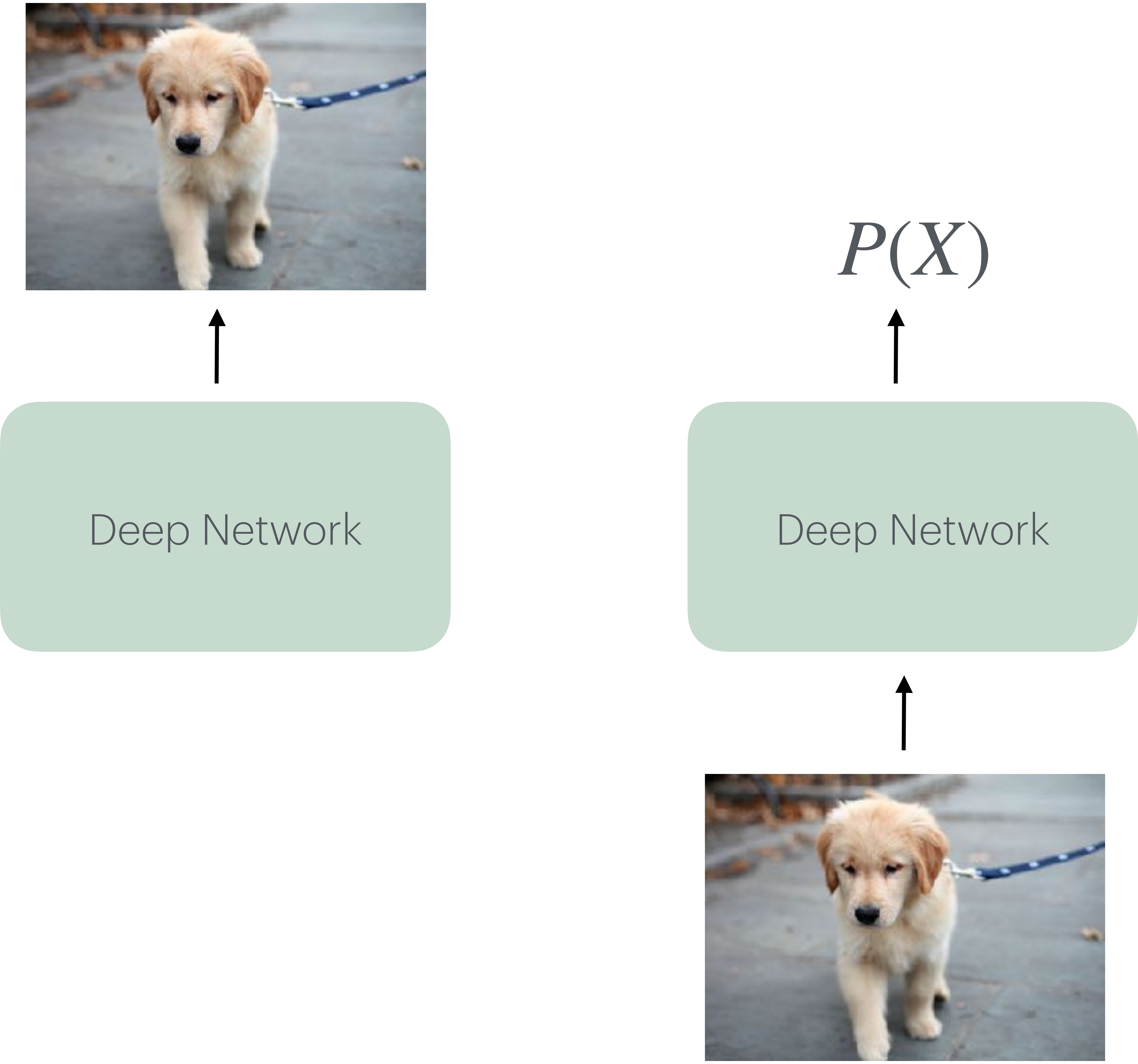
DALL-E

A case study

Philipp Krähenbühl, UT Austin

Generative models

- Two tasks of a generative model $P(X)$
  - Sampling: $x \sim P(X)$
  - Density estimation: $P(X = x)$



