Generative AI in Computer Vision:
multimodal understanding and generation

Vicky Kalogeiton

ESSAI & ACAI 2024

26/07/2024

About me

- *Assistant Professor*, 2020 –
  - VISTA Group, Ecole Polytechnique, France
  - Main genAI professor

- *Research Fellow,* 2019 – 2021
- *Post-doc,* 2018 – 2019
  - Visual Geometry Group, University of Oxford, UK
  - Andrew Zisserman

- *PhD*, 2013 – 2017
  - University of Edinburgh, UK, INRIA, Grenoble, France
  - Vittorio Ferrari, Cordelia Schmid

Today’s tutorial

Part I:
Introduction

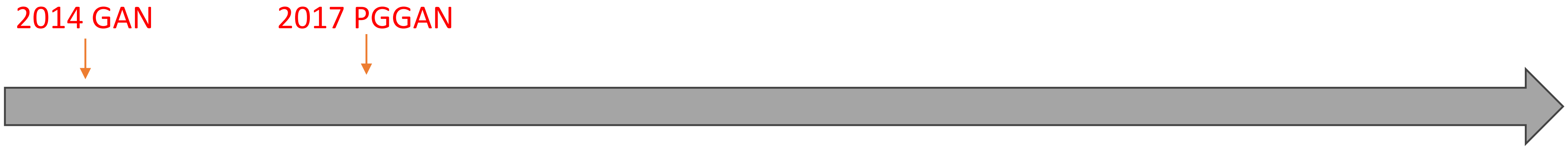
Part II:
Diffusion &
Guidance &
Control

Part III:
My research

[Slides by V. Kalogeiton, X. Wang]

Evolution of Multimodal Generative AI

2014 2015 2016 2017



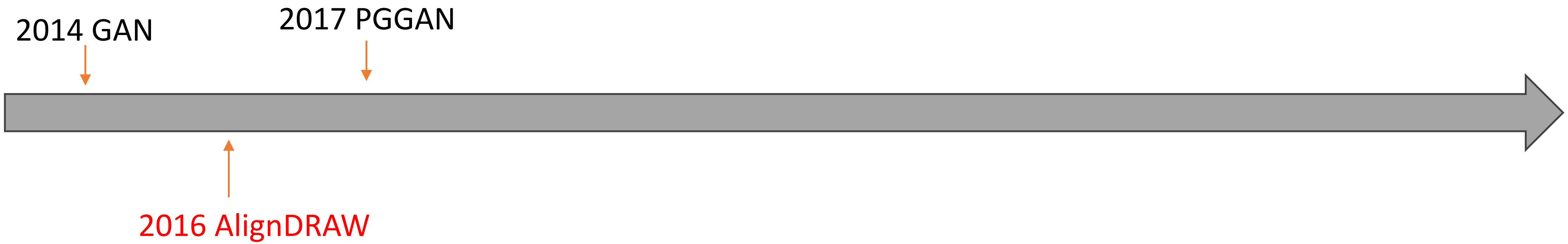
Evolution of Multimodal Generative AI

A stop sign is flying in blue skies.

A herd of elephants flying in the blue skies.

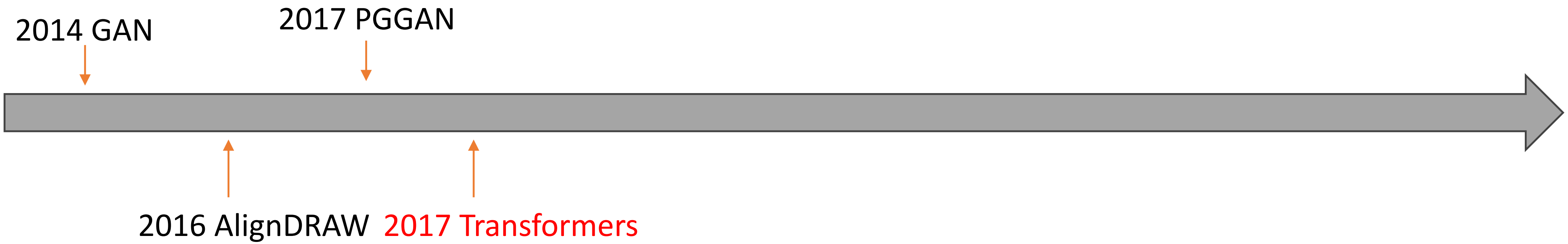
A toilet seat sits open in the grass field.

A person skiing on sand clad vast desert.

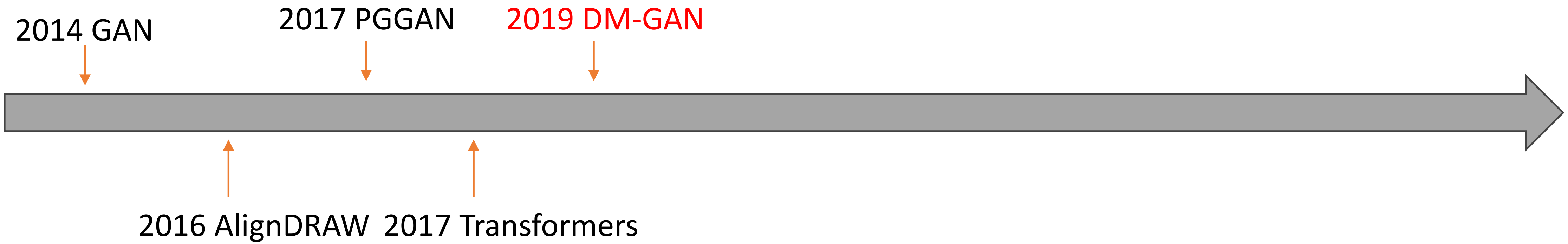


Evolution of Multimodal Generative AI

A stop sign is flying in blue skies. | A herd of elephants flying in the blue skies. | A toilet seat sits open in the grass field. | A person skiing on sand clad vast desert.



Evolution of Multimodal Generative AI



Evolution of Multimodal Generative AI

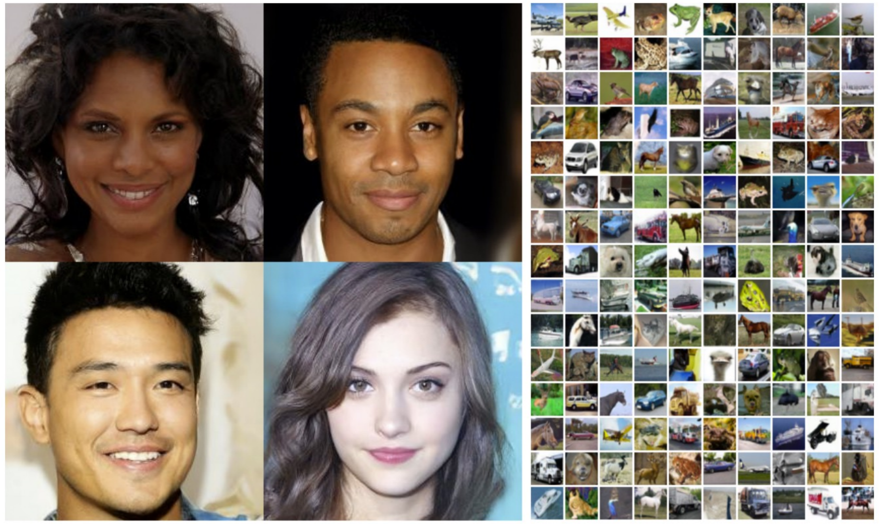


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

2014 GAN

2017 PGGAN

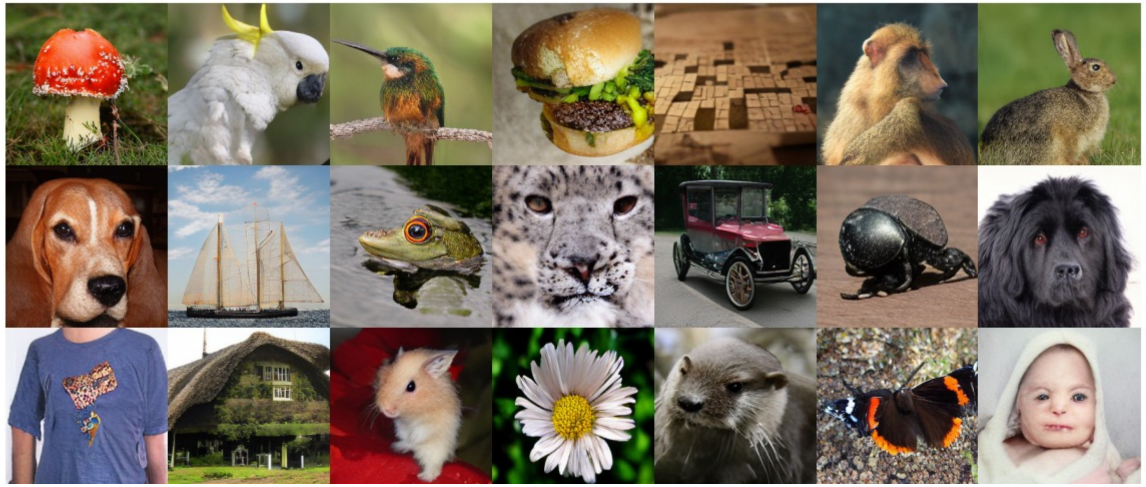
2019 DM-GAN

2016 AlignDRAW

2017 Transformers

2020 DDPM

Evolution of Multimodal Generative AI



2014 GAN

2016 AlignDRAW

2017 PGGAN

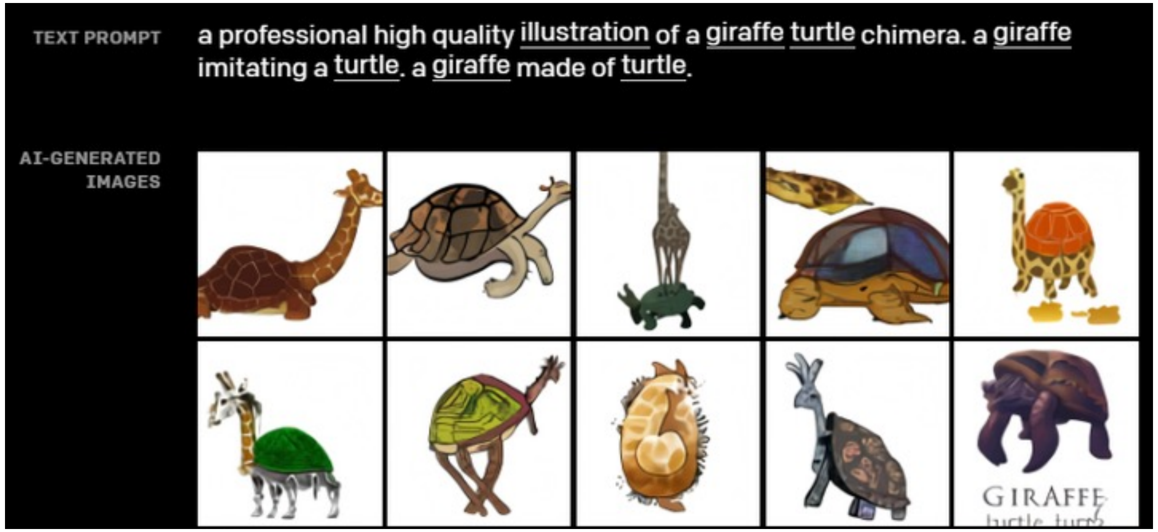
2017 Transformers

2019 DM-GAN(T2I)

2020 DDPM

2021 VQ-GAN

Evolution of Multimodal Generative AI



DALLE-1

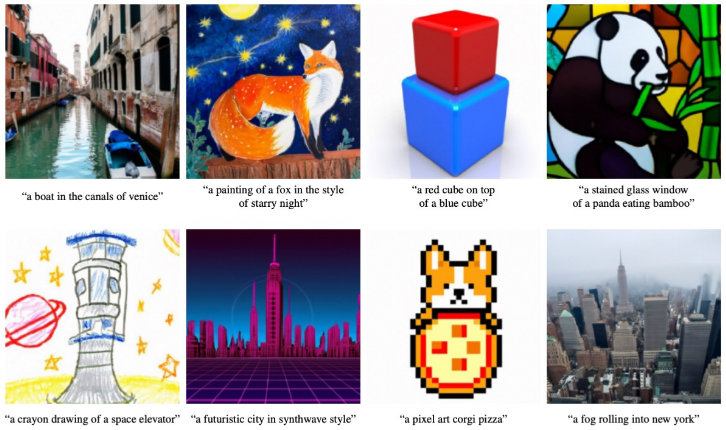


Figure 1. Selected samples from GLIDE using classifier-free guidance. We observe that our model can produce photorealistic images with shadows and reflections, can compose multiple concepts in the correct way, and can produce artistic renderings of novel concepts. For random sample grids, see Figure 17 and 18.

GLIDE

2014 GAN

2017 PGGAN

2019 DM-GAN(T2I)

2021 VQ-GAN

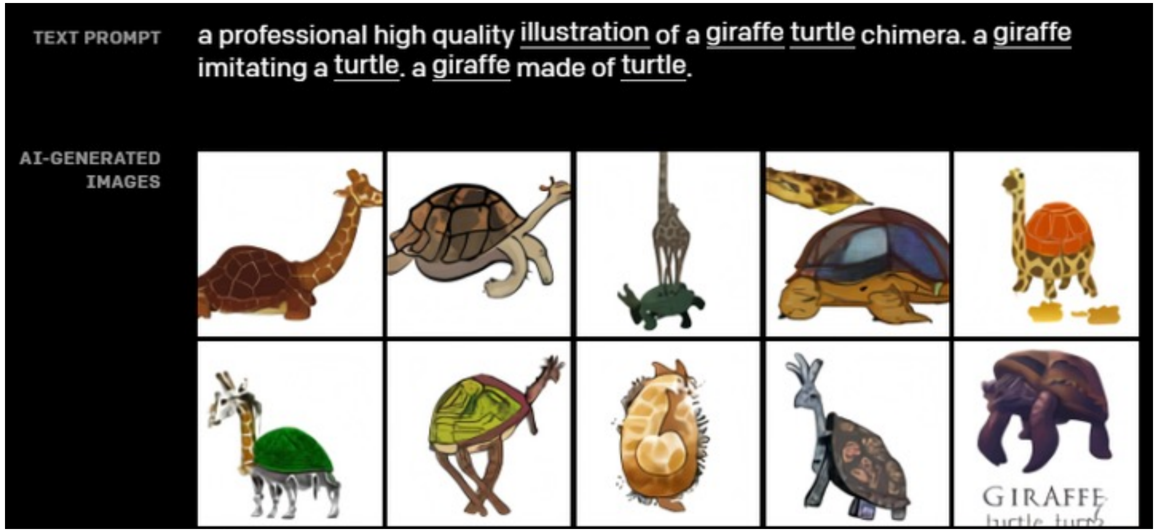
2022 DALLE, GLIDE

2016 AlignDRAW

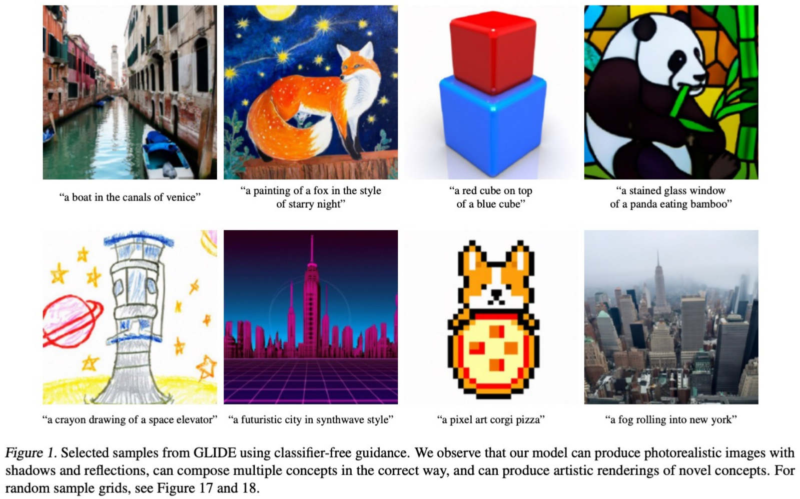
2017 Transformers

2020 DDPM

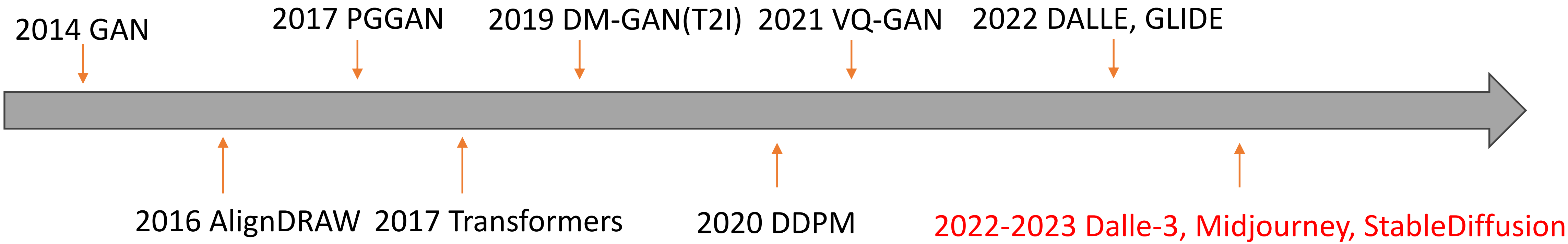
Evolution of Multimodal Generative AI



DALLE-1



GLIDE



2022 → 2023

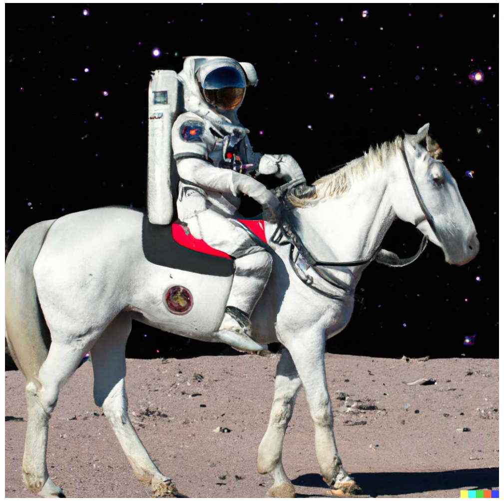
OpenAI: DALL-E3 (2023)

Midjourney (2023)

music, audio, animation, video, physical etc....

Dalle-2 (Text-to-Image)

A bowl of soup as a planet in the universe

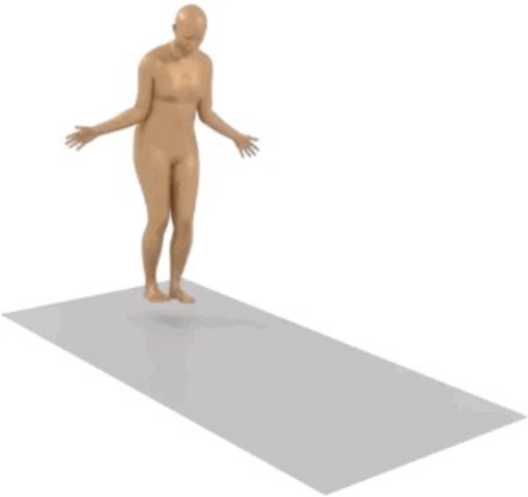


An astronaut riding a horse in a photorealistic style

Teddy bears mixing sparkling chemicals as mad scientists

Human Motion Diffusion (Text-to-Motion)

"A person punches in a manner consistent with martial arts."



"A person is skipping rope."

"a man kicks with something or someone with his left leg."

Make-A-Video (Text-to-Video)

A confused grizzly bear in a calculus class

A golden retriever eating ice cream on a beautiful tropical beach at sunset, high resolution

A panda playing on a swing set

SORA (Text-to-Video)

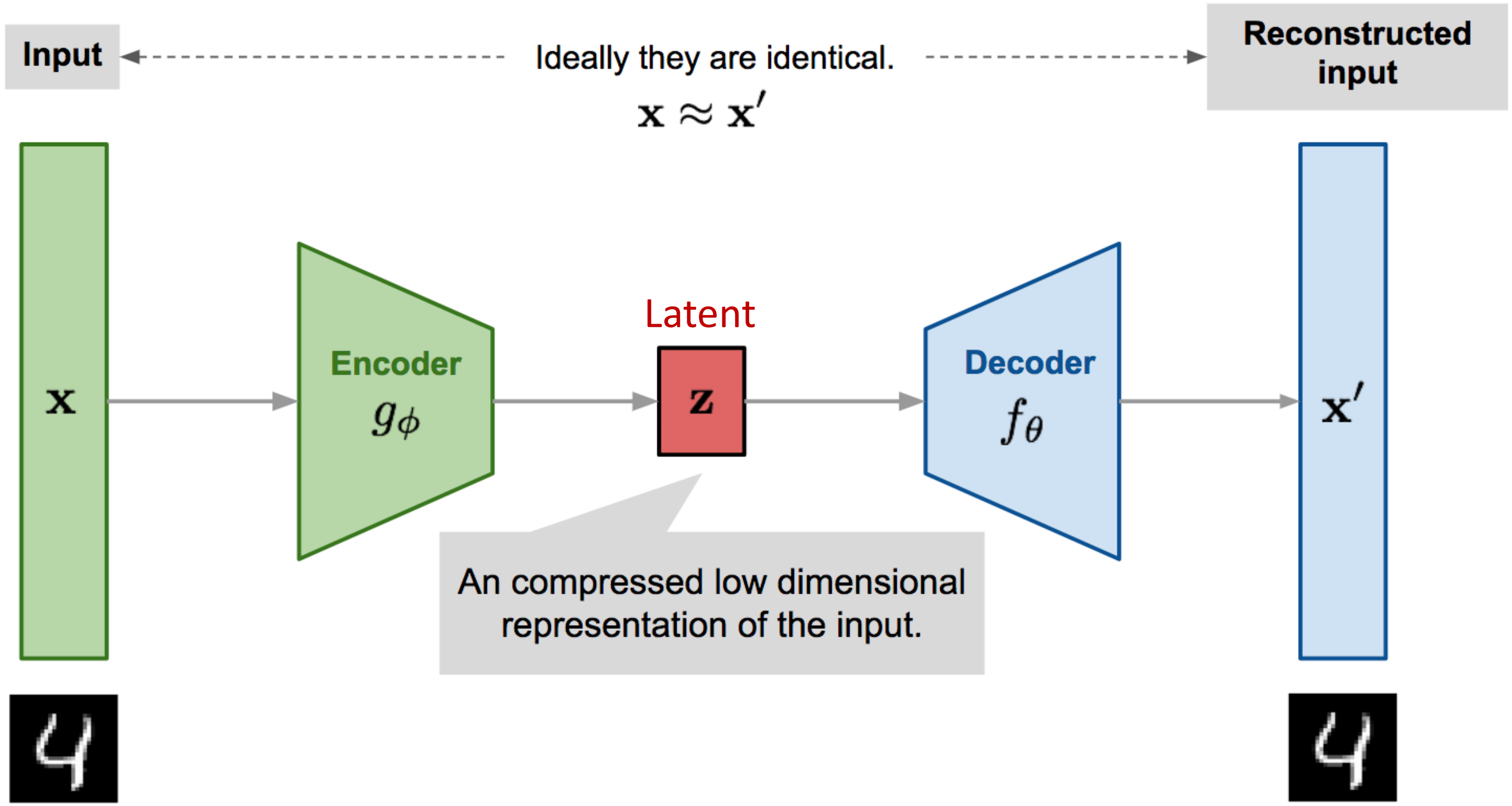
February 2024

Stable Diffusion 3

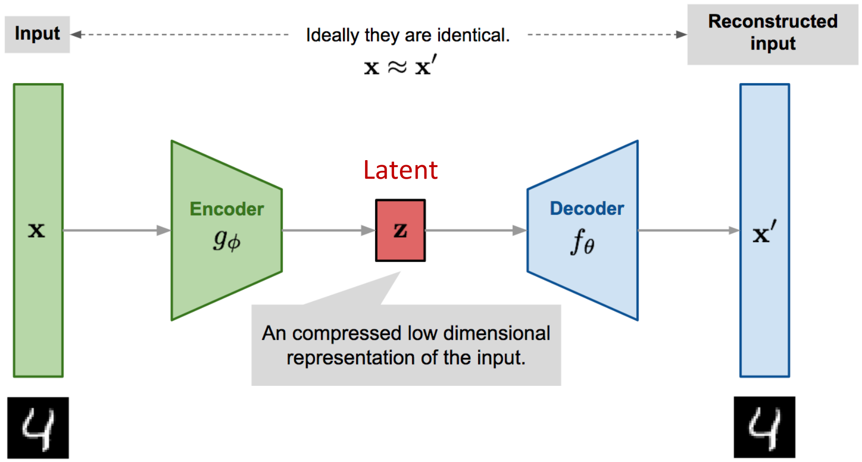
February 2024

AutoEncoder

AutoEncoder (AE)

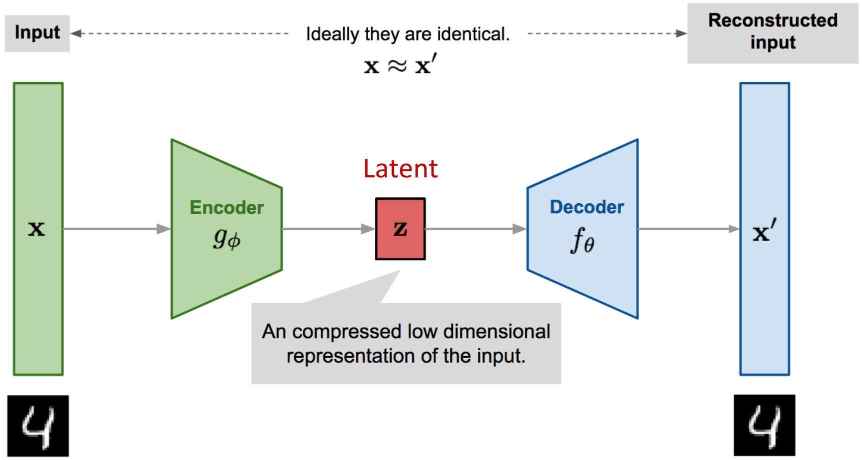


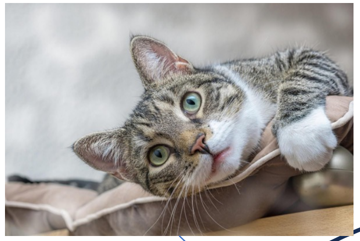
AutoEncoder (AE) Problems?

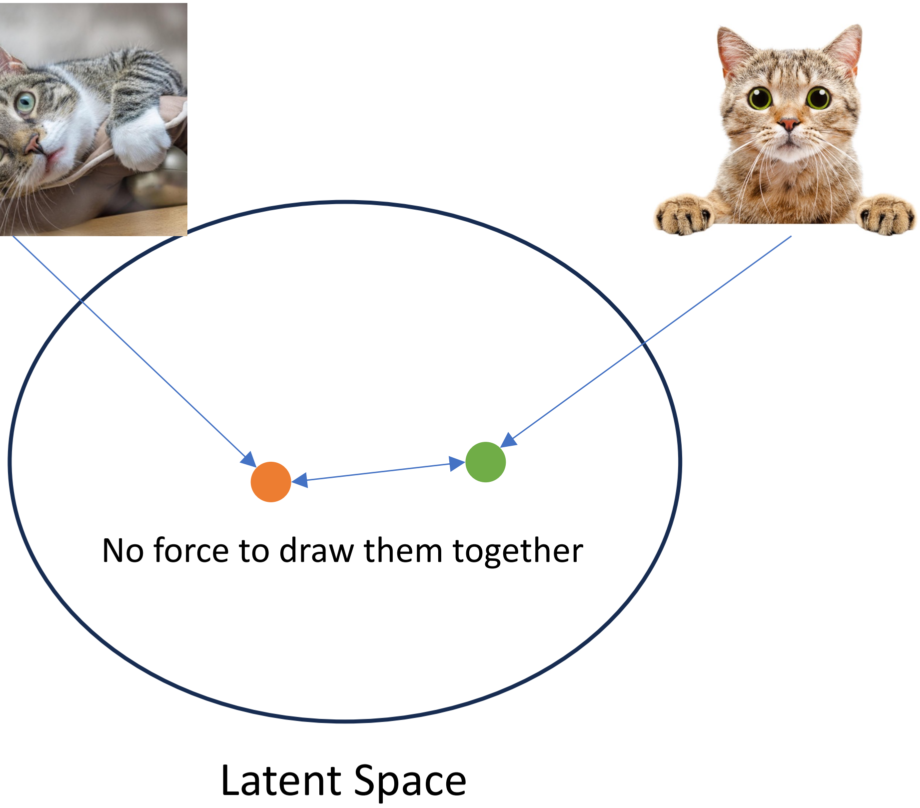


+ Easy to use, simple structure, fast to train

AutoEncoder (AE)

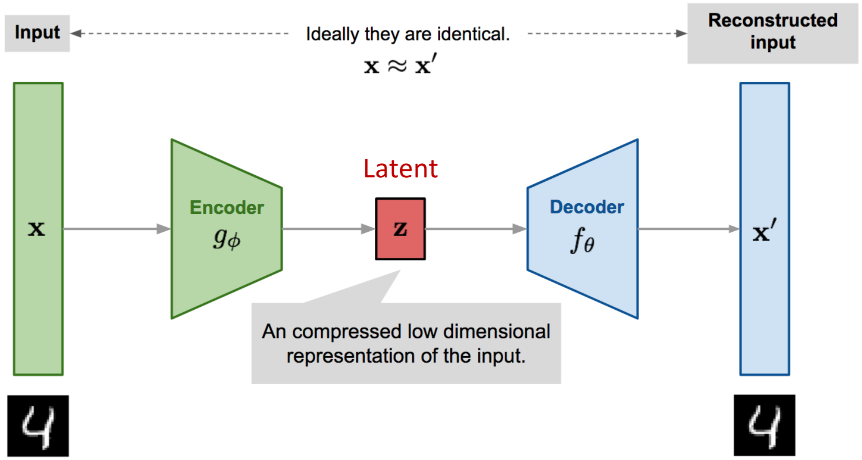


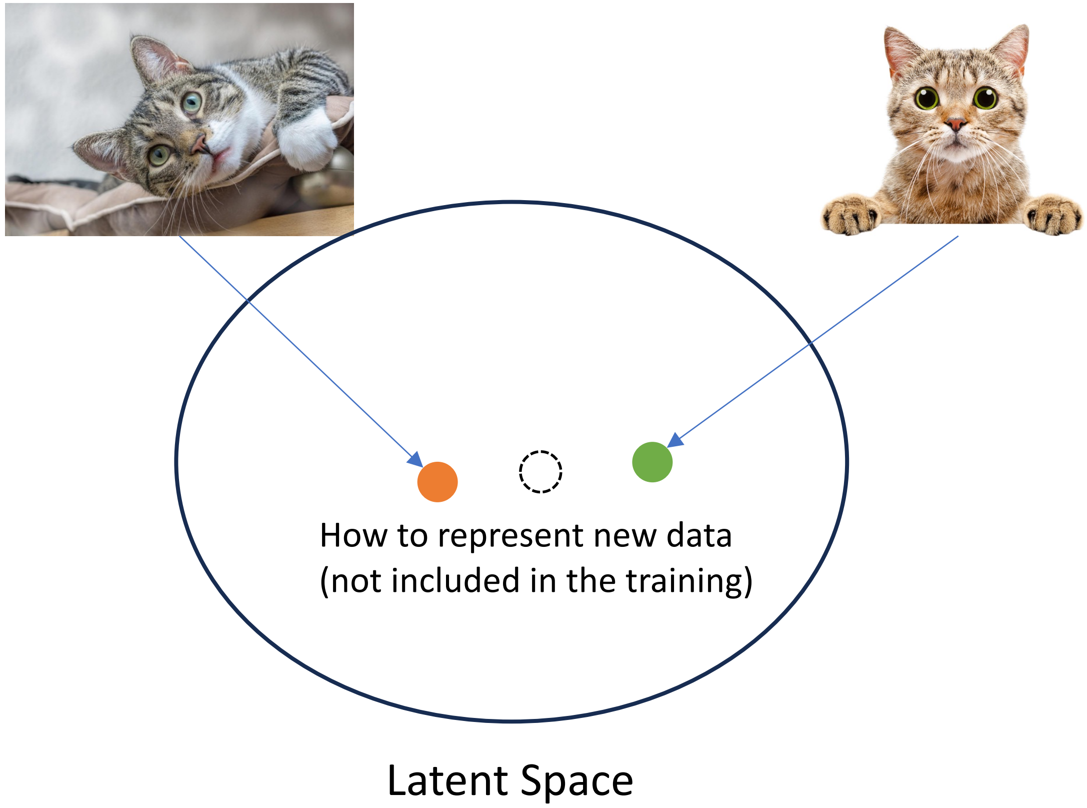




+ Easy to use, simple structure, fast to train
- Identity mapping is prone to overfit the data.
- non-interpolatable and non-smooth latent space
- Limited capacity of generating **<u>new</u>** data

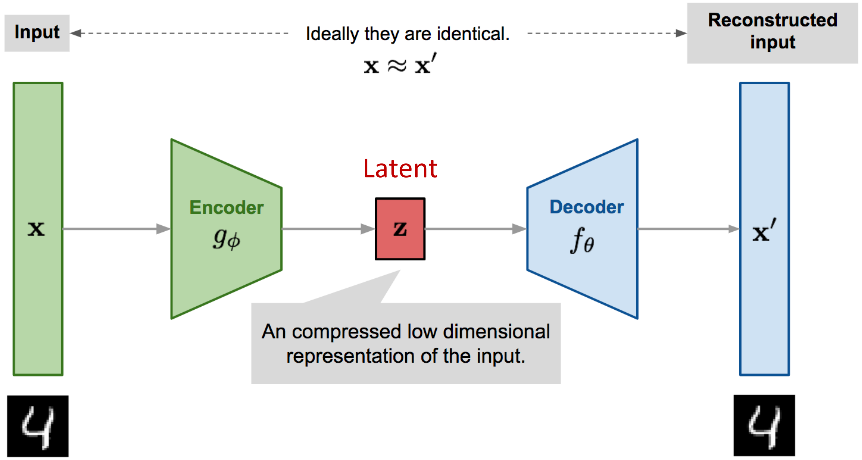
AutoEncoder (AE)

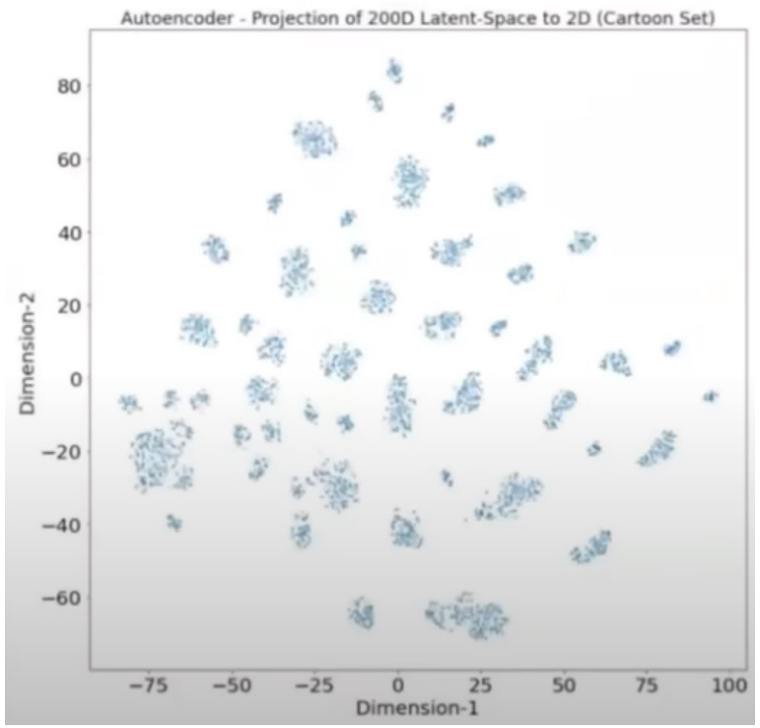




+ Easy to use, simple structure, fast to train
- Identity mapping is prone to overfit the data.
- non-interpolatable and non-smooth latent space
- Limited capacity of generating **new** data

AutoEncoder (AE)





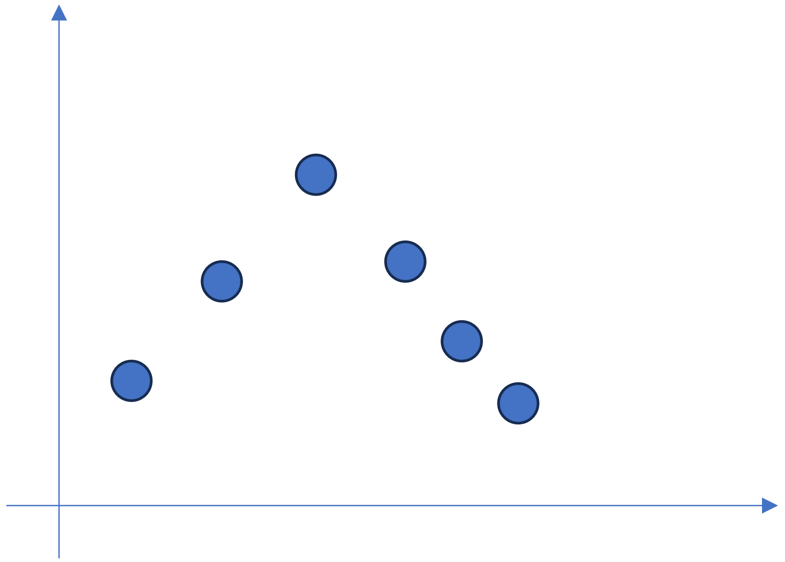
+ Easy to use, simple structure, fast to train
- Identity mapping is prone to overfit the data.
- non-interpolatable and non-smooth latent space
- Limited capacity of generating **new** data

Latent Space

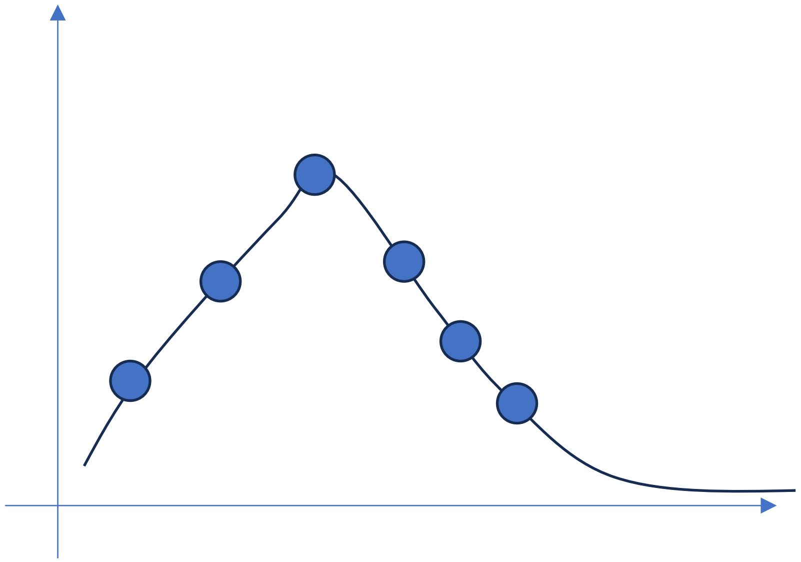
Variational AutoEncoder

Variational AutoEncoder (VAE)

Instead of learning data one by one:
Solution: Learn a distribution!



