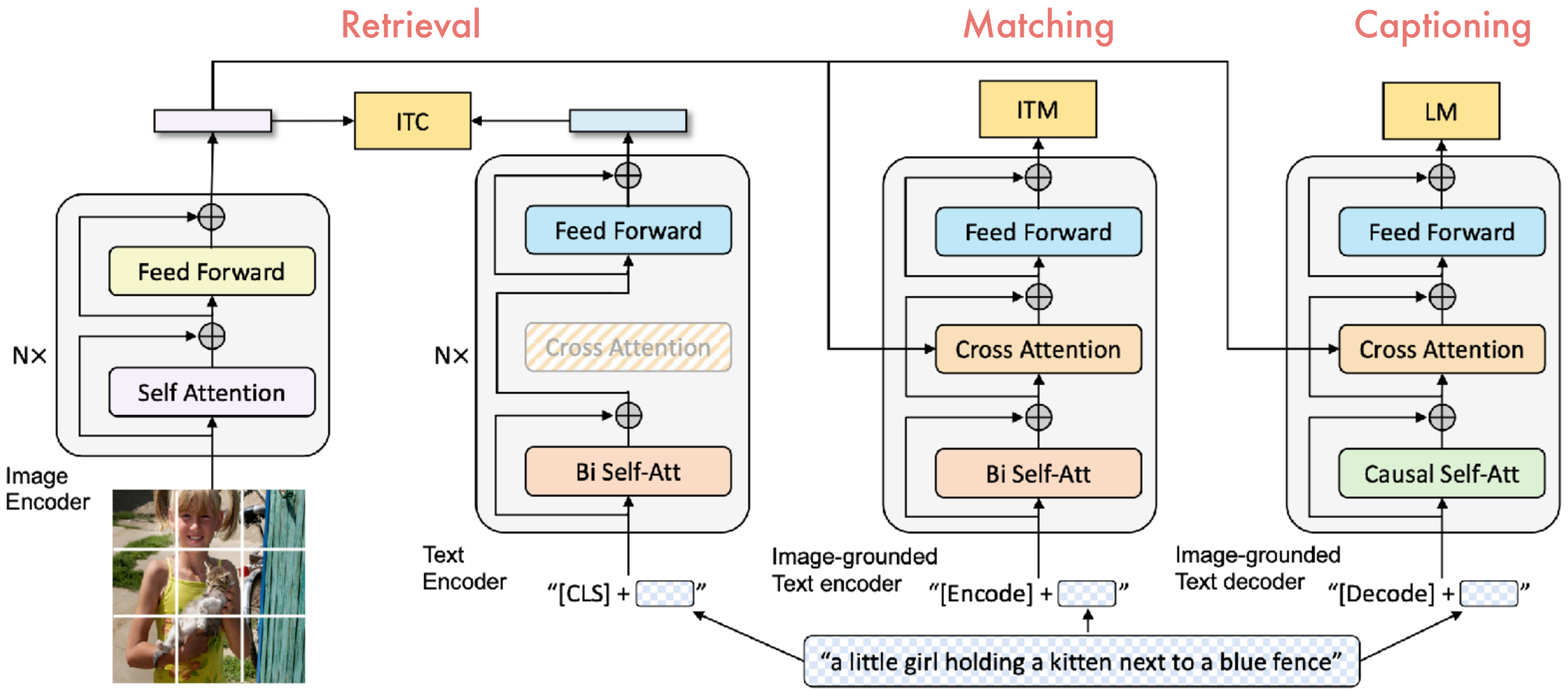
- Demo: https://huggingface.co/spaces/akhaliq/CLIP_prefix_captioning



Mokady et al., ClipCap: CLIP Prefix for Image Captioning, arXiv 2021

BLIP

Retrieval + Captioning

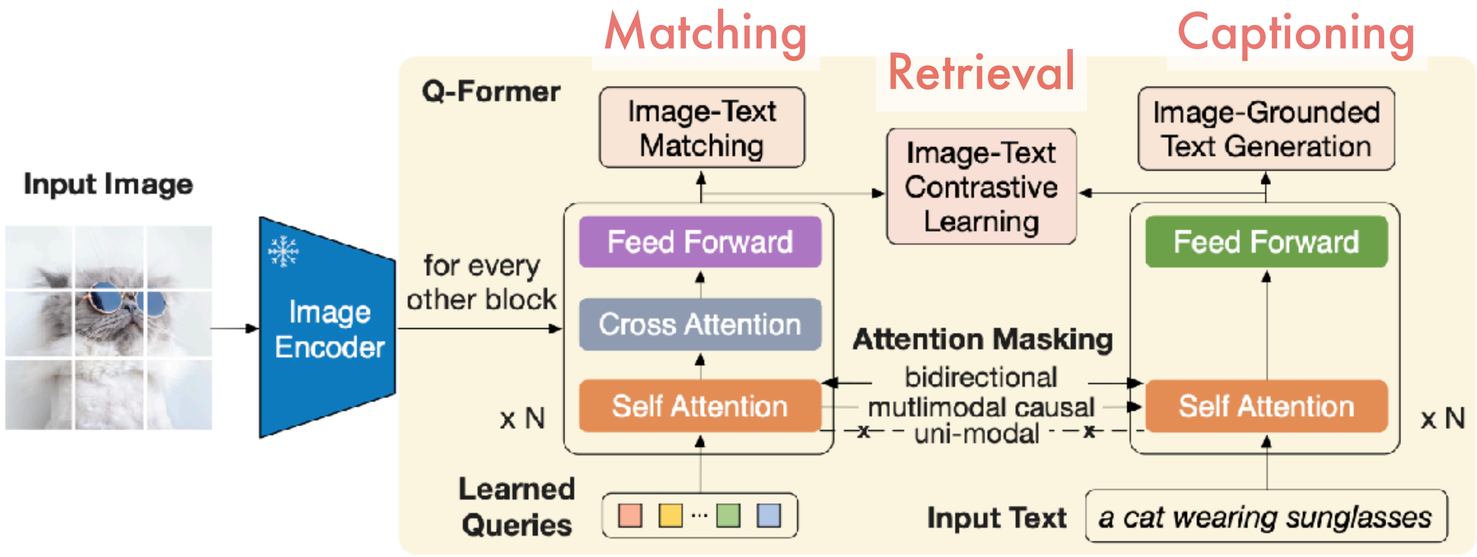


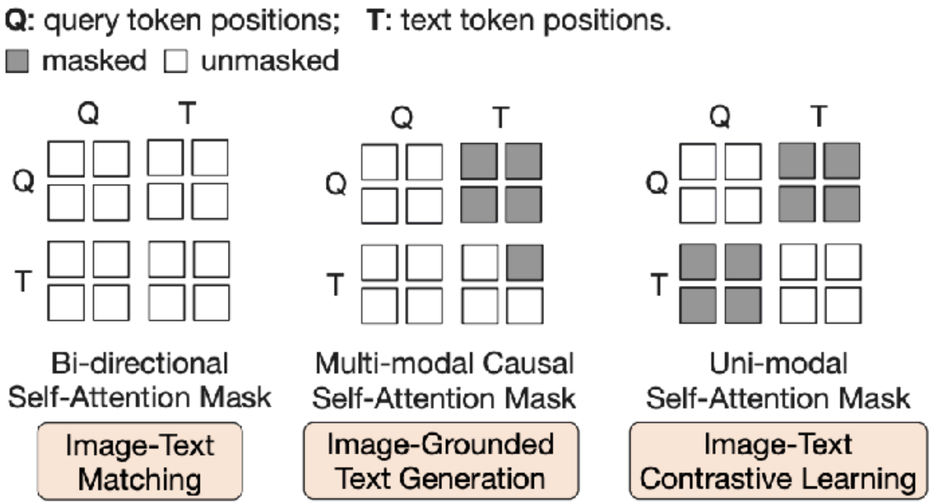
Li et al., BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, ICML 2022