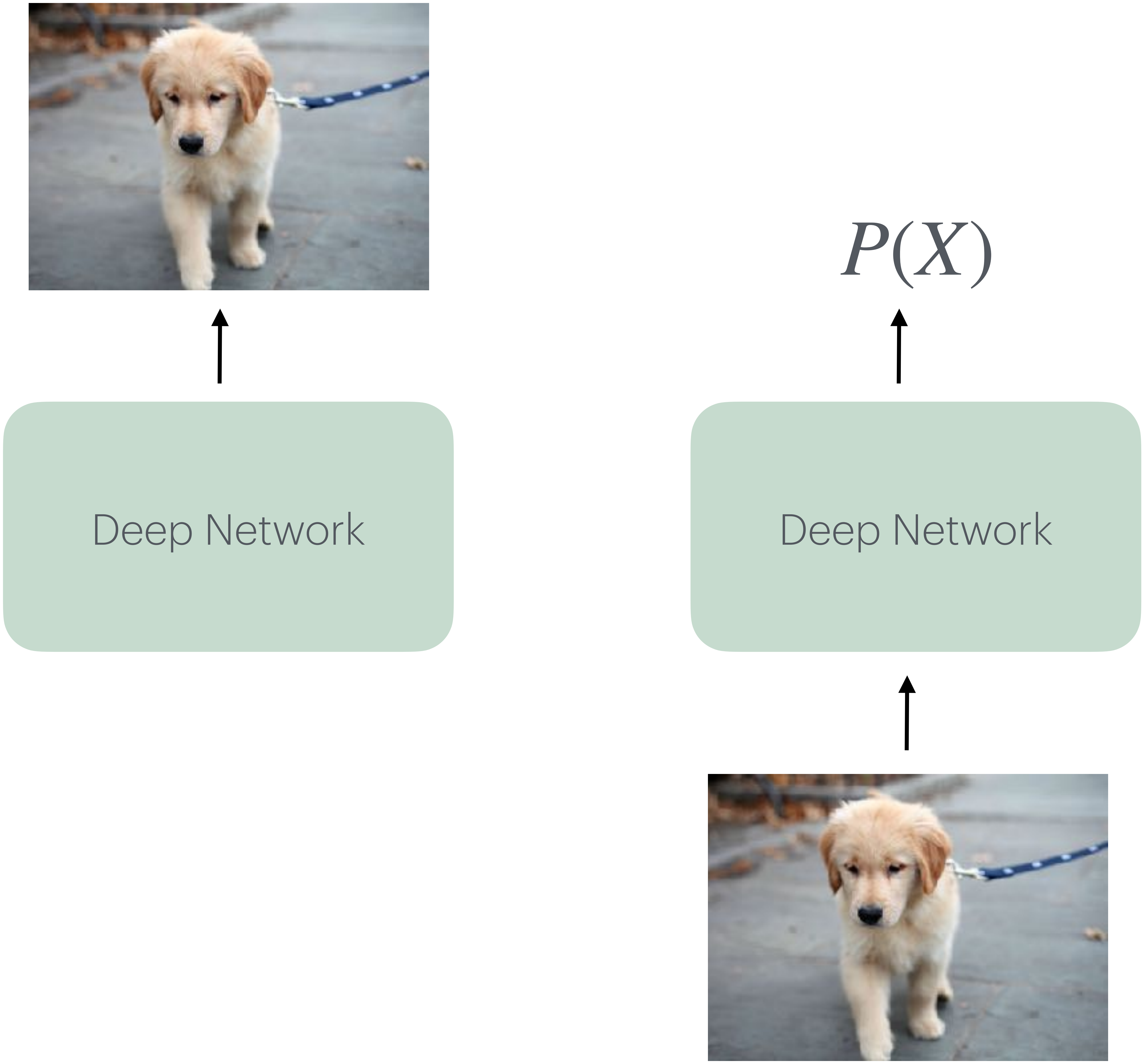
Deep Network

$P(X)$

Deep Network

Generative modeling is hard

- Density estimation $P(X = x)$
  - How to ensure $\sum_{x} P(x) = 1$ for all x
  - Impossible to compute (in general)
- Sampling $x \sim P(X)$
  - What is the input to the network?