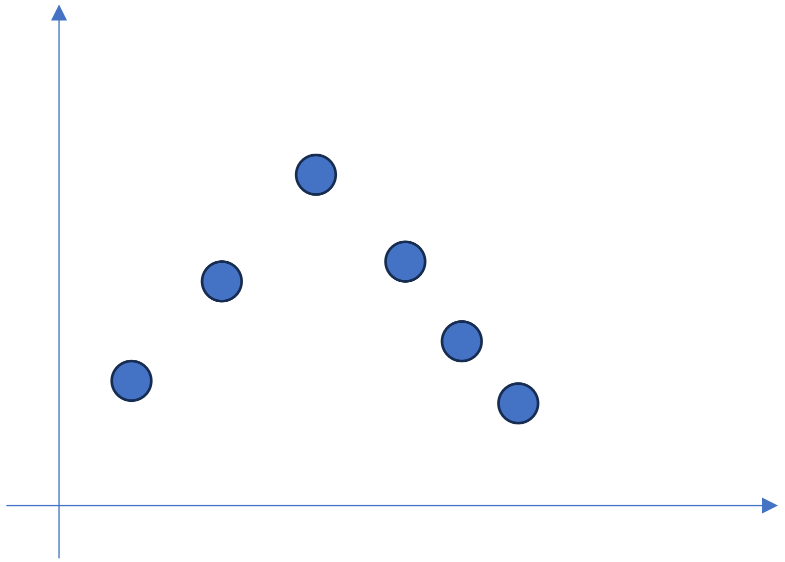
Data



Distribution of the Data
$p_\theta(X)$

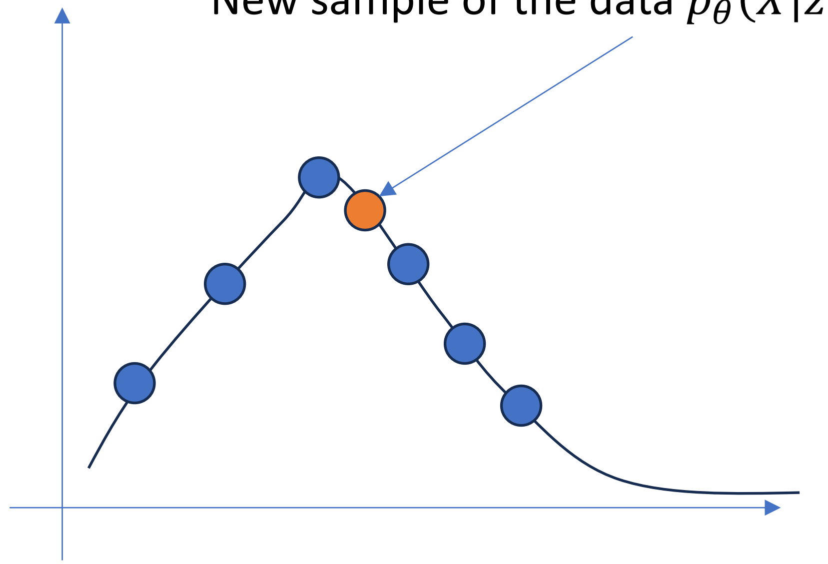
Variational AutoEncoder (VAE)

Instead of learning data one by one:
Solution: Learn a distribution with a hidden latent z!



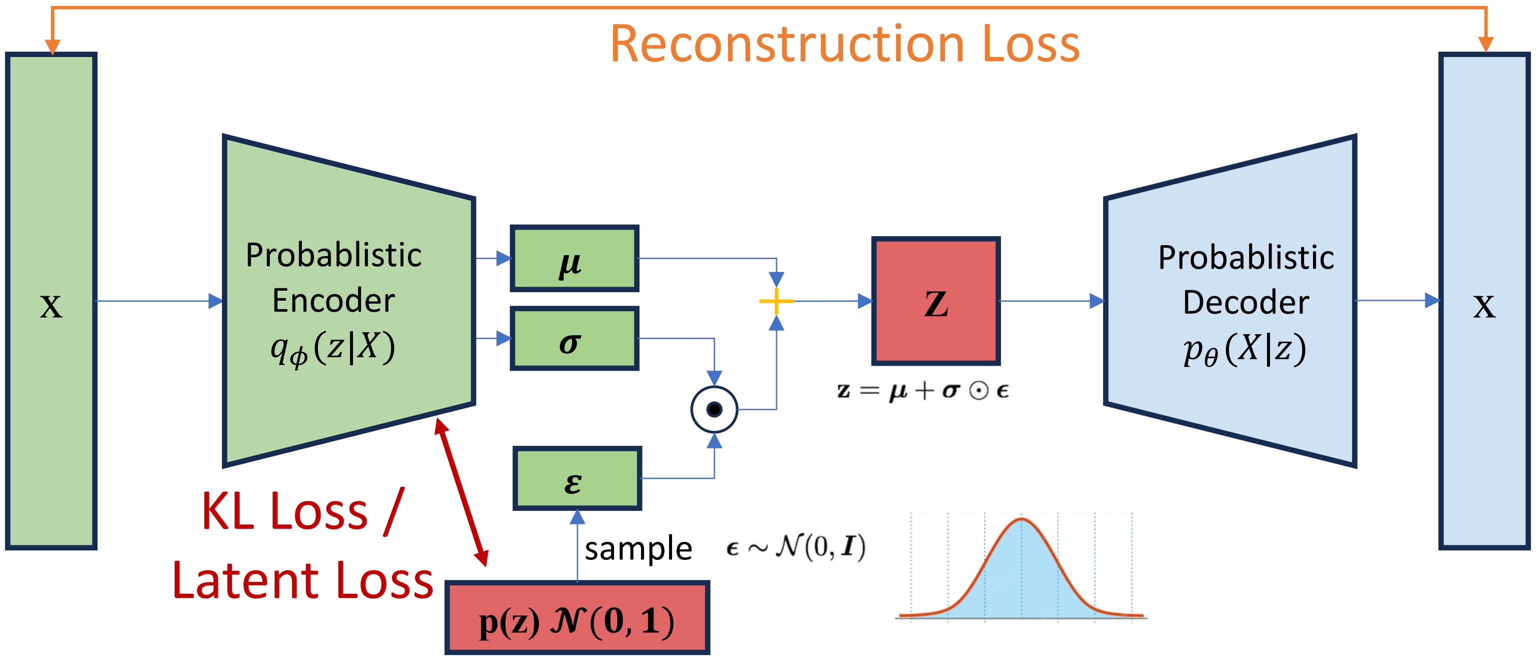
Data

New sample of the data $p_\theta(X|z); Z \sim p_\theta(z)$



Distribution of the Data with a hidden latent
$p_\theta(X|z)p_\theta(z)$

Variational AutoEncoder (VAE)

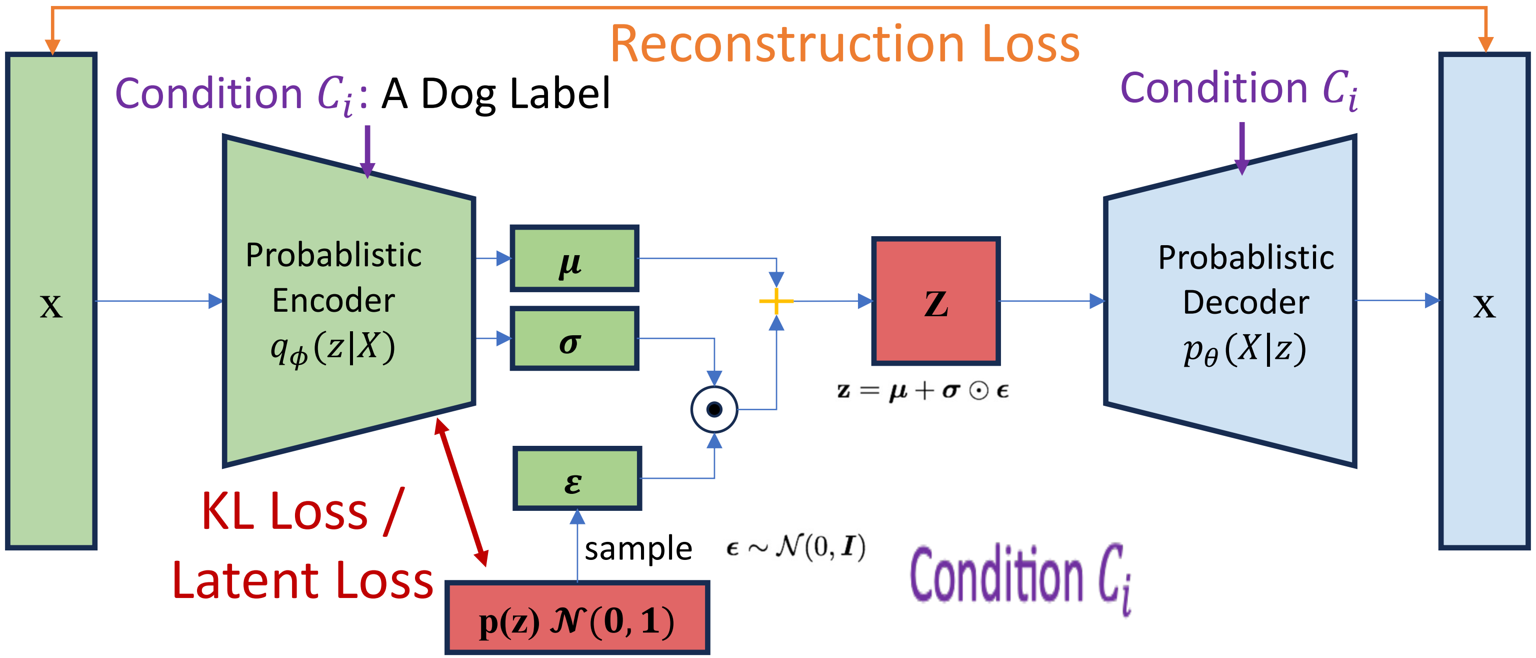


$$L_{\mathrm{VAE}}(\theta, \phi) = -\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) + D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}))$$

θ: Parameter of decoder
ϕ: Parameter of encoder

Kingma & Welling, 2014

Conditional VAE (CVAE)



$$L_{\mathrm{CVAE}}(\phi, \beta) = -\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c}_i)} \log p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{c}_i) + \beta D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{c}_i) \| p_\theta(\mathbf{z})\right)$$

θ: Parameter of decoder

ϕ: Parameter of encoder

Kingma & Welling, 2014

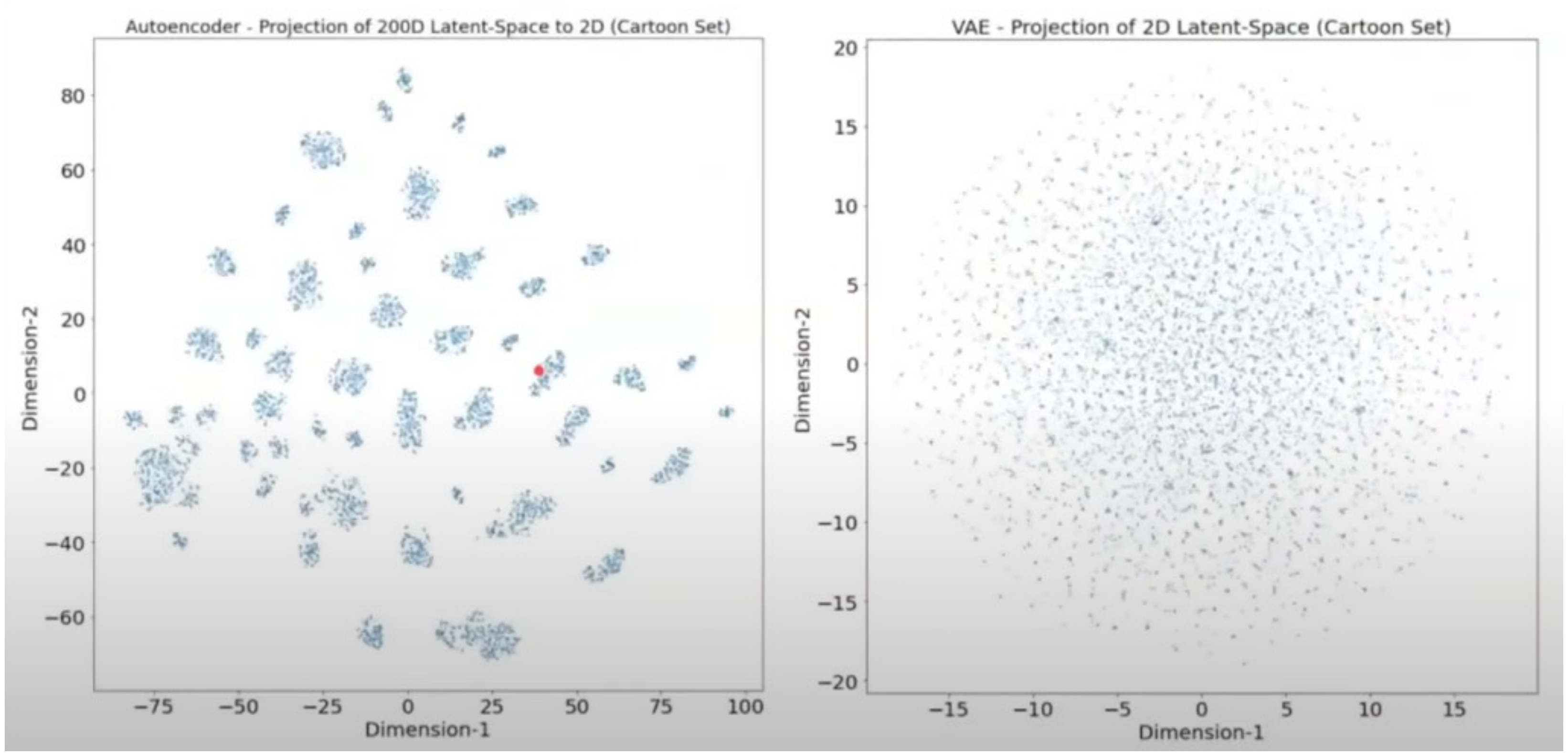
VAE Examples

Samples from Vanilla VAE (Kingma & Welling, 2014) on dataset CelebA

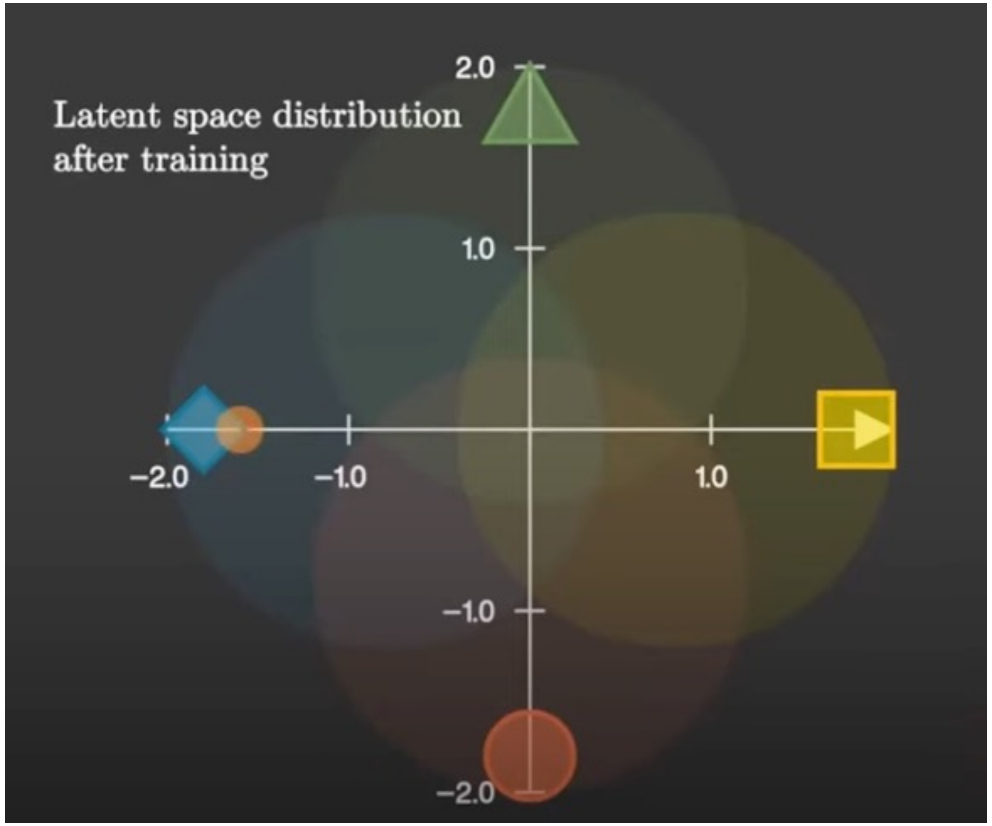
Problem with VAE

Guess?

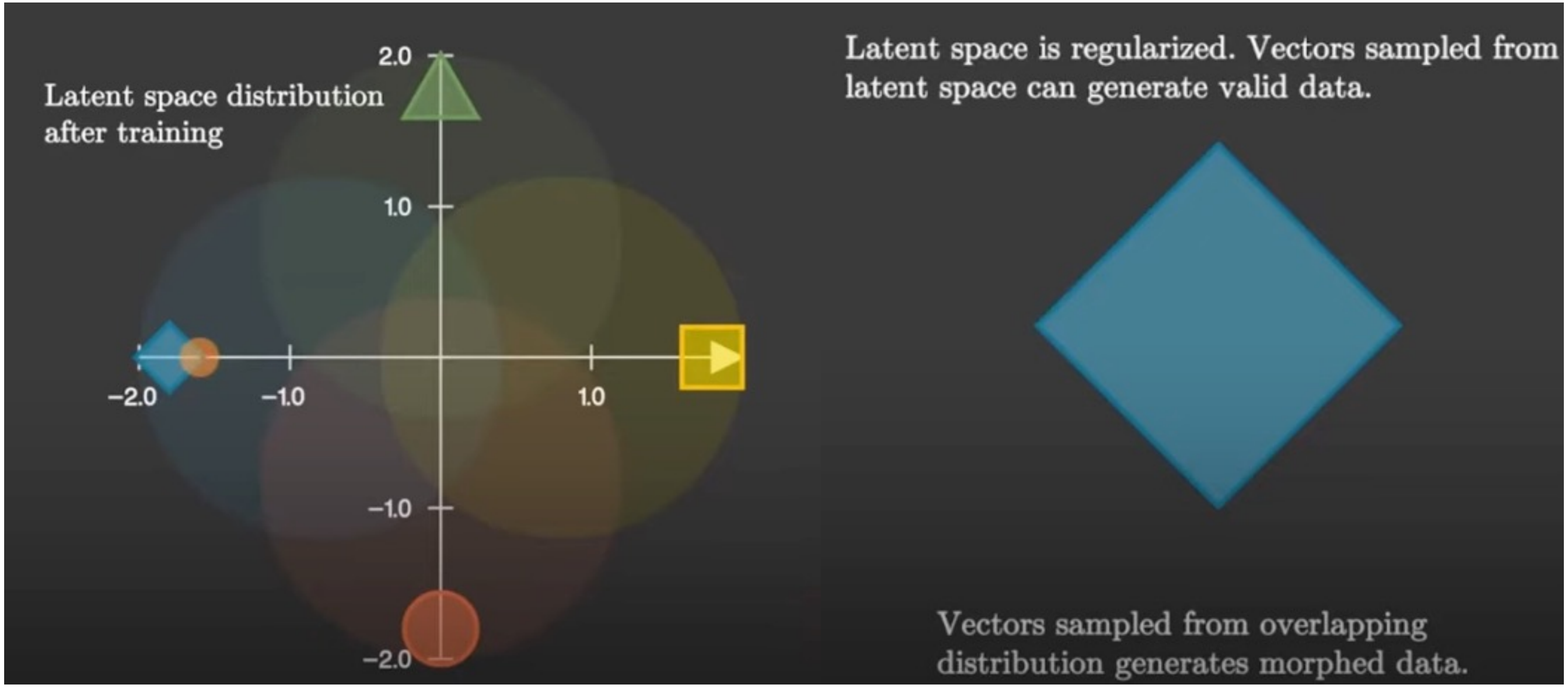
Problem with VAE



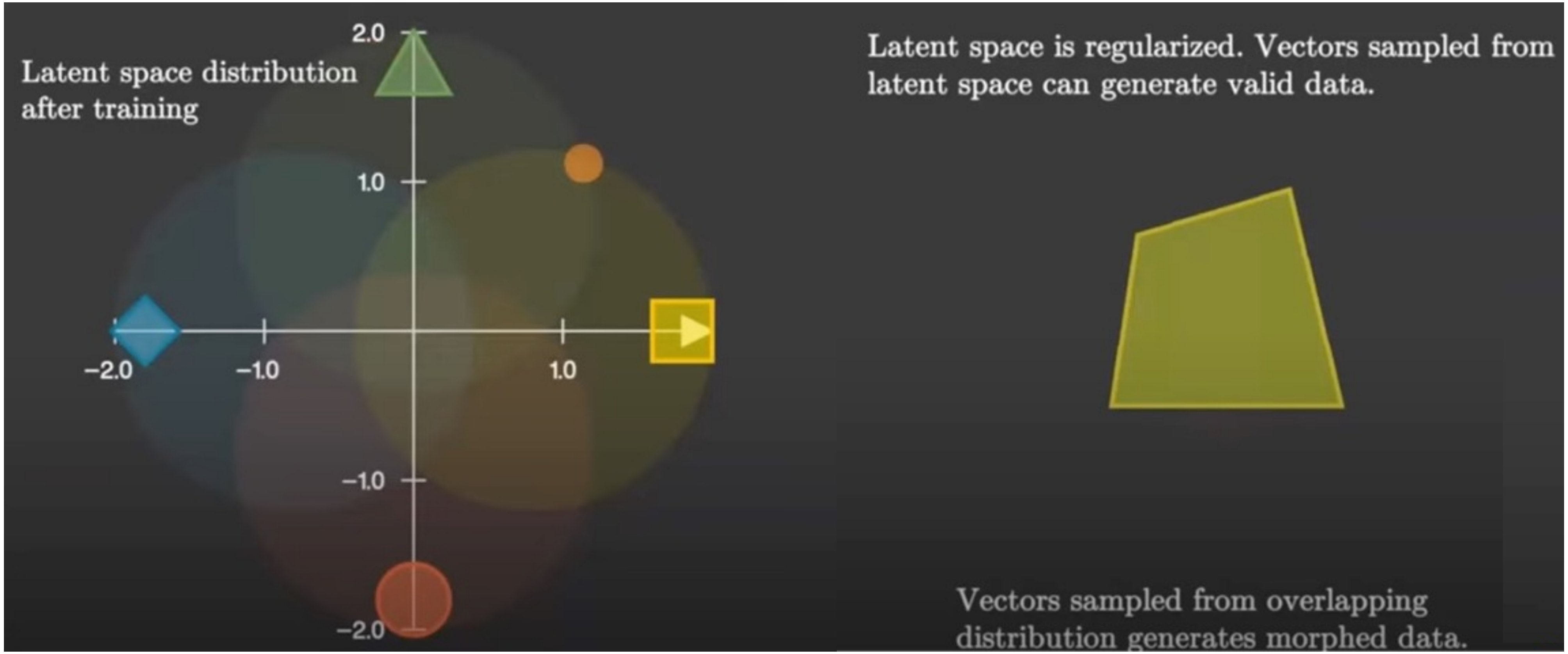
Problem with VAE



Problem with VAE

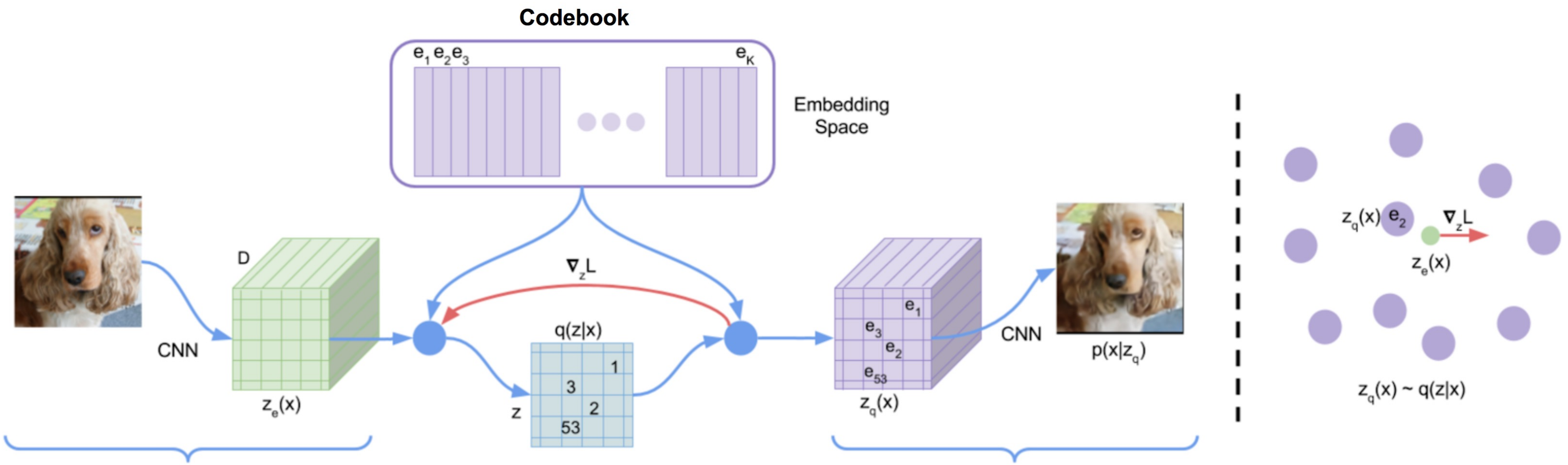


Problem with VAE



Vector Quantised-Variational AutoEncoder (VQVAE)

Vector Quantised-Variational AutoEncoder (VQVAE)



Encoder **Decoder**

Posterior categorical distribution:

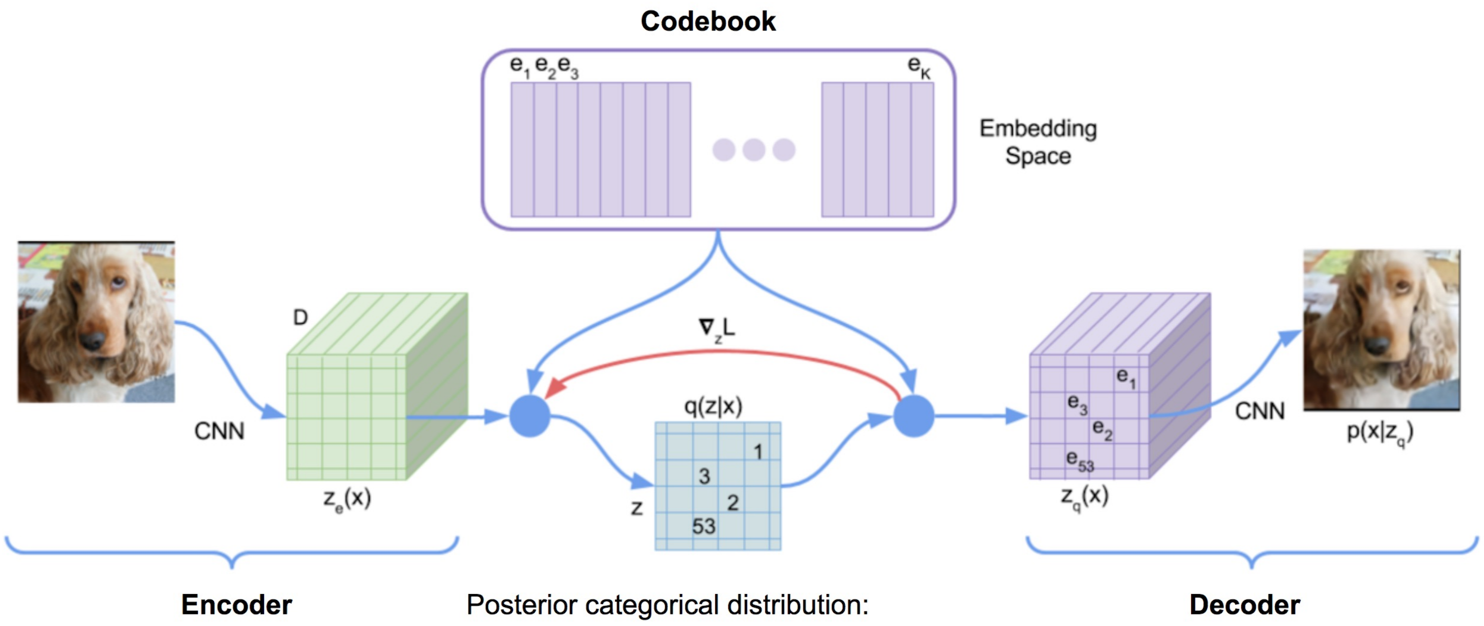
$$q(\mathbf{z} = \mathbf{e}_k|\mathbf{x}) = \begin{cases} 1 & \text{if } k = \arg\min_i \|\mathbf{z}_e(\mathbf{x}) - \mathbf{e}_i\|_2 \\ 0 & \text{otherwise.} \end{cases}$$

$$L = \underbrace{\|\mathbf{x} - D(\mathbf{e}_k)\|_2^2}_{\text{reconstruction loss}} + \underbrace{\|\text{sg}[E(\mathbf{x})] - \mathbf{e}_k\|_2^2}_{\text{VQ loss}} + \underbrace{\beta\|E(\mathbf{x}) - \text{sg}[\mathbf{e}_k]\|_2^2}_{\text{commitment loss}}$$

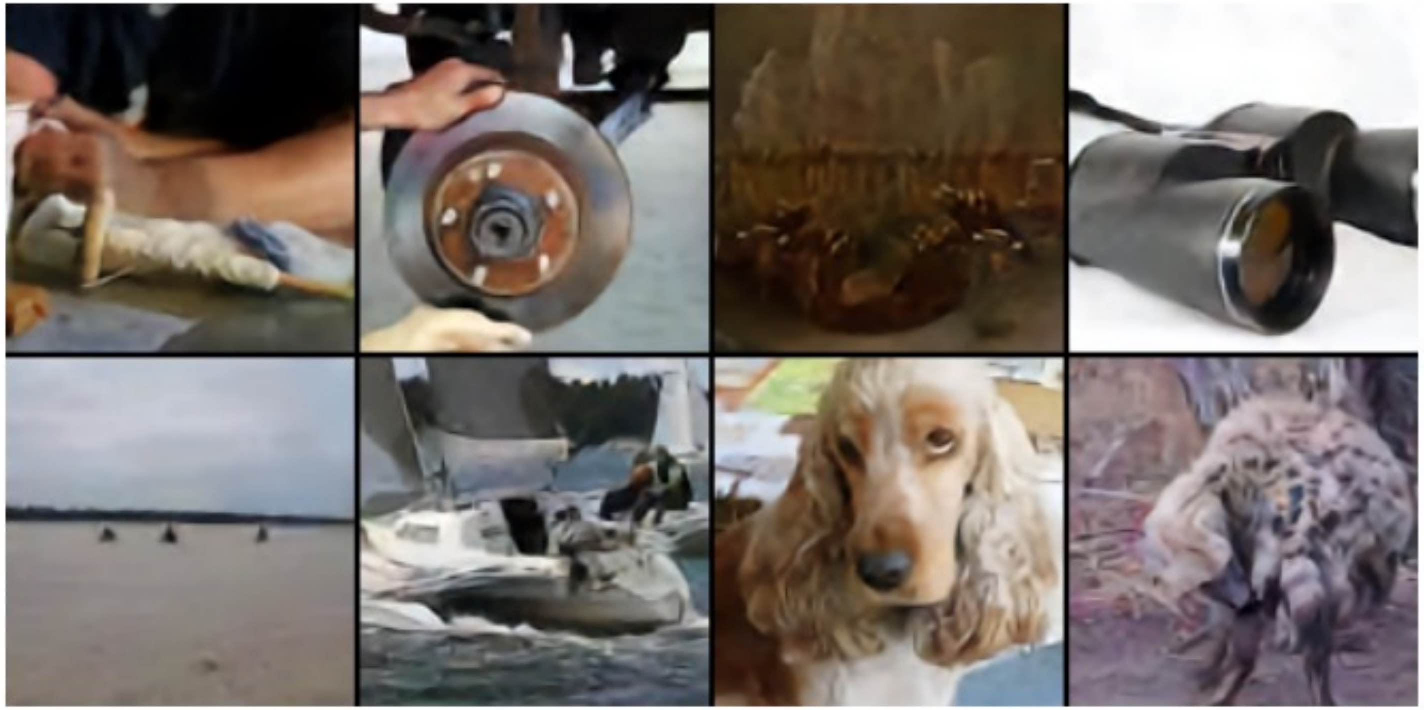
sg[.] for stop gradient

van den Oord, et al. 2017

Vector-Quantised-Variational AutoEncoder (VQVAE)

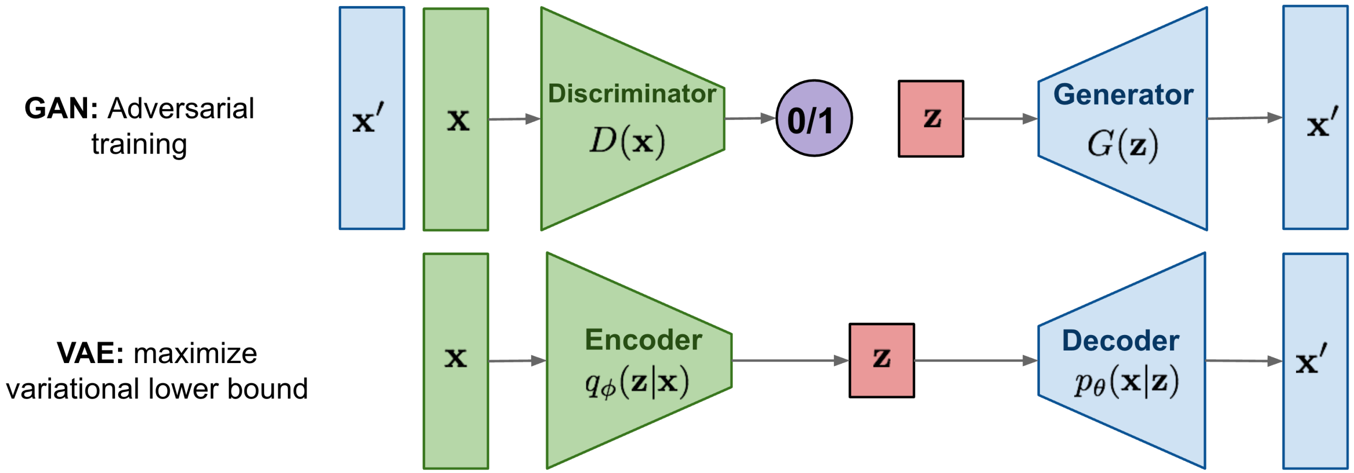


VQVAE



Sampled Results on ImageNet
VQVAE(Van den Oord, et al. 2017)

Generative Models



$$
\begin{aligned}
\min_G \max_D L(D, G) &= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \\
&= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1 - D(x)]
\end{aligned}
$$

unstable training
and mode collapse (learning data, instead of distribution)

$$
\begin{aligned}
L_{\text{VAE}}(\theta, \phi) &= -\log p_\theta(\mathbf{x}) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x})) \\
&= -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})) \\
\theta^*, \phi^* &= \arg\min_{\theta,\phi} L_{\text{VAE}}
\end{aligned}
$$

under-represtation of the distribution,
posteriori collapse (Gaussian Priori is not realistic)

Today's tutorial

Part I:
Introduction

Part II:
Diffusion &
Guidance &
Control

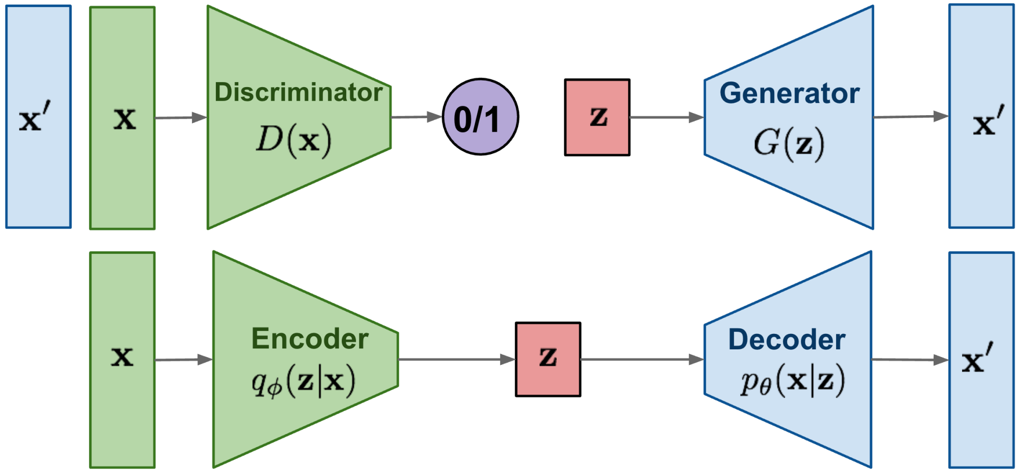
Part III:
My research

[Slides by V. Kalogeiton, X. Wang]

Generative Models

GAN: Adversarial training

VAE: maximize variational lower bound



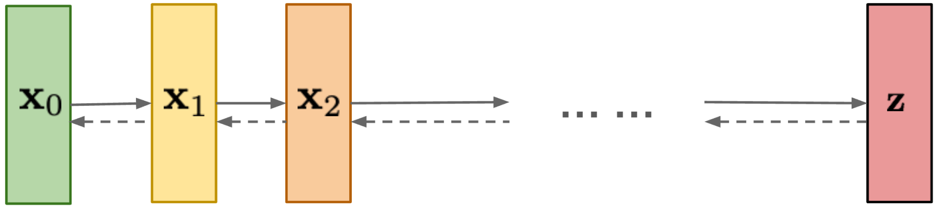
$$\min_G \max_D L(D,G) = \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$
$$= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1 - D(x))]$$

unstable training
and mode collapse (learning data, instead of distribution)

$$L_{\text{VAE}}(\theta,\phi) = -\log p_\theta(\mathbf{x}) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}))$$
$$= -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}))$$
$$\theta^*, \phi^* = \arg\min_{\theta,\phi} L_{\text{VAE}}$$

under-represtation of the distribution,
posteriori collapse (Gaussian Priori is not realistic)

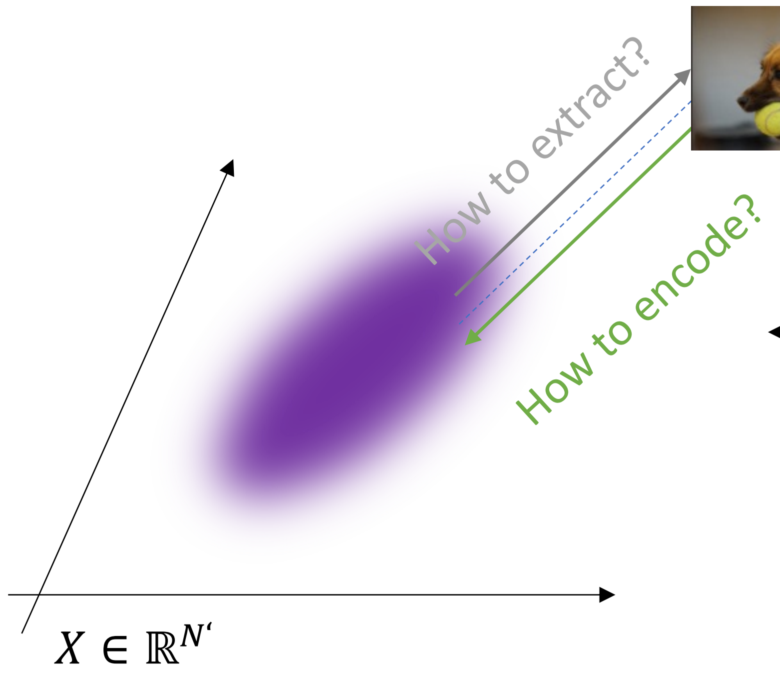
Diffusion models:
Gradually add Gaussian noise and then reverse



Better representation capacity,
and learn the whole distribution.