BLIP-2

Retrieval + Captioning with "Q-Former" architecture

Stage 1 training (similar to BLIP-1): **vision-text representation learning**





BLIP-2

Retrieval + Captioning with "Q-Former"

Stage 2 training with a frozen LLM to **generate text**

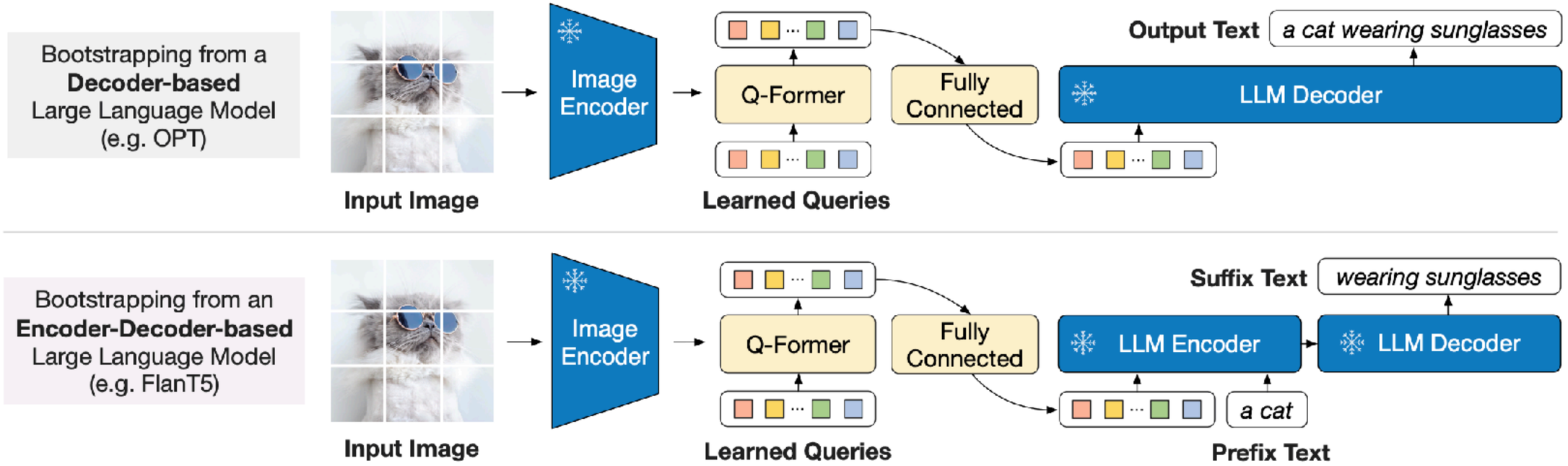


Figure 3. BLIP-2's second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). (**Top**) Bootstrapping a decoder-based LLM (e.g. OPT). (**Bottom**) Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.

BLIP1 & BLIP2 Training Data

Similar to CLIP
129M image-text pairs
with automatic captioning + filtering

- Visual Genome
- CC3M*
- CC12M*
- COCO
- LAION400M (115M)*
- SBU*

*web datasets

Li et al., BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, ICML 2022
Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, ICML 2023

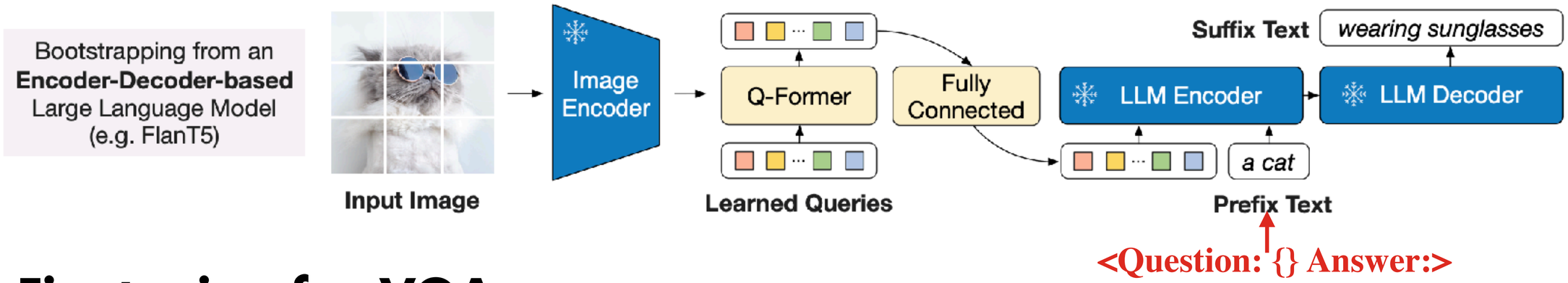
Bonus: VQA in 1 slide

Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning
- Visual question answering (VQA)

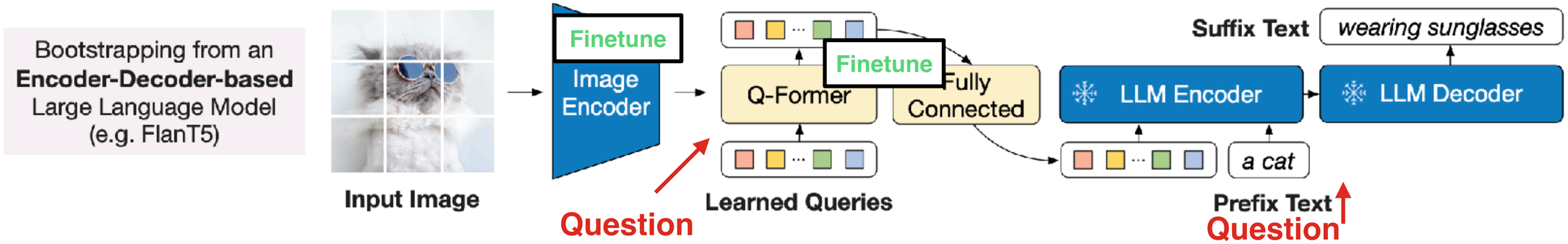
BLIP2 for Visual Question Answering

(a) Zero-shot (no finetuning for VQA)



(b) Finetuning for VQA

"Given annotated VQA data, we finetune the parameters of the Q-Former and the image encoder while keeping the LLM frozen. …
LLM receives Q-Former's output and the question as input, and is asked to generate the answer.
In order to extract image features that are more relevant to the question, we additionally condition Q-Former on the question."



Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, ICML 2023

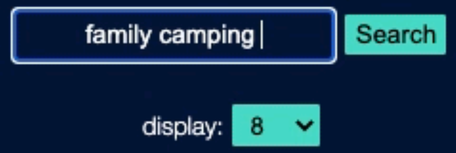
Bonus: Examples from our works

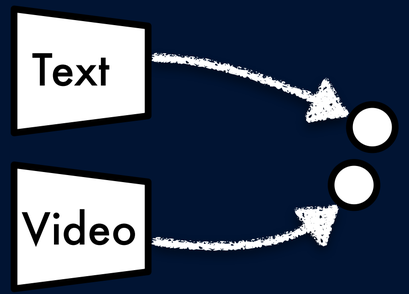
Retrieval tasks

Text Queries for Search:

[Bain, Nagrani, Varol, Zisserman, "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval", ICCV 2021]