Deep Network

$P(X)$

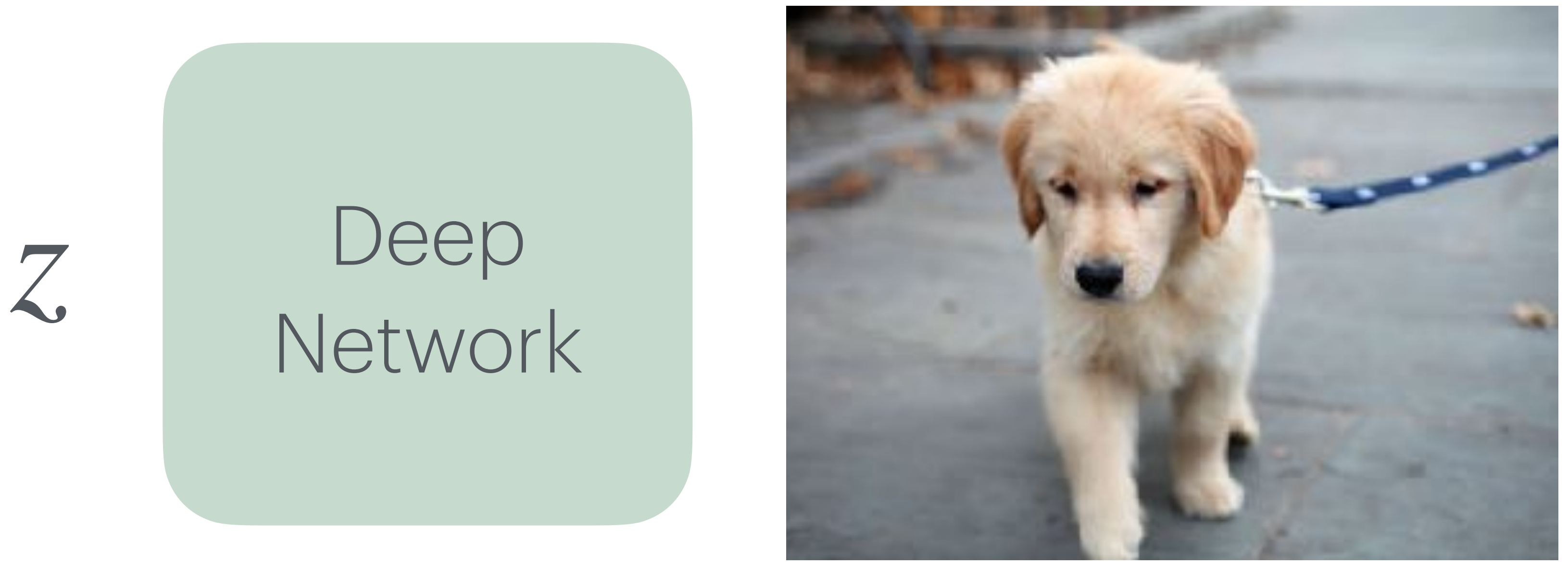
Deep Network

Generative models

Two kinds of models

Sampling based $x \sim P(X)$

- Sample $z \sim P(Z)$
- Learn transformation
  - $P(x|z)$ or $f : z \to x$

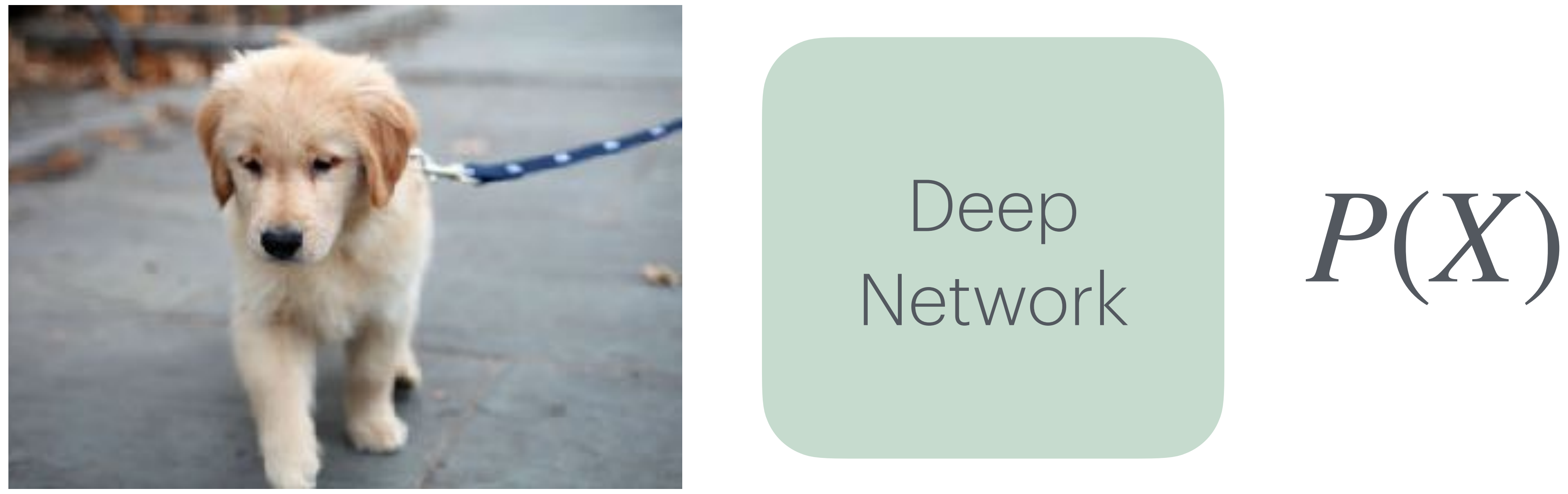


z

Deep Network

Density estimation based $P(X)$

- Learn special form of $P(X)$
- Model specific sampling / generation

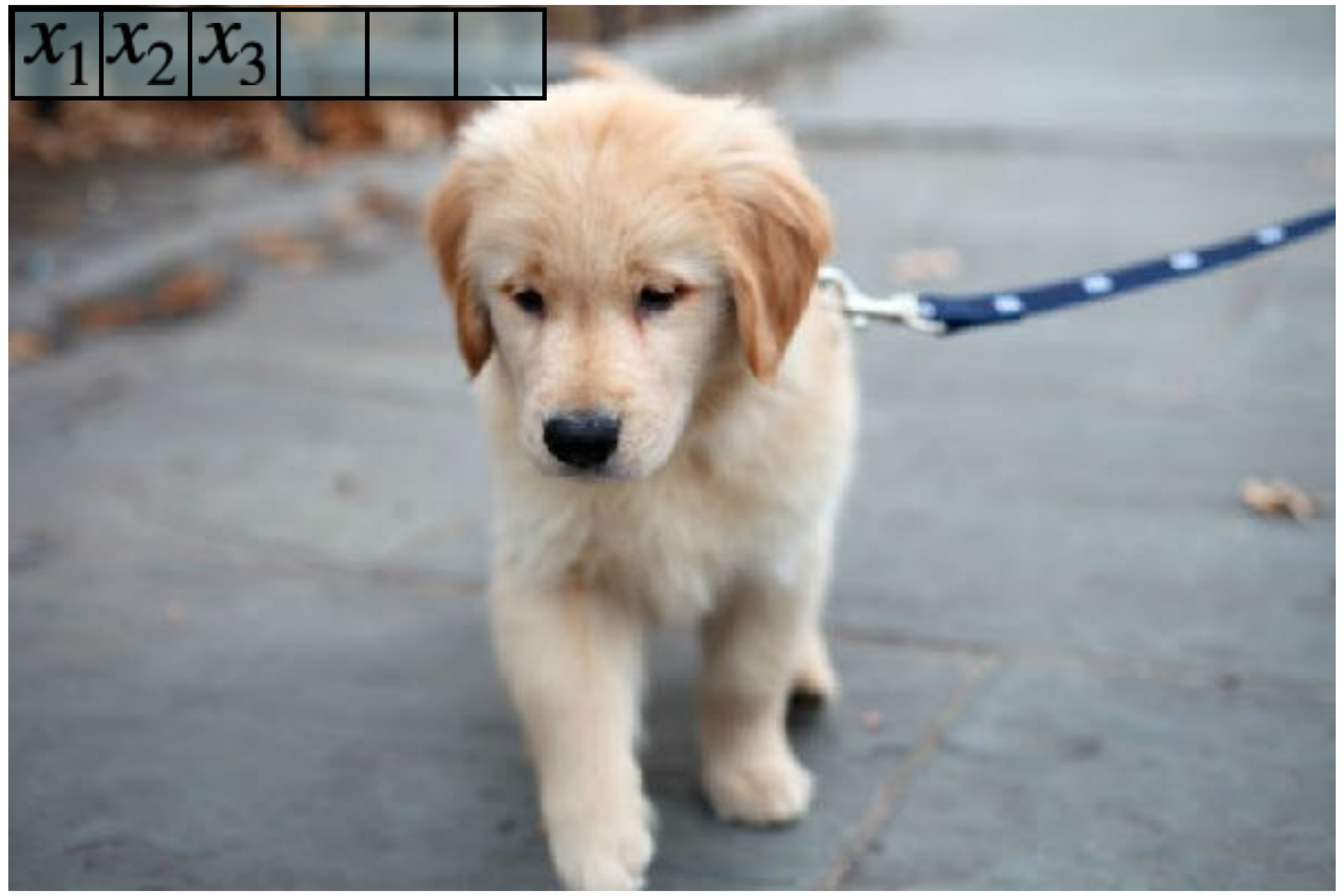


Deep Network

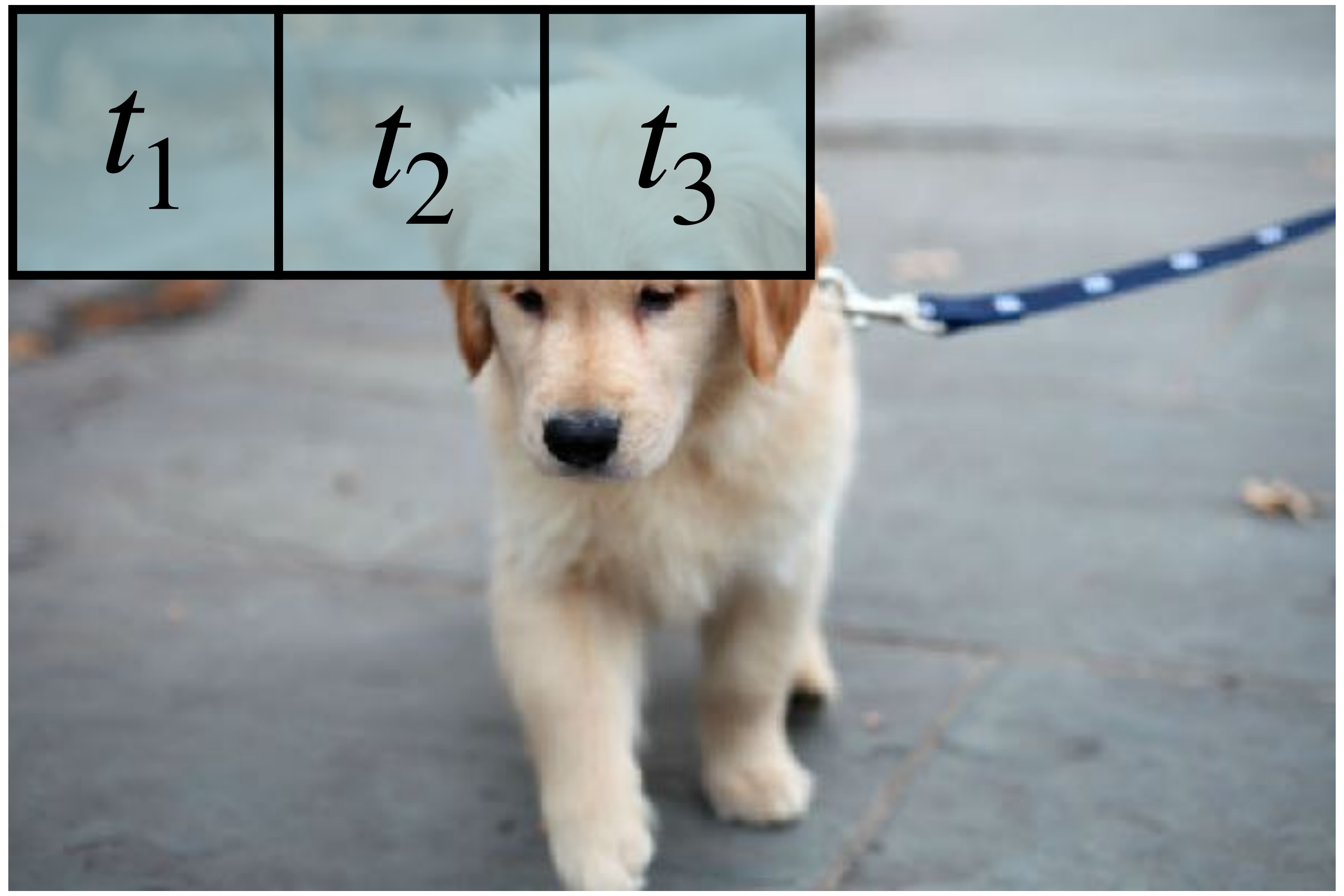
$P(X)$

Tokenization

- Image [1]
  - Convert patch p_i of pixels into token $t_i \in \{1, \ldots, K\}$
- Text [2]
  - Convert set of characters into token
- Protein-sequence [3]
  - Convert local protein structure to token



Vanilla auto-regressive model



Tokenized auto-regressive model

[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017
[2] Language models are unsupervised multitask learners. Alec Radford, et al. 2019
[3] Simulating 500 million years of evolution with a language model. Thomas Hayes, et al. 2024

Tokenization

A different view

- Convert
  - images ↔ streams of tokens
  - text ↔ streams of tokens
