Can Generative AI and Video Analysis Redefine the Future Vision? Unveiling the Potential and Challenges

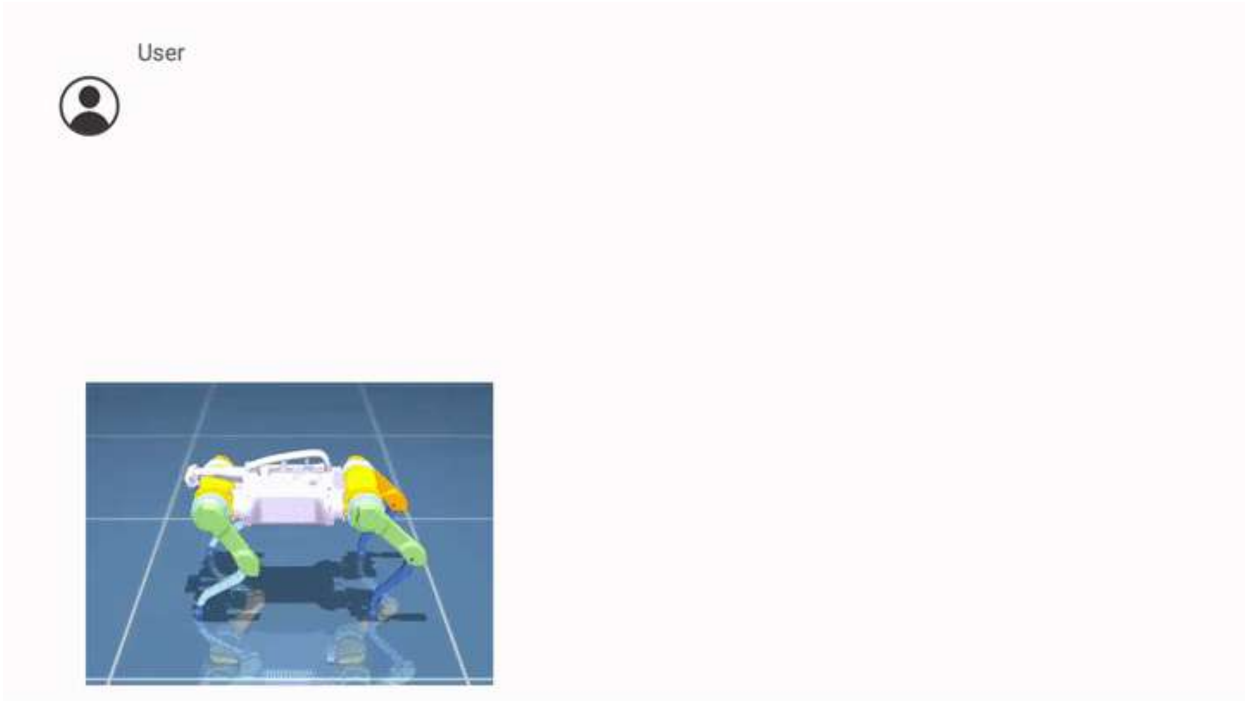
Nick Tai, NVIDIA AI Technology Center (NVAITC)

2023/12/11

- Introduction
- Video Action Anticipation
- Introduction to Diffusion Models
- Synthesis and Future Directions
- Conclusion and Discussion

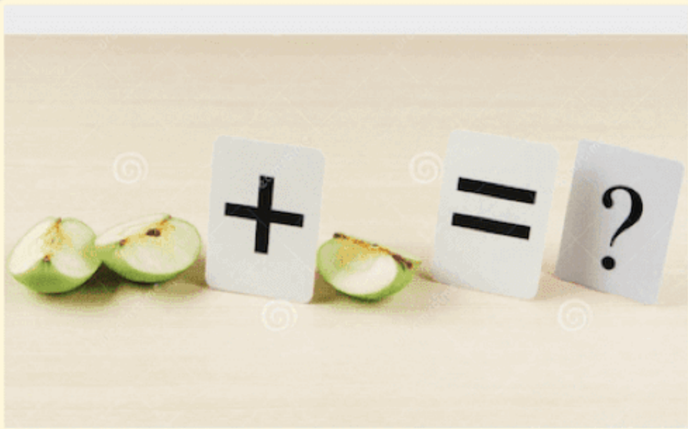
- **Introduction**
- Video Action Anticipation
- Introduction to Diffusion Models
- Synthesis and Future Directions
- Conclusion and Discussion

GEN-2 (Runway, 2023)



Language to Rewards for Robotic Skill Synthesis (Yu et al., 2023)

Visual Math Problem Solving



Q1: How to solve it?