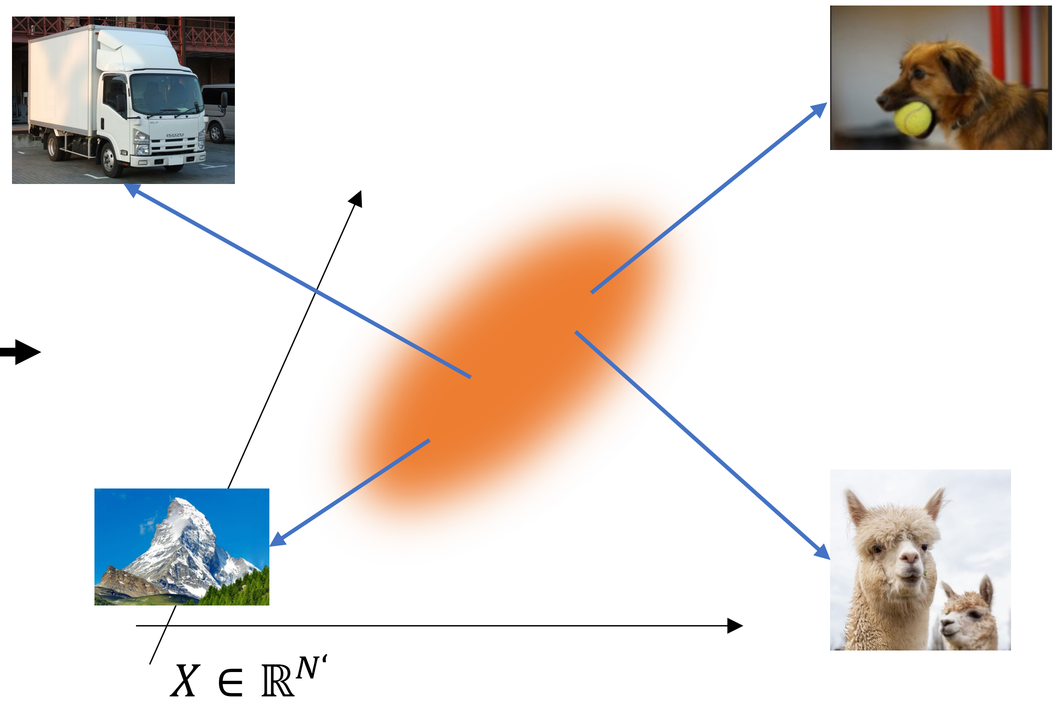
Diffusion Model

Generative Objective: Learn the distribution



find best θ

s.t. $P_\theta(X|z) \approx P(X|z)$



Distribution of $P(z)$ and we what to learn $P_\theta(X|z)$ with parameter $\theta \in \mathbb{R}^M$

Learning mapping of Real Data $P(X|z)$

Denoising Diffusion Probabilistic Models (DDPM): Forward Process

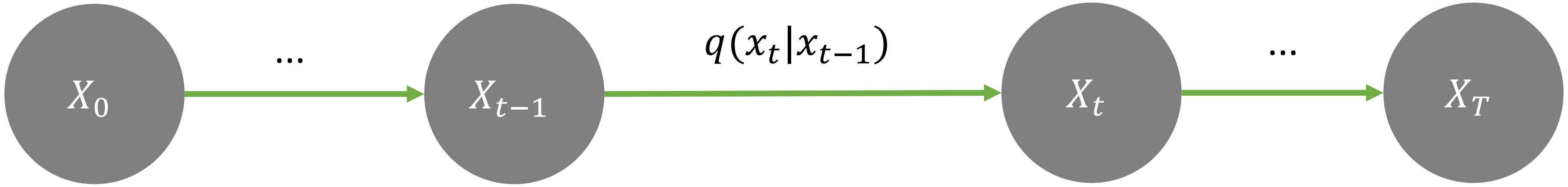
$$X_0 \xrightarrow{\dots} X_{t-1} \xrightarrow{q(x_t|x_{t-1})} X_t \xrightarrow{\dots} X_T$$

- Original image at X_0 and pure noise at X_T
- We repeat the noising T times
- $\beta_t \in (0,1)$ is a noise schedule

Forward:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

DDPM: Forward Process



- Original image at X_0 and pure noise at X_T
- We repeat the noising T times
- $\beta_t \in (0,1)$ is a noise schedule

Forward:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

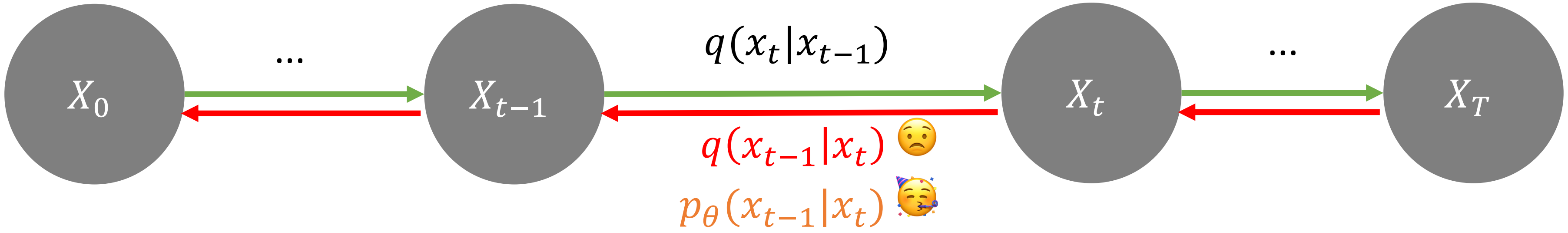
("Shortcut")
Sample any step using x_0 :

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$$

$$\alpha_t = 1 - \beta_t \qquad \bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$$

DDPM: Reverse (=Generative) Process



Generation:

$$\boxed{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underline{\underline{\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)}}, \underline{\underline{\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)}})}$$

Learnable parameters

A very nice property of Gaussian:
if $q(x_t|x_{t-1})$ is a Gaussian with small β (another reason we need many steps!)
→ then, $q(x_{t-1}|x_t)$ is also a Gaussian.
Therefore, we learn this Gaussian's mean and variance
by a network approximated $p_\theta(x_{t-1}|x_t)$

DDPM: Reverse/Generative Process

X_0 … X_{t-1} $\xrightarrow{q(x_t|x_{t-1})}$ X_t … X_T

$q(x_{t-1}|x_t)$ 😟

$p_\theta(x_{t-1}|x_t)$ 🥳

Forward

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

Bayes

$$q(x_{t-1} \mid x_t) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I})$$

$$\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_t\right)$$

$$\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$$

$\alpha_t = 1 - \beta_t$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Generation

DDPM: Reverse/Generative Process

X_0 … X_{t-1} $\xrightarrow{q(x_t|x_{t-1})}$ X_t … X_T

$q(x_{t-1}|x_t)$ 😟

$p_\theta(x_{t-1}|x_t)$ 🥳

Forward

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

Bayes

$$q(x_{t-1} \mid x_t) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I})$$

Minimize KL divergence

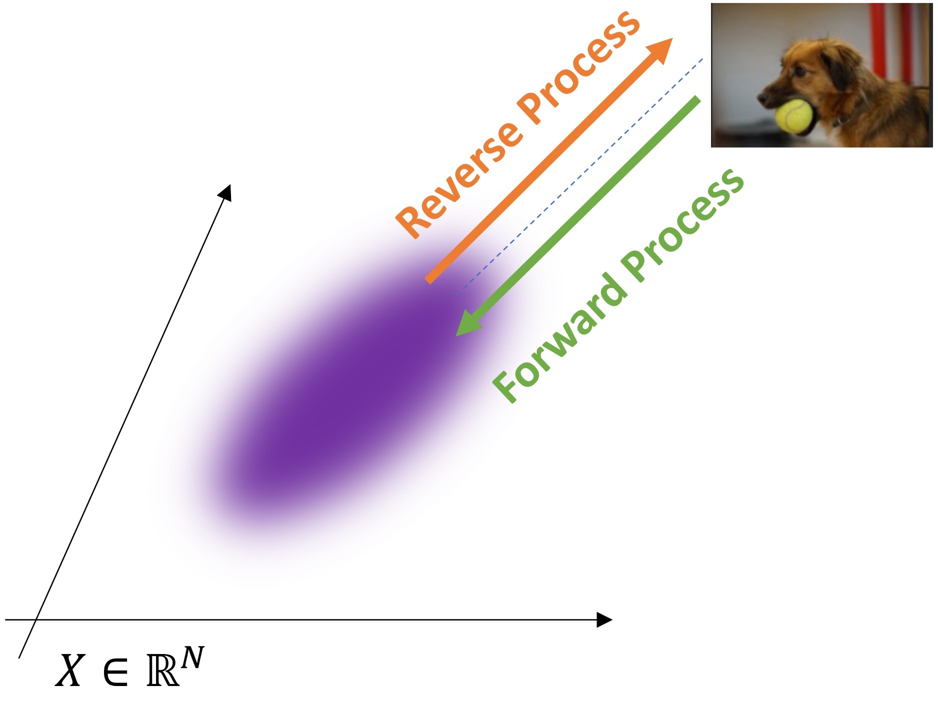
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Generation

Diffusion
Guidance

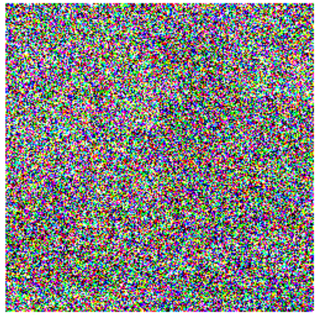
- Guided diffusion
- Control the diffusion
- Explicit condition
- Guided diffusion
- Why not guided diffusion?
- Classifier-free guidance
- Negative prompting

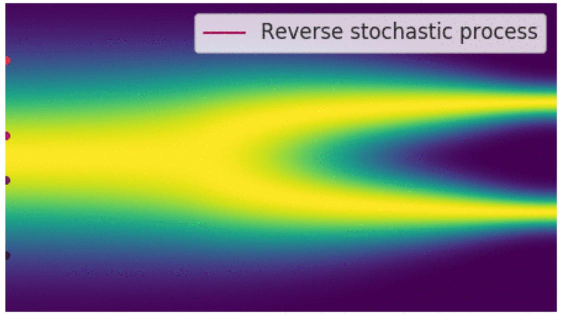
Control the Diffusion Model



Distribution of Learnt Data $P_\theta(X)$ with parameter $\theta \in \mathbb{R}^M$

Good, it means one noise gives me an image!

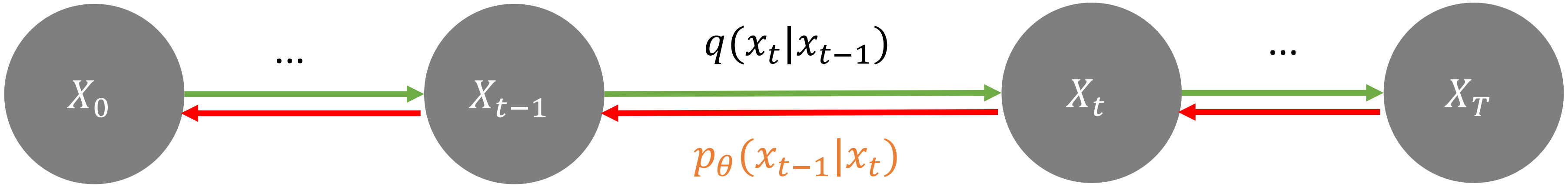




But how can I achieve **control** on this? For example, I want a dog image, rather than others.

Or even more complicated: "*A stained glass window of a panda eating bamboo.*" – text-to-image generation

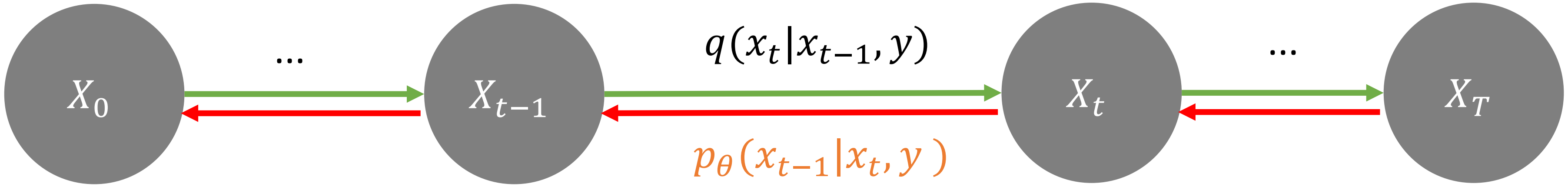
Control the Diffusion Model



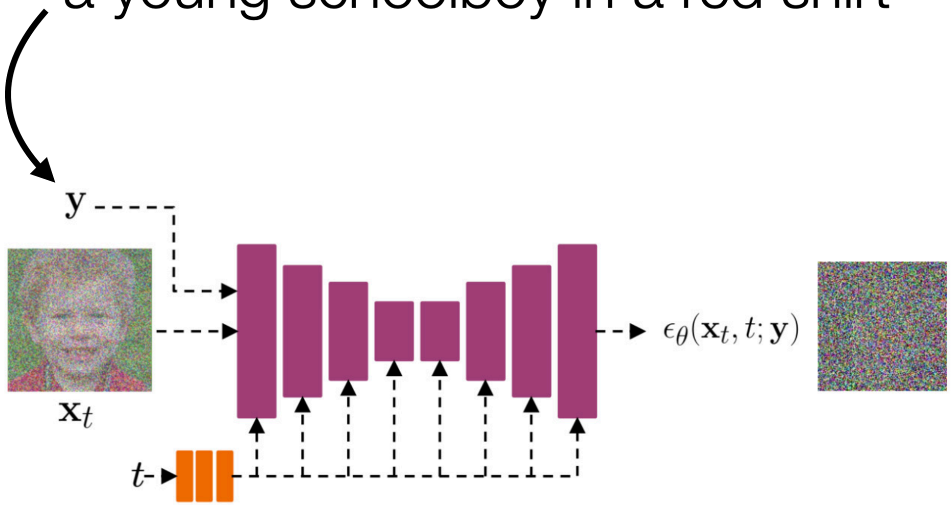
Where is the control?
How did we do with other models, e.g. VAE?

Explicit condition

Control the Diffusion Model: Explicit Condition

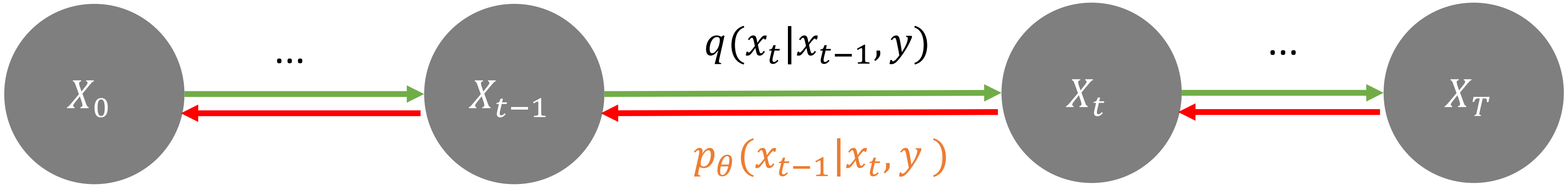


"a young schoolboy in a red shirt"

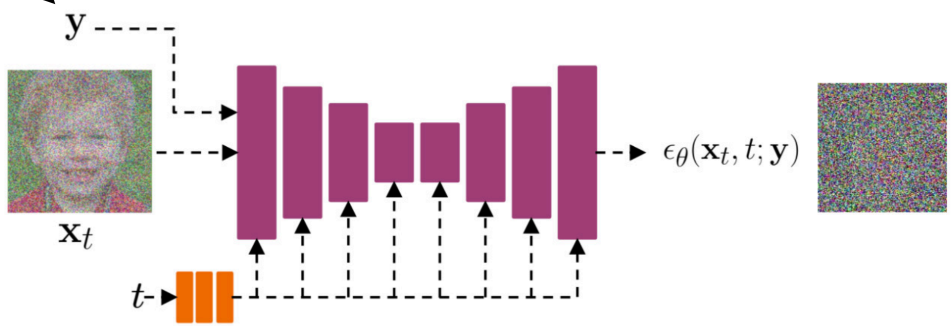


We can add it directly.

Control the Diffusion Model: Explicit Condition



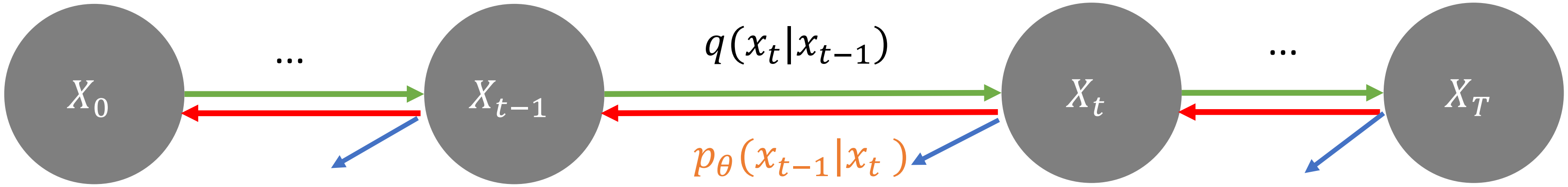
"a young schoolboy in a red shirt"



We can add it directly, but is this an effective way? Why?

Guided diffusion

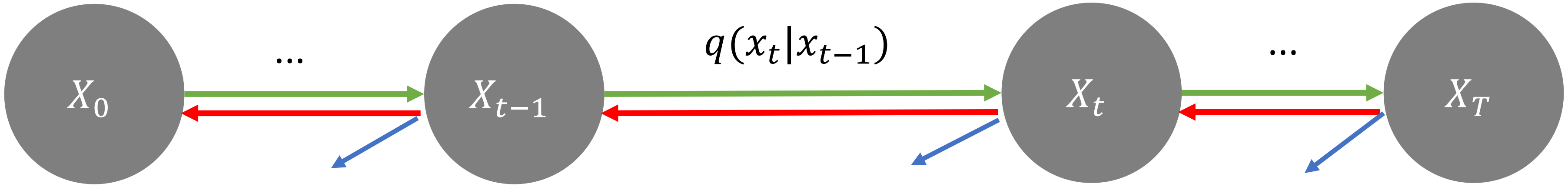
Control the Diffusion Model: Guided Diffusion



Shift to a dog label!

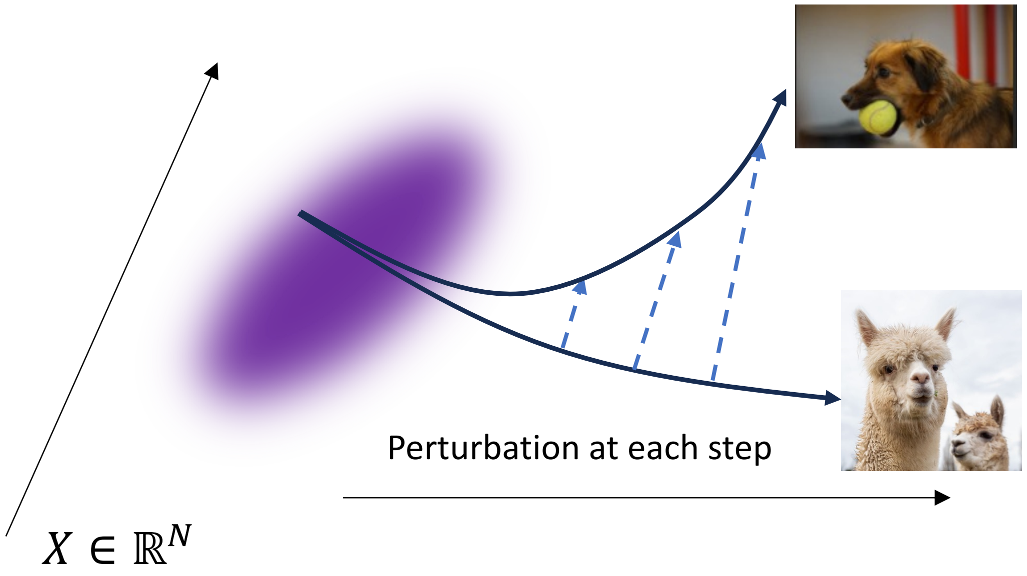
Let's perturb it step-by-step during the generation!

Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!



It's like a classifier: $p_\phi(y|x)$

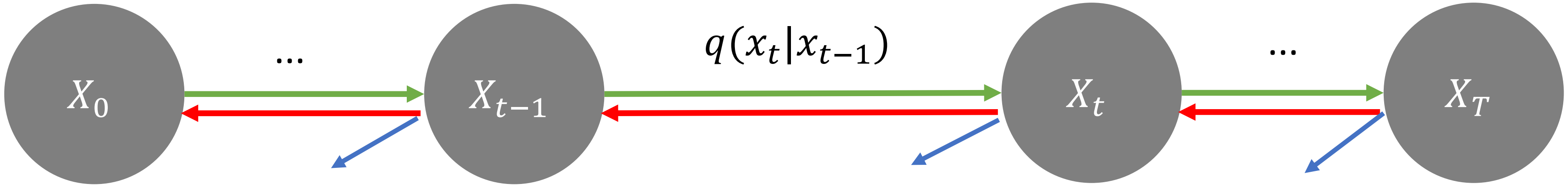
We just look at its gradient:

$\nabla_x \log p_\phi(y|x)$

Such that the generated x is similar to the condition label.

In sampling: $\epsilon_\theta(x_t, t) + \nabla_x \log p(y|x)$

Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

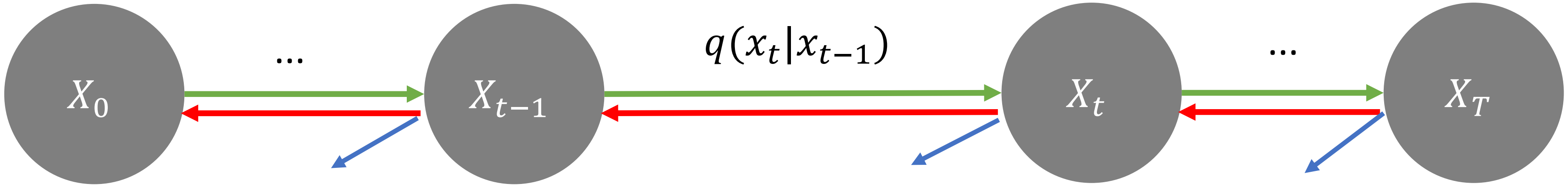
Shift to a dog label!

In sampling: $\epsilon_\theta(x_t, t) + \nabla_x \log p_\phi(y|x)$. ← **Guided Diffusion**

We need to train a classifier: $p_\phi(y|x)$, with the awareness of noise

$p_\phi(y|x)$ ⇒ A Dog

Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!

In sampling: $\epsilon_\theta(x_t, t) + \gamma \nabla_x \log p_\phi(y|x)$. ← **Guided Diffusion**

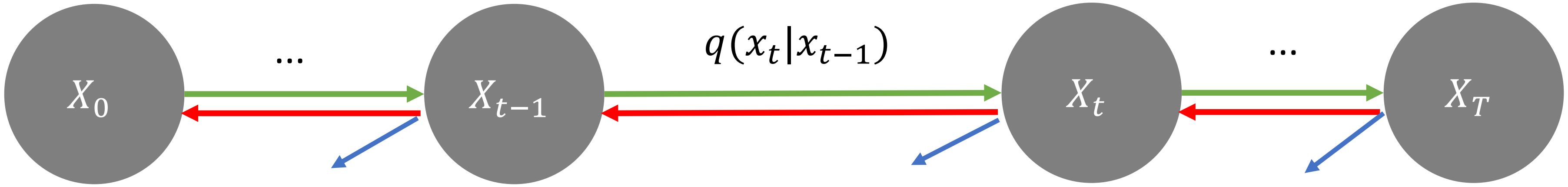
Label: Corgi

γ = 1

γ = 3

Dhariwal and Nichol, 2021

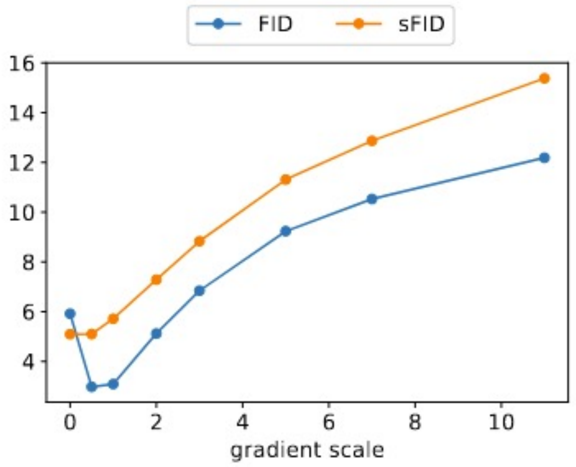
Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!

In sampling: $\epsilon_\theta(x_t, t) + \gamma\nabla_x \log p_\phi(y|x)$. ← **Guided Diffusion**



Dhariwal and Nichol, 2021

Guided Diffusion: Nearest Neighbors for Samples

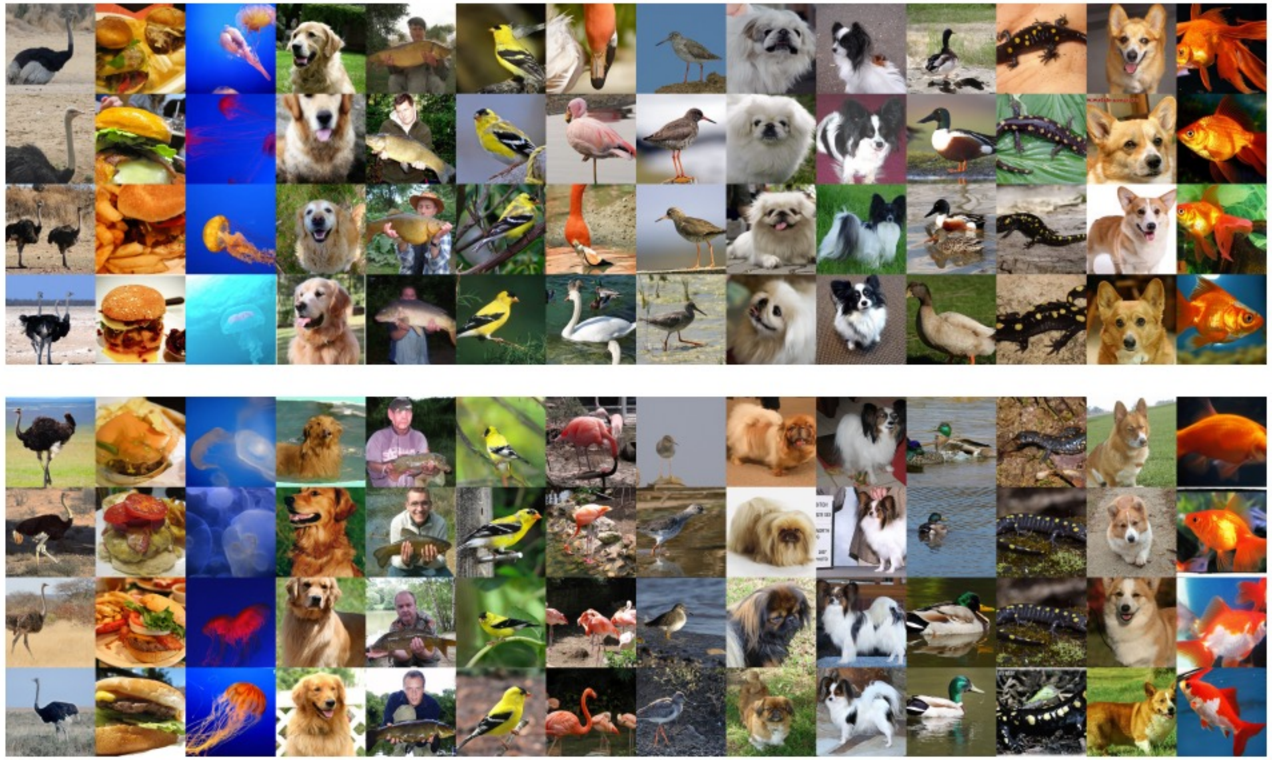


Figure 7: Nearest neighbors for samples from a classifier guided model on ImageNet 256×256. For each image, the top row is a sample, and the remaining rows are the top 3 nearest neighbors from the dataset. The top samples were generated with classifier scale 1 and 250 diffusion sampling steps (FID 4.59). The bottom samples were generated with classifier scale 2.5 and 25 DDIM steps (FID 5.44).

Guided Diffusion: Effect of Varying the Classifier Scale

Figure 8: Samples when increasing the classifier scale from 0.0 (left) to 5.5 (right). Each row corresponds to a fixed noise seed. We observe that the classifier drastically changes some images, while leaving others relatively unaffected.

Guided Diffusion: Examples

Figure 13: Samples from our best 512×512 model (FID: 3.85). Classes are 1: goldfish, 279: arctic fox, 323: monarch butterfly, 386: african elephant, 130: flamingo, 852: tennis ball.

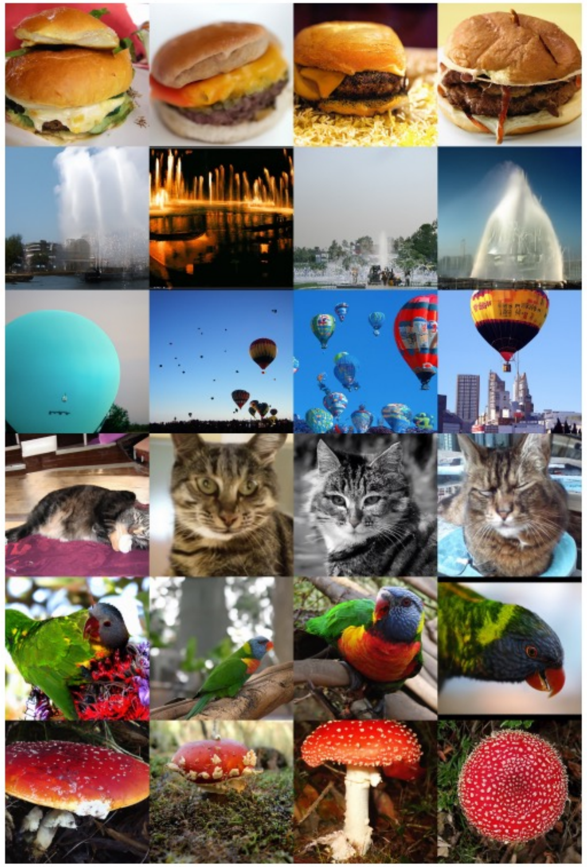
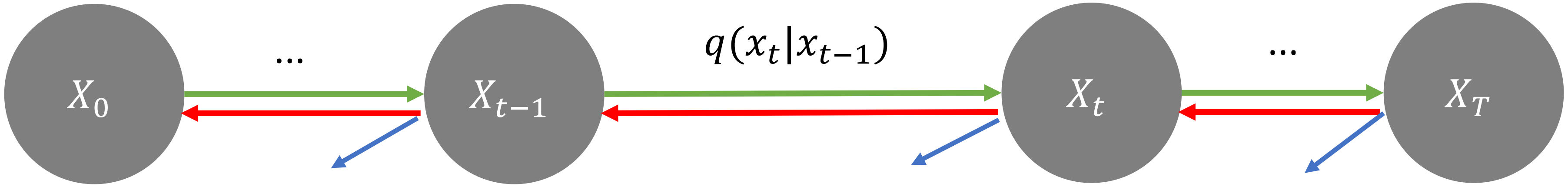


Figure 14: Samples from our best 512×512 model (FID: 3.85). Classes are 933: cheeseburger, 562: fountain, 417: balloon, 281: tabby cat, 90: lorikeet, 992: agaric.

Why not guided diffusion?

Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!

In sampling: $\epsilon_\theta(x_t, t) + \gamma \nabla_x \log p_\phi(y|x)$. ← **Guided Diffusion**

What do we NOT like in guided diffusion?

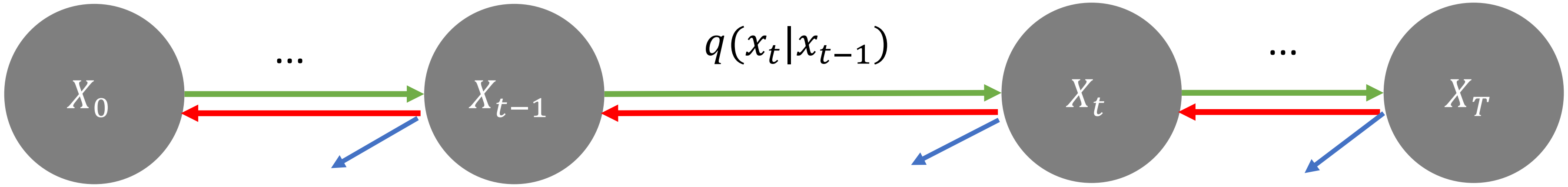
- Need to fine-tune and train a classifier
- Condition can only be label-based, hard to support other conditions like "text input"

Because for text, the classifier $p_\phi(y|x)$ ***does not*** exist.

Dhariwal and Nichol, 2021

Classifier-free guidance

Control the Diffusion Model: Classifier-Free Guidance



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!

At training: $p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

In sampling: $\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma \nabla_x \log p(y|x)$

$\nabla_x \log p(y|x) \propto \epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t)$

Finally: $\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma(\underbrace{\epsilon_\theta(x_t, t, y)}_{\text{conditional generation}} - \underbrace{\epsilon_\theta(x_t, t)}_{\text{unconditional generation}})$

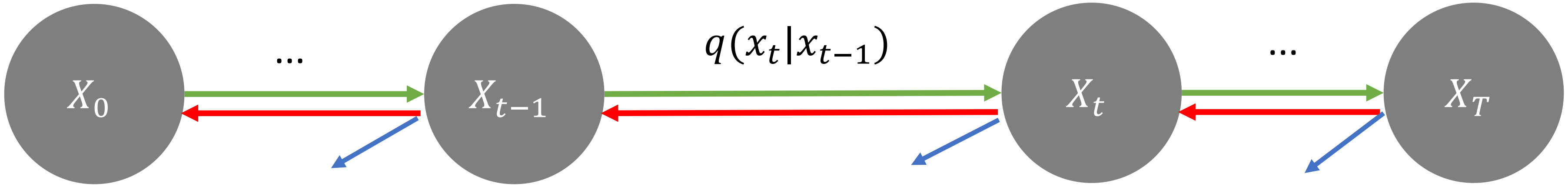
$$p(y|x) \propto \frac{p(x|y)}{p(x)}$$

$$\nabla_x \log p(y|x) \propto \nabla_x \log p(x|y) - \nabla_x \log p(x)$$

Thanks to Bayes

Ho and Salimans, 2022

Control the Diffusion Model: Classifier-Free Guidance



$$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$$

Shift to a dog label!