    - More in next section

$\leftrightarrow \boxed{t_1}\boxed{t_2}\boxed{t_3}$

A cute little dog fully focused on walking $\leftrightarrow \boxed{\hat{t}_1}\boxed{\hat{t}_2}\boxed{\hat{t}_3}$

DALL-E

- Let's learn a generative model over text and image tokens
  - $P(\mathbf{t} \mid \hat{\mathbf{t}}) = P(t_1 \mid \hat{\mathbf{t}})P(t_2 \mid t_1, \hat{\mathbf{t}})\ldots P(t_L \mid t_1, \ldots, t_{L-1}, \hat{\mathbf{t}})$
- Where do we get image-text data from?
- What architecture do we use?

$\leftrightarrow \boxed{t_1}\boxed{t_2}\boxed{t_3}$

A cute little dog fully focused on walking $\leftrightarrow \boxed{\hat{t}_1}\boxed{\hat{t}_2}\boxed{\hat{t}_3}$

Zero-Shot Text-to-Image Generation, Ramesh et al. 2021

DALL-E

Dataset

- Image captioning dataset
  - Conceptual Captions [1]
    - 3.3 million text-image
  - OpenAI Internal data (the internet)
    - 250 million text-images pairs
    - YFCC100M [2]
  - Lots of cleanup

IMG_9793: Streetcar (Toronto Transit) by Andy Nystrom

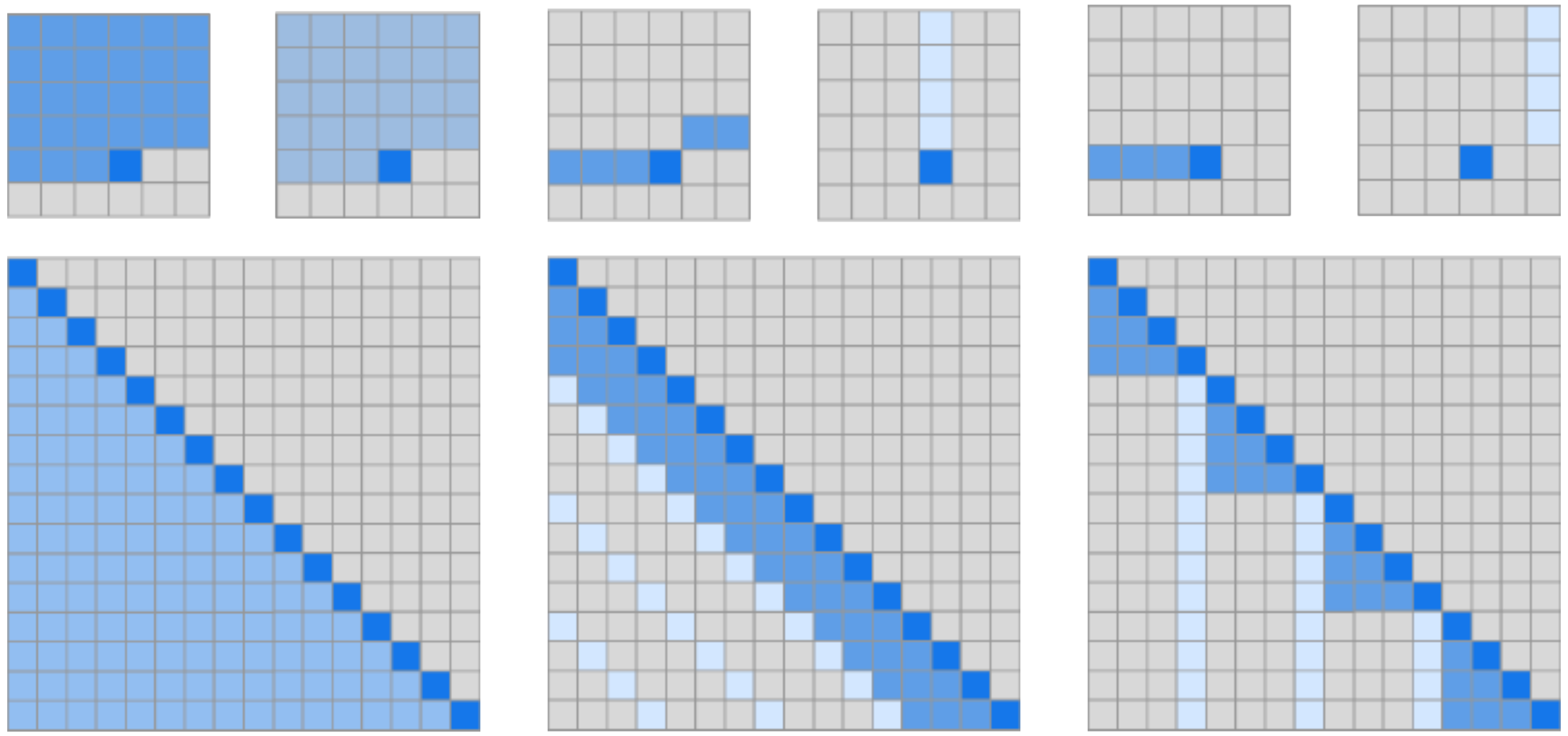
Celebrating our 6th wedding anniversary in Villa Mary by Rita & Tomek