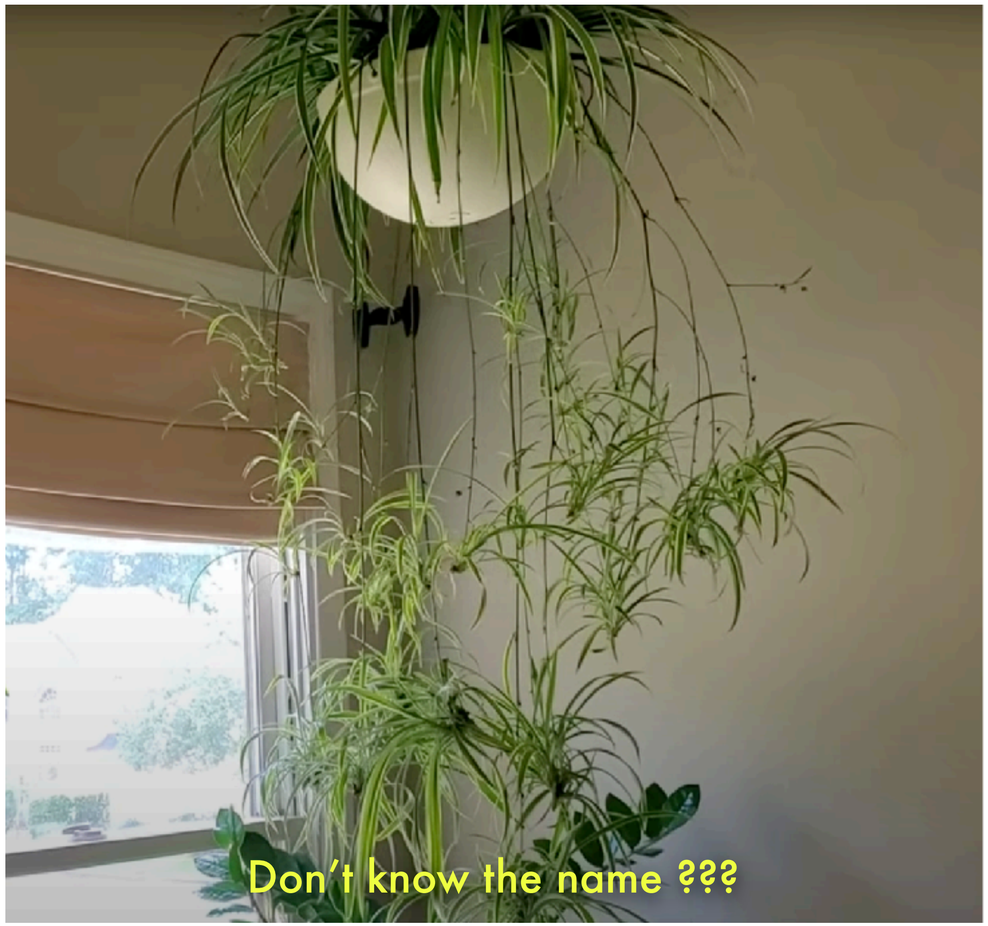
https://www.robots.ox.ac.uk/~vgg/research/frozen-in-time/

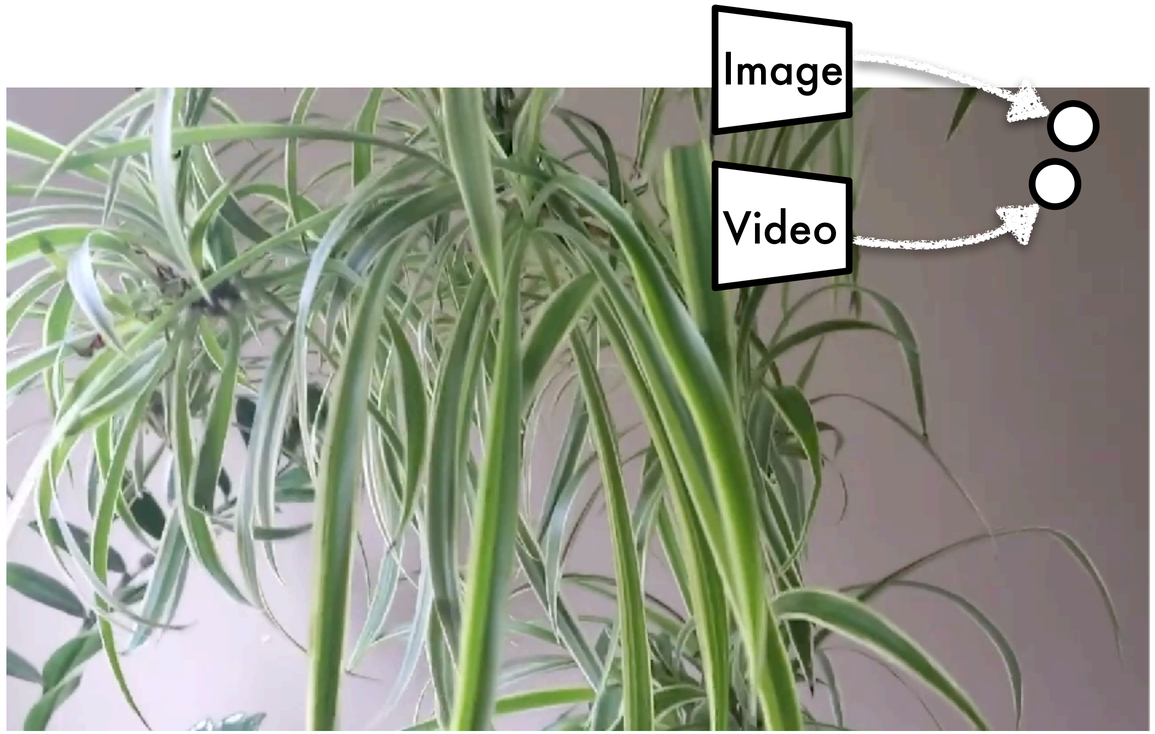




Efficient search with https://github.com/facebookresearch/faiss

Image Queries for Search:

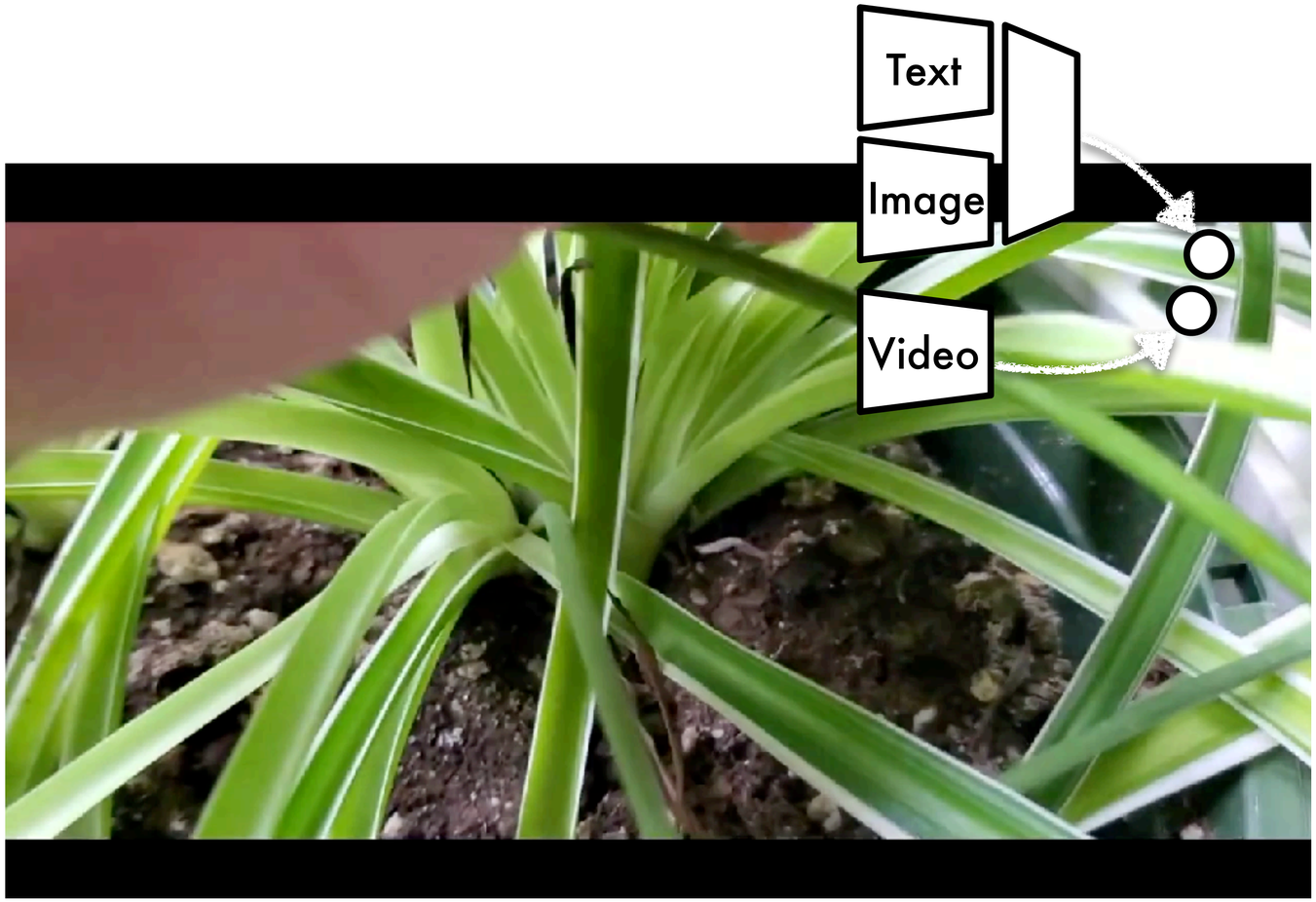




Spoiler: *Chlorophytum comosum* (aka “spider plant”)