CogVLM: The problem in the picture is a simple math problem, where a sliced green apple is equated to a question mark. Specifically, in the middle of the picture, there are three green apples placed side by side and three cardboard signs. The signs have black characters on them: a plus sign, an equal sign, and a question mark. The equation shown in the picture is: 2 + 1 = ?

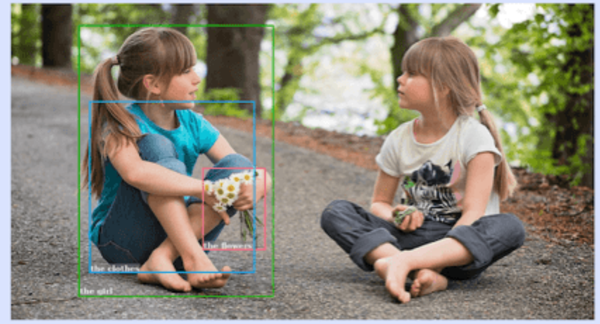
Q2: What's the final answer?

CogVLM: 2+1=3

Grounding Visual Question Answering

Q1: What color are the clothes of the girl whose hands are holding flowers? Let's think step by step.

CogVLM: Find the flowers [[328,519,436,780]]. Check the girl holding it, got the girl [[116,069,450,927]]. Check the clothes of it, got the clothes [[134,309,418,852]]. The clothes [[134,309,418,852]] are blue. So the answer is blue.