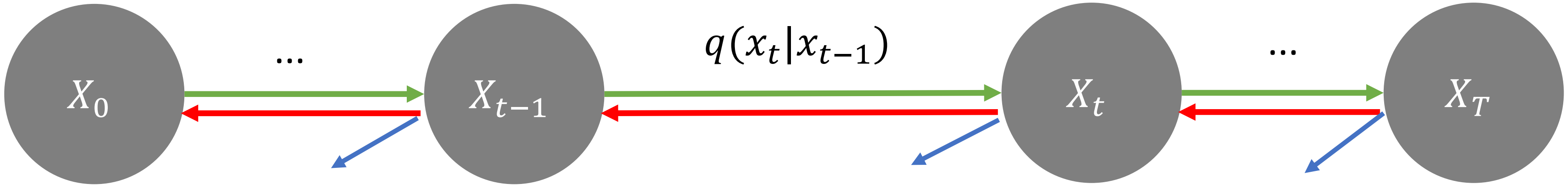
$$\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma \underbrace{(\underbrace{\epsilon_\theta(x_t, t, y)}_{\text{conditional generation}} - \underbrace{\epsilon_\theta(x_t, t)}_{\text{unconditional generation}})}_{\text{Implicit classifier}}$$

How to compute: $\epsilon_\theta(x_t, t, y)$ →
- explicit condition

How to compute: $\epsilon_\theta(x_t, t)$ →
- We randomly set the condition to null (drop-out condition)
- $\epsilon_\theta(x_t, t, y)$ → $\epsilon_\theta(x_t, t, \emptyset)$

Ho and Salimans, 2022

Control the Diffusion Model: Classifier-Free Guidance



$$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$$

Shift to a dog label!

$$\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$$

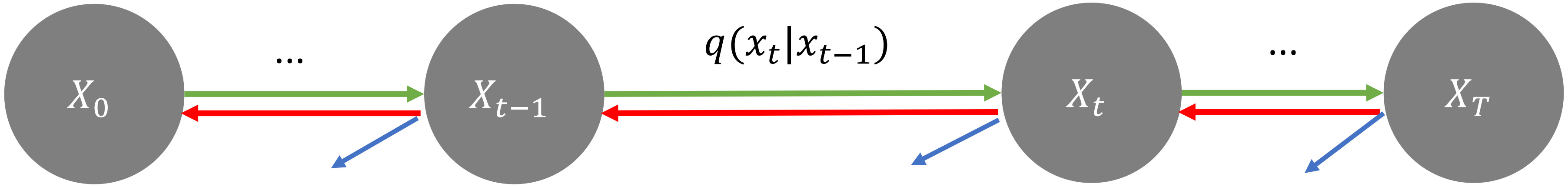
Label: Husky

γ=1

γ=3

Ho and Salimans, 2022

Control the Diffusion Model: Classifier-Free Guidance



$$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$$

Shift to a dog label!

$$\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$$

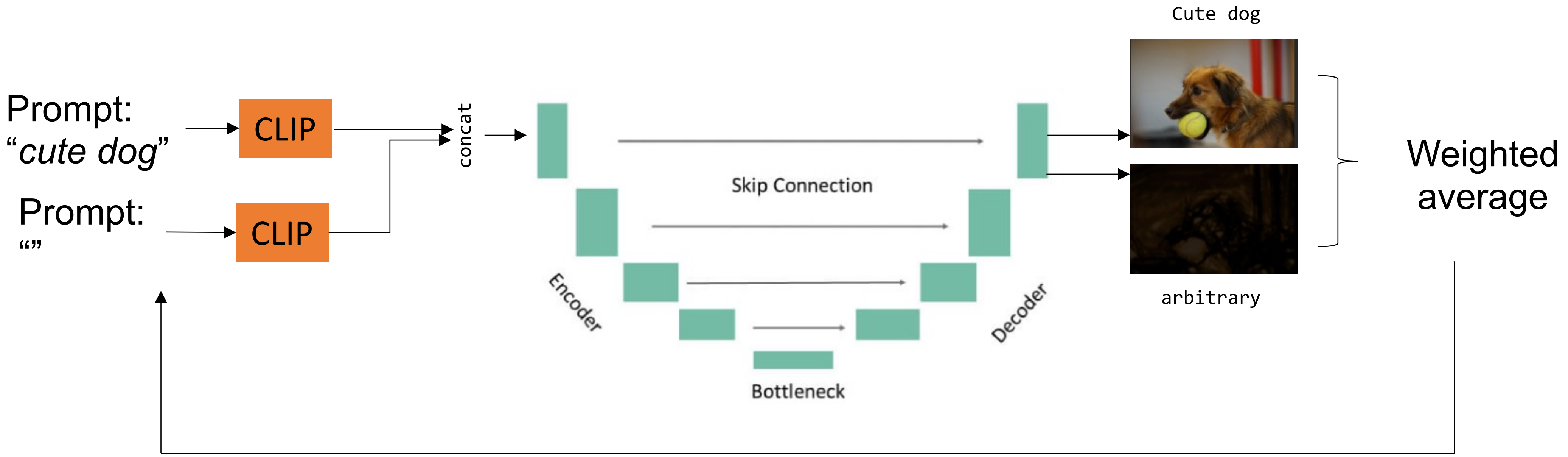
We do not need the explicit classifier: we can use text-encoder to condition on text.
Called : Classifier-Free Guidance (CFG)

“A panda is eating ice-cream” → CLIP/T5 → 768x1 dim

Condition Latent

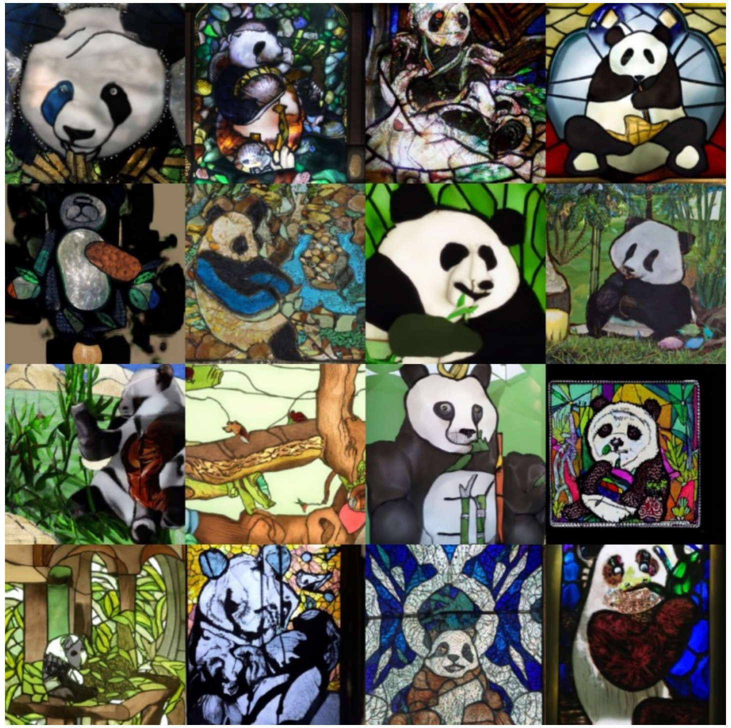
Ho and Salimans, 2022

Classifier free guidance



Classifier free guidance

Control the Diffusion Model: Classifier-Free Guidance



$\gamma = 1$

$\gamma = 3$

Caption: "*A stained glass window of a panda eating bamboo.*"

Control the Diffusion Model: Classifier-Free Guidance

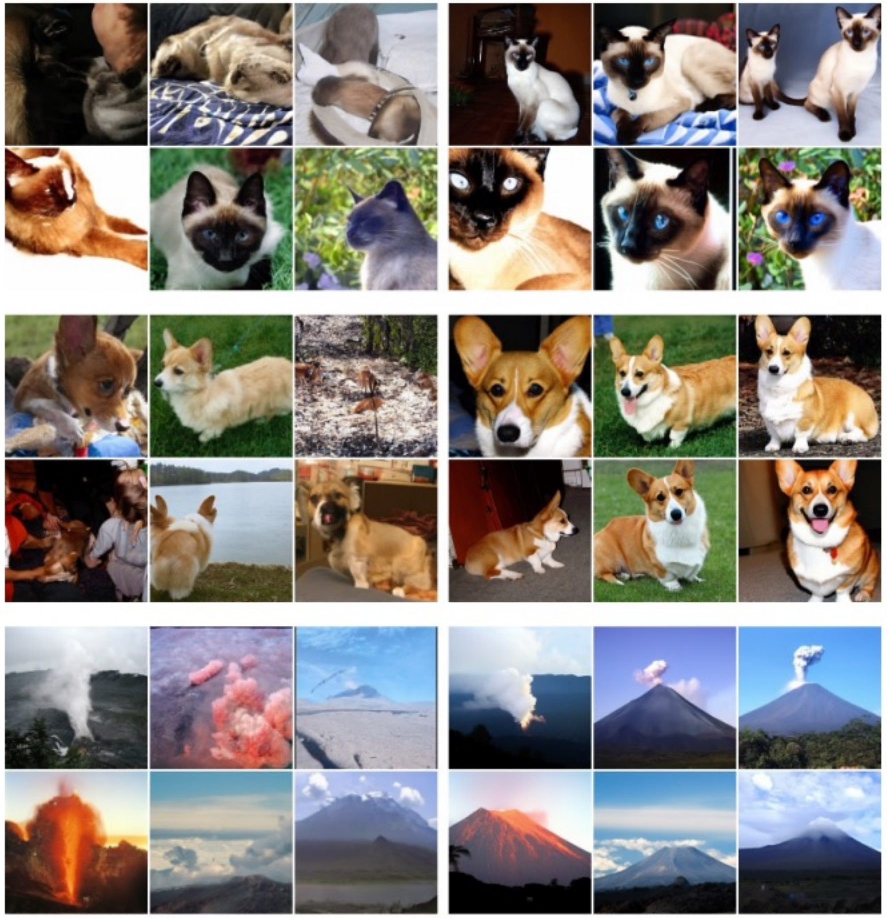


Figure 3: Classifier-free guidance on 128x128 ImageNet. Left: non-guided samples, right: classifier-free guided samples with $w = 3.0$. Interestingly, strongly guided samples such as these display saturated colors. See Fig. 8 for more.

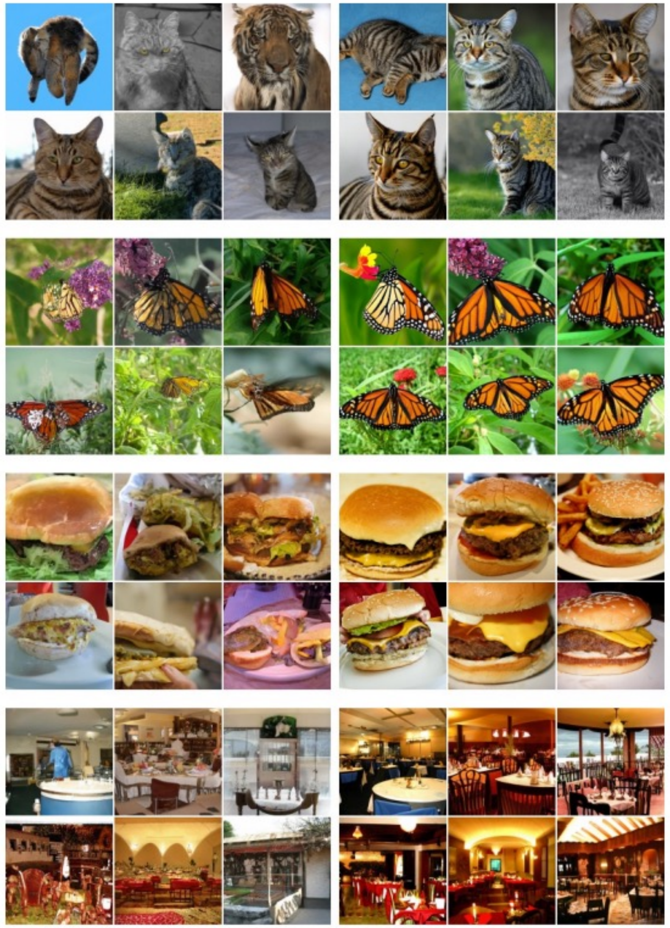
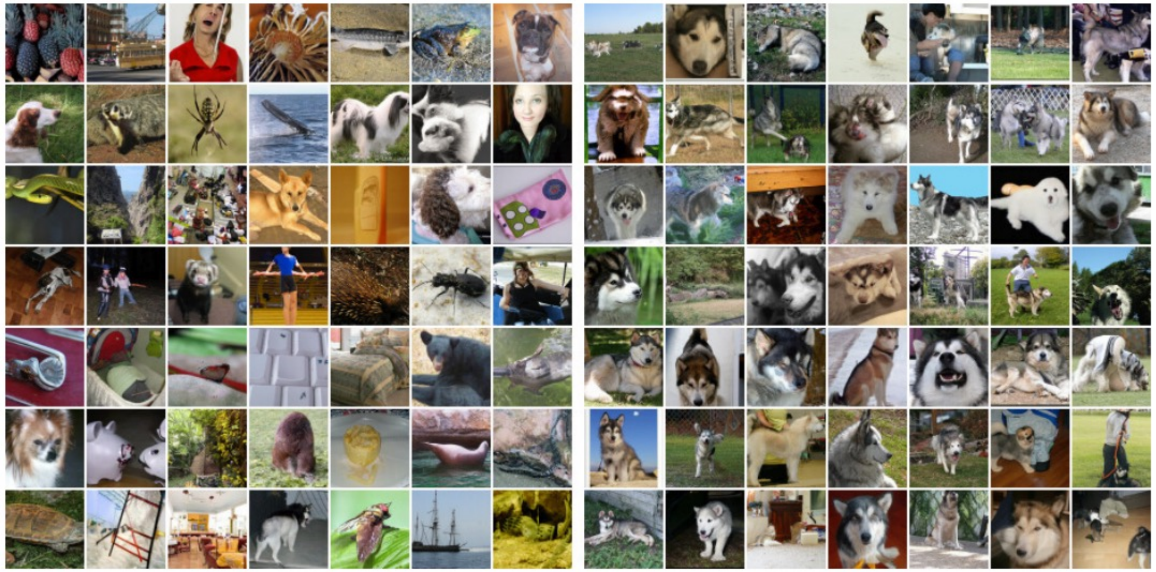
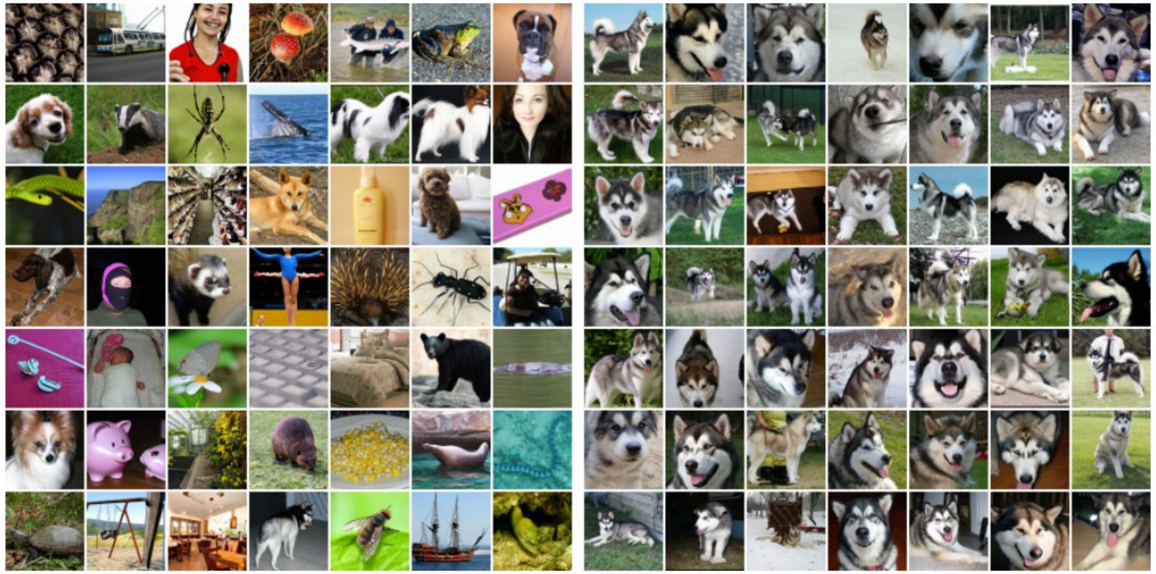


Figure 8: More examples of classifier-free guidance on 128x128 ImageNet. Left: non-guided samples, right: classifier-free guided samples with $w = 3.0$.

Control the Diffusion Model: Classifier-Free Guidance

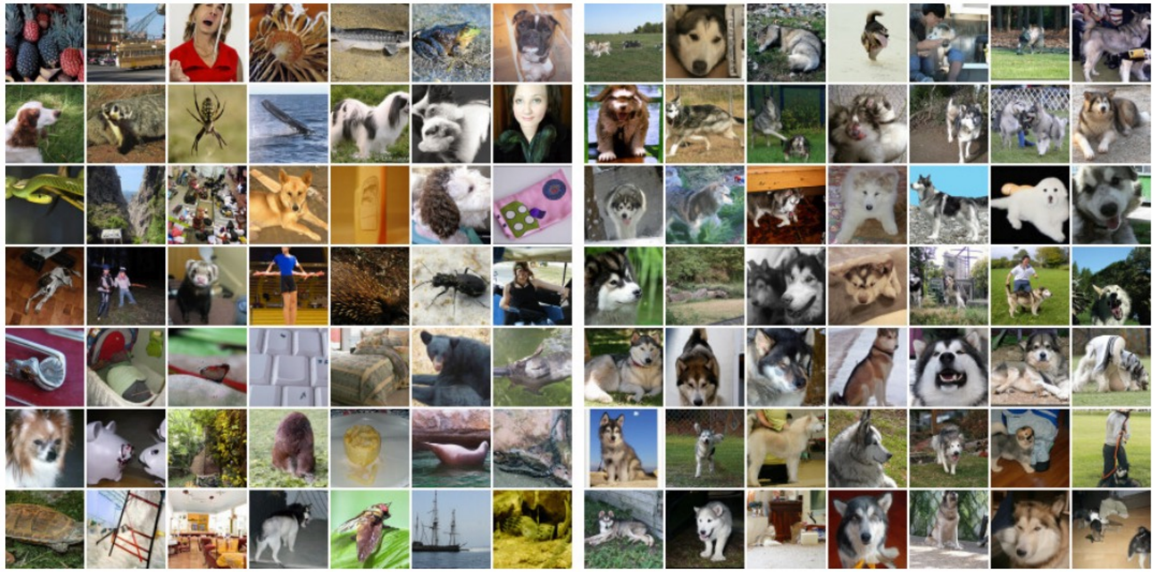


(a) Non-guided conditional sampling: FID=1.80, IS=53.71

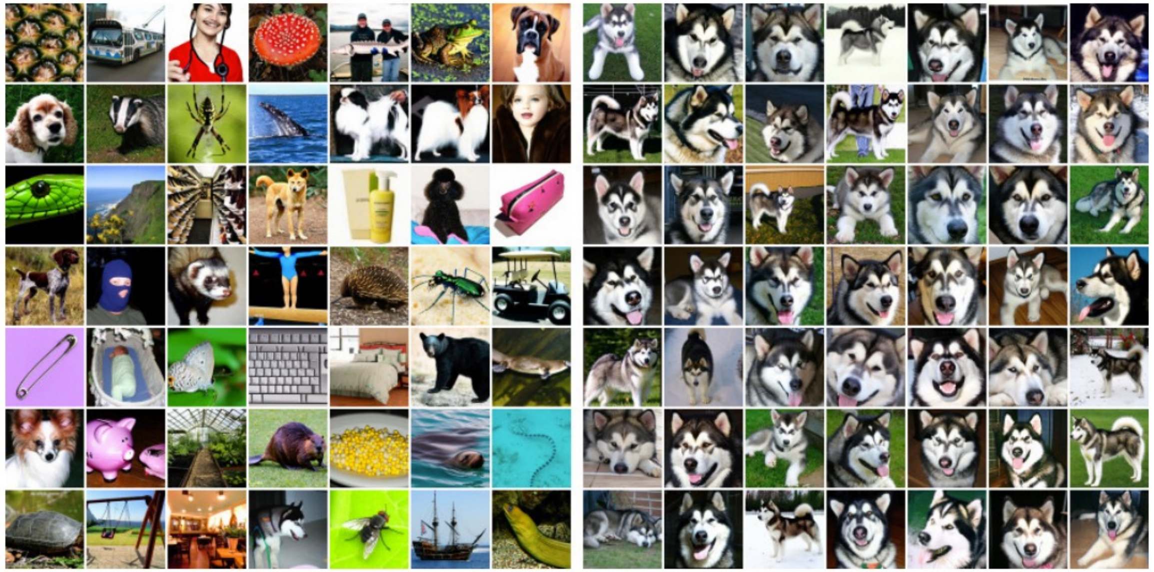


(b) Classifier-free guidance with $w = 1.0$: FID=12.6, IS=170.1

Control the Diffusion Model: Classifier-Free Guidance



(a) Non-guided conditional sampling: FID=1.80, IS=53.71



(c) Classifier-free guidance with $w = 3.0$: FID=24.83, IS=250.4

ControlNet

ControlNet: Introduction

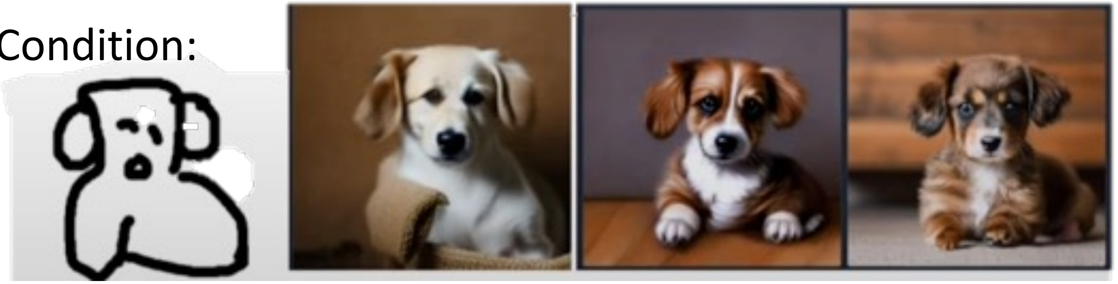
- Basic form of using diffusion models (e.g. Stable Diffusion) is text-to-image
  - Use text prompts as the conditioning to steer image generation so that you generate images that match the text prompt

Prompt: “Dog in a room”

- Control the output by giving more input conditions
  - Keep properties from text
  - Adhere to additional properties from condition

Prompt: “Dog in a room”

Condition:



ControlNet: Introduction

- NN architecture that helps you **control** pre-trained diffusion models (such as Stable Diffusion model) by adding extra conditions

Goodies:

✓End-to-end architecture

✓Robust on small dataset (<50k images)

✓As fast as fine-tuning

✓Can be trained on personal devices

✓Can scale to large amounts of data (millions to billions)

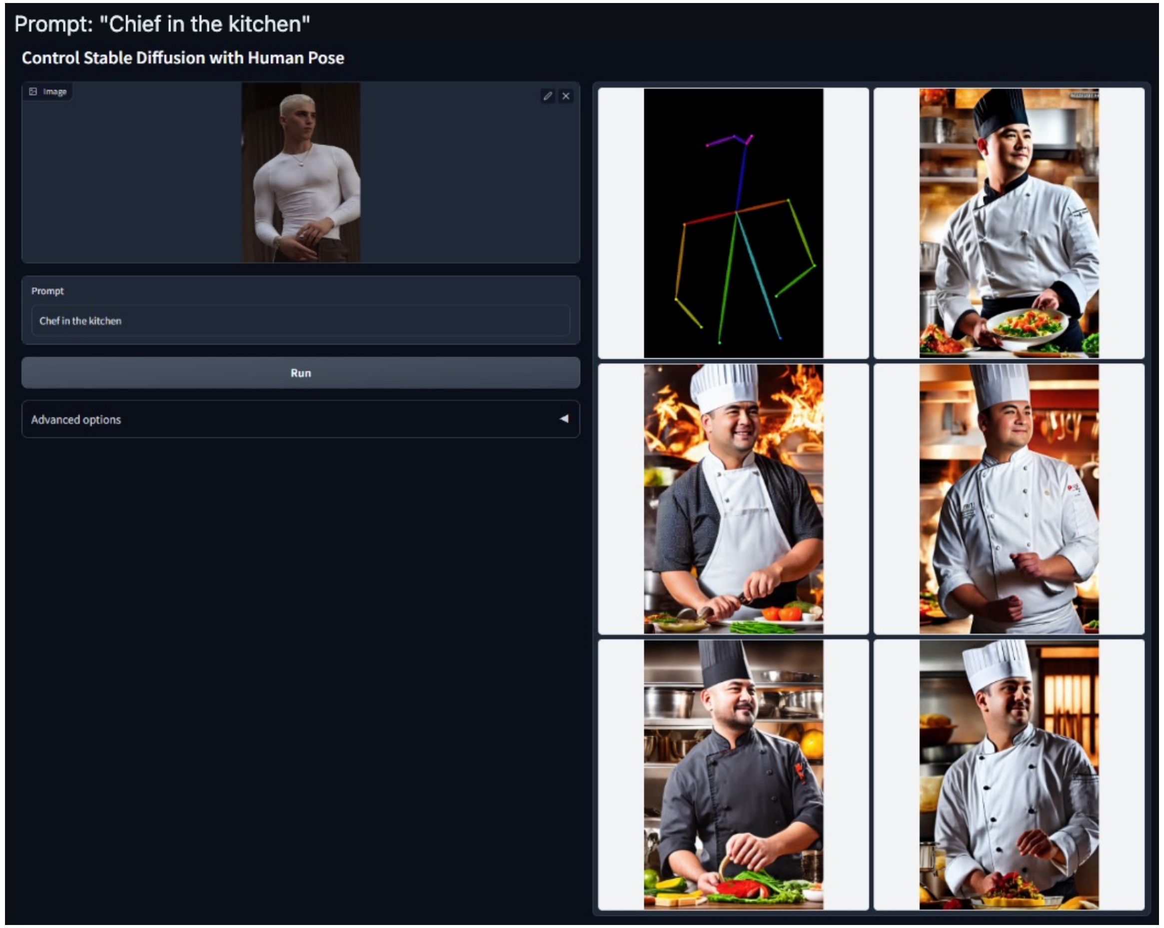
ControlNet: Introduction

- ControlNet adds one more conditioning in addition to the text prompt
  - allows for the manipulation of existing diffusion model architectures
  - think of it as a way to make slight changes to a neural network's structure and add desired properties or characteristics

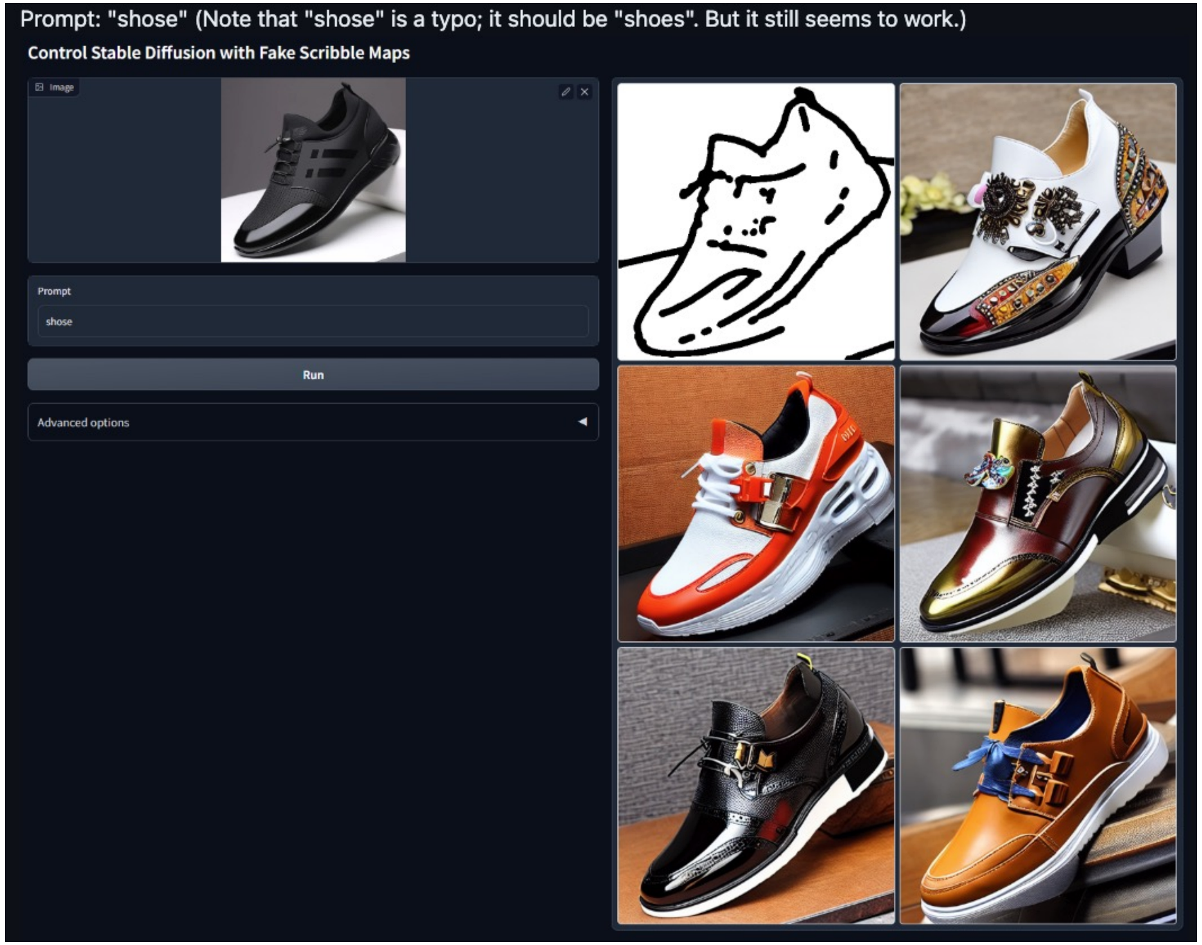
- The extra conditioning can take many forms
  - Segmentation map
  - Depth map
  - Pose
  - Infrared
  - HED map
  - Hough Line
  - Cartoon Line Drawing
  - …

Examples

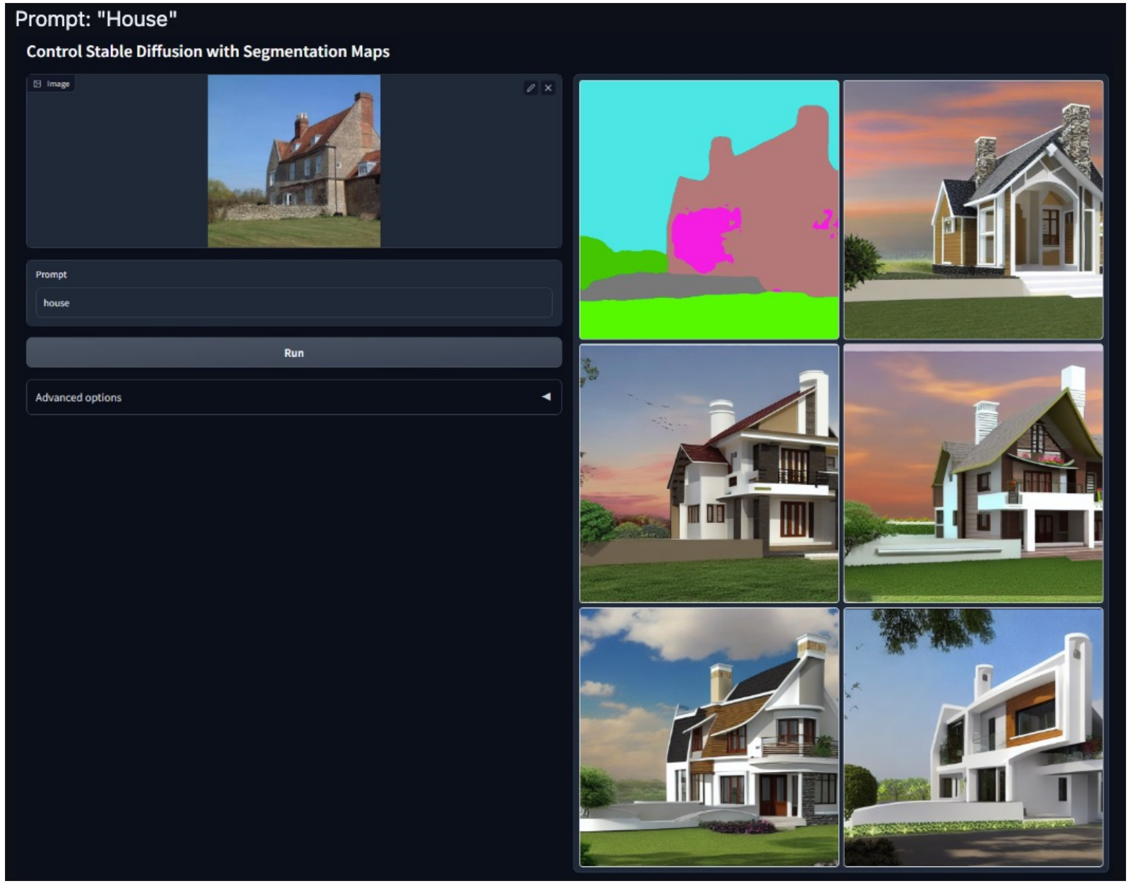
ControlNet examples
Condition: Pose

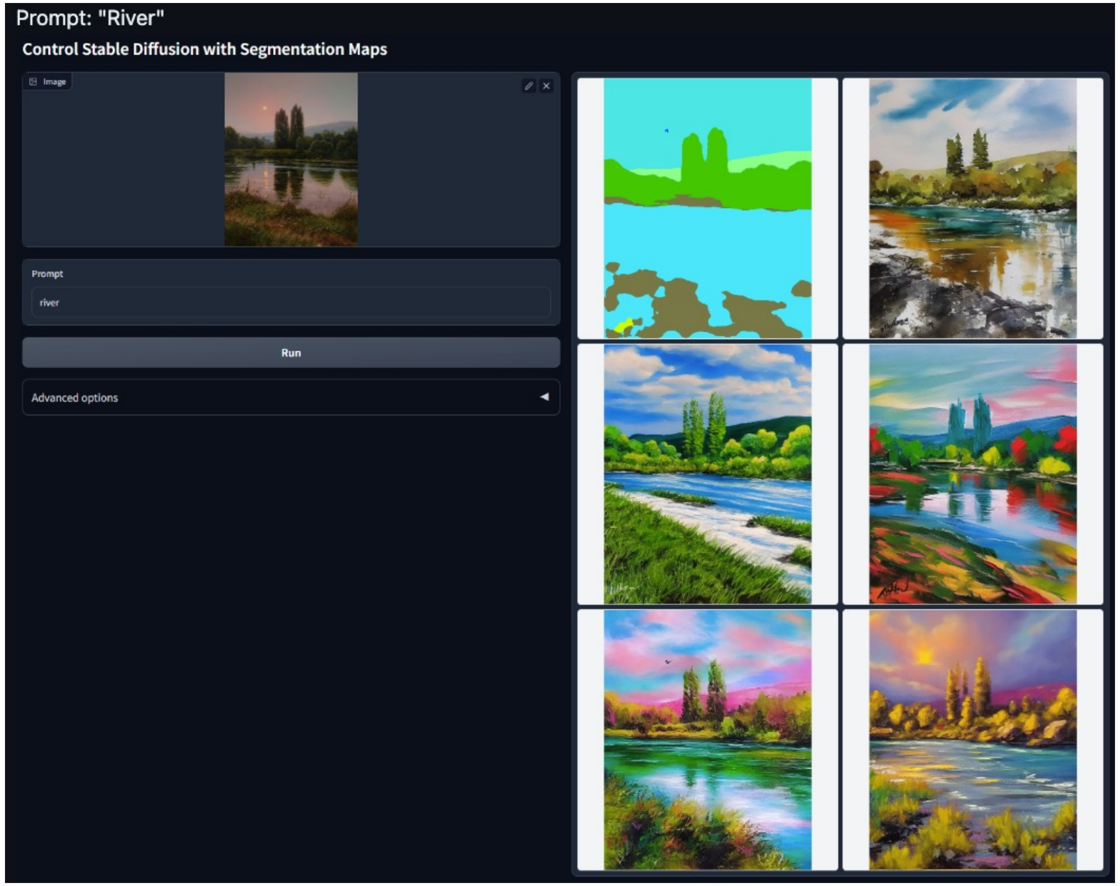


ControlNet examples
Condition: Scribble

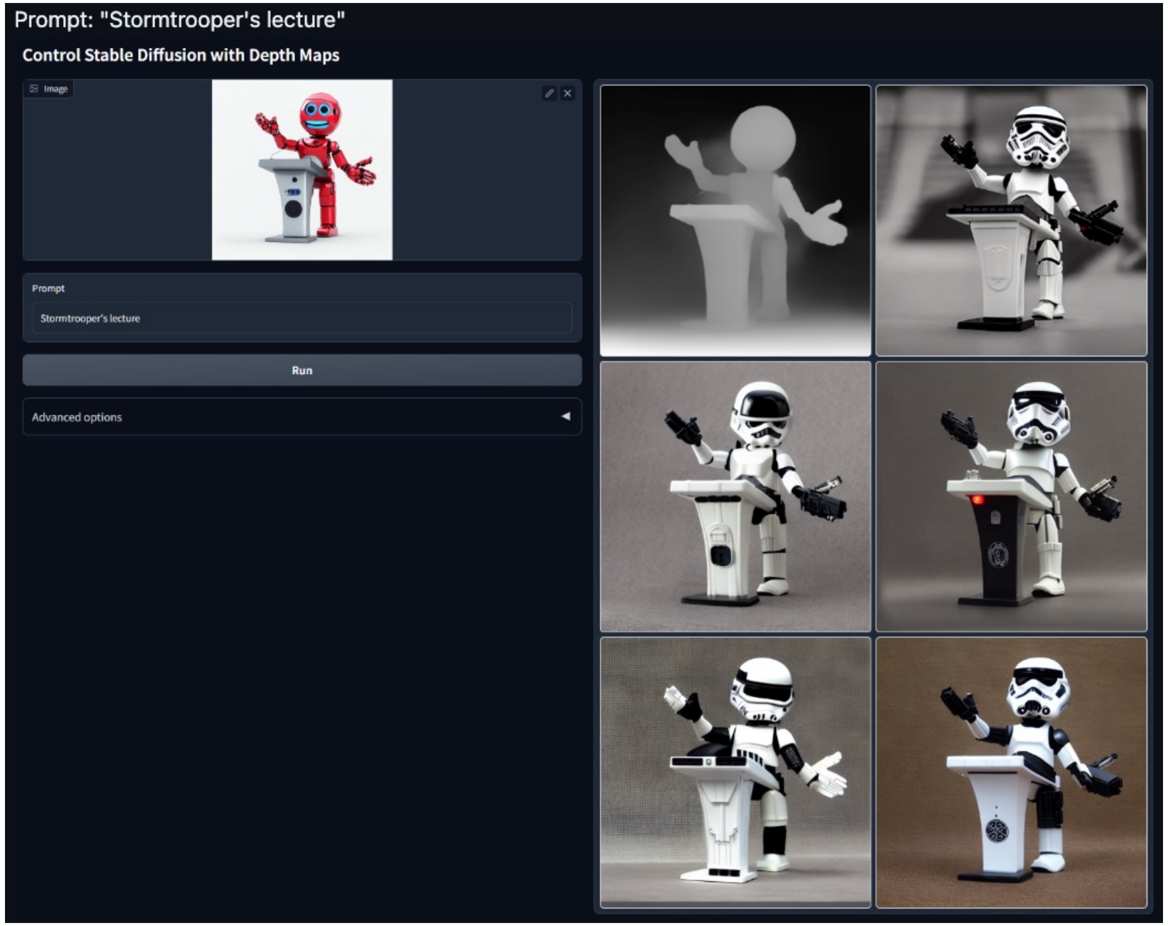


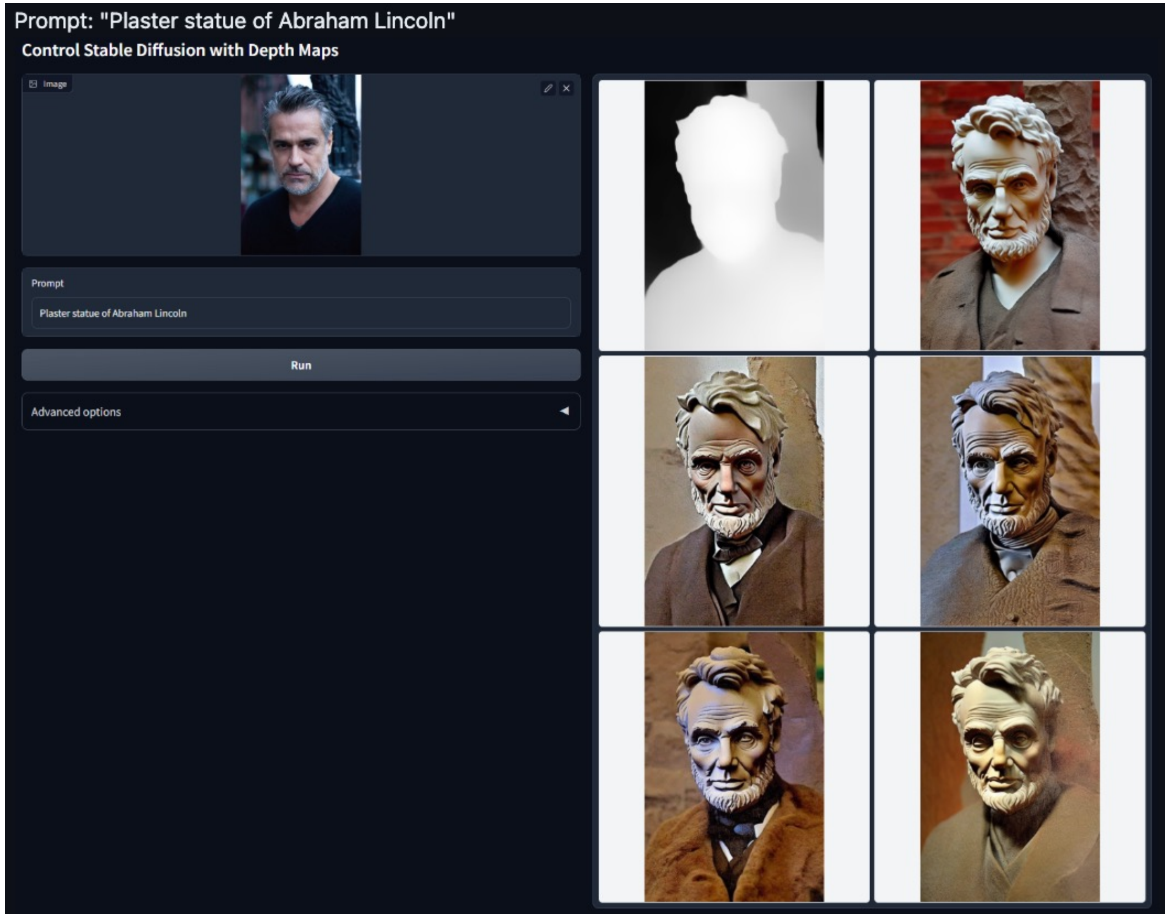
ControlNet examples
Condition: Segmentation map