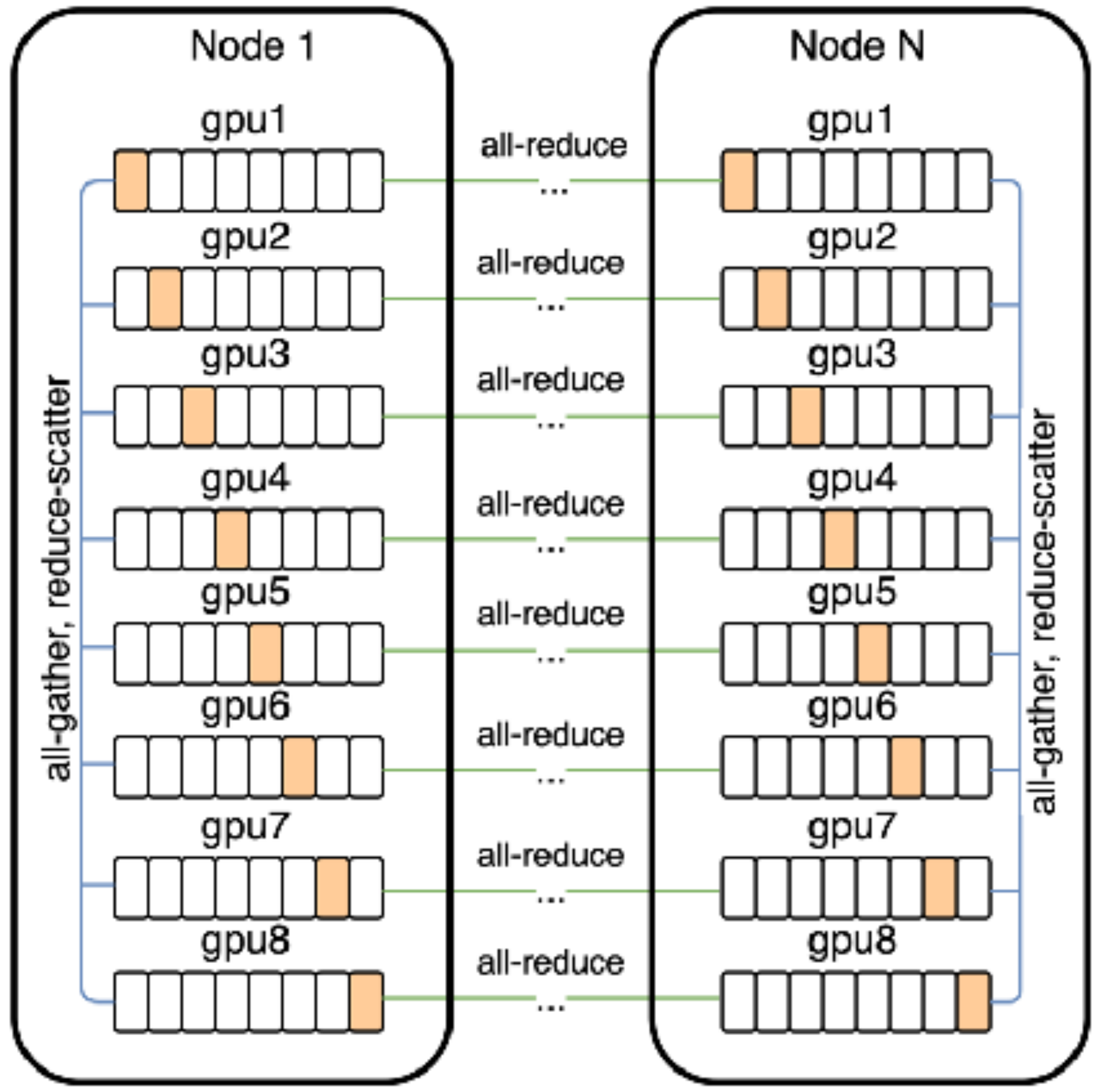
[1] Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning, Sharma et al. 2018
[2] YFCC100M: The New Data in Multimedia Research, Thomee et al. 2015

DALL-E

Architecture

- Sparse transformer [1]
- Mixed-precision training
- Sharded Multi-GPU training
    - Pre-cursor to FSDP