Image+Text Queries for Search:

"prune this plant"



Ventura, Yang, Schmid, Varol, CoVR: Learning Composed Video Retrieval from Web Video Captions, AAAI 2024 (TPAMI 2024).

Problem: Data? No annotated image-text-video triplets.

CoVR: Learning Composed Video Retrieval from Web Video Captions

Large video-caption database


Moscow, russia - april 1, 2017: customers watch quadrocopters at the opening of dji authorized store


Path in alpine meadows


Futuristic modern magic dynamic animation with dark blue background with rotation...


Kyiv, ukraine – august 20, 2018. fans at the stadiumKyiv, ukraine – august 20, 2018. fans at the stadium

Happy chinese new year, 2020. new year festival for chinese people all over the world.


Caucasian joyful little girl in festive mood jumping on sofa holding xmas gift in hands exciting about holiday present...


Hacker and security hd animation


Novosibirsk, russia, march 18, 2018: fitness festival "zumba marathon". instructors and athletes from siberia...


Happy couple posing in the park on a sunny day


Lake hubsugul, mongolia - june 12, 2015: shaman obo and animal remains as sacrifices on stones...


Noosa, queensland / australia - 2 mar 2018: a young man and a young woman and kids children...


Close-up of kohlrabi vegetable on wooden table 4k


Damneon saduak, thailand, - december, 25, 2019: vendors selling their products to tourists in the famous floating markets, 100 kilometres...


Fancy-dress mature man raising arms and jumping on the beach


Circa 1944 - a secretary hits her typewriter and, later, takes a screwdriver to it and an office...

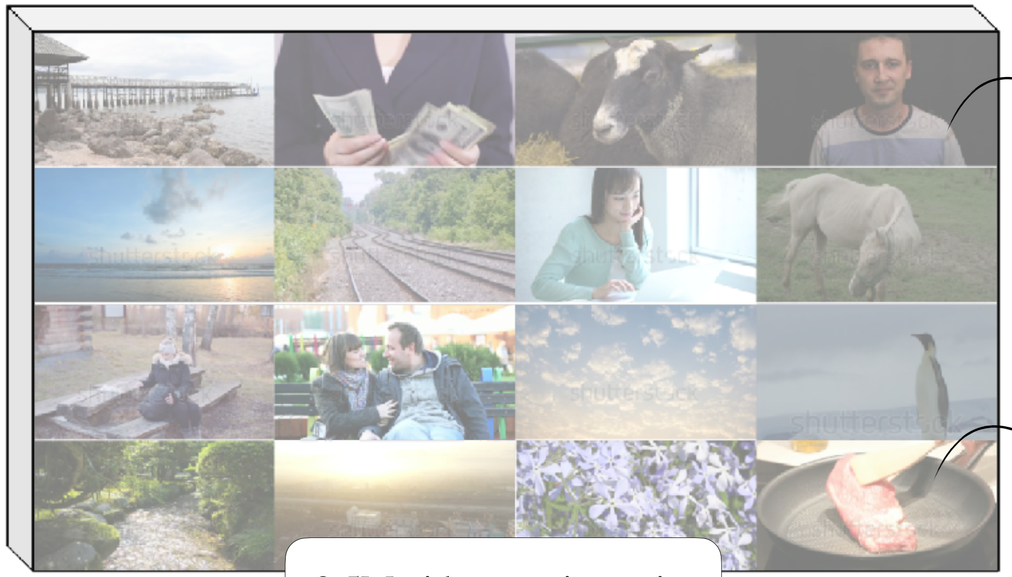

Cute baby boy touch fountain streams and turn the satisfied face slow motion


Berlin - august 15: the friedrichstrasse station this crossing station is one of the most important hubs in berlin. on august 15,2017Berlin - august 15...


Porto, portugal. aerial view of the old city with promenade of the douro river at sunset

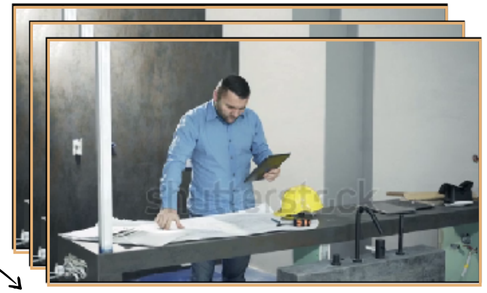
Ventura, Yang, Schmid, Varol, CoVR: Learning Composed Video Retrieval from Web Video Captions, AAAI 2024 (TPAMI 2024).

Mining similar caption pairs & Modification text generation (MTG)



2.5M video-caption pairs

WebVid



q

Young man working with **tablet** and blueprints at his new home

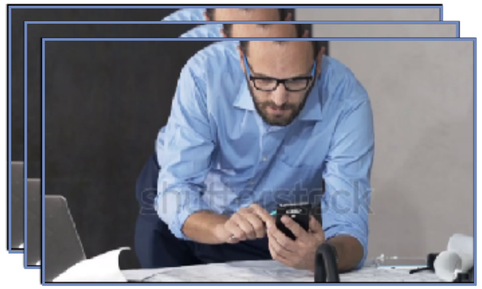
Young man working with **smartphone** and blueprints at his new home