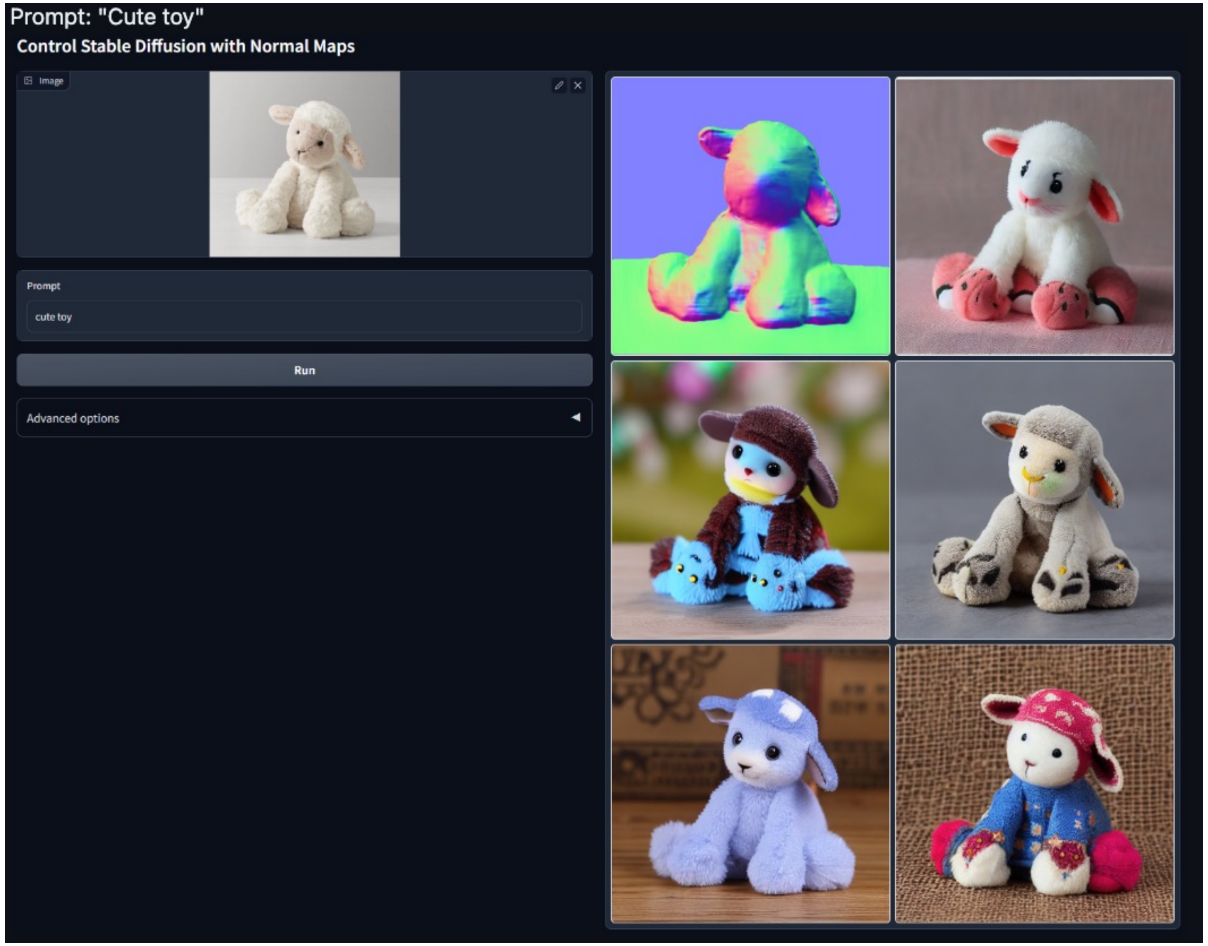


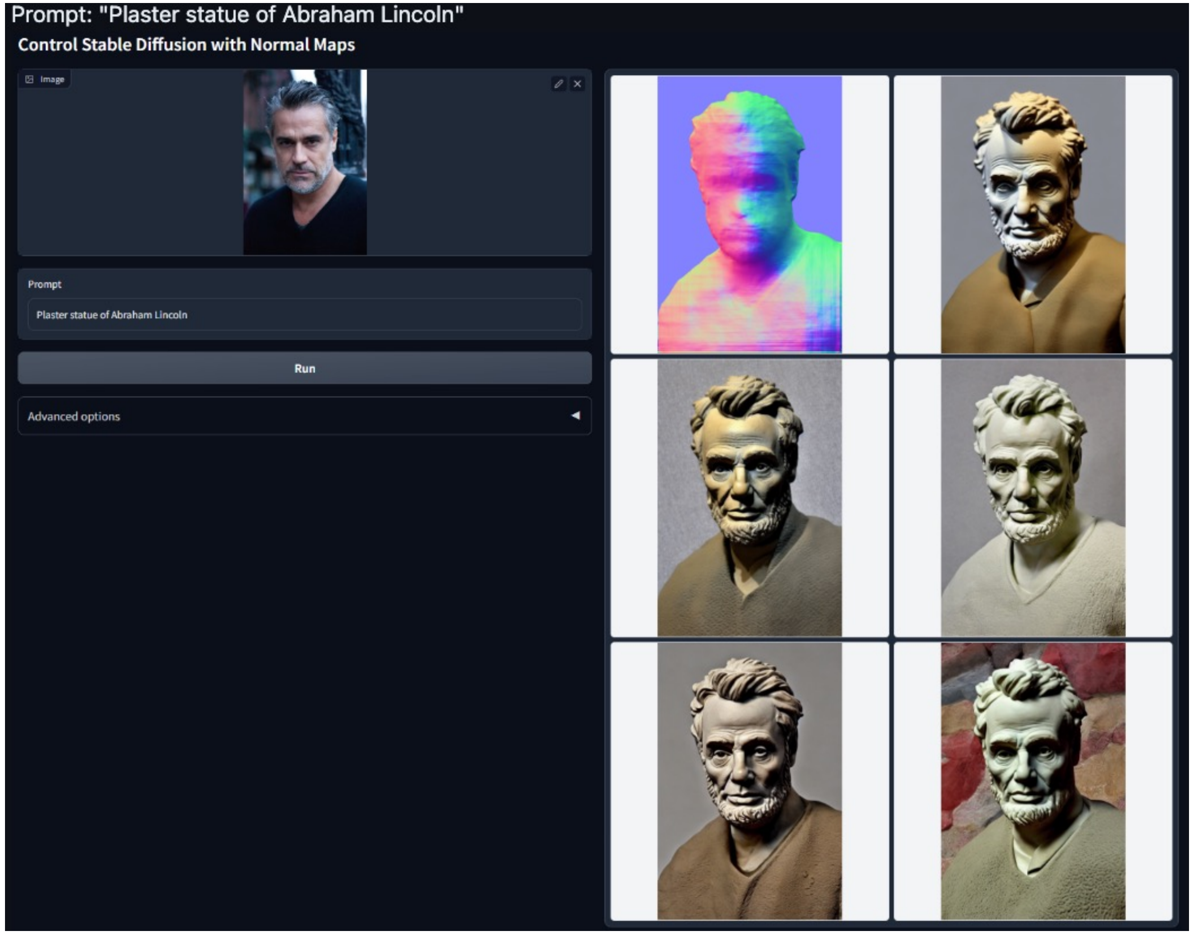
ControlNet examples
Condition: Depth map



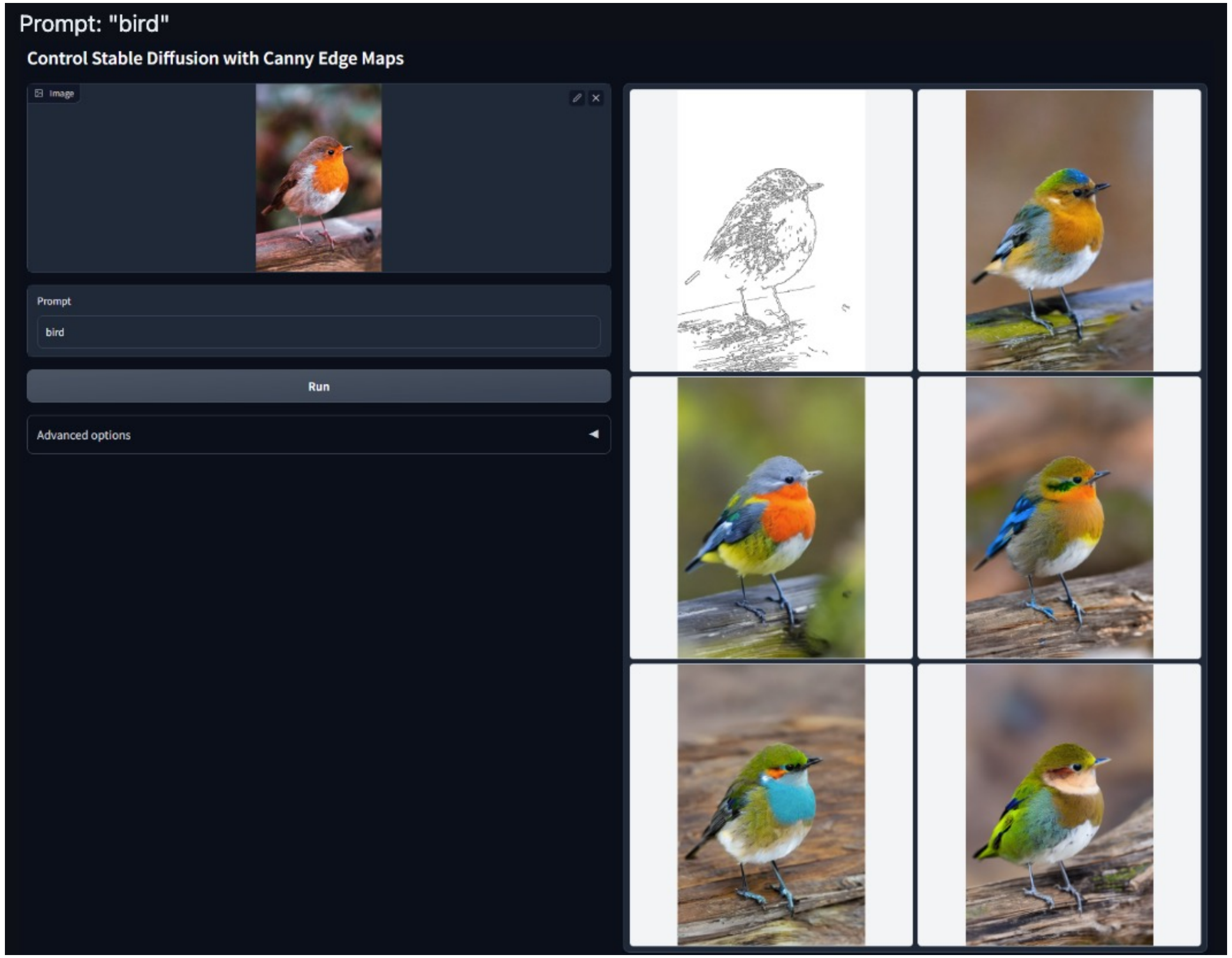


ControlNet examples
Condition: Normal map

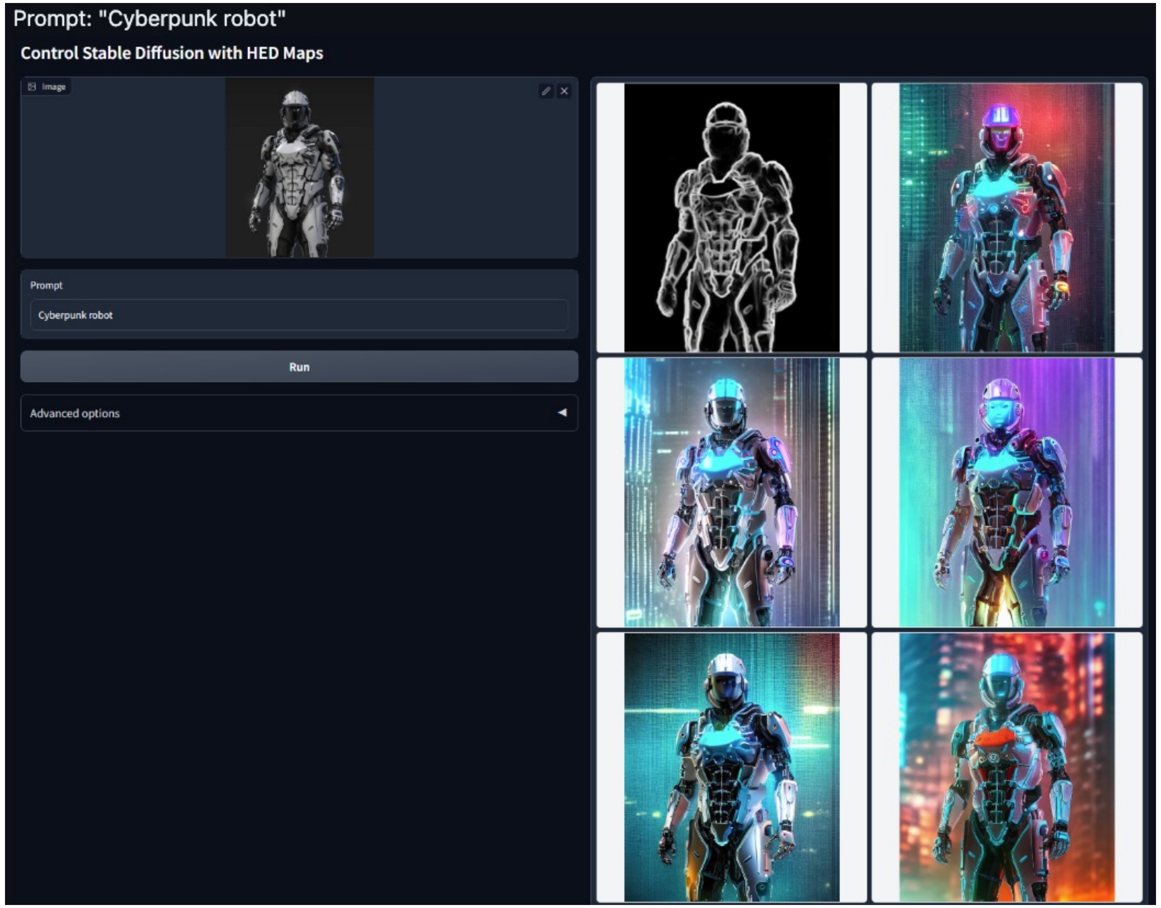




ControlNet examples
Condition: Canny Edge map



ControlNet examples
Condition: HED Map



Motivation

Motivation

- Can large models be applied to facilitate specific tasks?
- What kind of framework should we build to handle the wide range of problem conditions and user controls?

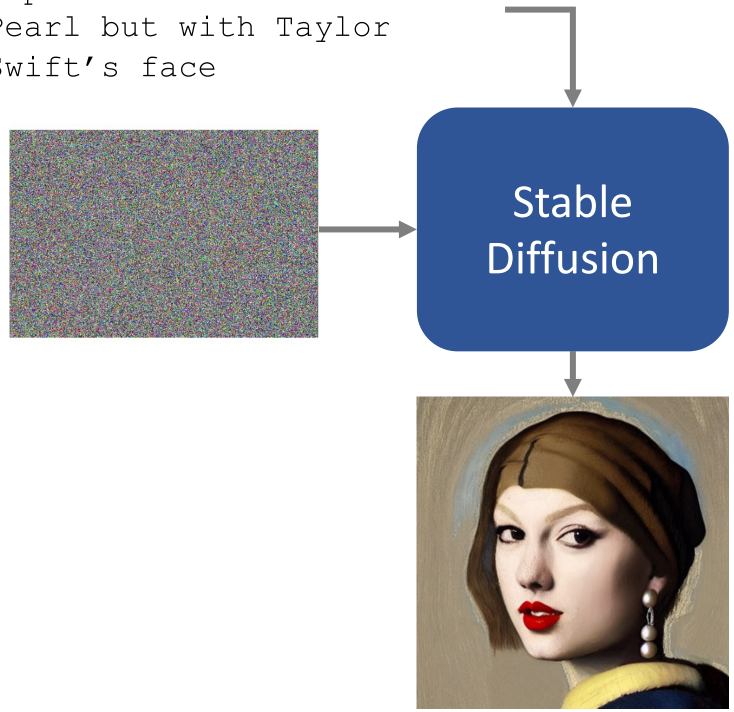
Motivation

- Can large models be applied to facilitate specific tasks?
- What kind of framework should we build to handle the wide range of problem conditions and user controls?
- Three findings:
  - The available data scale in a task-specific domain is not always as large as that in the general image-text domain
  - Large computation clusters are not always available
  - Various image processing problems have diverse forms of problem definitions, user controls, or images annotations

ControlNet

ControlNet

Prompt
`a portrait of Girl With A Pearl but with Taylor Swift's face`



ControlNet

Prompt
`a portrait of Girl With A Pearl but with Taylor Swift's face`

Stable Diffusion

Prompt
`a portrait of Girl With A Pearl but with Taylor Swift's face`

Stable Diffusion

ControlNet

Prompt
`a portrait of Girl With A Pearl but with Taylor Swift's face`

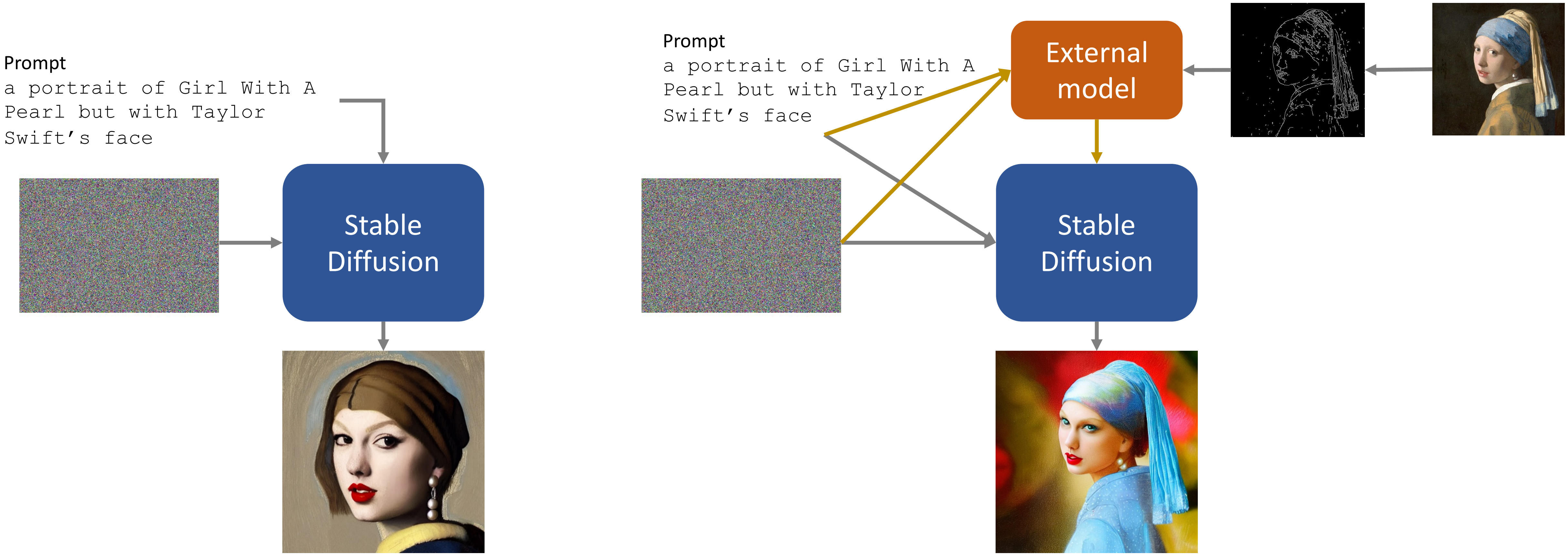
Stable Diffusion

Prompt
`a portrait of Girl With A Pearl but with Taylor Swift's face`

Stable Diffusion

What is the most efficient way to train a model to take in additional conditioning inputs?

ControlNet


Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face
Stable Diffusion
Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face
External model
Stable Diffusion

Today’s tutorial

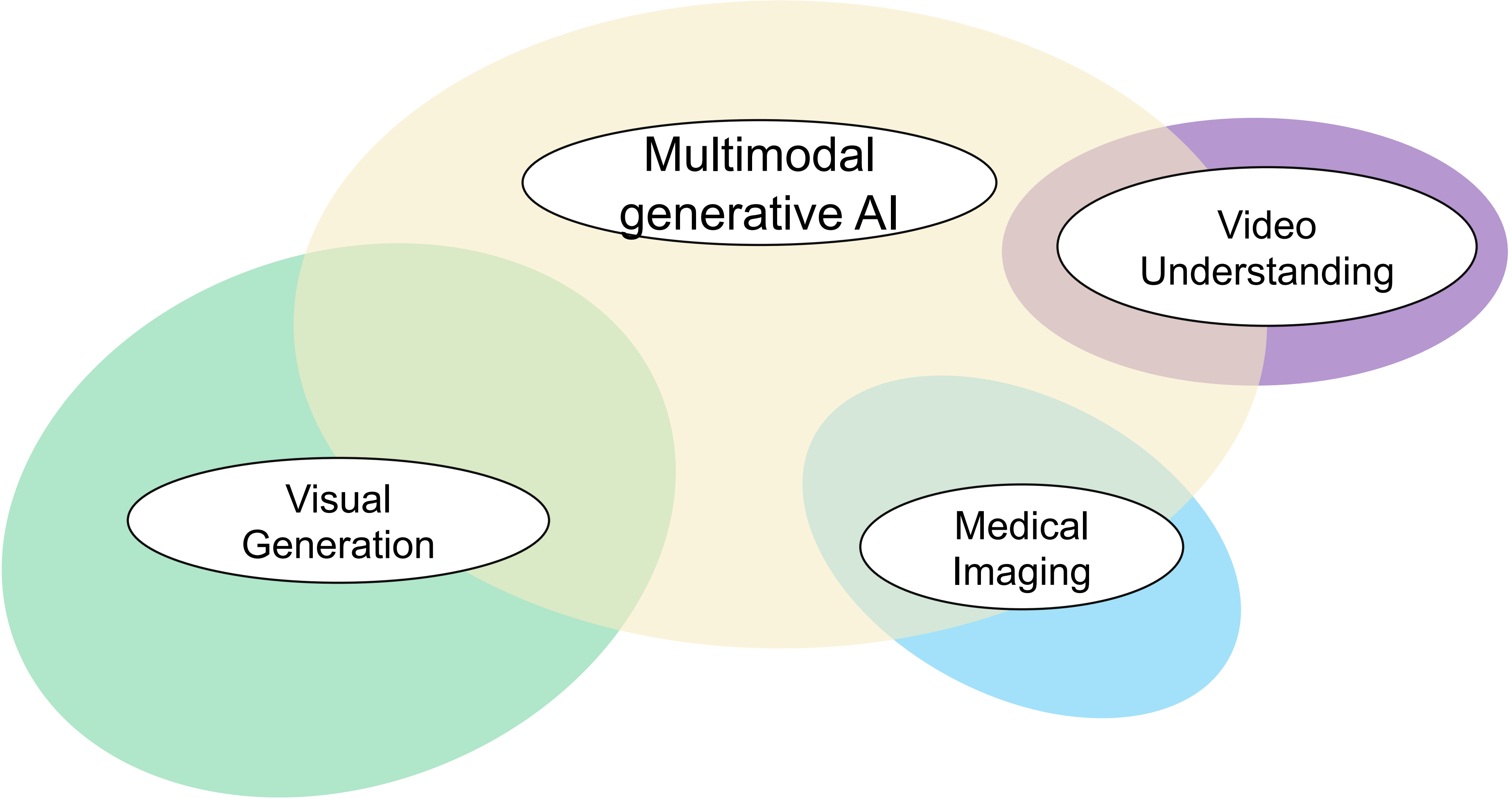
Part I:
Introduction

Part II:
Diffusion &
Guidance &
Control

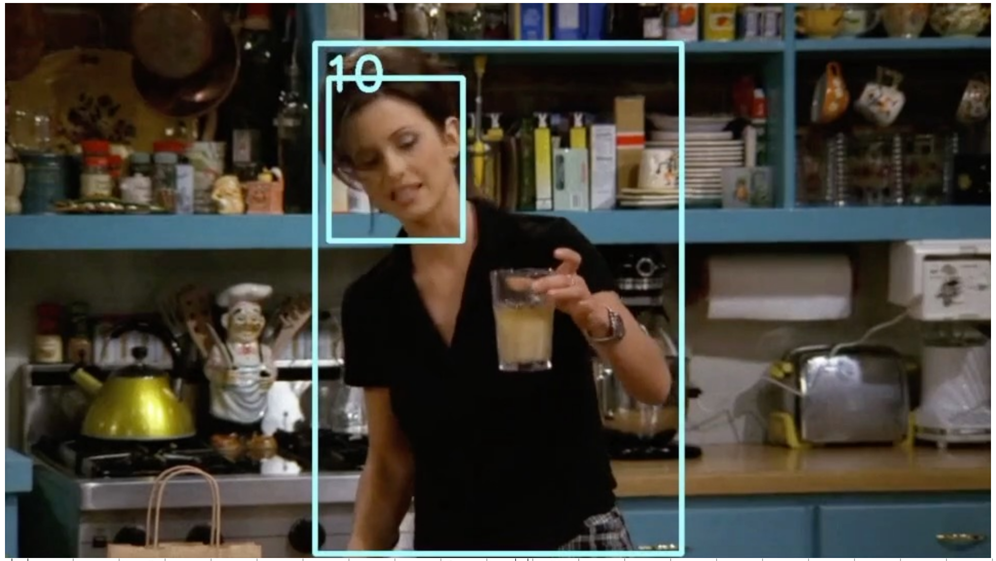
Part III:
My research

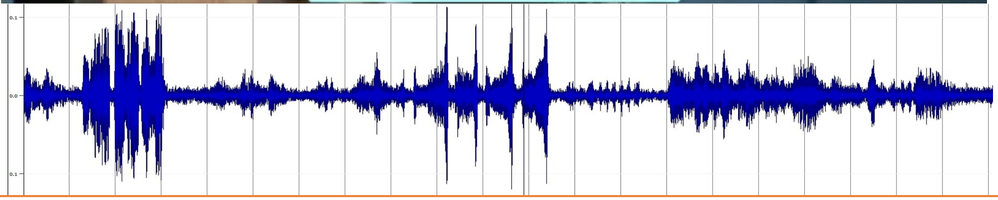
[Slides by V. Kalogeiton, X. Wang]

Research agenda



Multimodal generative AI: video, audio, text





Subtitles

- **Rachel:** You guys, do this look like something the girlfriend of a paleontologist would wear?
- **Phoebe:** I don't know. You might be the first one

Low-level understanding

- Characters (Rachel, Phoebe, …)
- Located in an apartment
- Winter (clothing)

High-level reasoning

- Anxious (what to wear, …)
- Interactions
- Joke to diffuse the situation

Challenges

Manual collection of training samples → Prohibitive

Multimodal cues can help (phone ringing, vacuum sound, …)

Challenges

Vocabulary → Not well defined

Open

umbrella

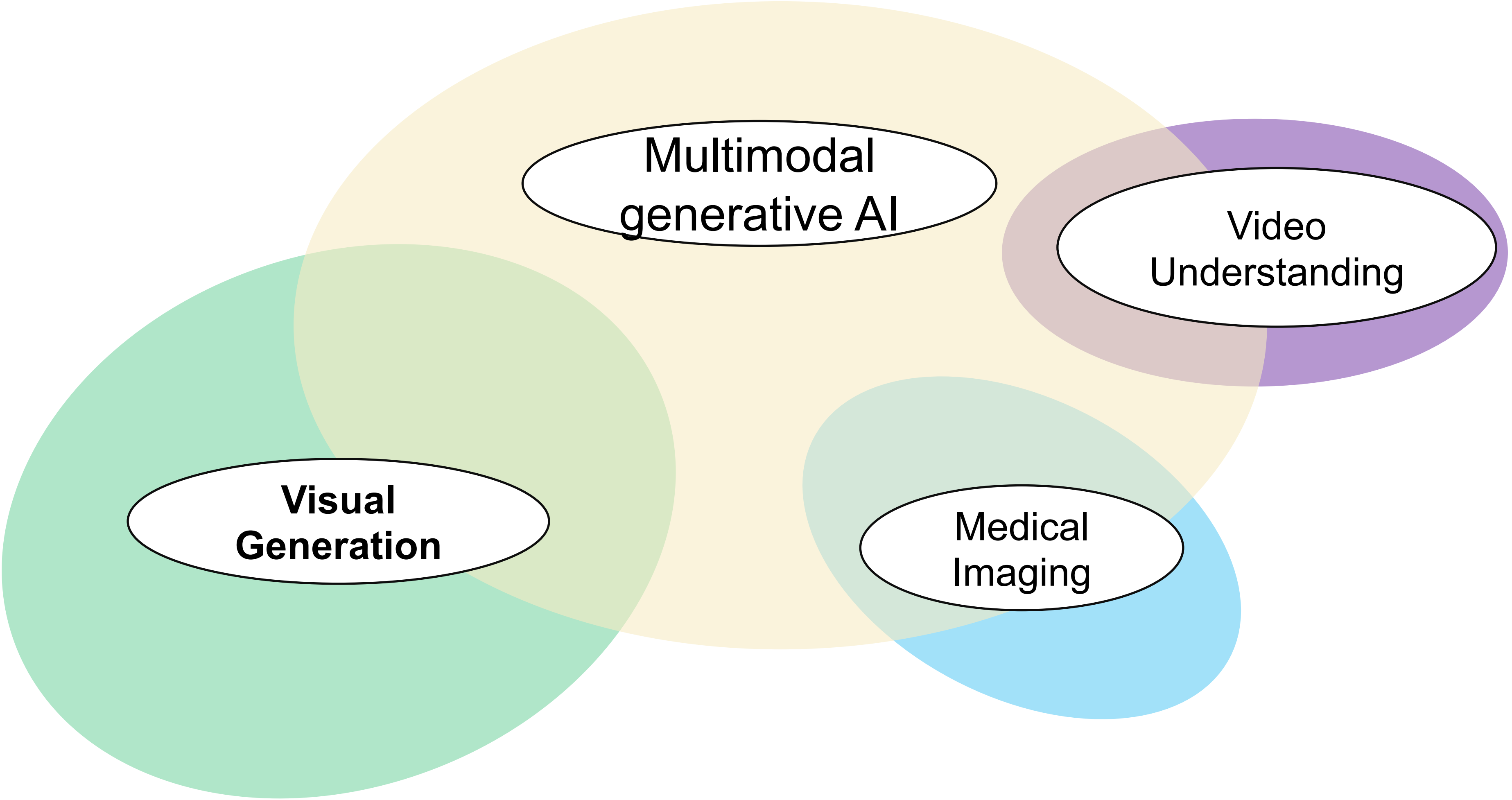
door

handle

bottle

Language or audio can help!

Research agenda



Analysis of Classifier-Free Guidance Weight Schedulers






Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernandez Abrevaya, David Picard, Vicky Kalogeiton, submission 2024

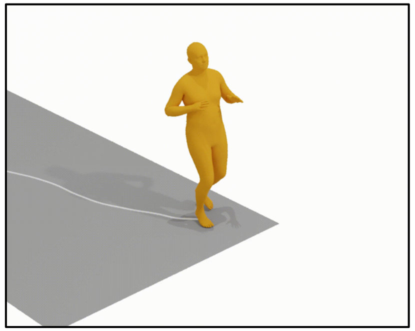
Introduction

- Classifier-Free Guidance is the key method for conditioning diffusion models based on various input modalities (label, text, etc.)

- $\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \omega\,(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$

- CFG consists of generation term + guidance term and ω is used to control the conditioning magnitude

prompt condition: Darth Vader is surfing on the waves. [From SVD]



prompt condition: A person is running backwards quickly. [From MDM]

prompt condition: An astronaut is riding a green horse. [From SDXL]

Label condition: "Corgi"[From CFG]

[Ho et al., 2022]

Introduction

- Classifier-Free Guidance is the key method for conditioning diffusion models based on various input modalities (label, text, etc.)