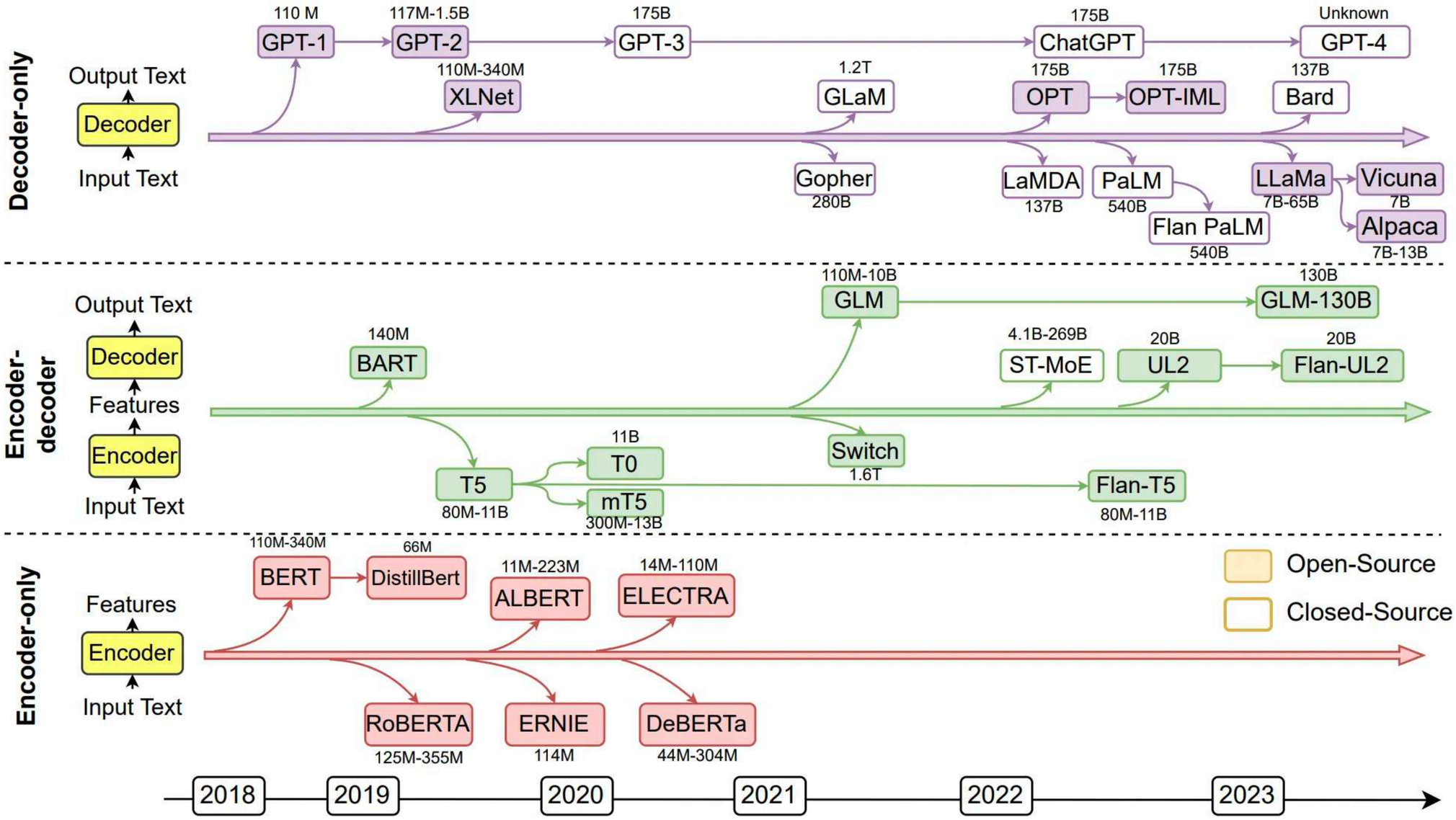


Q2: What color are the jeans of the girl whose hands are not holding flowers? Let's think step by step.

CogVLM: Find the not(flowers) hands [[654,637,710,728]]. Check the girl with it, got the girl [[526,109,880,946]]. Check the jeans of it, got the jeans [[530,613,880,871]]. The jeans [[530,613,880,871]] are gray. So the answer is gray.

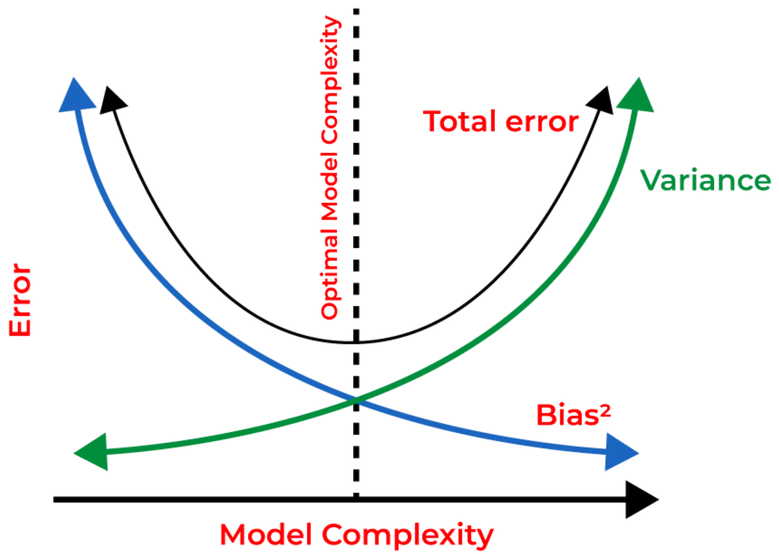
CogVLM (Wang et al., 2023)