MTG-LLM

Trained with 715 text triplet annotations

change tablet for smartphone

t

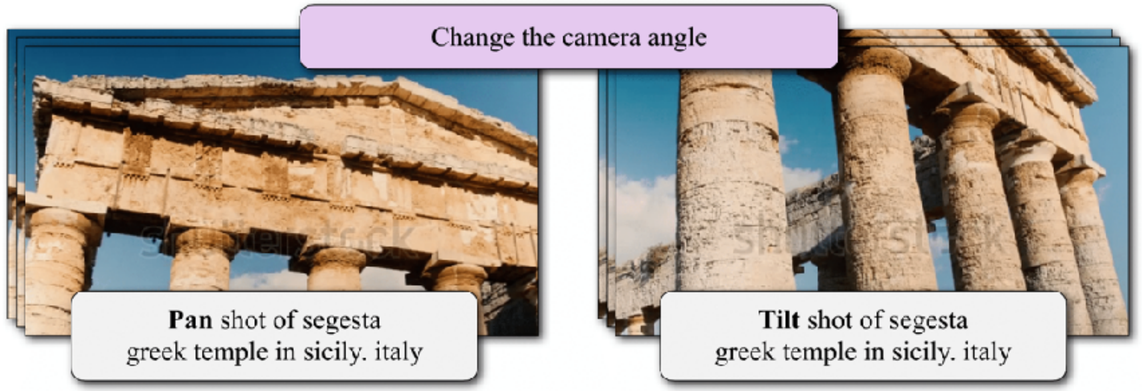


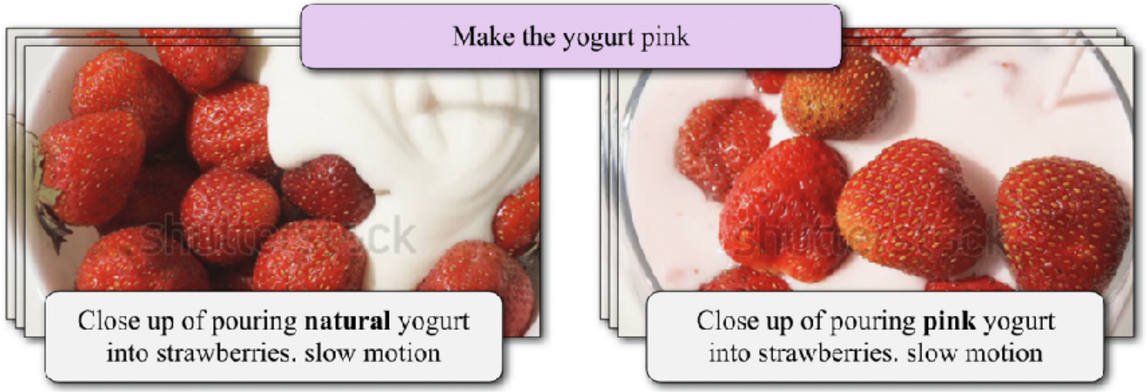
v

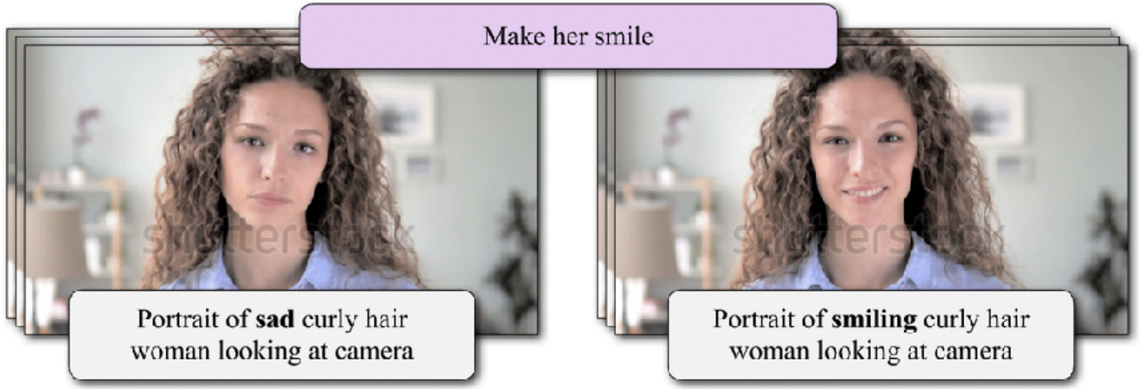
["WebVid", Bain, Nagrani, Varol, Zisserman, ICCV 2021]

Ventura, Yang, Schmid, Varol, CoVR: Learning Composed Video Retrieval from Web Video Captions, AAAI 2024 (TPAMI 2024).

WebVid-CoVR: 1.6M generated triplets

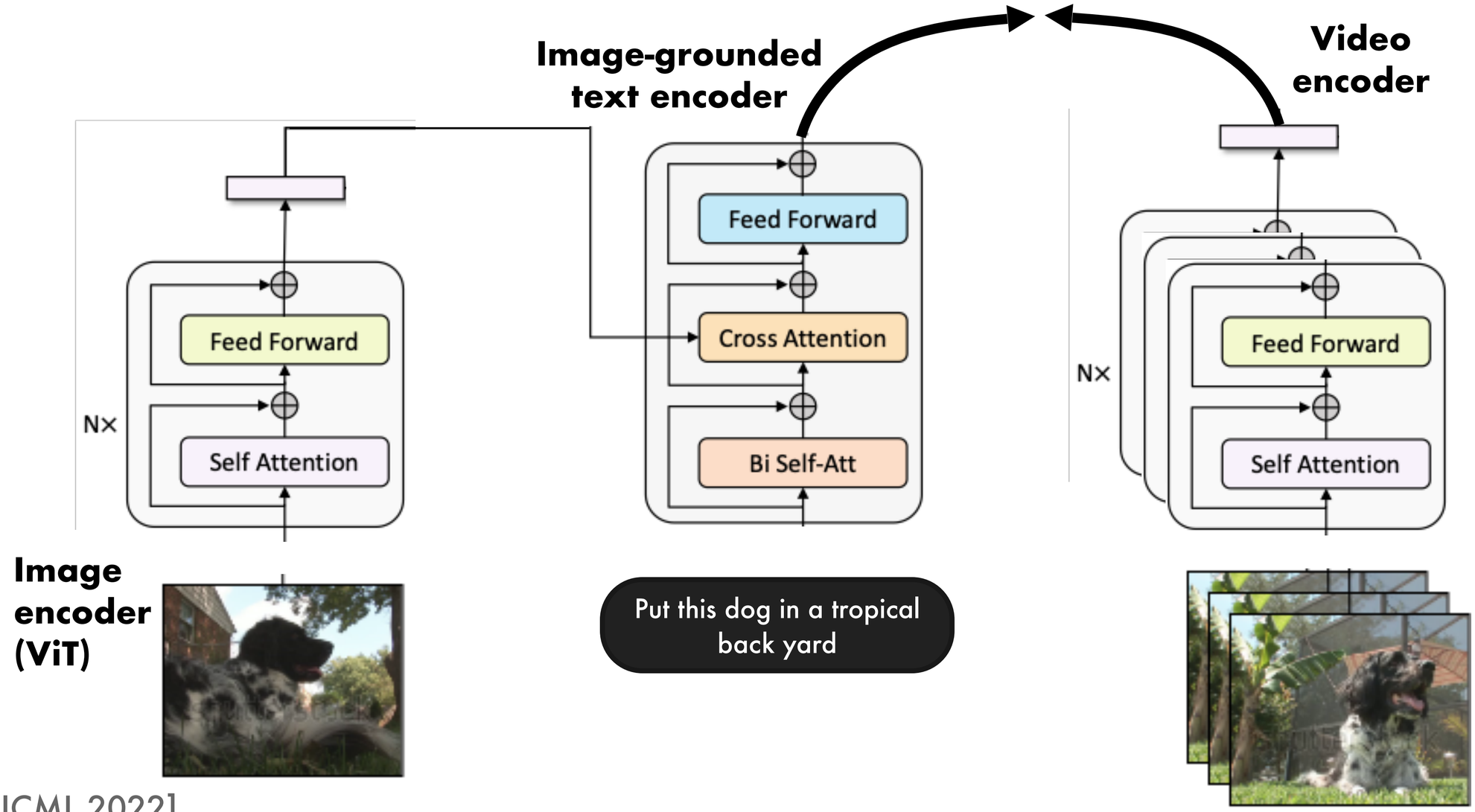






Ventura, Yang, Schmid, Varol, CoVR: Learning Composed Video Retrieval from Web Video Captions, AAAI 2024 (TPAMI 2024).

Model overview based on BLIP



["BLIP", Li et al., ICML 2022]

Ventura, Yang, Schmid, Varol, CoVR: Learning Composed Video Retrieval from Web Video Captions, AAAI 2024 (TPAMI 2024).

Training Data Augmentation with Synthetic Renderings

Caption

This scene contains a gift wrapping and two humans. They are in a street with grey floor, green plants on the side, and houses around. The first human is to the right of the gift wrapping. The first human is walking. The first human wears a red shirt and solid black pants. The first human has brown hair. The second human stands straight. The second human has brown hair. The second human is wearing blue jeans pants, brown shoes, and a white shirt. The second human is male.