- $\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \omega\,(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$

- CFG consists of generation term + guidance term and ω is used to control the conditioning magnitude

- As a hyperparameter, tuning guidance scale ω is important to balance the generation quality, textual adherence and generation diversity

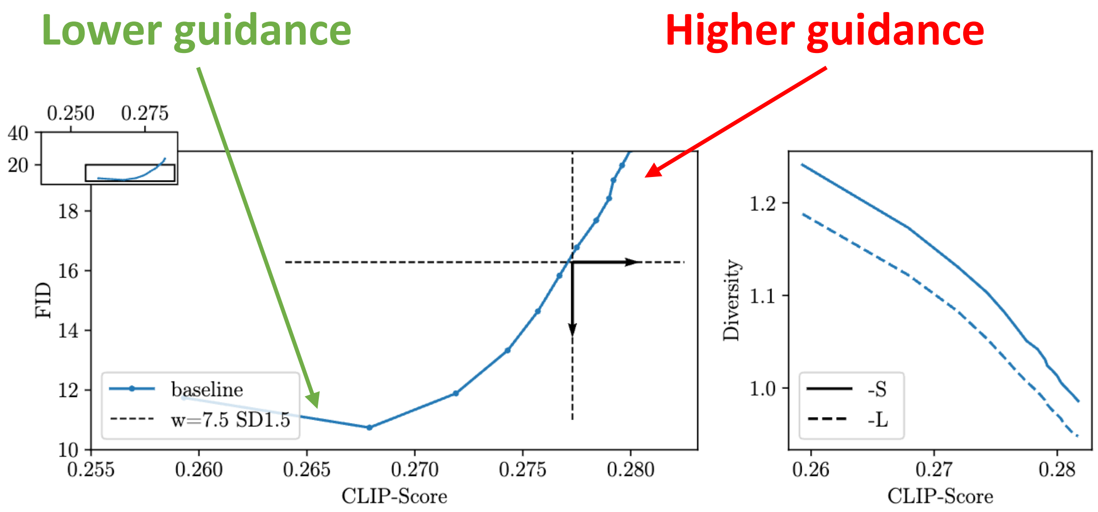
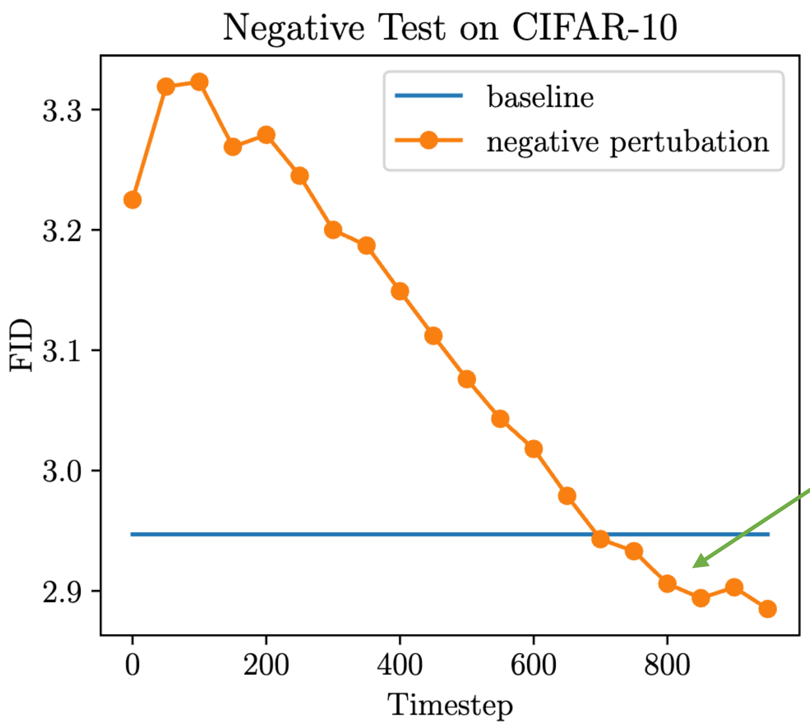


Figure. FID vs. CLIP-Score and Diversity vs. CLIP-Score on different guidance scale

[Ho et al., 2022]

Negative Perturbation Experiment

- Remove varying intervals of guidance scale with respect to the timestep of the generation



Observation:
Removing the initial stage of Classifier-Free Guidance
→
improves generation quality (FID)
→
constant guidance: not effective design

Solution

$$\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \omega(t)\,(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$$

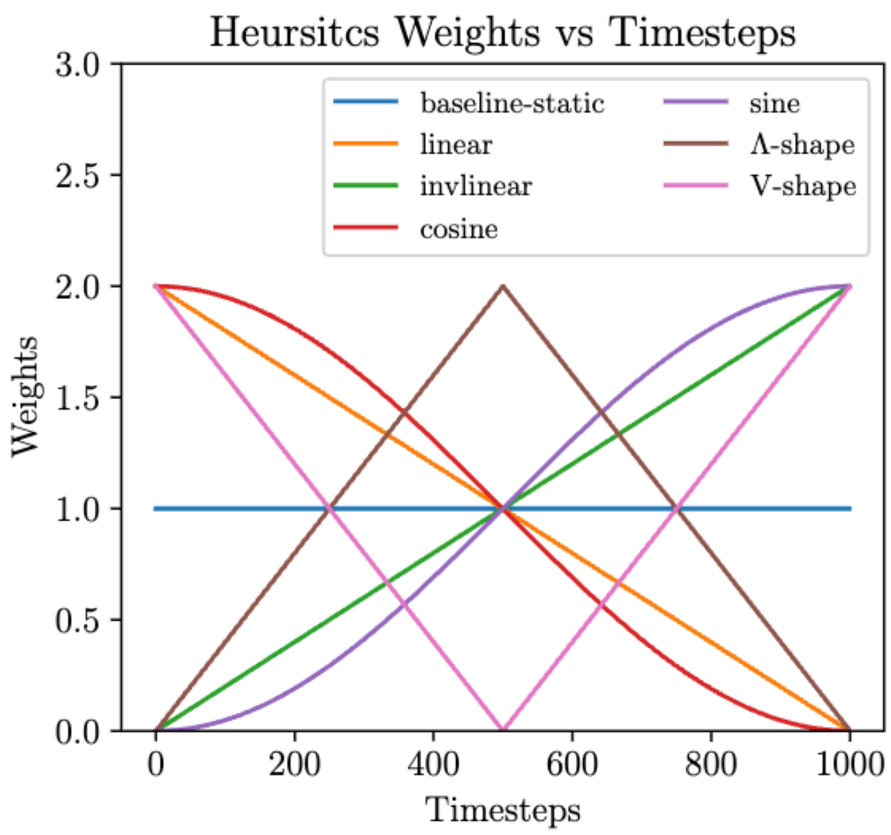
Replace constant guidance, we with guidance schedulers $\omega(t)$ that vary according to generation timesteps

- Two families:
  - Heuristic functions
  - Parametrized functions

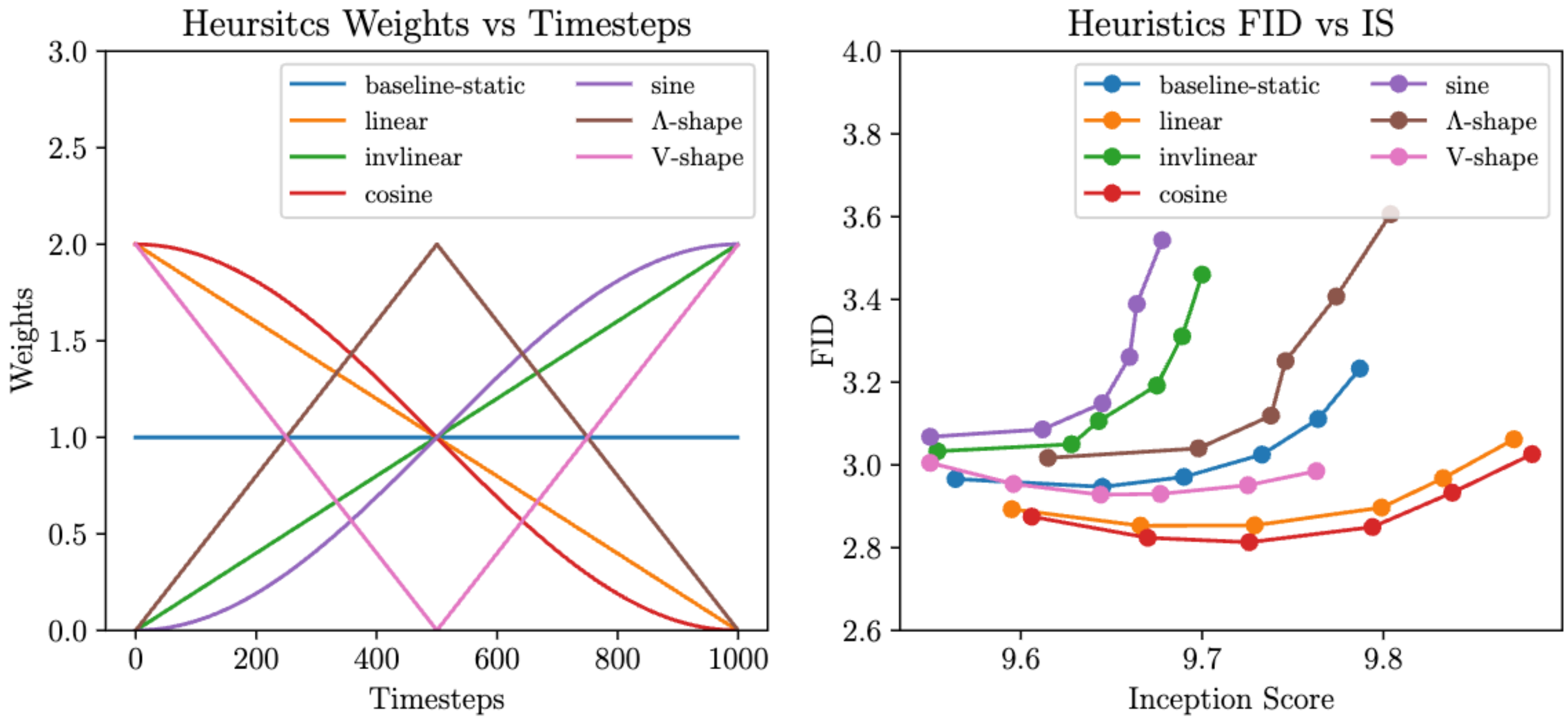
- Analyze results

Replace static by Heuristic functions

$$
\begin{aligned}
\text{linear: } \omega(t) &= 1 - t/T, \\
\text{invlinear : } \omega(t) &= t/T, \\
\text{cosine: } \omega(t) &= \cos(\pi t/T) + 1, \\
\text{sine: } \omega(t) &= \sin(\pi t/T - \pi/2) + 1, \\
\text{V-shape: } \omega(t) &= \text{invlinear}(t) \ \textit{if } t < T/2, \ \text{linear}(t) \ \textit{else}, \\
\Lambda\text{-shape: } \omega(t) &= \ \text{linear}(t) \ \textit{if } t < T/2, \ \text{invlinear}(t) \ \textit{else}.
\end{aligned}
$$

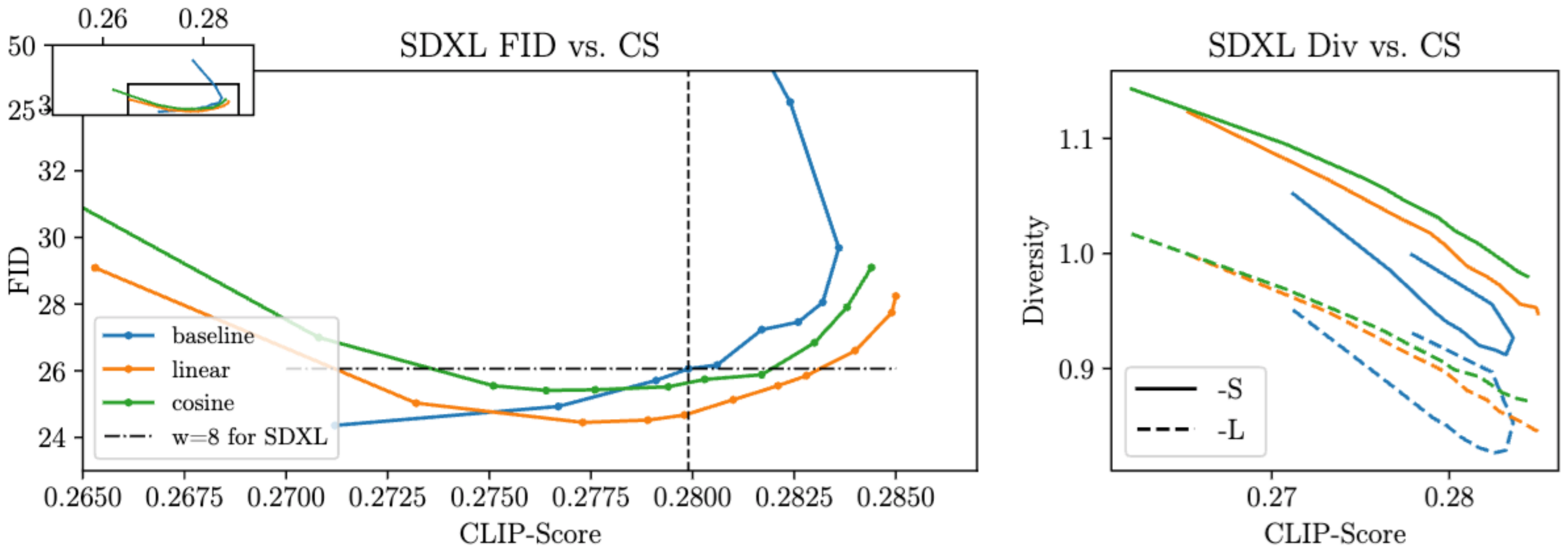


Quantitative Results: Heuristic functions
Class-conditional generation



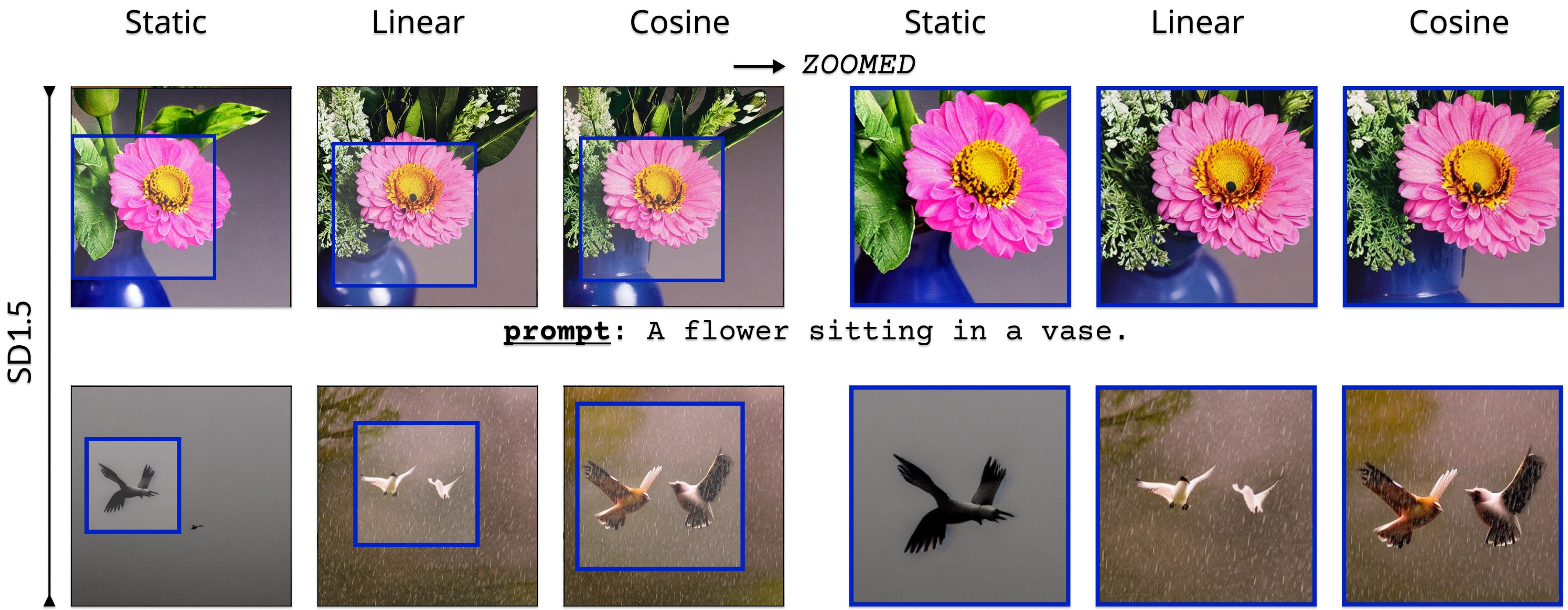
Monotonically increasing shape heuristic performs the best

Quantitative Results: Heuristic functions Text-to-image generation



Monotonically increasing shape guidance schedulers achieve a better balance *of quality, conditional adherence, and diversity*

Qualitative Results: Heuristic functions Text-to-image generation

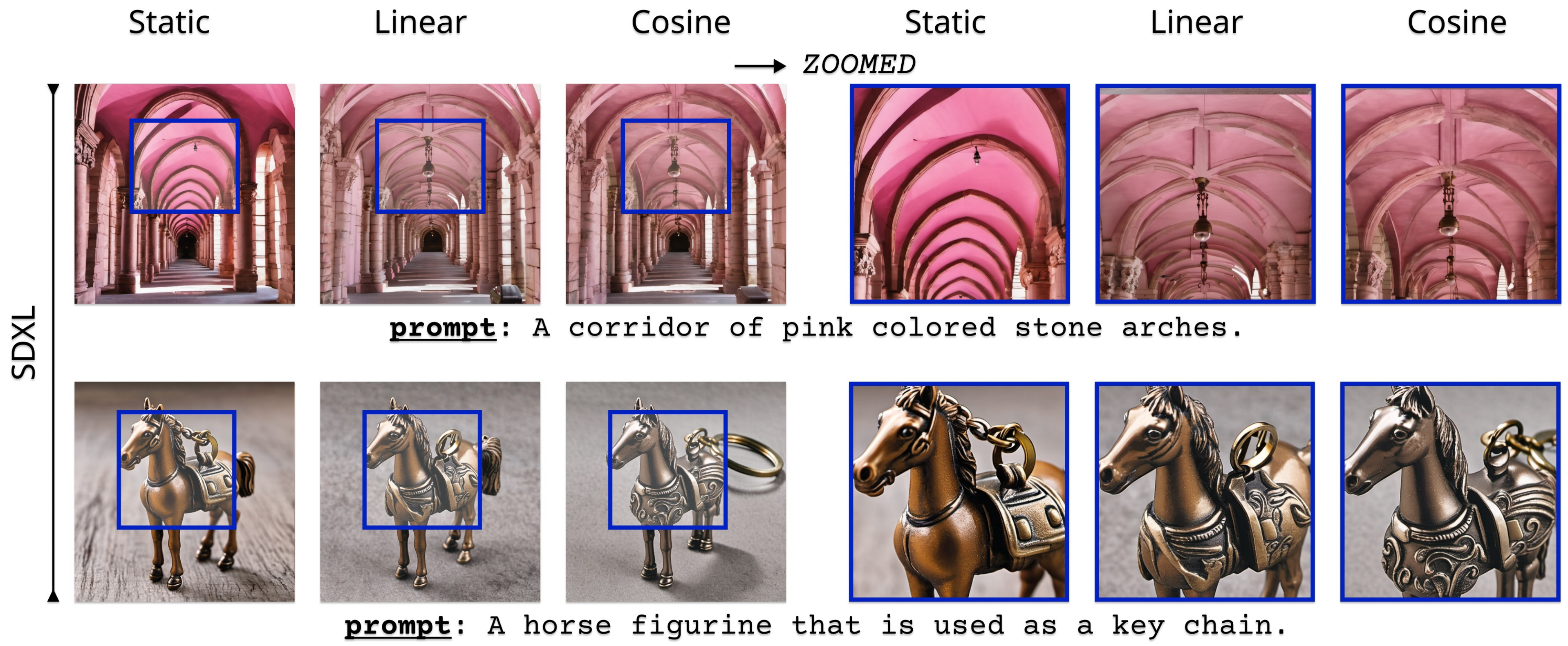


prompt: A flower sitting in a vase.

prompt: Two birds flying in the sky in a bad weather.

Better quality

Qualitative Results: Heuristic functions
Text-to-image generation

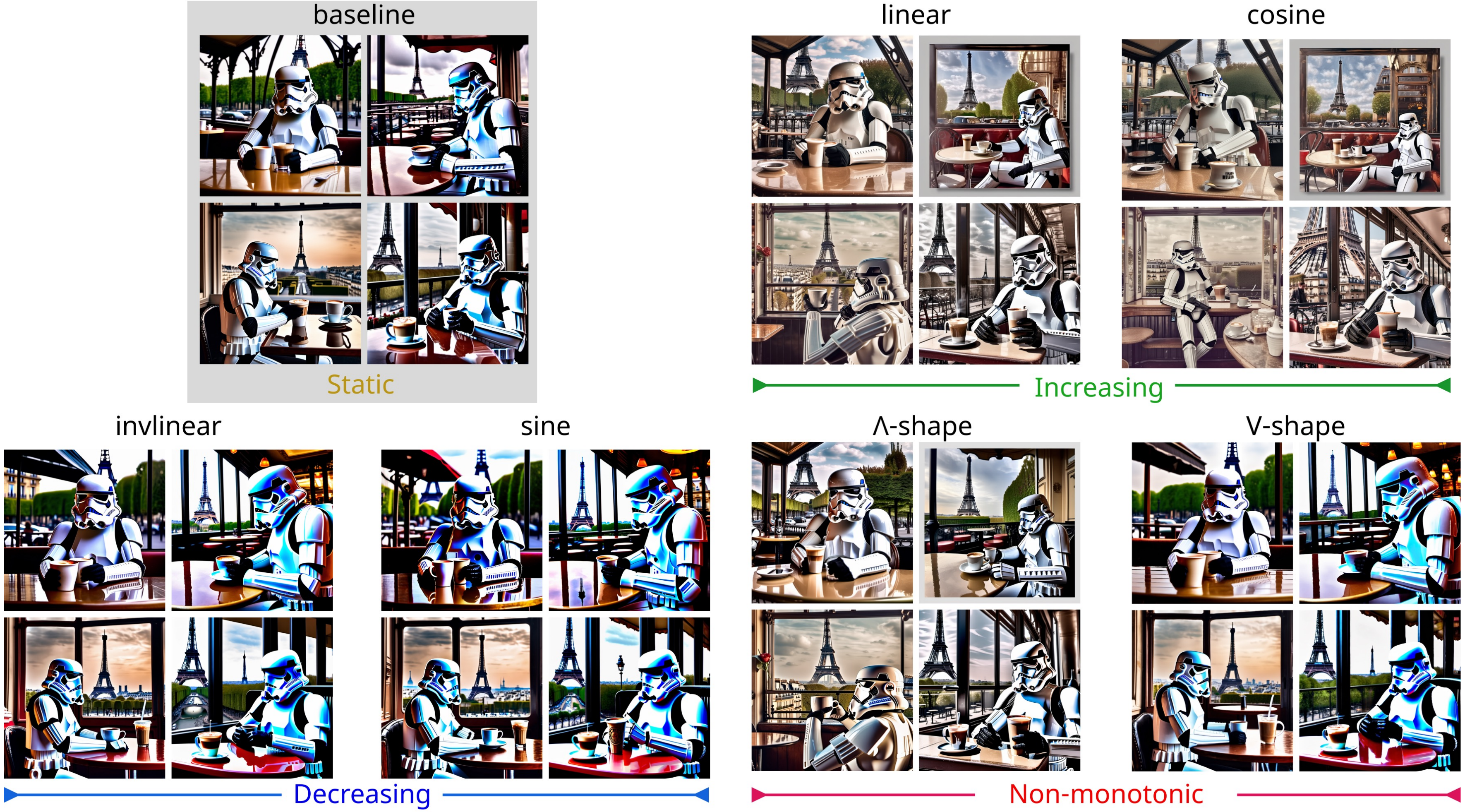


Better quality

Qualitative Results: Heuristic functions Text-to-image generation

Better diversity

Qualitative Results: Heuristic functions
Text-to-image generation

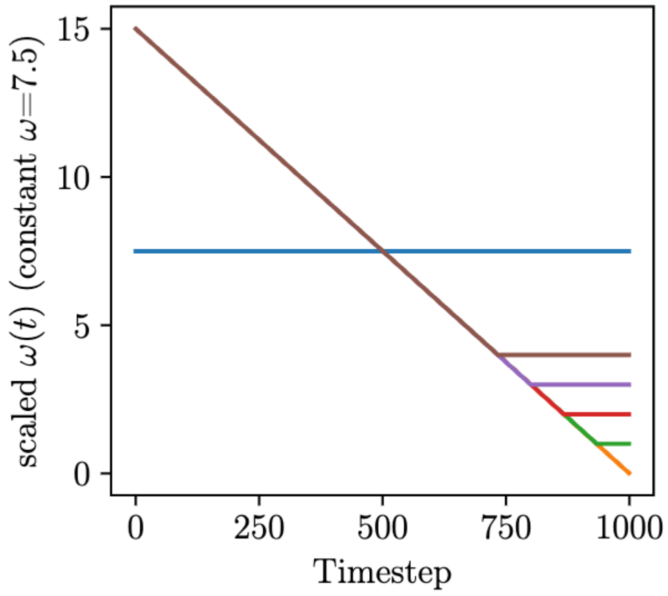


SDXL

Prompt: Stormtrooper drinking coffee in a Paris cafe bar, with Eiffel Tower in the background.

Replace static by Parametrized functions

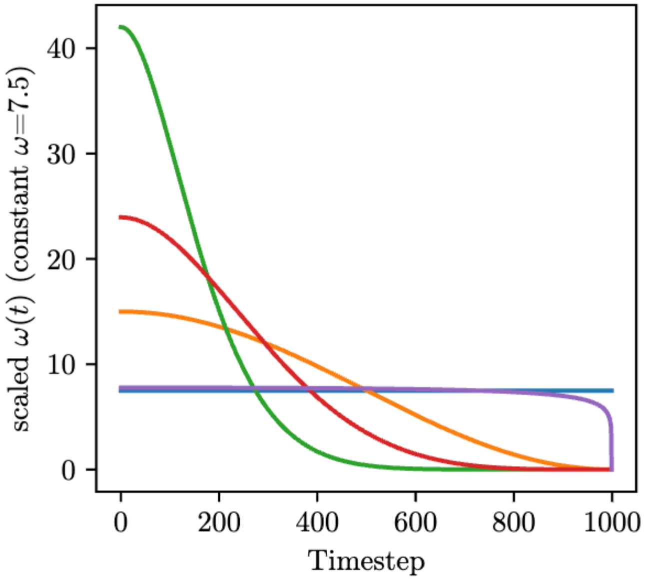
Clamping



$$\omega(t, c) = \max(\omega(t), c)\,;$$

$\omega(t)$ = linear, cosine, etc.

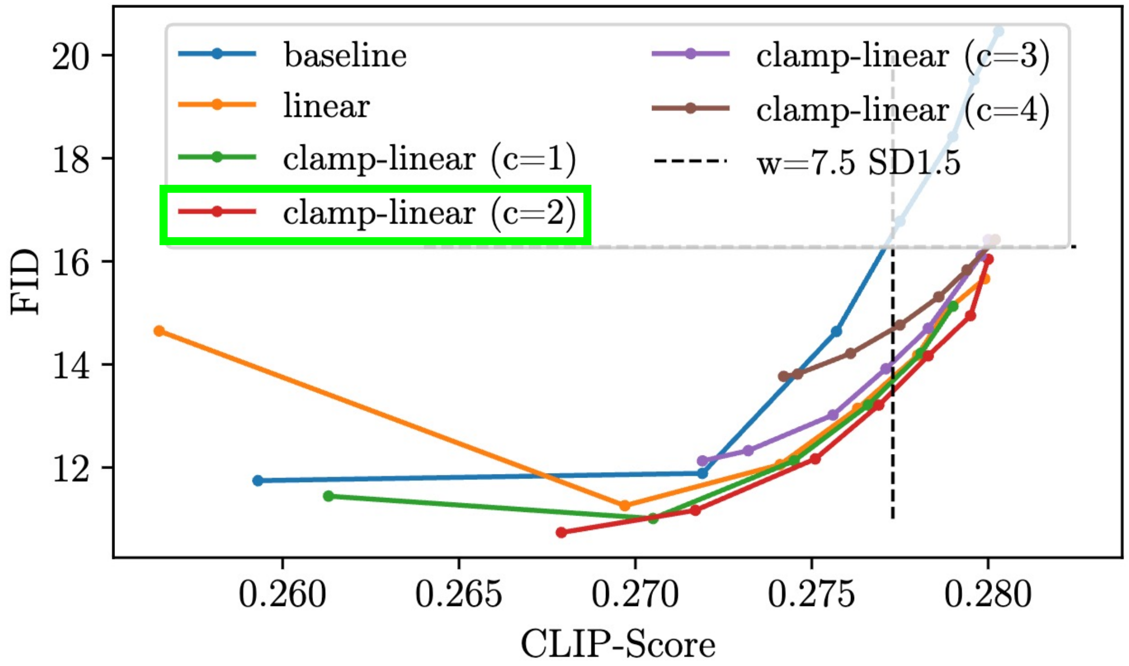
Parametrized cosine



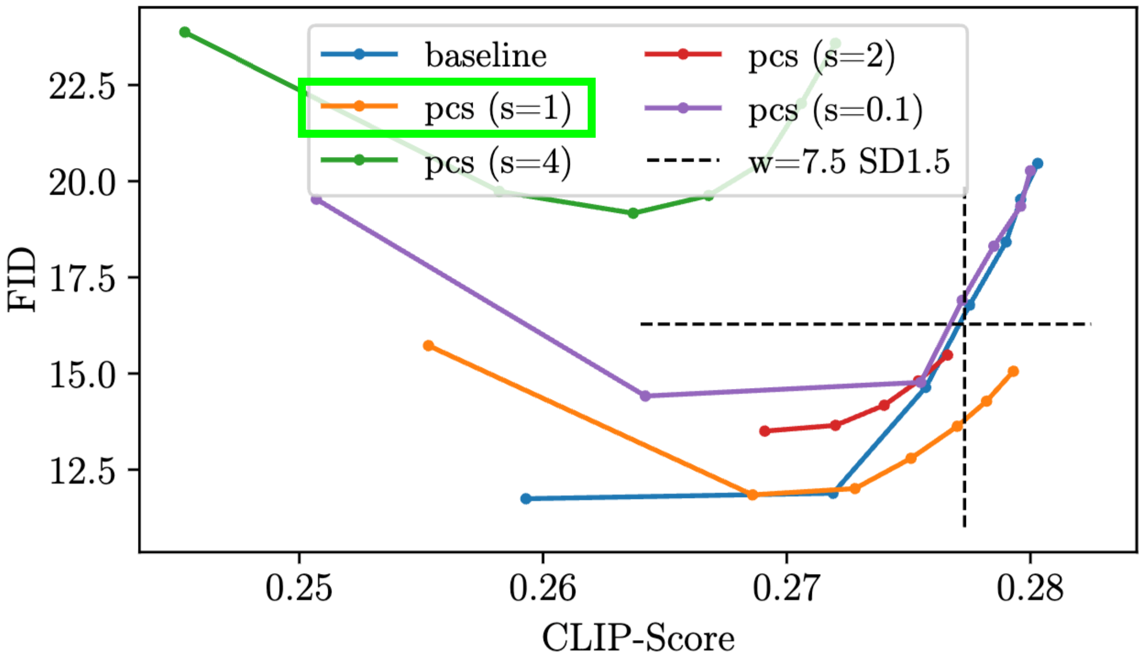
$$\omega(t, s) = \frac{1 - \cos \pi (\frac{T - t}{T})^{s}}{2} \omega$$

Quantitative Results: Parametrized functions

Clamping

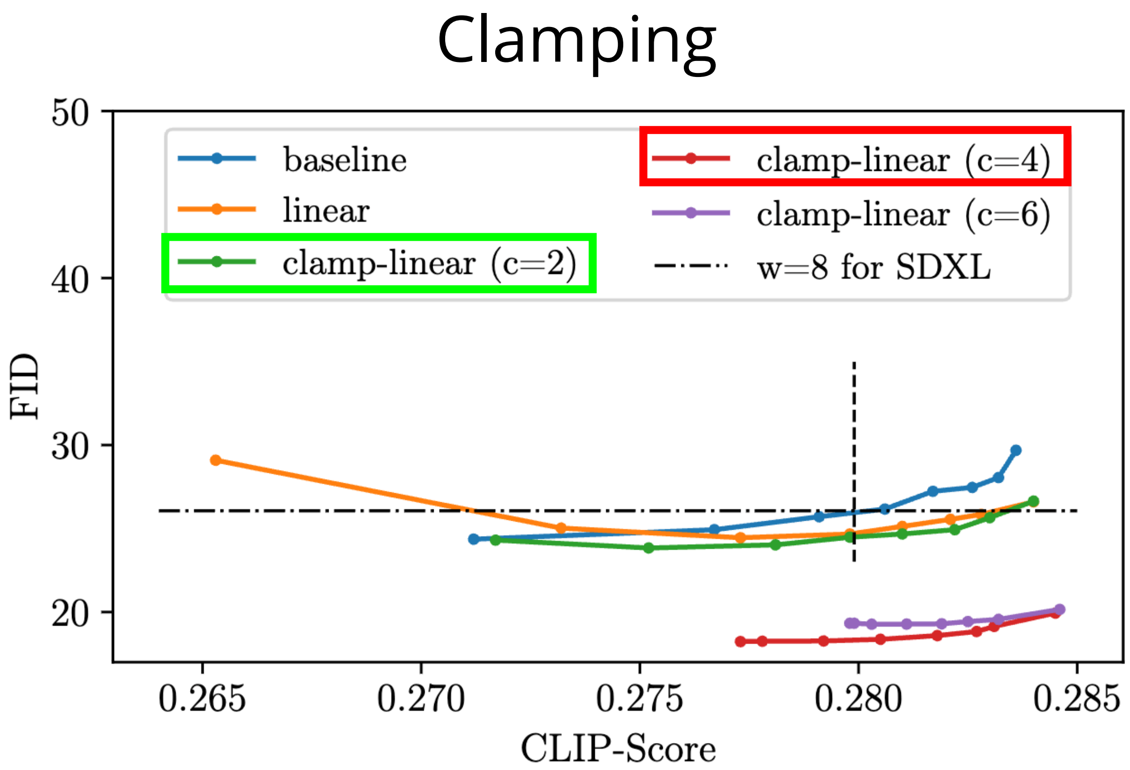


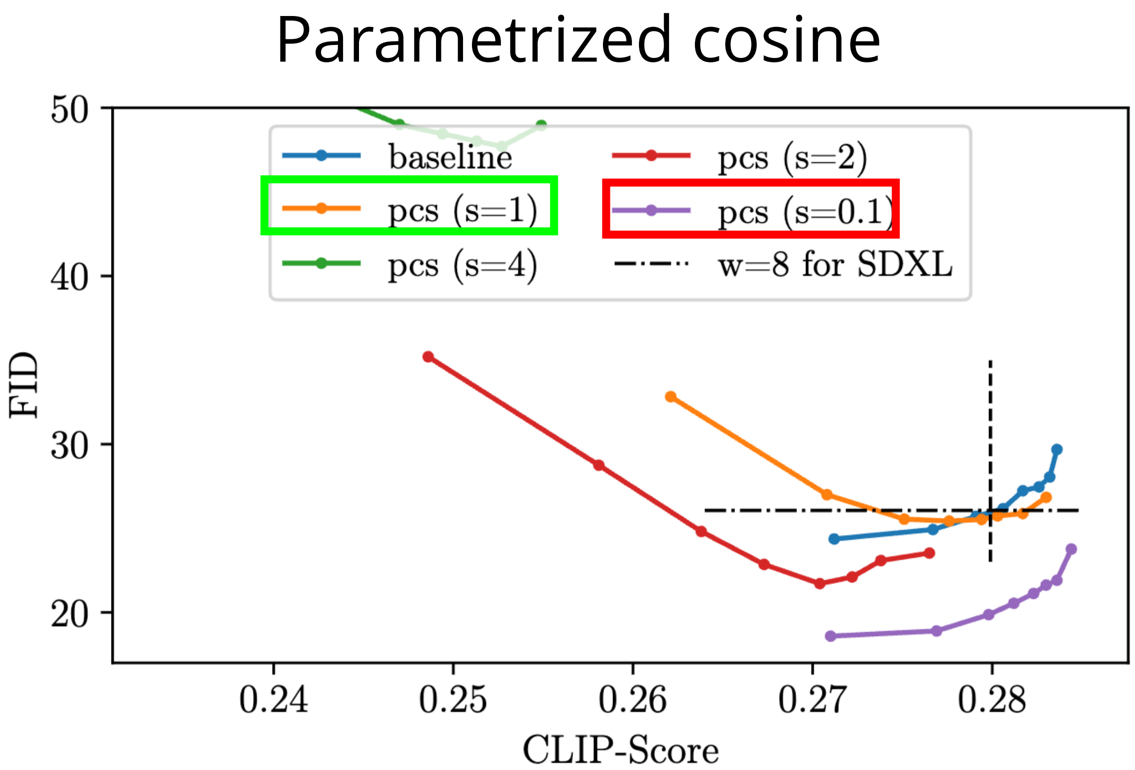
Parametrized cosine



SD1.5

Quantitative Results: Parametrized functions

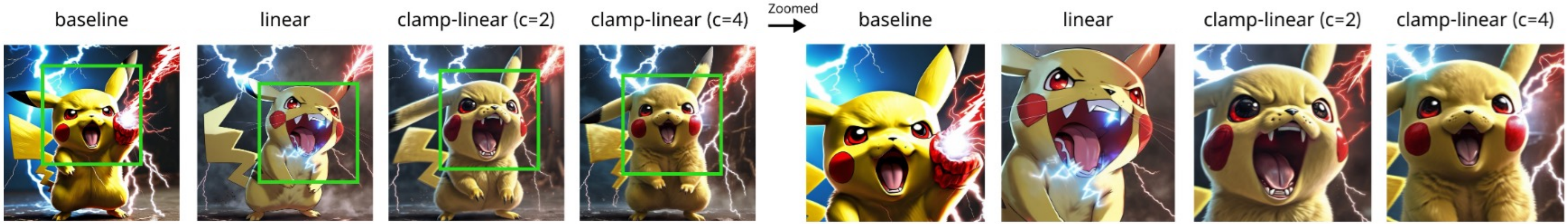




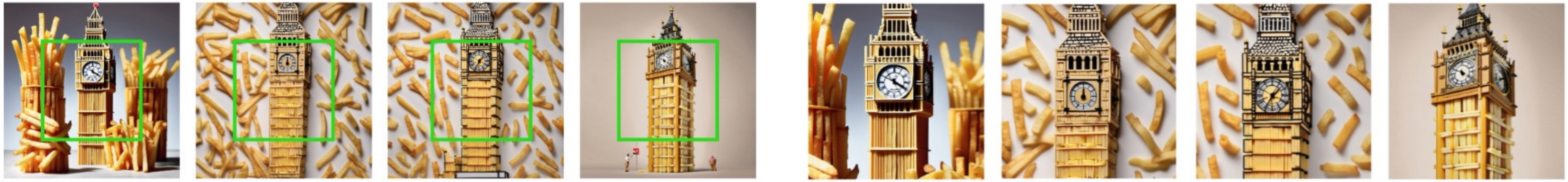
Observation: tuning correctly can improve the performance, but the tuning is *not generalizable*

SDXL

Qualitative Results: Parametrized functions



prompt: A Pikachu with an angry expression and red eyes, with lightning around it, hyper realistic style.