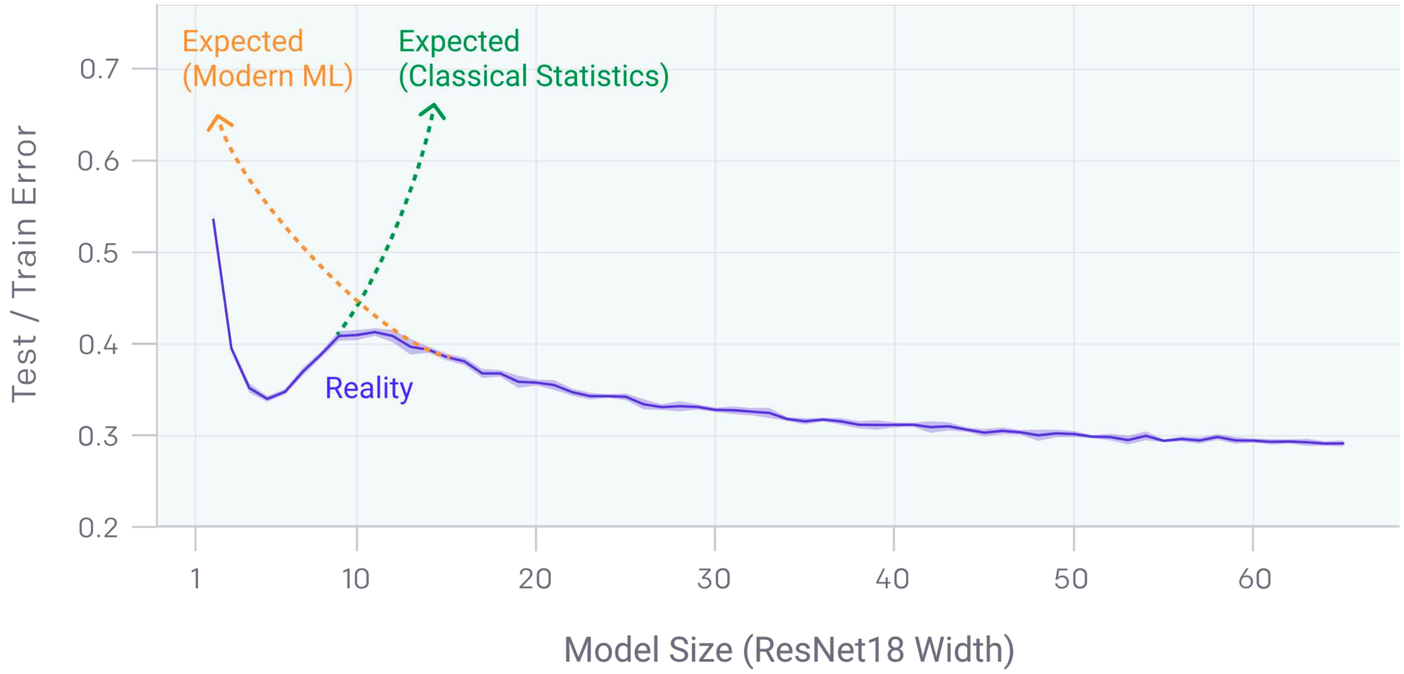
The Evolution of Generative AI



Why Large Models?

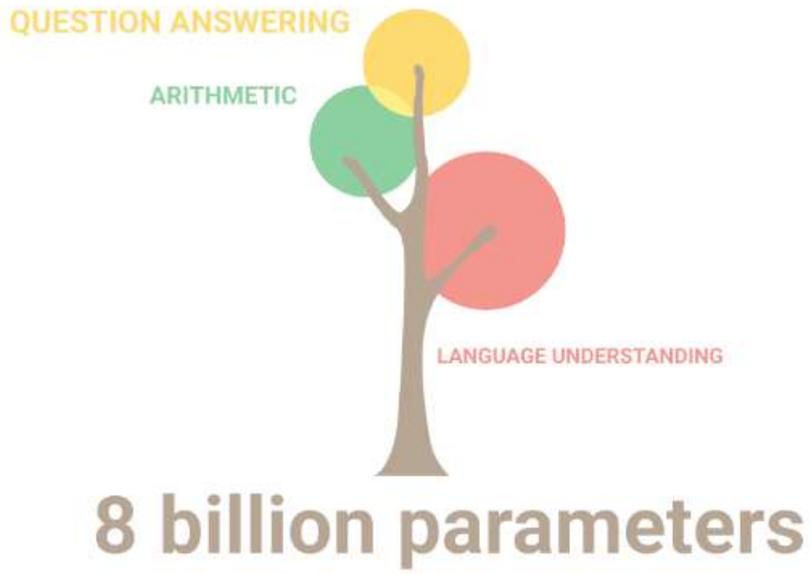
Double Decent Theory; (OpenAI 2019: https://openai.com/research/deep-double-descent)





Why Large Models?