Cascante-Bonilla, Shehada, Smith, Doveh, Kim, Panda, Varol, Oliva, Ordonez, Feris, Karlinsky,
Going Beyond Nouns With Vision & Language Models Using Synthetic Data, ICCV 2023

This scene contains a car tire, a television set, a stool, a microwave, a car tire, and one human. They are in a street with grey floor, green plants on the side, and houses around. The human is to the right of the stool. The car tire is in front of the television set. The microwave is to the left of the television set. The television set is behind the stool. The human is to the right of the car tire. The car tire is behind the stool. The car tire is in front of the television set. The stool is in front of the microwave. The human is behind the stool. The microwave is to the left of the stool. The car tire is to the right of the microwave. The human is to the right of the television set. The car tire is to the left of the human. The stool is in front of the car tire. The car tire is to the right of the car tire. The human is in front of the television set. The television set is to the right of the car tire. The car tire is to the left of the stool. The microwave is in front of the television set. The human is to the right of the microwave. The car tire is to the right of the stool. The human straight jump with full twist. The human is bald. The human is male. The human wears dark blue jeans. The human is clothed in a blue hoodie with a white logo on the front.

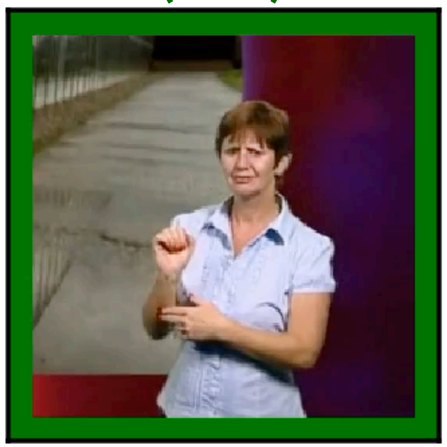
Cascante-Bonilla, Shehada, Smith, Doveh, Kim, Panda, Varol, Oliva, Ordonez, Feris, Karlinsky,
Going Beyond Nouns With Vision & Language Models Using Synthetic Data, ICCV 2023

Text-to-Sign Video Retrieval

Query Sentence: We're growing tomatoes almost all year round now.

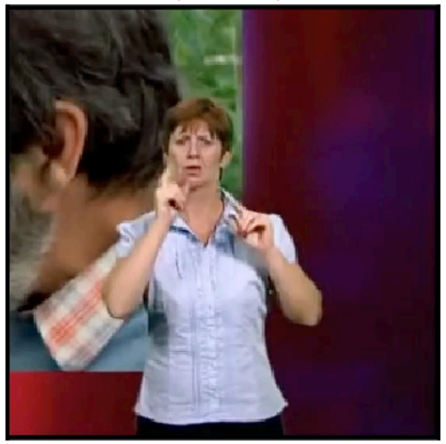
Top 5 Retrieved Videos

(0.81)



We're growing tomatoes almost all year round now.

(0.78)



To get the best performance from the **fruit** and the **plant**, we need to give the **plant** everything that it needs.

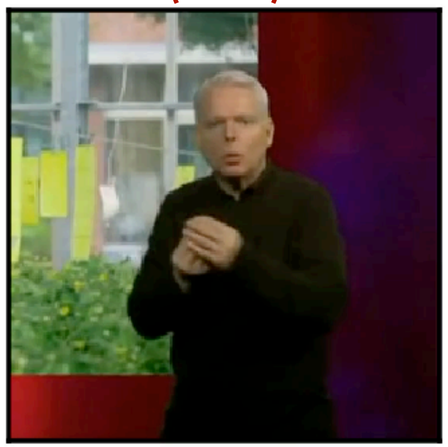
(0.78)

Because you need extra heat to **grow tomatoes** out of **season**, it makes them more expensive.

(0.76)

July to October were always the **months** to eat **tomatoes** but now vast heated greenhouses mean we can grow them between February and November.

(0.75)

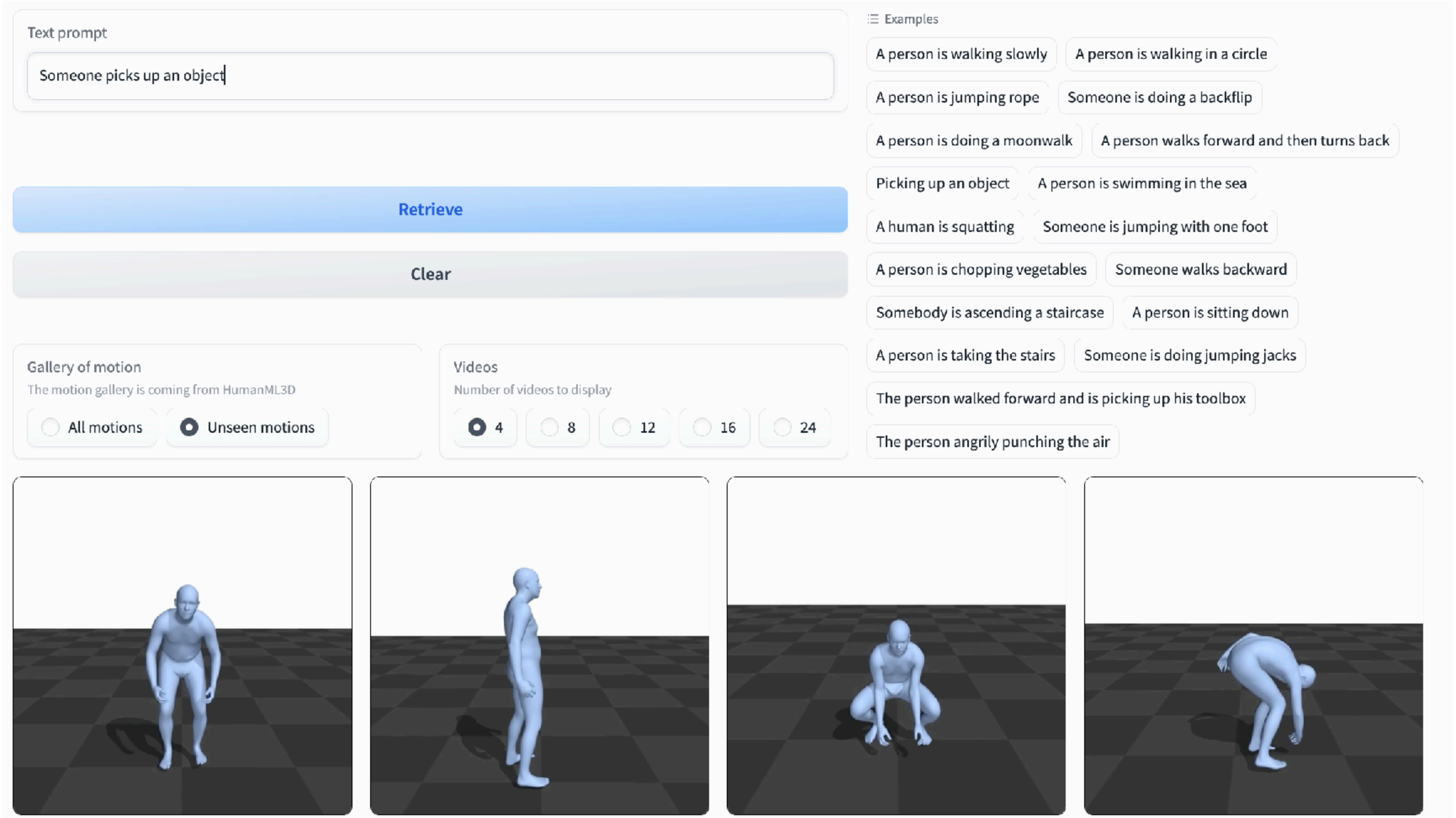


So, why did you choose **tomatoes** as the main medium for this gene?

(Unused)

[Raude*, Prajwal*, Momeni*, Bull, Albanie, Zisserman, Varol, **arXiv 2024**]

Words in the lexical field of the query video subtitle are in blue.