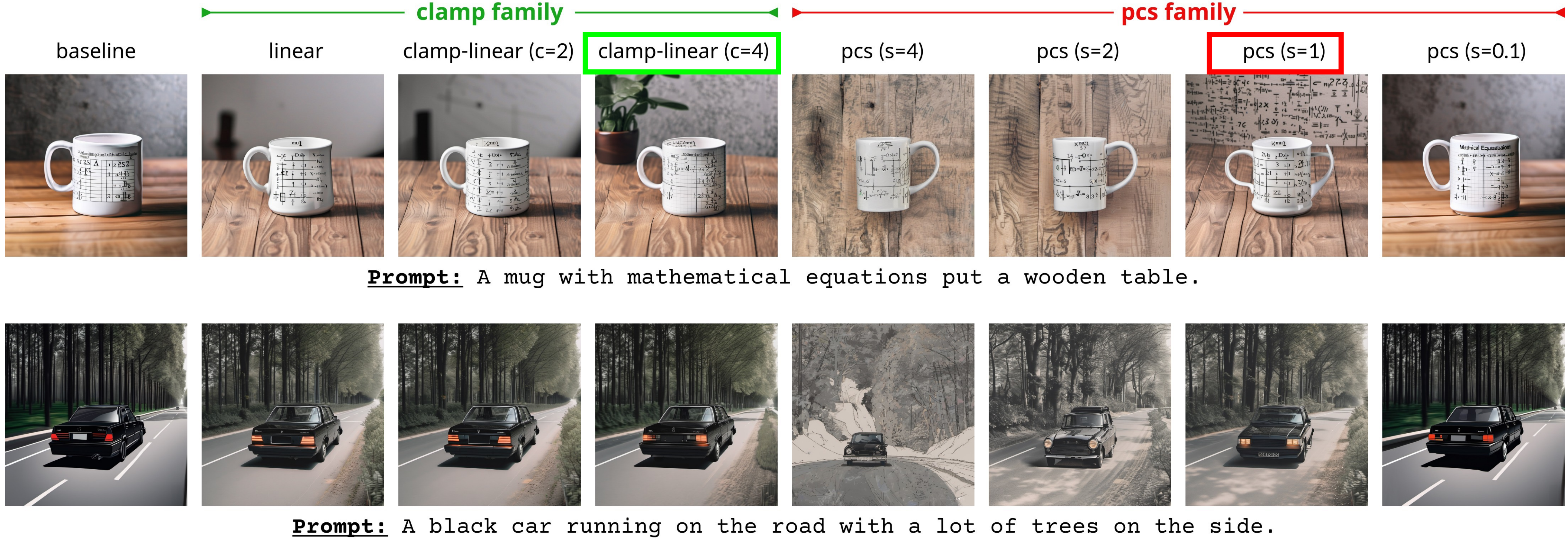


prompt: Big Ben made of French fries.

Textual comprehension, fidelity, attention to detail

SDXL

Qualitative Results: Parametrized functions



+ better details (mug)
+ more realistic (car)
+ better textured background (mug)

Conclusion

- Among heuristic functions, monotonically increasing guidance schedulers enhance both performance and diversity

- Well-tuned parameterized functions can achieve *better performa*nce but risk overfitting and require additional time and computational resources for tuning

- The implementation code is 1-line, w/o retraining the model

Low static guidance:

```
w = 2.0
for t in range(1, T):
  eps_c = model(x, T-t, c)
  eps_u = model(x, T-t, 0)
  eps = (w+1)*eps_c - w*eps_u
  x = denoise(x, eps, T-t)
```

✗ Fuzzy images, but many details and textures

High static guidance:

```
w = 14.0
for t in range(1, T):
  eps_c = model(x, T-t, c)
  eps_u = model(x, T-t, 0)
  eps = (w+1)*eps_c - w*eps_u
  x = denoise(x, eps, T-t)
```

✗ Sharp images, but lack of details and solid colors

Dynamic guidance:

```
w0 = 14.0
for t in range(1, T):
  eps_c = model(x, T-t, c)
  eps_u = model(x, T-t, 0)
  # clamp-linear scheduler
  w = max(1, w0*2*t/T)
  eps = (w+1)*eps_c - w*eps_u
  x = denoise(x, eps, T-t)
```

✓ Sharp images with many details and textures, without extra cost.

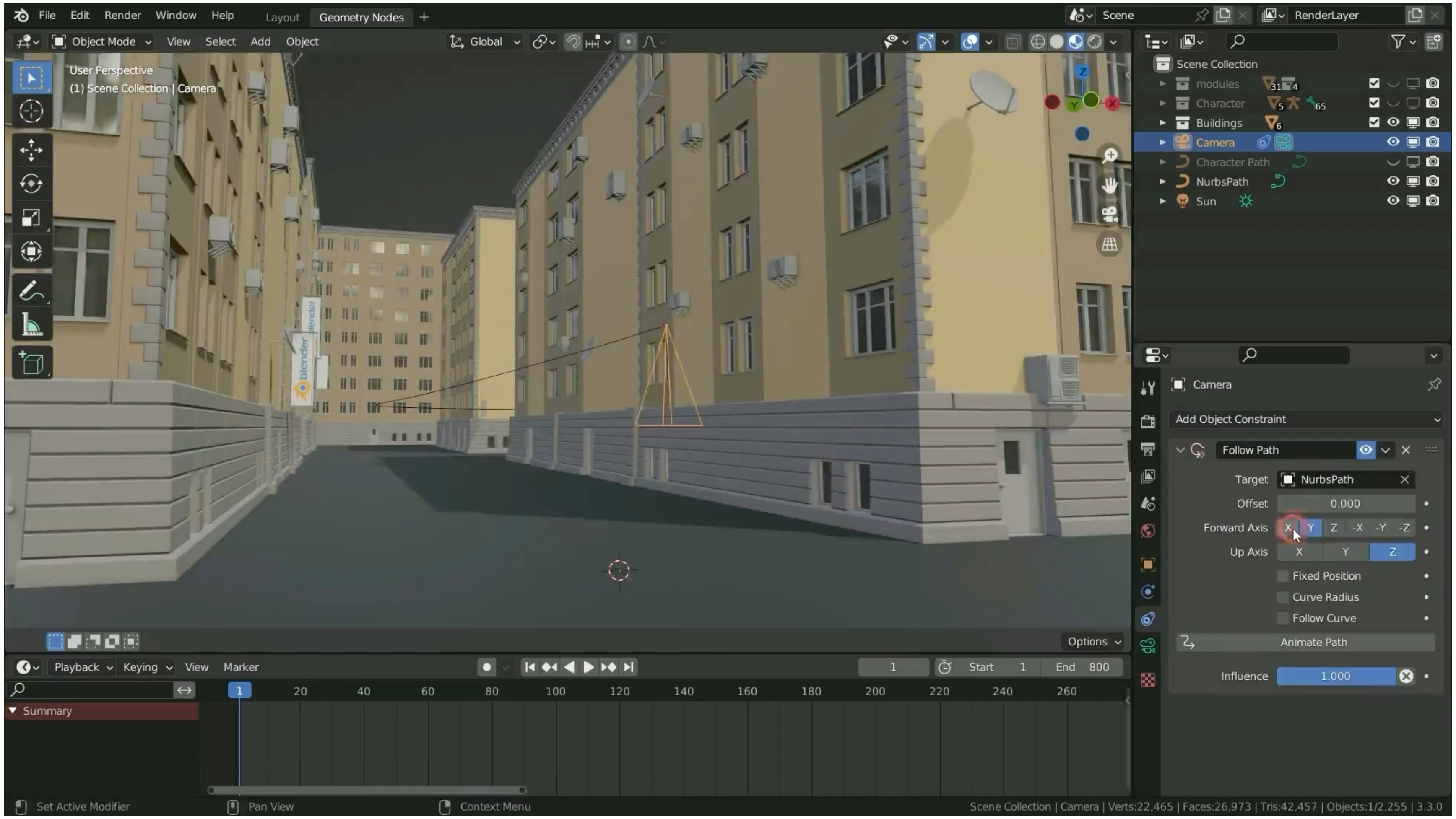
"full body, a cat dressed as a Viking, with weapons in his paws, on a Viking ship, battle coloring, glow hyper-detail, hyper-realism, cinematic, trending on artstation"

E.T. the Exceptional Trajectories: Text-to-camera-trajectory generation with character awareness

Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, Vicky Kalogeiton
ECCV 2024

Introduction

Challenges: Democratization



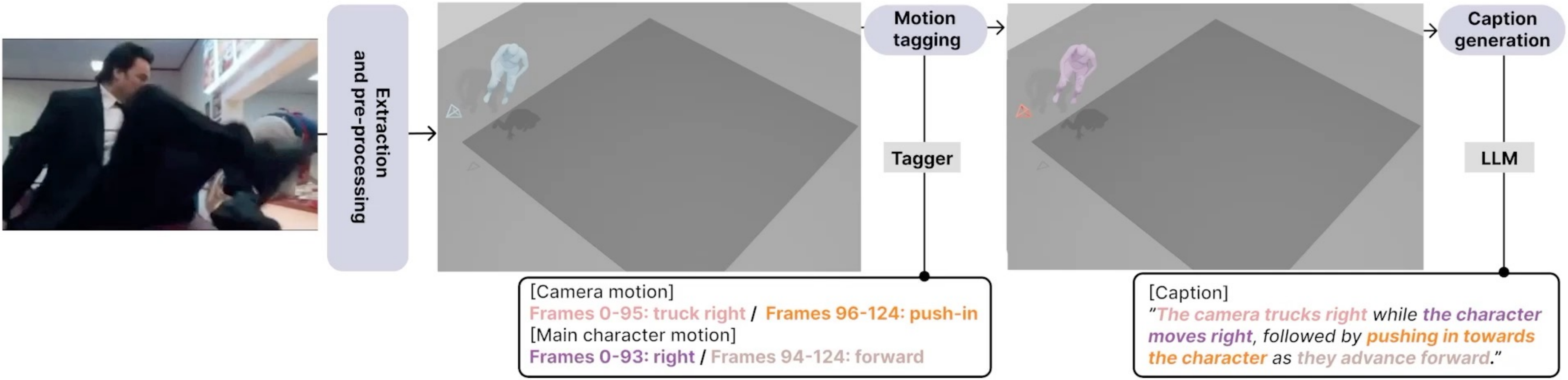
Challenges: Film grammar

No editing

Edited

The importance, motivations, intentions, emotions, … conveyed by a scene depending on the filmographer's style

The Exceptional Trajectories dataset Creation pipeline



The Exceptional Trajectories dataset (E.T.)

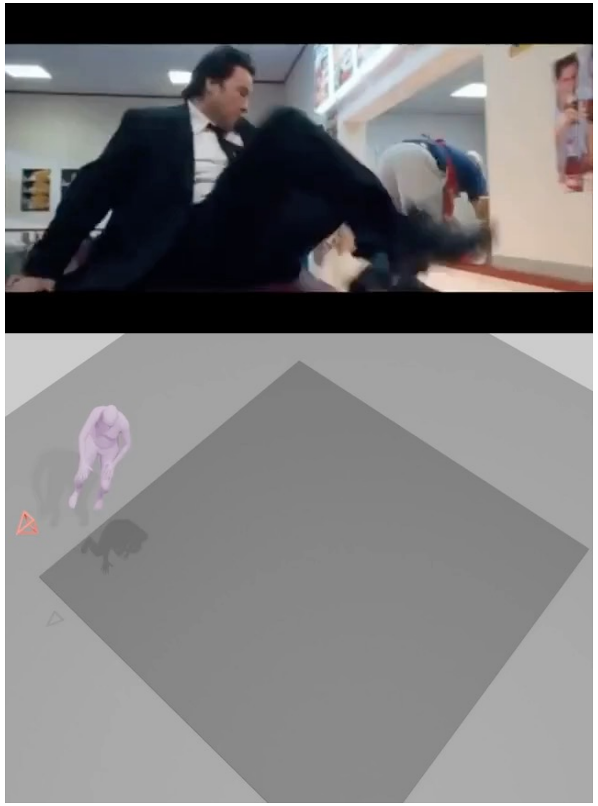
Dataset	#Samples	#Frames	#Hours	Domain	Character Traj	Character #Captions	Camera Traj	Camera #Captions	#Vocabulary
KIT Motion-Language [30]	4K	0.8M	11.23	Mocap	✓	6K		-	1,623
HumanML3D [10]	14K	2M	28.59	Mocap	✓	45K		-	**5,371**
RealEstate10k [47]	79K	**11M**	**121**	Youtube		-	✓	-	-
CCD [18]	25K	4.5M	50	Synthetic		-	✓	25K	48
E.T. (Ours)	**115K**	**11M**	120	**Movie**	✓	**115K**	✓	**230K**	1,790

→ **Cinematic content:** extracted from real-world movies

→ **Scale:** 120+ hours of content

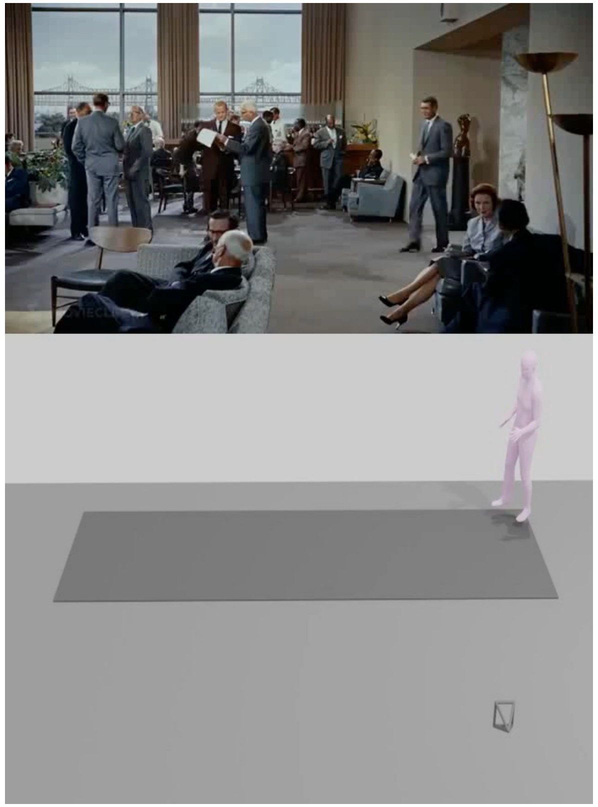
→ **Controlability:** camera AND character trajectories w/ captions

The Exceptional Trajectories dataset (E.T.)



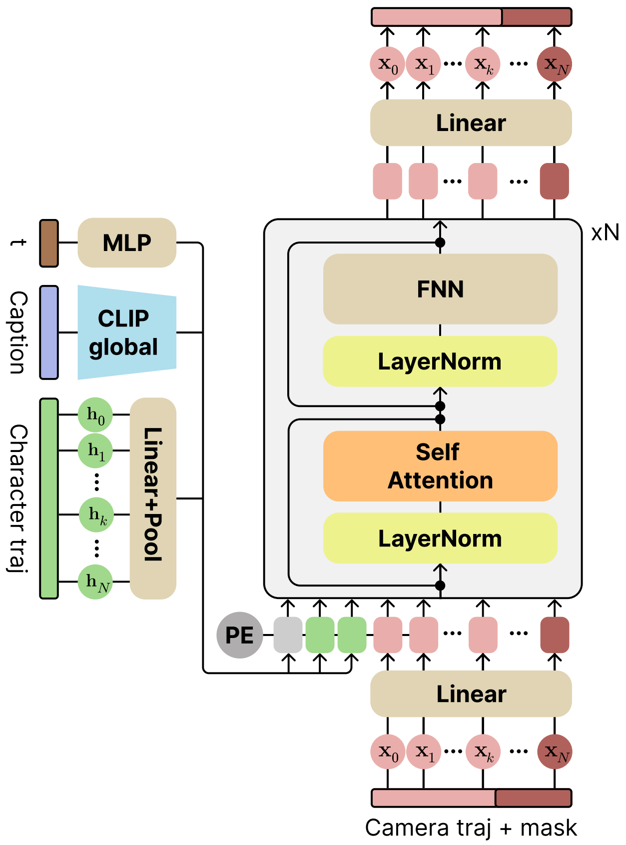
The camera trucks right while the character moves right, followed by pushing in towards the character as they advance forward

While the character moves right and then forward, the camera trucks right to follow their motion



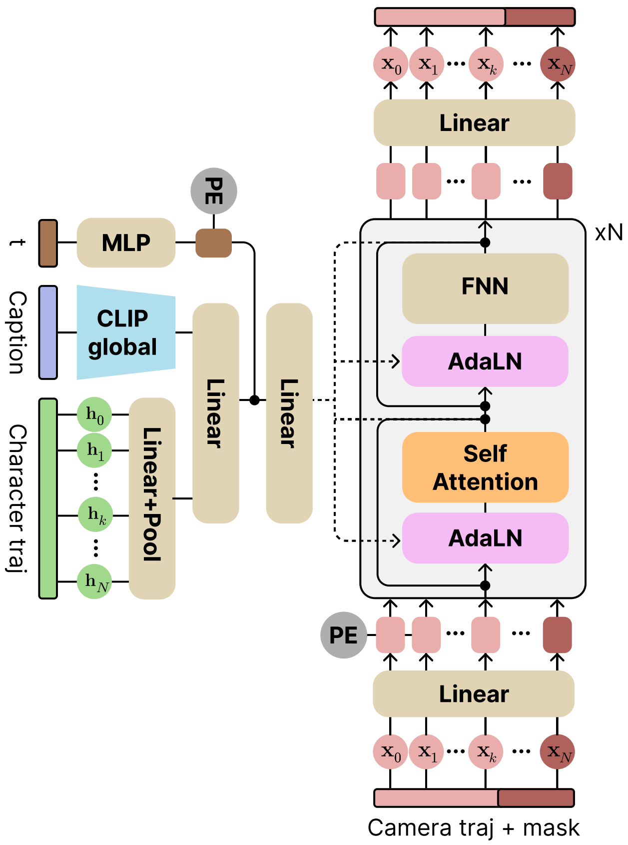
As the character moves left, the camera trucks left to maintain a consistent framing

Diffusion tRansformEr Camera TrajecORy (DIRECTOR)



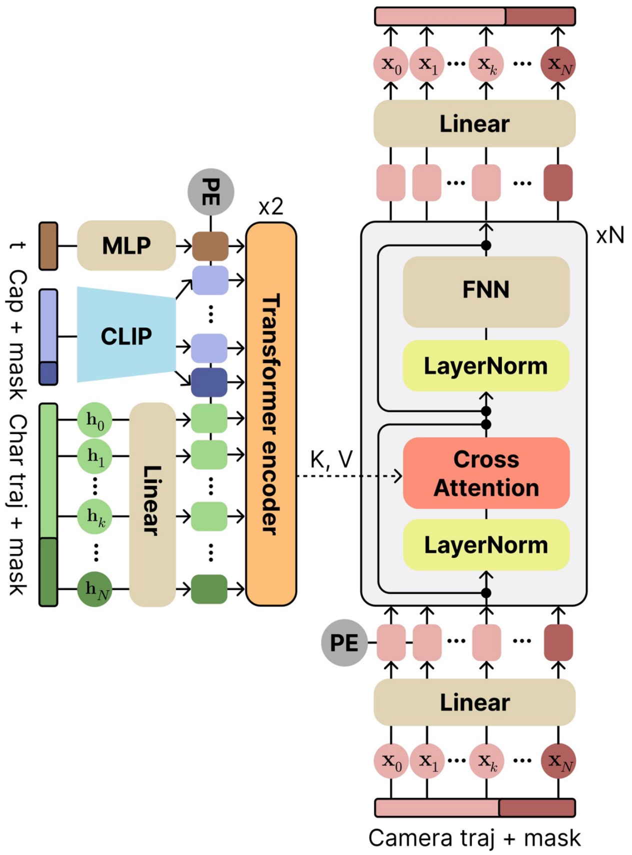
Config A - incontext

conditioning is added to the context of Transformer input



Config B - adaLN

DiT-like: Conditionings concat into single token; AdaLN instead of layer-norm

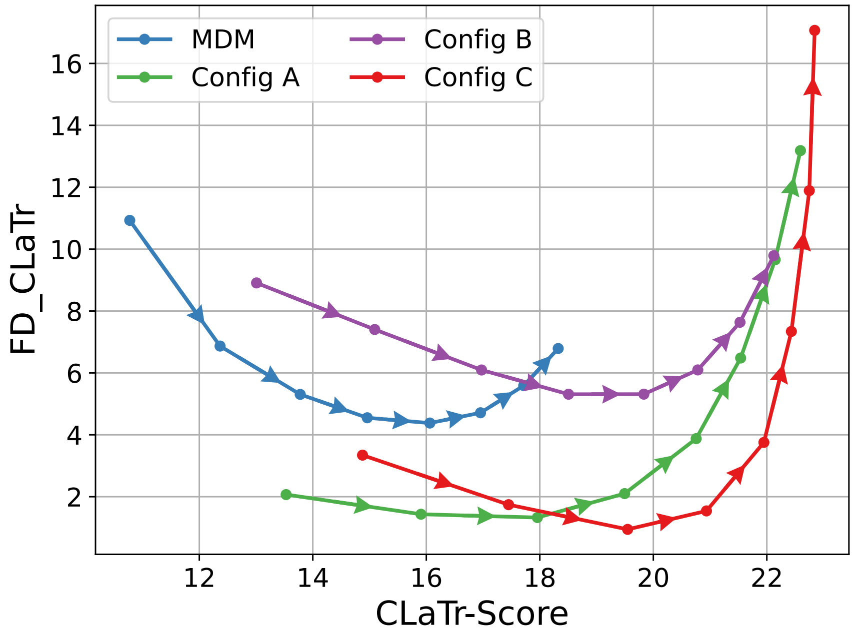


Config C – cross-attention

leverage the full sequence length of conditioning via Transformer Enc

Quantitative results

Set	Methods	ω	Camera trajectory quality: FD_{CLaTr} ↓	P ↑	R ↑	D ↑	C ↑	Text-camera coherence: CS ↑	C-P ↑	C-R ↑	C-F1 ↑
Pure traj	MDM	1.8	6.10	0.77	0.68	0.89	0.80	21.26	0.81	0.75	0.76
	Config A	1.6	5.16	0.82	0.67	1.00	0.86	21.88	0.84	0.78	0.80
	Config B	1.8	6.61	0.80	0.72	0.92	0.82	23.10	0.85	0.80	0.86
	Config C	1.6	4.57	0.83	0.65	1.00	0.87	21.49	0.83	0.78	0.80
Mixed traj	MDM	2.0	6.79	0.78	0.65	0.85	0.76	18.32	0.36	0.36	0.34
	Config A	1.4	3.88	0.82	0.68	0.98	0.85	20.76	0.43	0.43	0.42
	Config B	1.6	6.10	0.78	0.74	0.85	0.78	20.78	0.41	0.40	0.39
	Config C	1.4	3.76	0.83	0.67	1.00	0.86	21.95	0.49	0.49	0.48

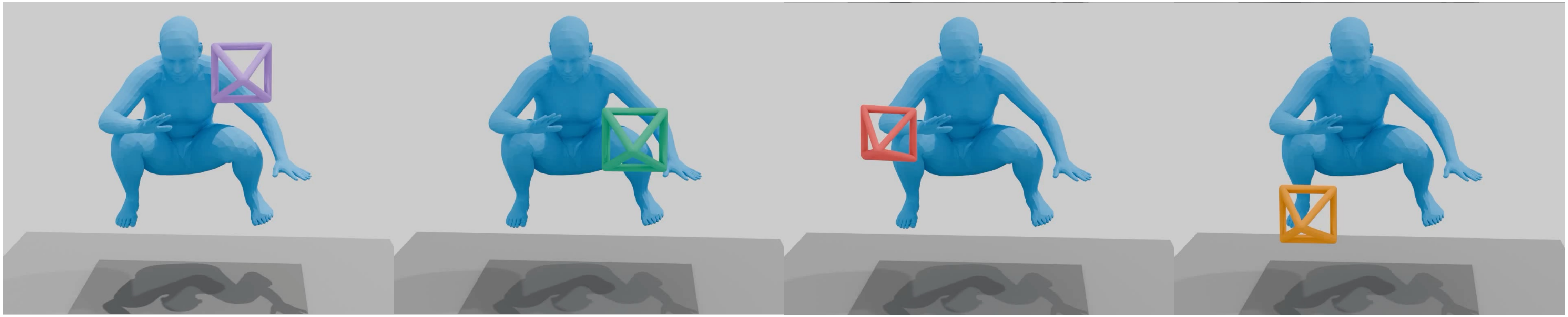


→ Config A (in-context): has a good tradeoff efficiency/performances

→ Config B (adaLN): fails to handle complex sequential conditions

→ Config C (CA): performs the best thanks to CA conditioning

Qualitative results: controlability



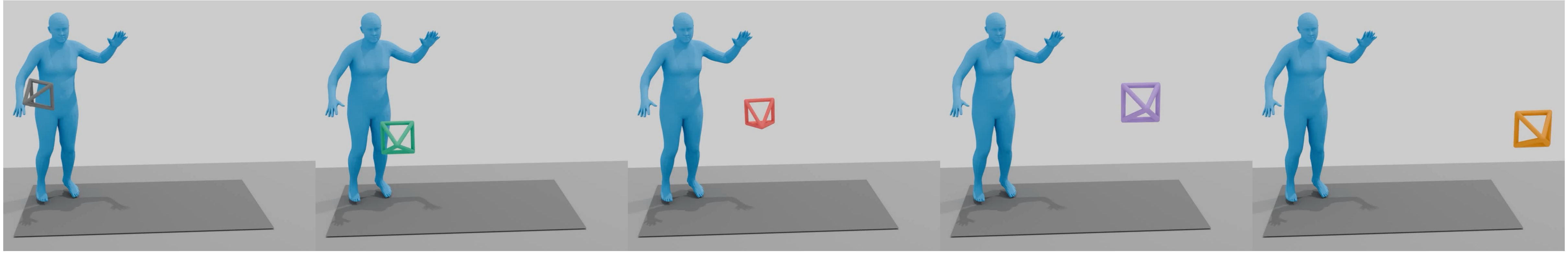
The camera **trucks right** while the character remains stationary

The camera **trucks left** while the character remains stationary

The camera **booms top** while the character remains stationary

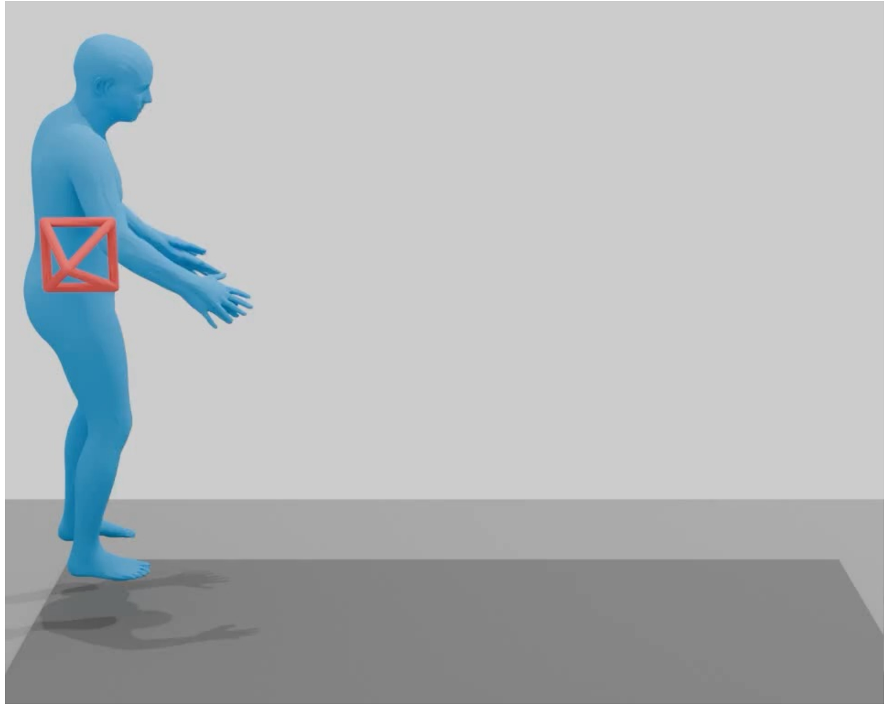
The camera **booms bottom** while the character remains stationary

Qualitative results: diversity

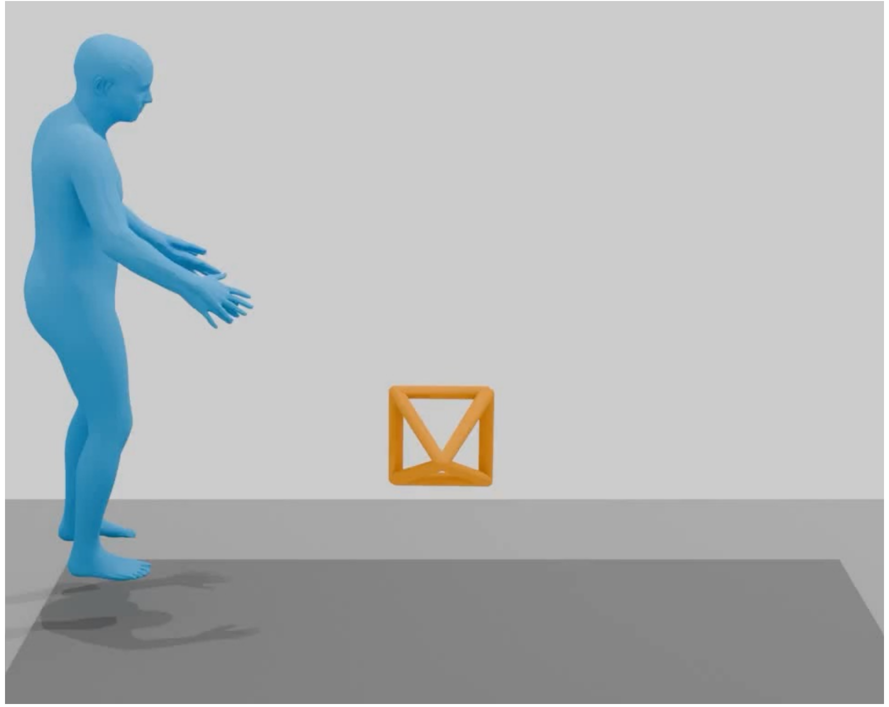


While the character moves right, the camera performs a **boom bottom**

Qualitative results: complexity



While the character moves to the right,
the camera stays static and pushes-in
once the character stops

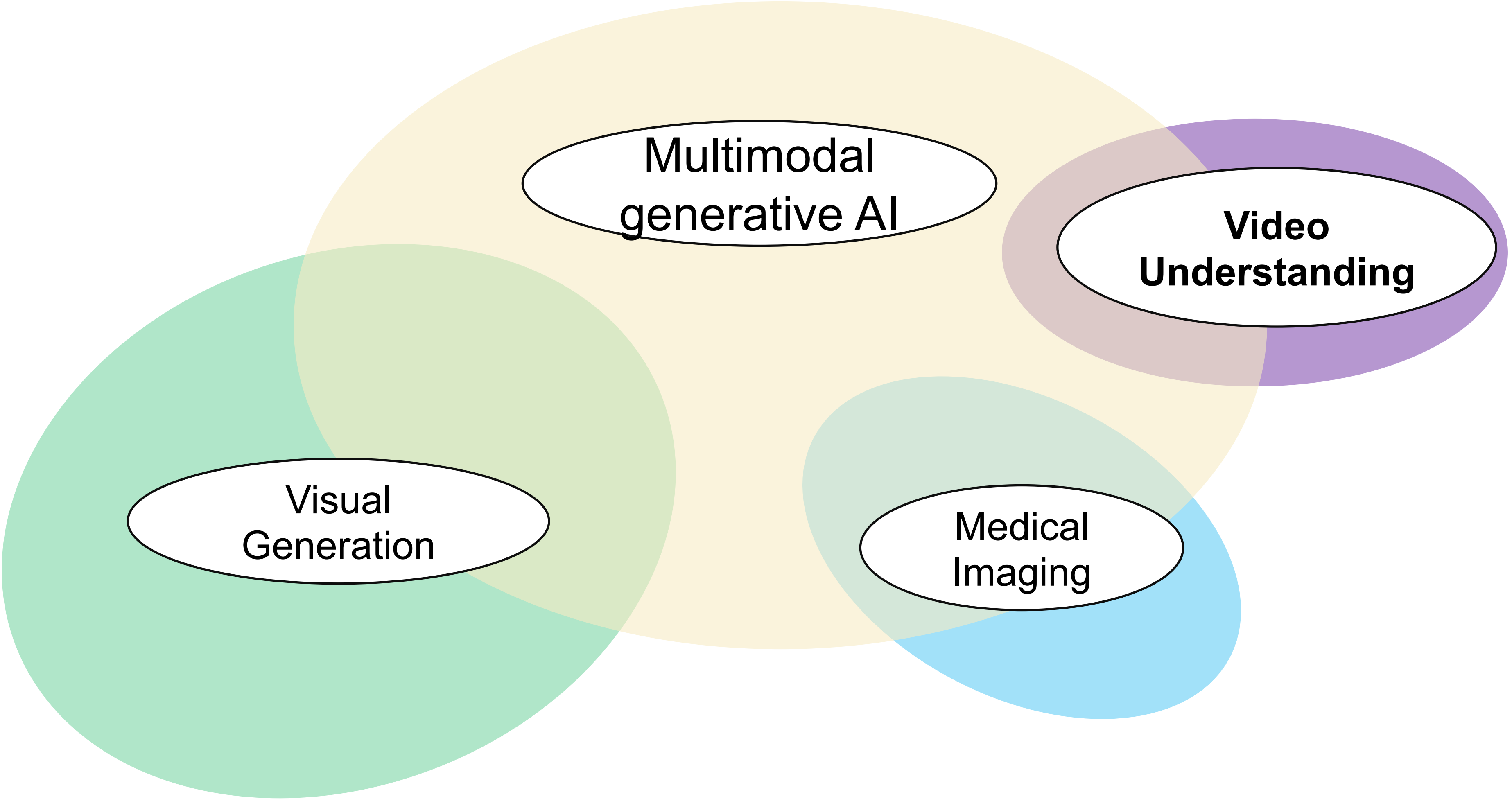


While the character moves to the right,
the camera trucks right and remains static
once the character stops

Qualitative results: character-awareness

The camera remains static as the character moves to the
[right / left]

Research agenda



Is this funny?

Audiovisual Learning of Funny Moments in Videos

Mia: Three tomatoes are walking down the street -- a poppa tomato, a momma tomato, and a little baby tomato. Baby tomato starts lagging behind.
Poppa tomato gets angry, goes over to the baby tomato, and squishes him... and says, ... "Catch up."

[Video scene from Pulp fiction, 1994, source video: https://www.youtube.com/watch?v=4L5LjjYVsHQ]

FunnyNet-W: Multimodal Learning of Funny Moments in Videos

[IJCV 2024, ACCV 2022, Oral, Honorable Mention Award
Z.S. Liu, R. Courant, V. Kalogeiton]

Code & demo:
https://www.lix.polytechnique.fr/vista/projects/2022_accv_liu/

CIHE Caritas Institute of Higher Education 明 愛 專 上 學 院

VISTA Visual Worlds: Temporal Analysis, Animation and Authoring

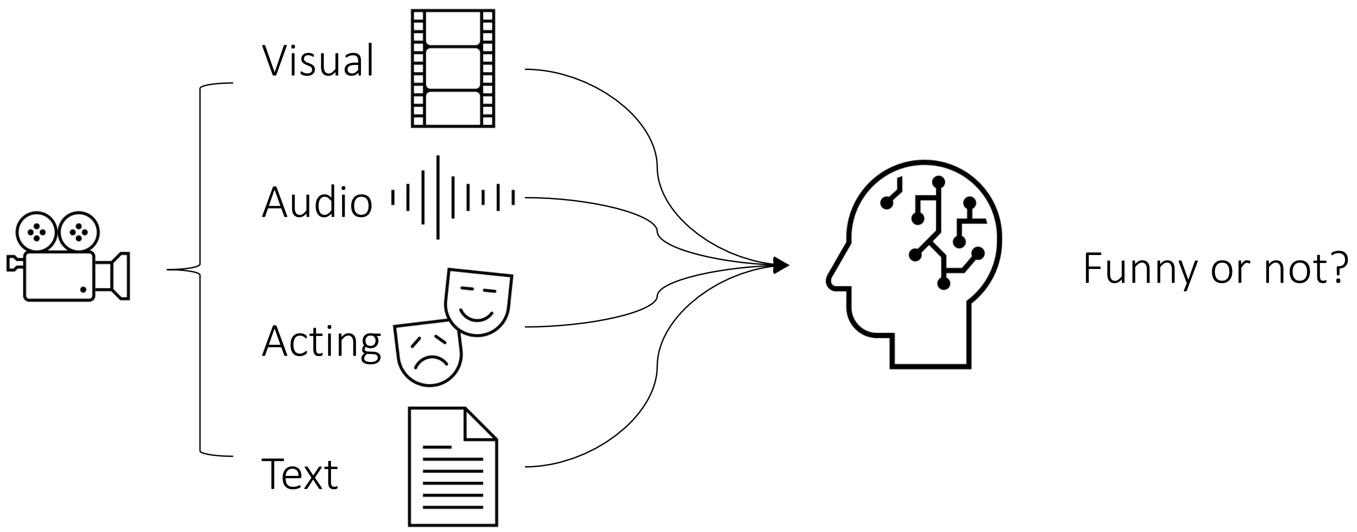
Why does it matter ?