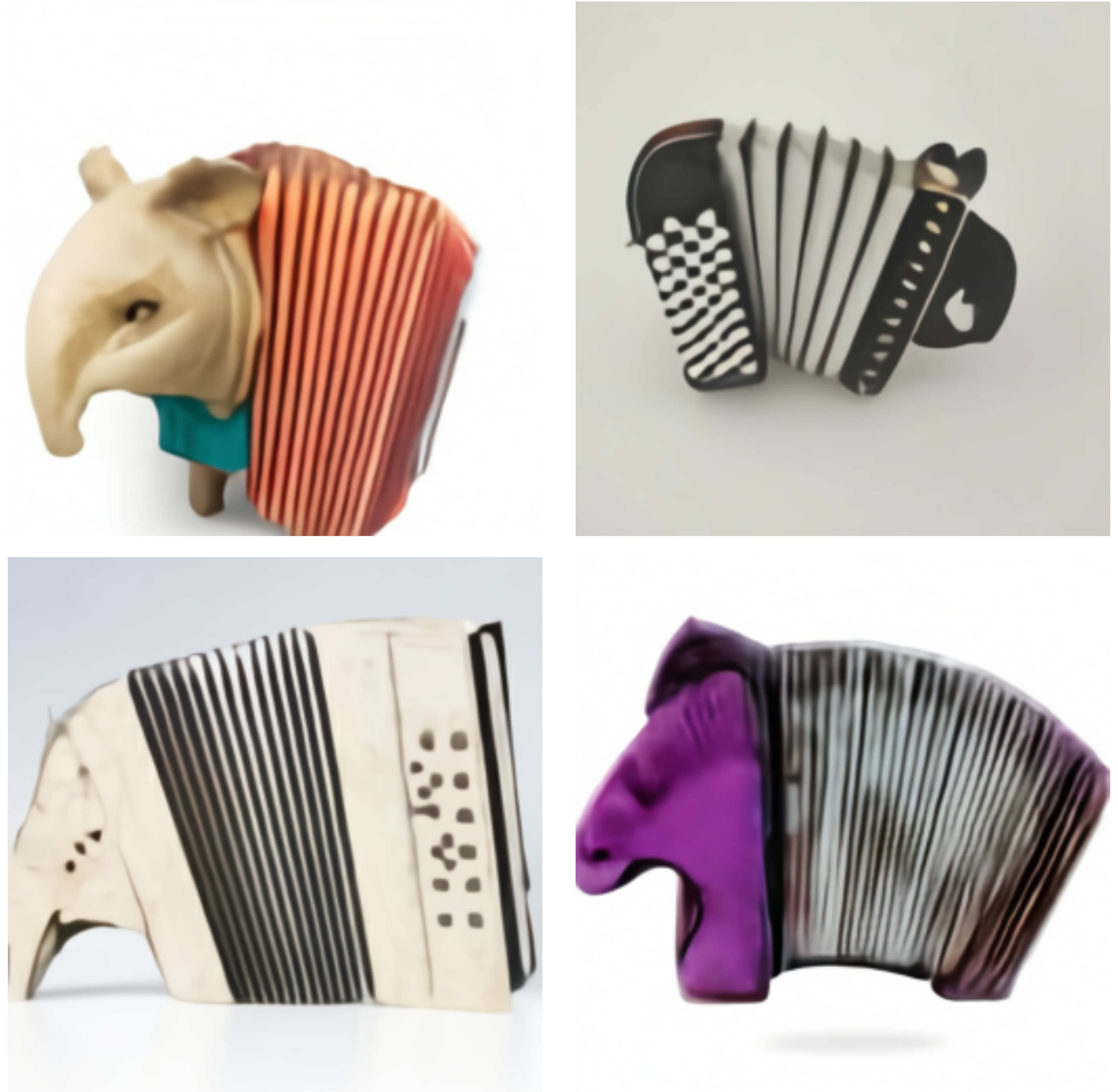




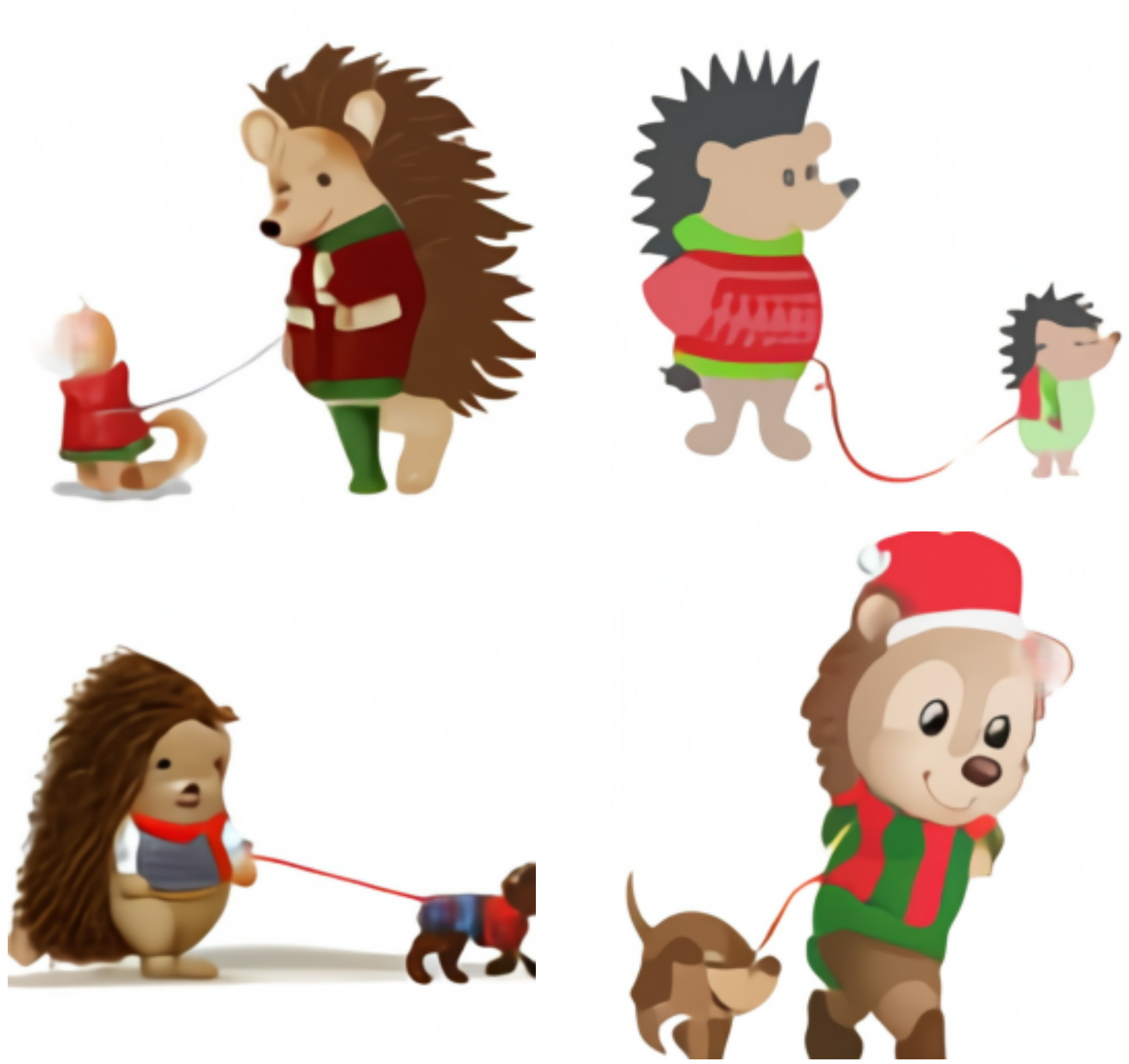
[1] Generating Long Sequences with Sparse Transformers, Child et al. 2019