[Petrovich, Black, Varol. TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis, ICCV 2023]

Vision-to-text tasks

Movie Audio Description Generation

Audio Description (AD) = Narration describing visual elements in the movie to aid the visually impaired

AD Movie clip from `Out of Sight' (1998) **with Audio Description**

Han, Bain, Nagrani, Varol, Xie, Zisserman, AutoAD: Movie Description in Context, CVPR 2023

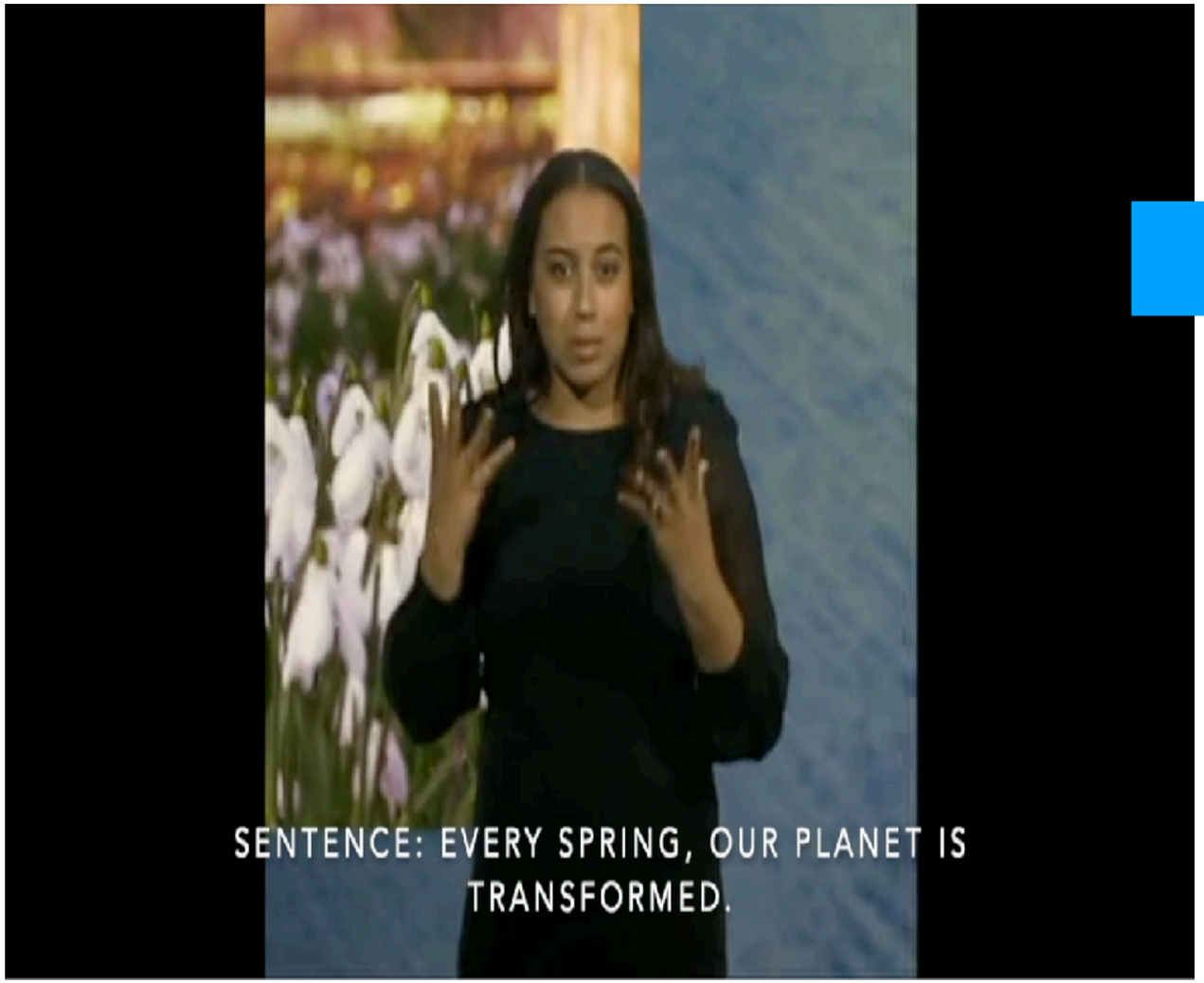
Han, Bain, Nagrani, Varol, Xie, Zisserman, AutoAD II: The Sequel – Who, When, and What in Movie Audio Description, ICCV 2023

Han, Bain, Nagrani, Varol, Xie, Zisserman, AutoAD III: The Prequel – Back to the Pixels, CVPR 2024

Xie, Han, Bain, Nagrani, Varol, Xie, Zisserman, AutoAD-Zero: A Training-Free Framework for Zero-Shot Audio Description, ACCV 2024

Qualitative examples of AutoAD-III

Sign Language Translation: Video in, Text out



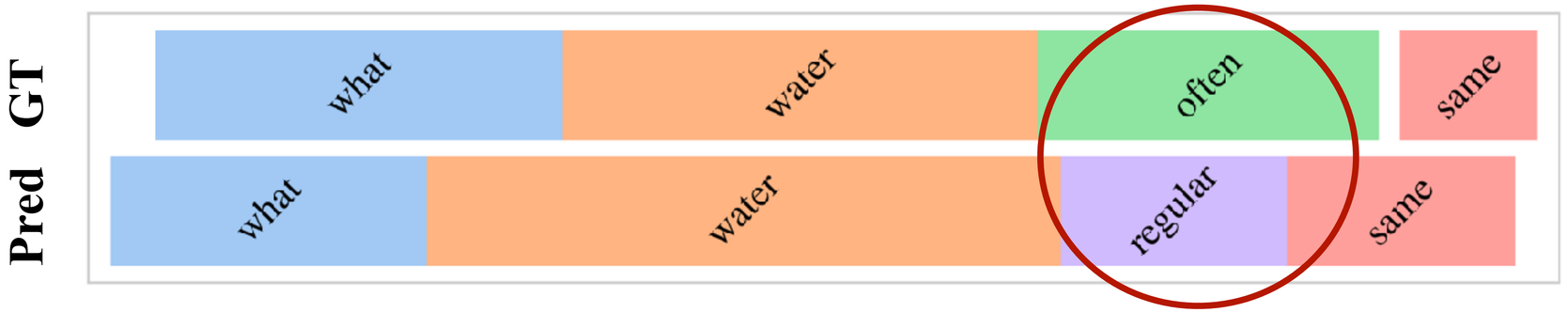
Ground-truth translation:

"Every Spring, our planet is transformed"

Ground-truth sign transcription:

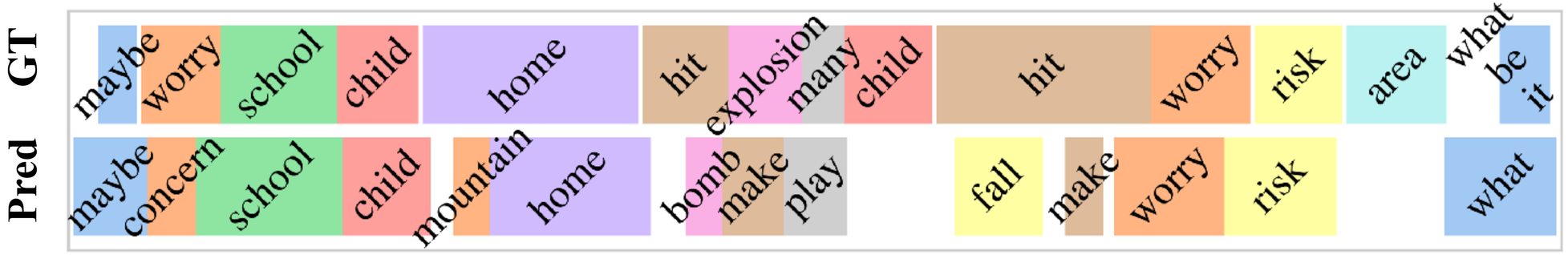
(EVERY; SPRING; OUR; PLANET; **HAPPEN; WHAT;** *TRANSFORM)*

Sign Language Transcription (aka Continuous SL Recognition)



Word Error Rate: 25.0

Subtitle (unused): The majority of sweat is just straightforward water.