In-the-wild funny moment detection: detecting funny moments in any content.

Human-machine interactions: making conversational AI more spontaneous.

Make computers funny! comprehending what is funny.

Background and introduction

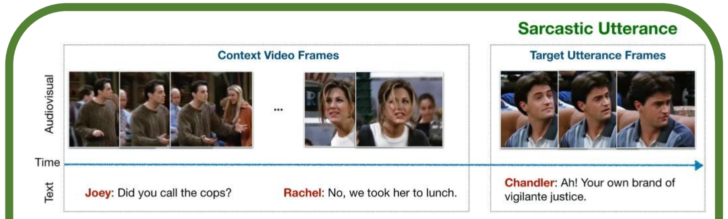


Understanding funnyness: complex → Purely visual / auditory / mix both

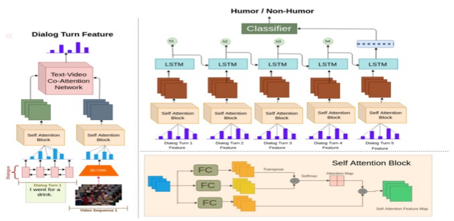
Multimodality

No recipe for the perfect joke!

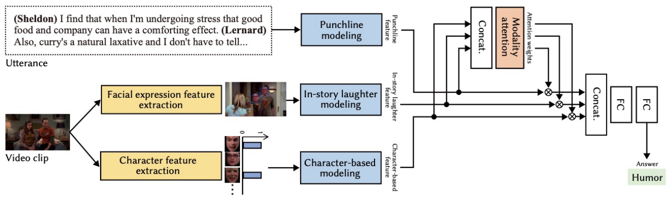
Related work



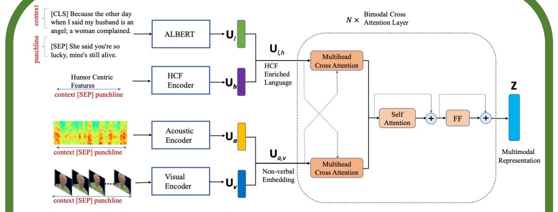
MUStARD
[Castro et al., ACL, 2019]



MHD
[Patro et al., WACV, 2021]



Laughing Machine
[Kayatani et al., WACV, 2021]



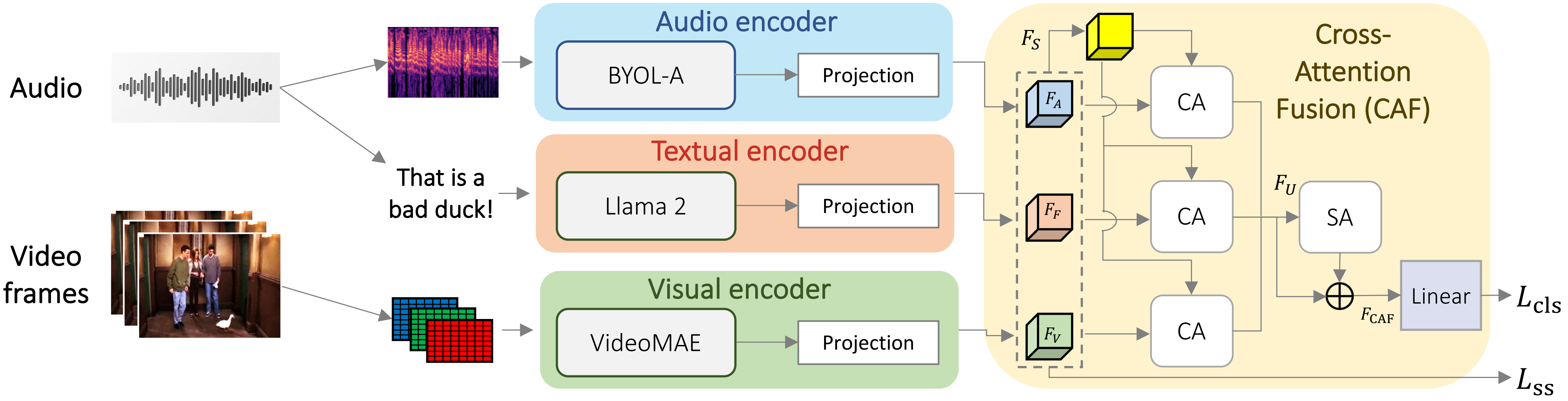
HKT
[Hasan et al., AAAI 2021]

Focus mainly on textual modality
- ➢ Limited, imperfect
- ➢ Not flexible

Ours

+ Exploit only raw video modalities: audio, visual, text
+ Self-supervised: use canned laughter for supervision

Method: FunnyNet-W



Audio encoder: $F_A \in \mathbb{R}^{512}$

Visual encoder: $F_V \in \mathbb{R}^{512}$

Text encoder: $F_T \in \mathbb{R}^{512}$

$\oplus \rightarrow$ Fused vector: $F_U \in \mathbb{R}^{3\times512}$

Cross Attention (CA) for across-cue correlations

$$F_S = \sum_{i\in\{A,V,T\}} \sigma\left(\frac{Q_U K_i^T}{\sqrt{d}}\right) V_i$$

Self Attention (SA) for inter correlations

$$F_{CAF} = F_S + \sigma\left(\frac{Q_S K_S^T}{\sqrt{d}}\right) V_S$$

Ablations

Visual

A	V	T^a	F1	Acc
BYOL-A	Timesformer	Bert	84.2	80.9
BYOL-A	VideoMAE	Bert	85.3	82.3
A	**V**	**T^{gt}**		
BYOL-A	Timesformer	Bert	84.9	80.8
BYOL-A	VideoMAE	Bert	87.2	83.8

Text

A	V	T^a	F1	Acc
BYOL-A	VideoMAE	Bert	85.3	82.3
BYOL-A	VideoMAE	GPT2	85.2	82.3
BYOL-A	VideoMAE	LlaMa-2	88.2	85.6
A	**V**	**T^{gt}**	**F1**	**Acc**
BYOL-A	VideoMAE	Bert	87.2	83.8
BYOL-A	VideoMAE	GPT2	88.1	85.6
BYOL-A	VideoMAE	LlaMa-2	89.3	86.8

Audio

A	V	T^a	F1	Acc
BEATS	VideoMAE	LlaMa-2	78.2	65.1
CAV-MAE	VideoMAE	LlaMa-2	87.3	83.8
BYOL-A-v2	VideoMAE	LlaMa-2	87.6	84.7
BYOL-A	VideoMAE	LlaMa-2	88.2	85.6

V	A	T	F1	Acc
✓	–	–	73.2	64.1
–	✓	–	73.7	66.6
–	–	✓	77.8	68.1
✓	✓	–	84.3	79.3
–	✓	✓	84.5	80.3
✓	–	✓	74.9	64.3
✓	✓	✓	**88.2**	**85.6**

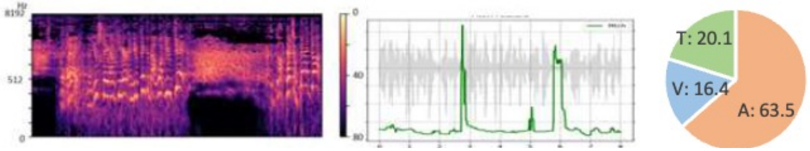
CAF Self	CAF Cross	A+V F1	A+V Acc	A+T F1	A+T Acc	V+T F1	V+T Acc	A+V+T F1	A+V+T Acc
-	-	80.1	76.5	81.0	76.9	73.5	63.8	82.4	77.8
✓	-	81.1	77.3	81.4	77.5	74.4	64.4	85.7	81.8
-	✓	83.6	78.7	82.3	78.7	74.6	64.2	85.4	81.4
✓	✓	**84.3**	**79.3**	**84.5**	**80.3**	**74.9**	**64.3**	**88.2**	**85.6**
MMCA [95]		83.1	78.3	83.4	79.8	73.6	63.8	87.0	84.5
CoMMA [85]		83.5	78.5	83.9	80.3	74.2	64.1	87.6	85.1

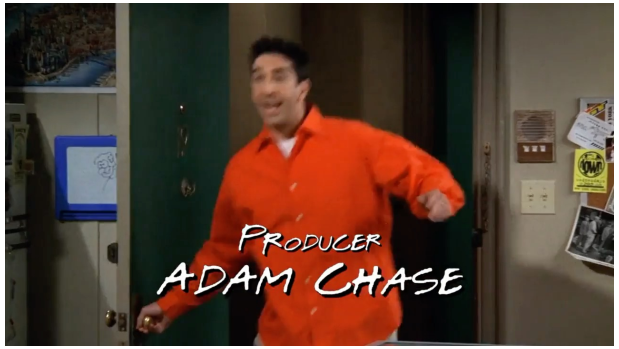
Qualitative results: Modality impact

Funny predictions

Chandler: Okay! Now you stay out here and think about what you did!!

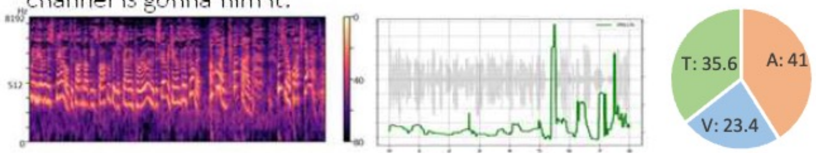
Ross: That is a duck.
Chandler (high pitch): That is a bad duck!!





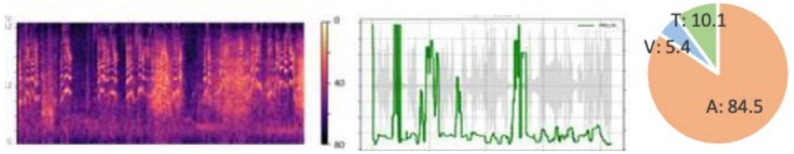
Ross: they are putting together this panel to talk about fossils they just found in Peru and the Discovery channel is gonna film it.

Chandler: Oh my God! (pause) Who is gonna watch that?



Phoebe: I am setting the phone down. Don't go anywhere, I am still here.

Phoebe (speech rate change): One sec! One second! Wait! One second! Just!



- Positive: high pitch, pause, speech rate change indicate punchline for laughter

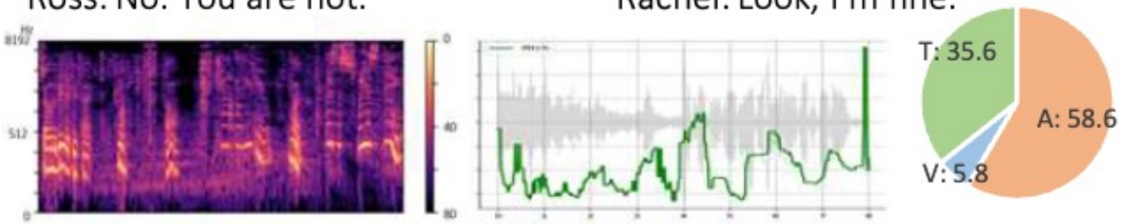
FunnyNet-W: Text | FunnyNet-W: Visual | FunnyNet-W: Audio

Qualitative results: Modality impact

Not-funny predictions

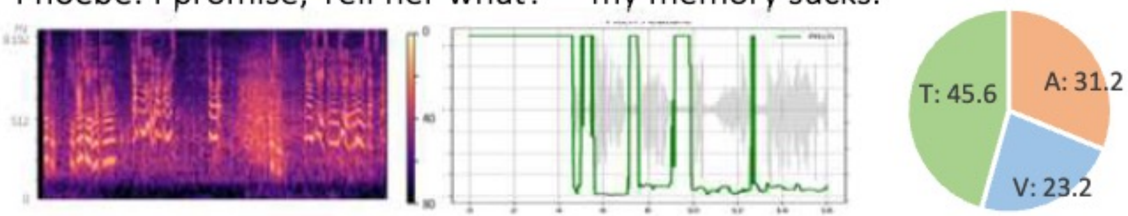
Ross: Is that still…
Rachel: I'm fine. I'm fine.
Ross: No. You are not.

Rachel: Yes, I am.
Ross: Rach
Rachel: Look, I'm fine.



Pete: Can you promise you won't tell he though?
Phoebe: I promise, Tell her what?

Pete: Thanks a lot.
Phoebe: No. I'm intuitive but my memory sucks.



- Positive: high pitch, pause, speech rate change indicate punchline for laughter
- Negative: neutral voice and facial expressions

FunnyNet-W: Text | FunnyNet-W: Visual | FunnyNet-W: Audio

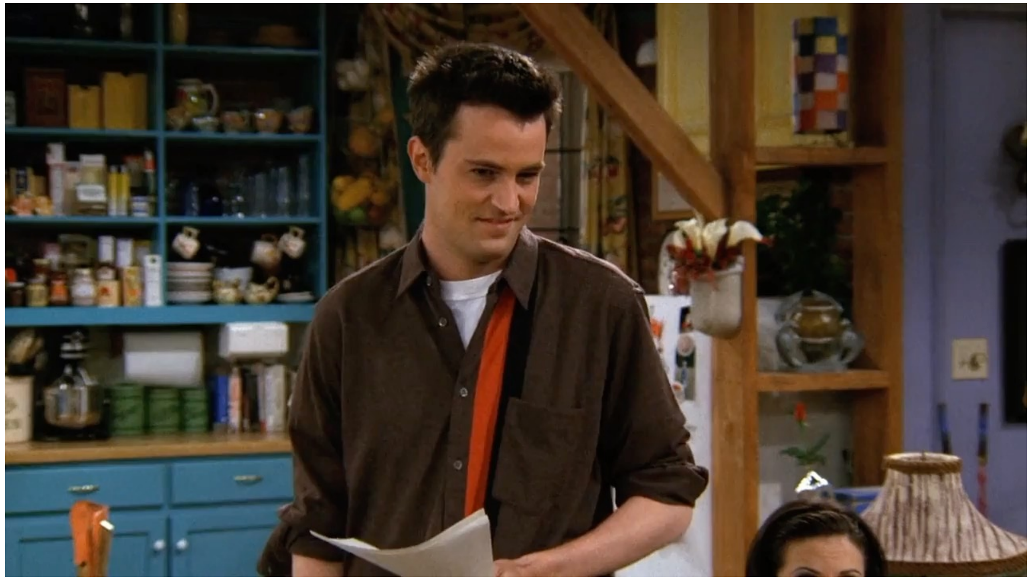
Quantitative results: Comparison to chatbot

Prompt engineering	Prompt training	F1	Accuracy
Generic	-	14.5	41.8
	✓	44.3	46.5
Specific	-	64.1	53.2
	✓	71.1	55.9
FunnyNet-W (T)		77.8	68.1
FunnyNet-W (A+V+T)		88.2	85.6

Models		LlaMa-2		FunnyNet-W
		w/o PT	w PT	
Funny (positive)	They are putting together this panel to talk about fossils they just found in Peru and the Discovery channel is gonna film it. Oh my god, who's gonna watch that?	No	Yes	Yes
	I didn't wear this suit for a year because you hated it. You're not my girlfriend anymore, Now that you're on your own, you're free to look as stupid as you'd like.	No	Yes	Yes
Not funny (Negative)	I hope it won't be too weird. will it? Rache? No, not at all. I'm actually gonna bring someone myself	No	Yes	No
	Let me walk you home and stop by every newsstand and burn every copy of The Times and The Post.	Yes	Yes	No

Failure cases: False Positives

Strong emotional responses expressed by single wording



Chandler: Something else I just said?
Rachel: I don't know. Weren't you the guy who told me to quit my job when I had absolutely nothing else to do? Ha! Ha! Ha! Ha!

Gunther: Rachel, I just made you cocoa.
Rachel: OMG, you are so nice.
Monica: (screaming) Ah!!
Phoebe: Are you guys OK?

Failure cases: False Negatives

Subtle sarcastic comments with straight face and no follow-up indications or inside jokes that require long-term understanding

Ross: I made a mistake.
Rachel: A mistake? Where were you trying to put it in? Her purse?
Phoebe: Where? Where did he put in?

Joey: You know, they call it "The Ross".
Joey: People like, huh, he's got a Ross.
Ross: Yeah, that would be cool.

Sitcom w/ canned laughter

Sitcom w/ canned laughter

Examples of well classified
funny moments

Movies

Sitcoms w/o canned laughter

Stand-up comedy

Audio-only

Conclusion

- FunnyNet-W: self-supervised audio-visual model for funny moment detection
- Exploit only raw video modalities: *audio, visual, textual*
- Audio is the *dominant* cue
- Outperforms the state of the art
- Future work: *other languages, other types of humor*

Movie Question-Answering

Time scale

Motion planning

Task planning

Event planning

e.g. vacation planning

e.g. planning education; building a house…

seconds — minutes — hours — days — years → time

Time scale

Text resources

Cook boks, travel guides, blogs,

documentaries, history books

seconds

minutes

hours

days

years

time

Image/Video resources

Instructional videos, timelapses

Movies

SFD: Short Film Dataset Ghermi et al., ArXiv 2024

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

Comedy

The Beast, the Phantom and the Hunchback use a dating app to find love.

Romance

A couple on a first date clash over astrology.

Horror

*A young woman trapped in a bathroom stall during an active sho*ting.*

Action

The biggest boxing fight of 1960 takes an unexpected turn.

Drama

A mother struggles with bullies who torment her disabled daughter.

Animation

A young boy gets lost in a strange forest. Then his father tries to rescue him.

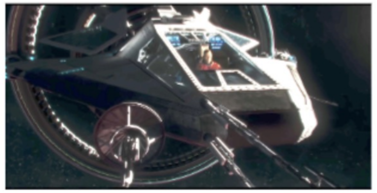
Documentary

How Movie Sounds are Made. An Inside Look at the World of Foley Artist.

Experimental

A stranded soul searches an endless desert for his purpose.

Sci-Fi



A lone astronaut testing the first faster-than-light spacecraft travels farther than he imagined

AI movie

After their father dies, two brothers turn to a traditional battle method to decide who will rule the kingdom.

Dance

Testament to timidity and enthusiasm; the dancers shed their hardened pandemic-built exteriors

Western

A young woman in the Old West sets off on a journey of revenge after her sister's murder

SFD is a VideoQA dataset, containing 1,078 movies and 4,885 questions. Videos last 13 minutes on average.

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

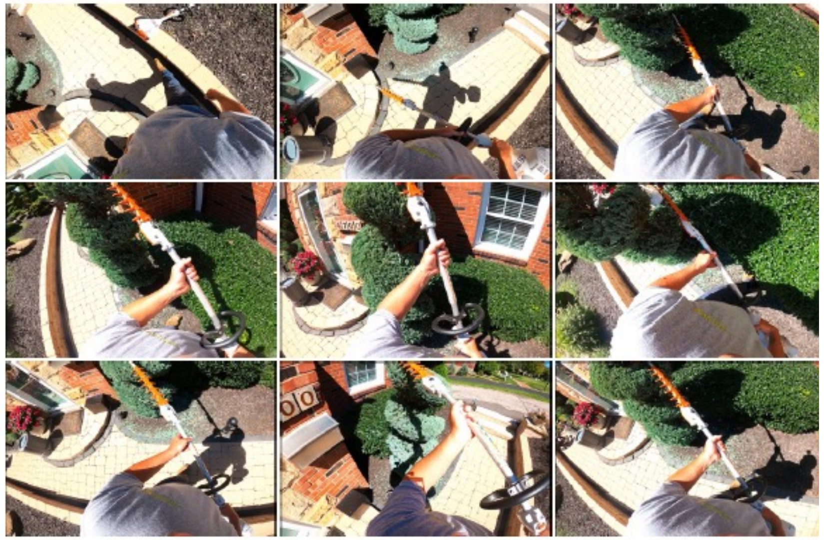
Movies

A young girl lives in a bunker, protected from the monsters. Today she goes outside...

Instructional videos

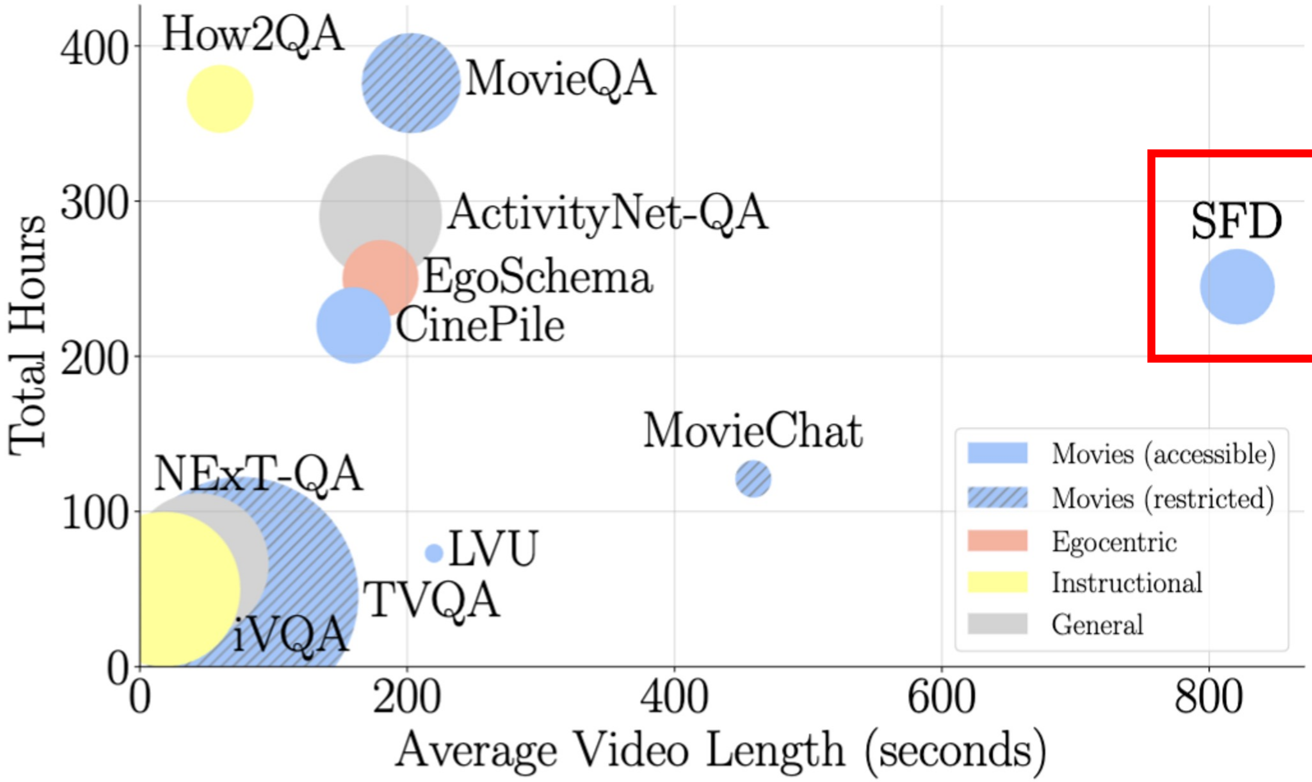
How to tie a necktie?

Egocentric videos



C was in the front yard, pulled the starter string and trimmed the tree with a hedge trimmer

Why another VideoQA dataset?

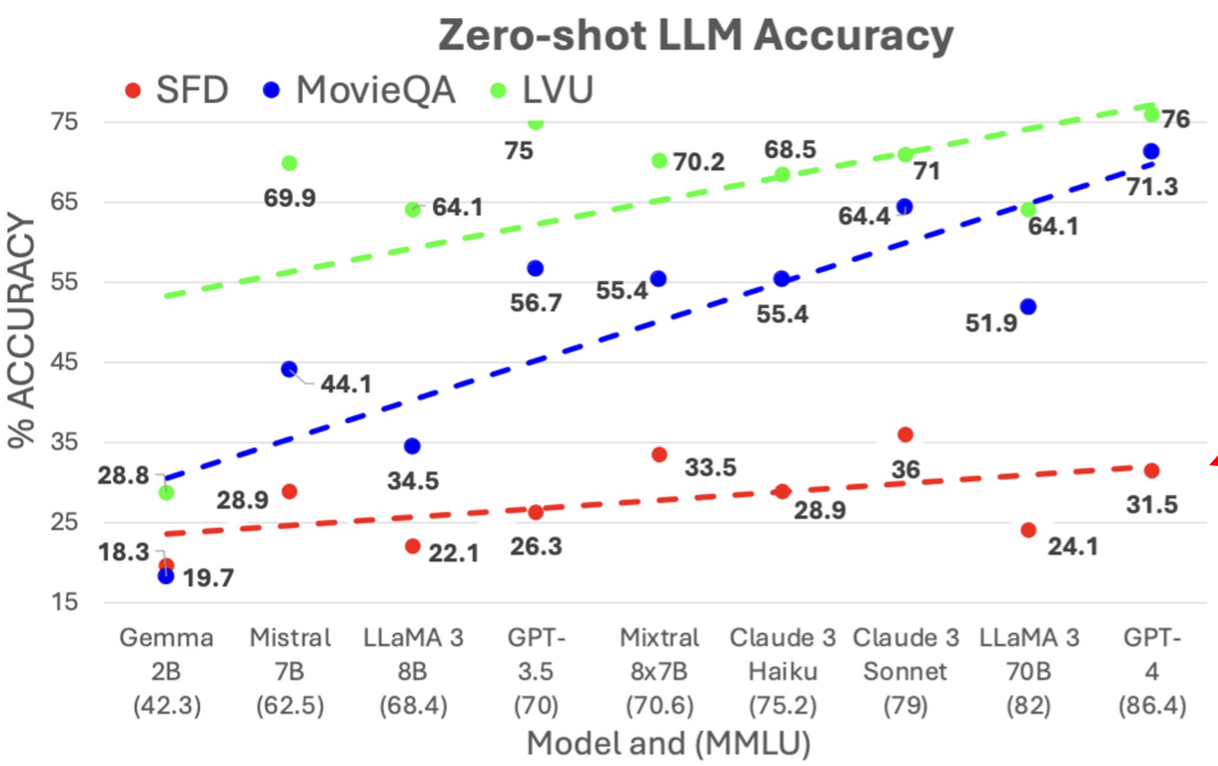


- Story-level QAs
- Publicly available videos

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/



Modern LLMs memorize common movies and can answer
Questions in in LVU and MovieQA given movie names

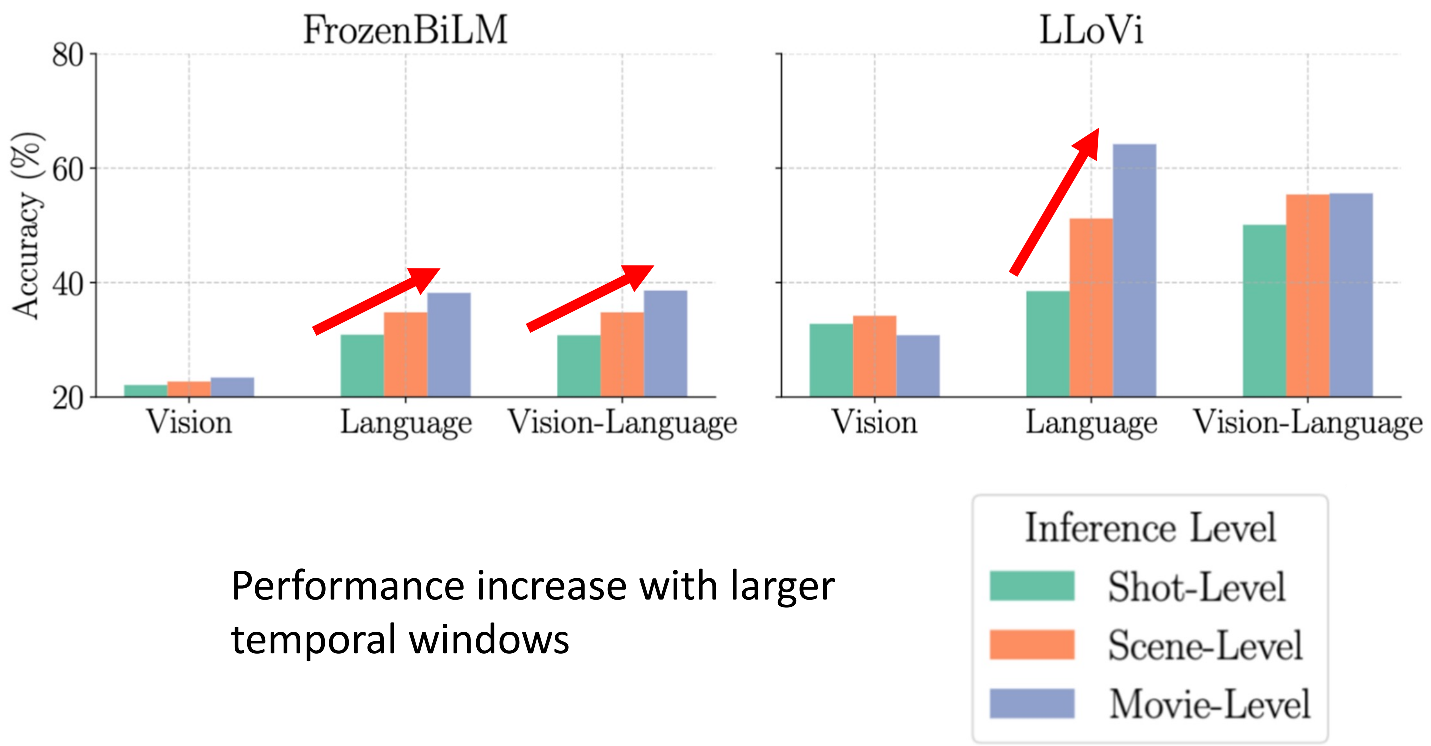
Why another VideoQA dataset?

- Story-level QAs
- Publicly available videos
- Limited/No data leakage

Short Film Dataset (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/



Performance increase with larger temporal windows

Why another VideoQA dataset?

- Story-level QAs
- Publicly available videos
- Limited/No data leakage
- Questions with long temporal context

SFD Short Film Dataset (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

Why another VideoQA dataset?

Method	Venue	% Accuracy Multiple-Choice QA V	L	VL	Open-Ended QA V	L	VL
Random		20.0	20.0	20.0	-	-	-
FrozenBiLM [5]	NeurIPS 2021	23.4	38.2	38.6	-	-	-
mPLUG-Owl2 [74]	CVPR 2024	38.3	20.7	21.3	22.1	1.8	1.6
Video-LLaVA [33]	arXiv 2023	34.2	21.3	24.7	19.2	6.4	8.0
LLoVi [79]	arXiv 2023	30.8	64.2	55.6	16.2	40.3	24.7
LangRepo [26]	arXiv 2024	29.0	32.1	31.0	3.5	10.4	9.5
MovieChat [54]	CVPR 2024	8.4	6.4	8.0	14.0	15.7	11.8
TimeChat [47]	CVPR 2024	25.5	6.4	31.8	26.4	9.4	5.9
Human		59.0	70.9	89.8	-	-	-

V: Only visual input

L: Only text input (speech transcripts)

VL: Visual+text input

- Story-level QAs
- Publicly available videos
- Limited/No data leakage
- Questions with long temporal context
- Finding #1: Transcript-only performance of best LLMs is approaching human

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

Why another VideoQA dataset?

Method	Venue	% Accuracy Multiple-Choice QA V	L	VL	Open-Ended QA V	L	VL
Random		20.0	20.0	20.0	-	-	-
FrozenBiLM [5]	NeurIPS 2021	23.4	38.2	38.6	-	-	-
mPLUG-Owl2 [74]	CVPR 2024	38.3	20.7	21.3	22.1	1.8	1.6
Video-LLaVA [33]	arXiv 2023	34.2	21.3	24.7	19.2	6.4	8.0
LLoVi [79]	arXiv 2023	$\boxed{30.8}$	64.2	55.6	16.2	40.3	24.7
LangRepo [26]	arXiv 2024	29.0	32.1	31.0	3.5	10.4	9.5
MovieChat [54]	CVPR 2024	8.4	6.4	8.0	14.0	15.7	11.8
TimeChat [47]	CVPR 2024	25.5	6.4	31.8	26.4	9.4	5.9
Human		$\boxed{59.0}$	70.9	89.8	-	-	-

V: Only visual input
L: Only text input (speech transcripts)
VL: Visual+text input

- Story-level QAs
- Publicly available videos
- Limited/No data leakage
- Questions with long temporal context
- Finding #1: Transcript-only performance of best LLMs is approaching human
- Finding #2: Vision-only performance of best VLMs is **18% below human**

Short Film Dataset (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

Why another VideoQA dataset?

Method	Venue	% Accuracy Multiple-Choice QA V	L	VL	Open-Ended QA V	L	VL
Random		20.0	20.0	20.0	-	-	-
FrozenBiLM [5]	NeurIPS 2021	23.4	38.2	38.6	-	-	-
mPLUG-Owl2 [74]	CVPR 2024	38.3	20.7	21.3	22.1	1.8	1.6
Video-LLaVA [33]	arXiv 2023	34.2	21.3	24.7	19.2	6.4	8.0
LLoVi [79]	arXiv 2023	30.8	$\boxed{64.2}$	$\boxed{55.6}$	16.2	40.3	24.7
LangRepo [26]	arXiv 2024	29.0	32.1	31.0	3.5	10.4	9.5
MovieChat [54]	CVPR 2024	8.4	6.4	8.0	14.0	15.7	11.8
TimeChat [47]	CVPR 2024	25.5	6.4	31.8	26.4	9.4	5.9
Human		59.0	$\boxed{70.9}$	$\boxed{89.8}$	-	-	-

V: Only visual input
L: Only text input (speech transcripts)
VL: Visual+text input

- Story-level QAs
- Publicly available videos
- Limited/No data leakage
- Questions with long temporal context
- Finding #1: Transcript-only performance of best LLMs is approaching human
- Finding #2: Vision-only performance of best VLMs is **18% below human**

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

SONGKRAN

A coffee machine salesman falls for a boutique cafe owner on a business trip to Thailand.

- Hello, what can I get you?
- Do you speak English?
- Yes I do! How can I help you?

- So let me get this right, you don't drink coffee like at all?
- No, it makes me jittery, I don't like the taste.

- Don't worry!
- Should we go back?

[MUSIC]

What problem does Pete encounter on his way to the hotel?

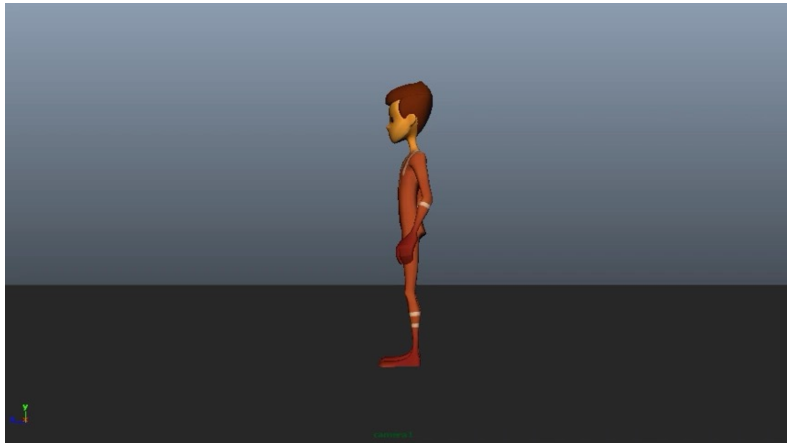
A) He loses his passport and must navigate Bangkok's bureaucracy to get a temporary one.
B) He is pickpocketed in a crowded market and loses his money and phone.
C) He gets stuck in Bangkok's traffic and decides to walk, getting lost in the process. ✔
D) He mistakenly takes the wrong bus and ends up in a distant part of the city.
E) He finds that his hotel reservation has been mistakenly cancelled.

- Story-level QAs
- Publicly available videos
- Limited/No data leakage
- Questions with long temporal context
- Finding #1: Transcript-only performance of best LLMs is approaching human
- Finding #2: Vision-only performance of best VLMs is **18% below human**

What's next?

Future work: Multimodal Self-supervision

- Inherit motion characteristics
  - Audio
  - Speech
  - Language
  - 3D dynamics
  - Temporal correlation of frames



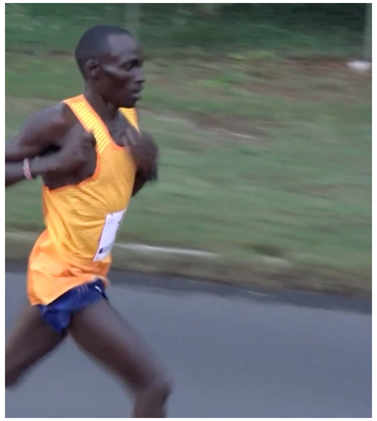
Running

High-five

Future work:
Open set with Large Language Models

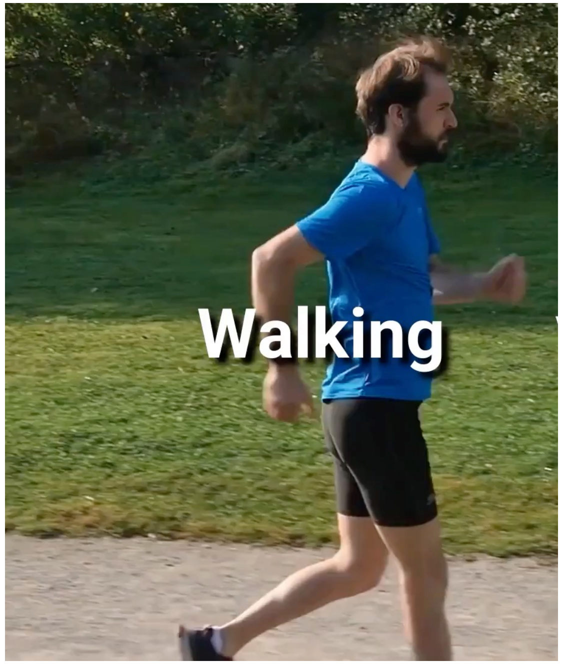
Sprinting

Running



Marathon

Walking



Long-term goal

- "`Once upon a time in a faraway land, there was a dragon called Zoe and she was different…`" started the mother's bedtime story
  - Video → visual illustration
  - Colors, sounds, characters → story
  - Characters and music → personalized

Story-level generation

Long-term story level understanding

- Multimodality
- Long-term reasoning

→

Text-to-video generation

- Dynamic storytelling techniques
- Text to visual content alignment

Future work: Towards General AI

Thank you