DALL-E

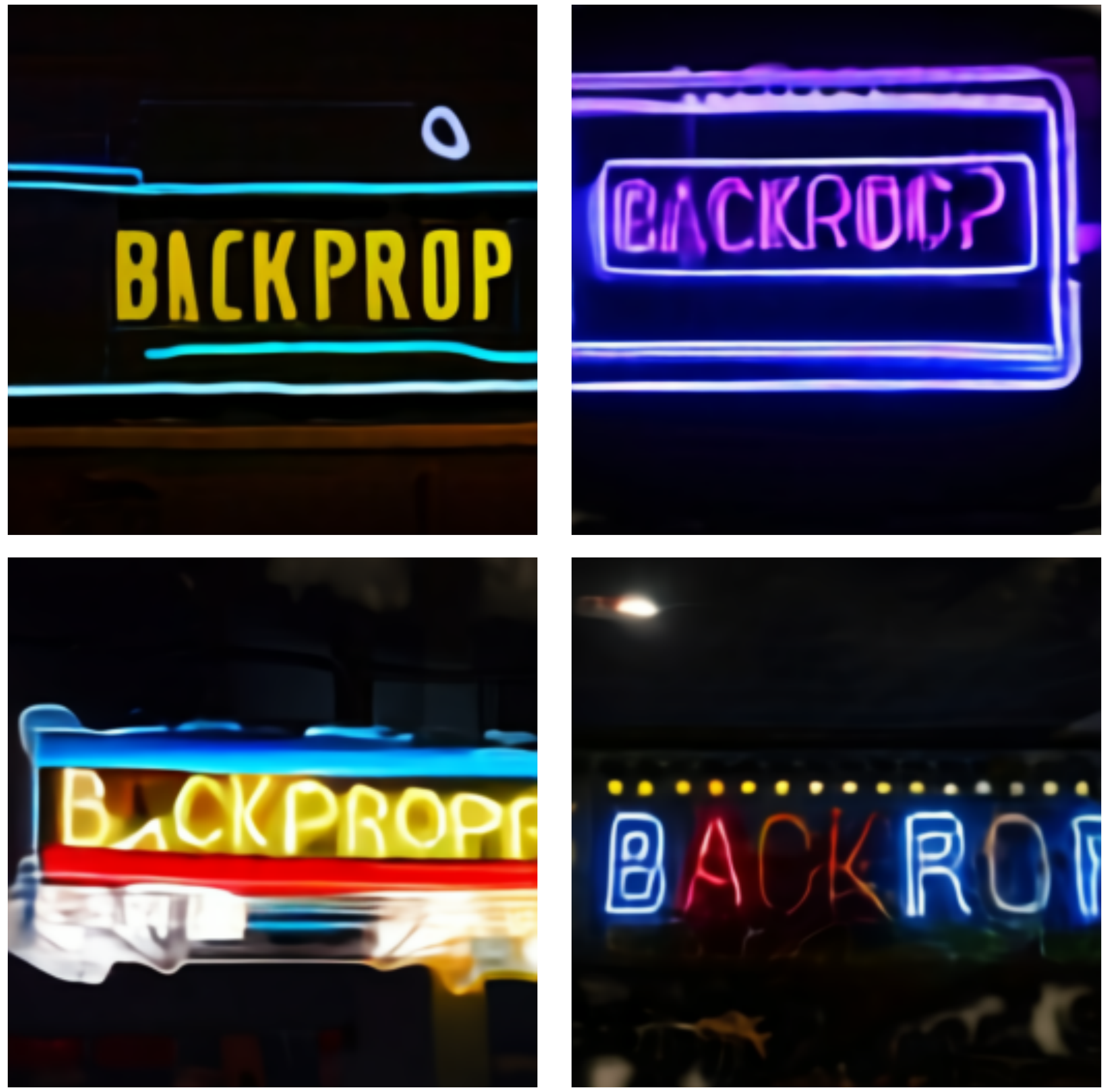
Results



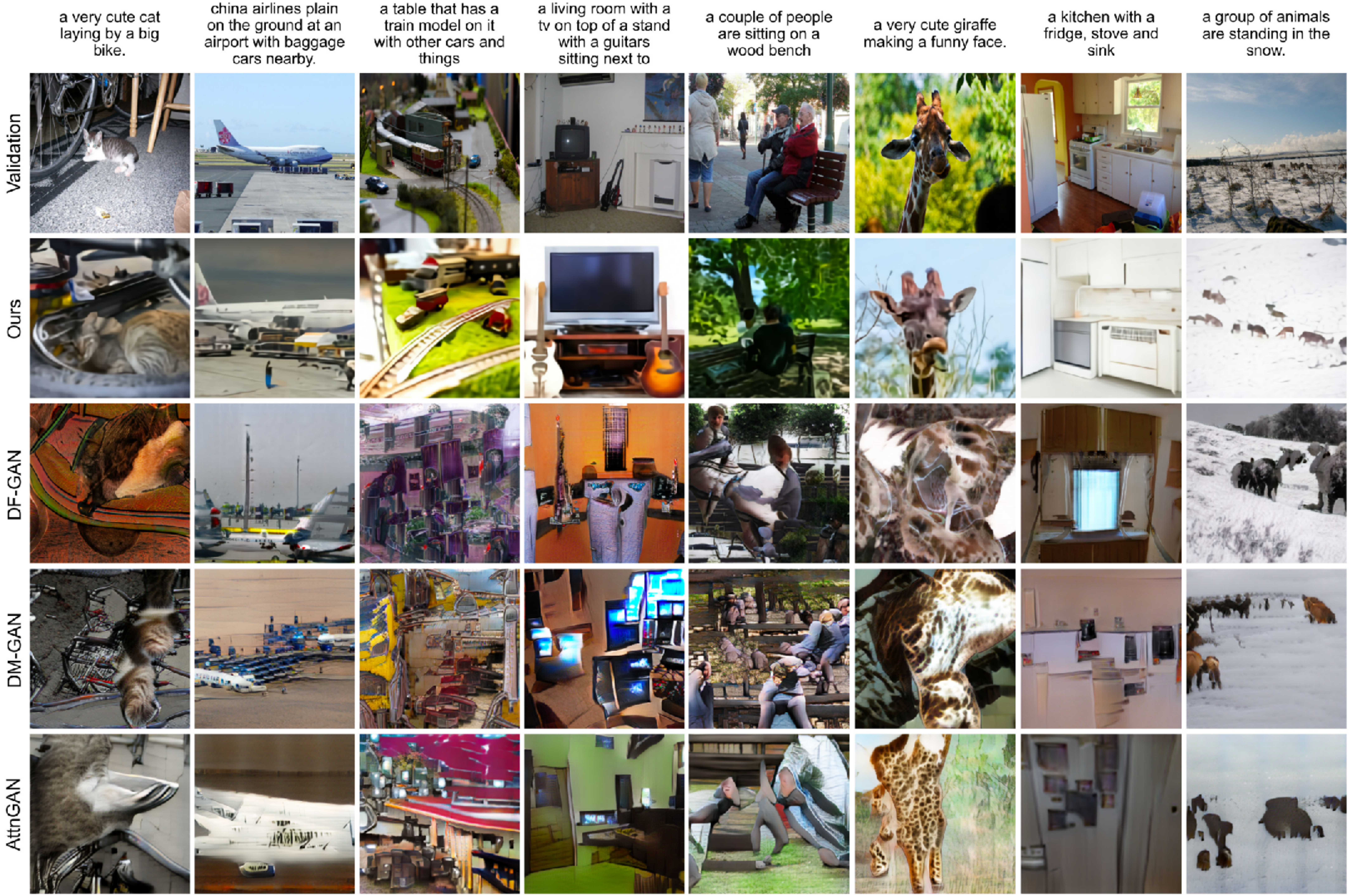
a tapir made of accordion. a tapir with the texture of an accordion.