Word Error Rate: 42.9

Subtitle (unused): We were worried that it might happen again, falling on a primary school, someone's home, or a playground.

Raude, Prajwal, Momeni, Bull, Albanie, Zisserman, Varol, A Tale of Two Languages: Large-Vocabulary Continuous Sign Language Recognition from Spoken Language Supervision, arXiv 2024

Sign Language Translation with Contextual Cues

Background Caption

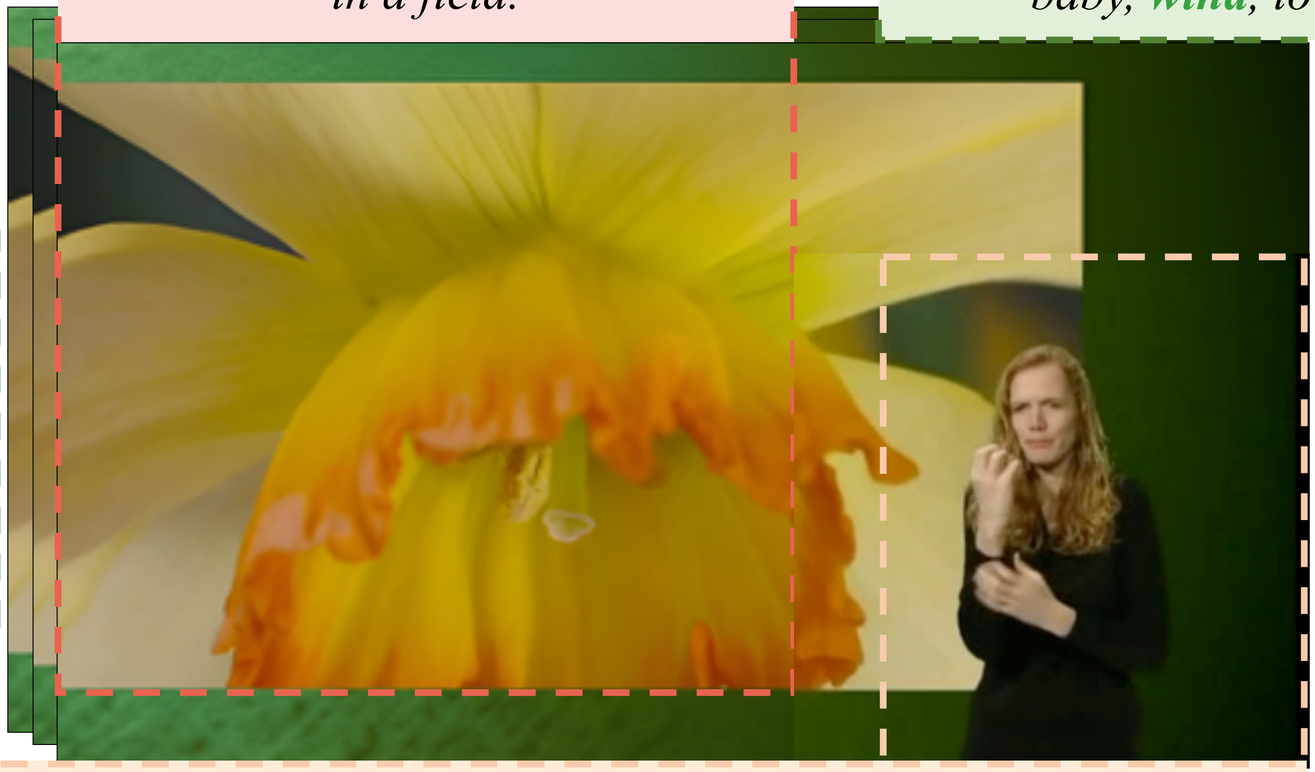
*Yellow **daffodils** growing outdoors in a field.*

Pseudo-Glosses

*steam, pepper, **beautiful**, **flower**, baby, **wind**, lovely…*

Previous Sentence

*The **wind** is really starting to pick up*



Signer


petals


flower


beautiful


wind


beautiful

GT Translation: *Beautiful petals that wave about in the wind.*
Prediction: *They're beautiful flowers and they're wafting in the wind.*

Raajesh, Jang, Momeni, Varol, Zisserman. (WIP)

Sign Language Translation with Contextual Cues

BG Image

Sign

PG : good, new, tree, jungle, national, country, climate, garden, grow, pop, time, in, Christmas, use, when, price, two, twenty, five

Prev : I've got an evening with the dark skies of the UK to look forward to.

BG : satellite, image, woman, green, dress, united, states, uk, new, forest, national, park, map, lymburts, showing, location

GT : The New Forest National park was created in 2005.

Vid : The new national forest was born in 2005.

Vid+PG : This is a new National Nature Reserve, established in 2005.

Vid+PG+Prev : The New jungle was set up in 2005.

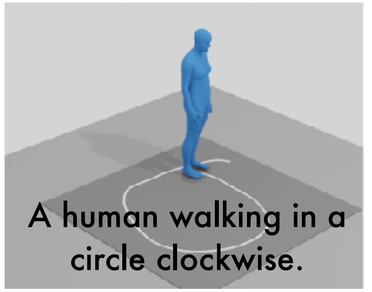
Vid+PG+Prev+BG: The New Forest National Park was only established in 2005.

Raajesh, Jang, Momeni, Varol, Zisserman. (WIP)

Text-to-vision tasks

Text to 3D Human Motion Generation

Pickup



1) ACTOR

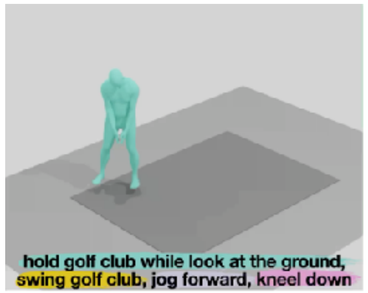
✦ Action-conditioned synthesis

[Petrovich, Black, Varol. ICCV'21]

2) TEMOS

✦ Text-conditioned synthesis

[Petrovich, Black, Varol. ECCV'22]



3) TEACH

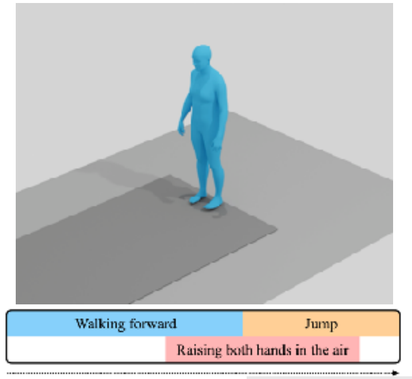
✦ Temporal compositionality for sequential actions

[Athanasiou, Petrovich, Black, Varol. 3DV'22]

4) SINC

✦ Spatial compositionality for simultaneous actions

[Athanasiou*, Petrovich*, Black, Varol. ICCV'23]



5) STMC

✦ Multi-track timeline control for motion synthesis

[Petrovich, Litany, Iqbal, Black, Varol, Peng, Rempe. CVPRW'24]

6) TMR

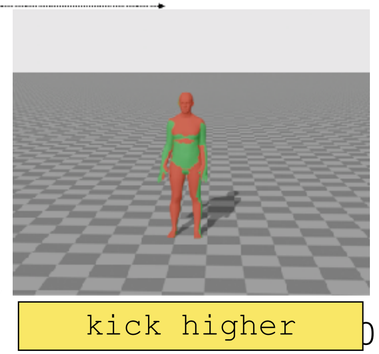
✦ Text-to-motion retrieval

[Petrovich, Black, Varol. ICCV'23]

7) MotionFix

✦ Text-based editing

[Athanasiou, Cseke, Diomatoris, Black, Varol. Siggraph Asia'24]



Agenda

1. Generative neural networks

- VAE: Variational autoencoders
- GAN: Generative adversarial networks
- Diffusion models

2. Vision & language

- Text-to-image retrieval
- Text-to-image generation
- Image captioning
- Bonus: Visual question answering (VQA)
- Bonus: Examples from our works