an illustration of a baby hedgehog in a christmas sweater walking a dog



a neon sign that reads “backprop”. a neon sign that reads “backprop”. backprop neon sign

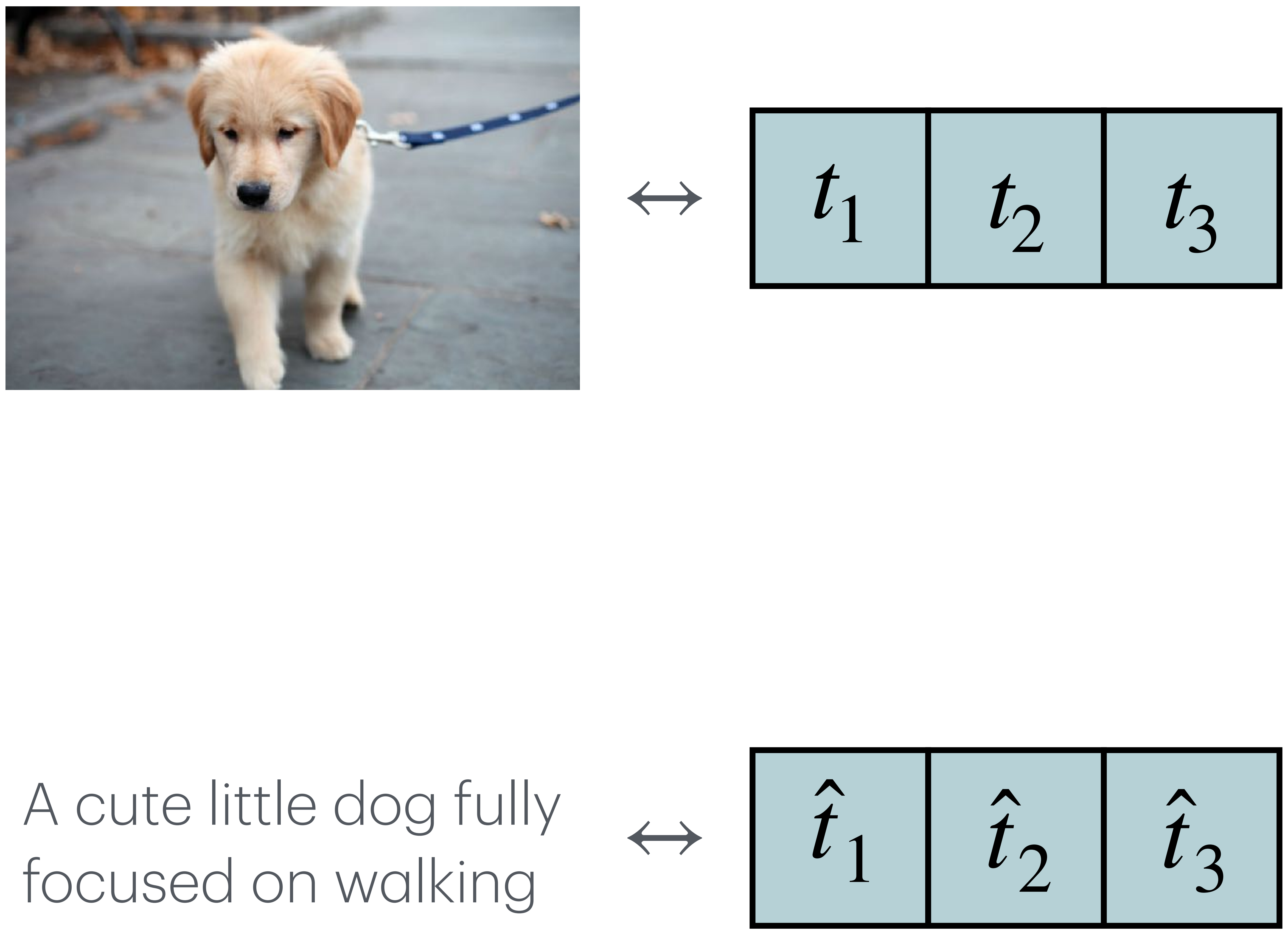


a very cute cat laying by a big bike.
china airlines plain on the ground at an airport with baggage cars nearby.
a table that has a train model on it with other cars and things
a living room with a tv on top of a stand with a guitars sitting next to
a couple of people are sitting on a wood bench
a very cute giraffe making a funny face.
a kitchen with a fridge, stove and sink
a group of animals are standing in the snow.
Validation
Ours
DF-GAN
DM-GAN
AttnGAN

DALL-E

Lessons learned

- Data is king
- Scaling matters
- Models can be simple



A cute little dog fully focused on walking

Zero-Shot Text-to-Image Generation, Ramesh et al. 2021

References

- [1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017
- [2] Language models are unsupervised multitask learners. Alec Radford, et al. 2019
- [3] Simulating 500 million years of evolution with a language model. Thomas Hayes, et al. 2024
- [4] Zero-Shot Text-to-Image Generation, Ramesh et al. 2021
- [5] Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning, Sharma et al. 2018
- [6] YFCC100M: The New Data in Multimedia Research, Thomee et al. 2015
- [7] Generating Long Sequences with Sparse Transformers, Child et al. 2019