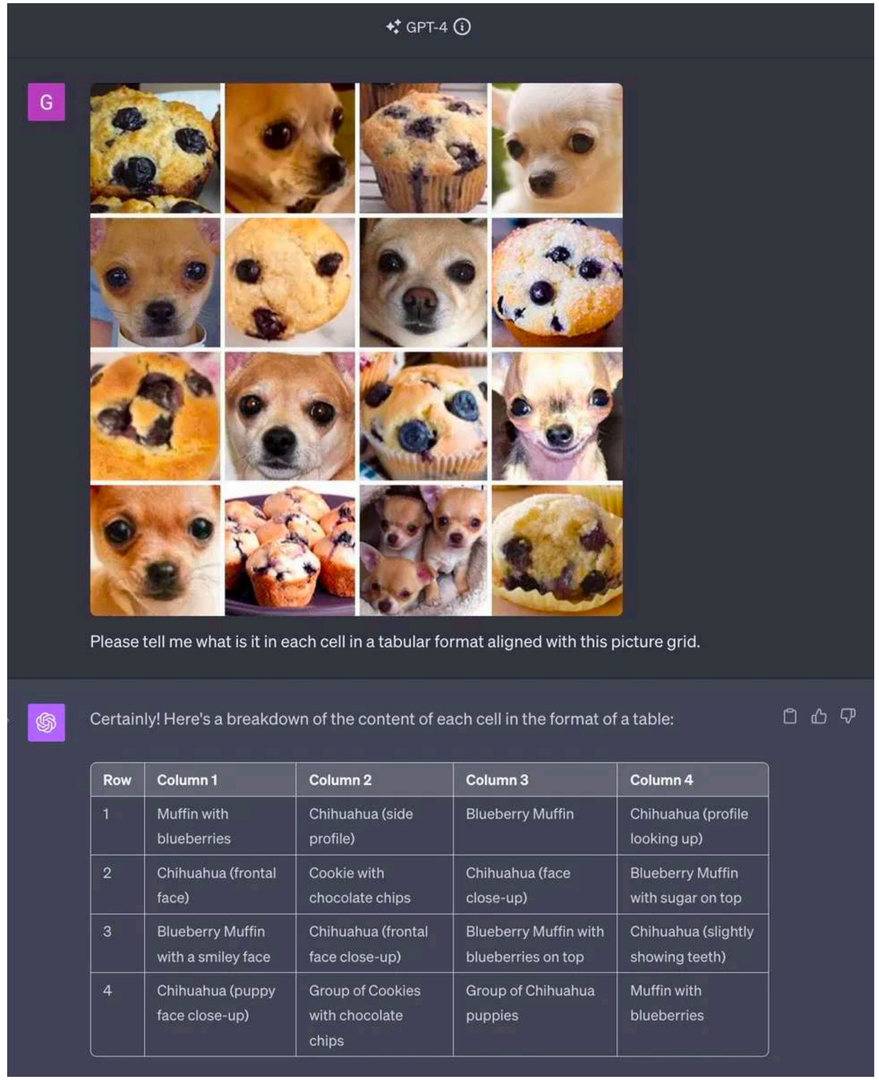
More weights, More strength



Vision-Language Model (VLM)

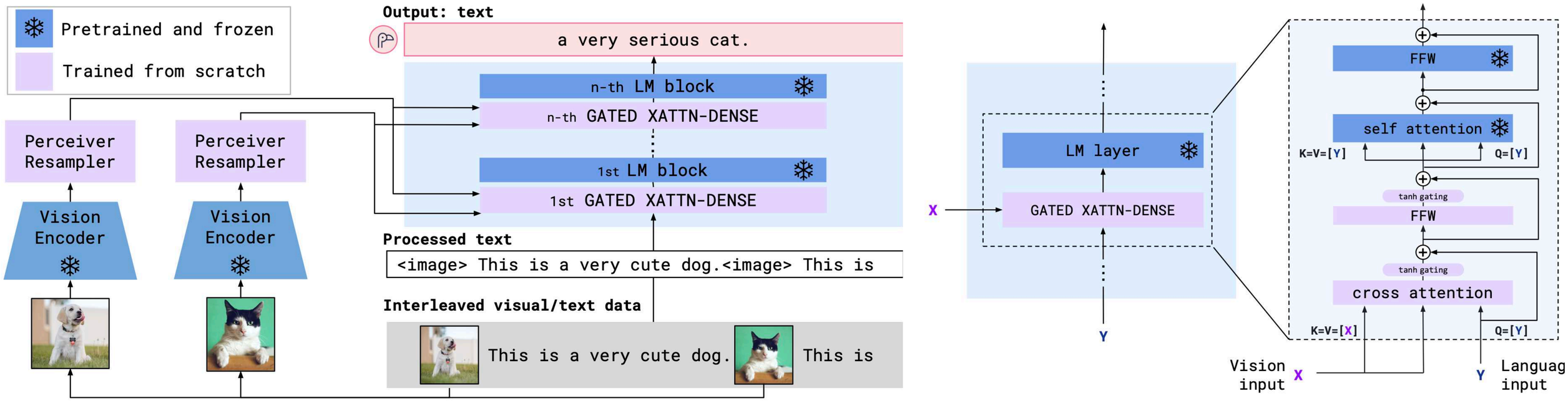
- Vision-Language Models (VLMs) are an emerging class of Generative AIs that aim to understand and generate content that combines both visual and textual information. They are designed to capture the intricate relationships between visual data, such as images or videos, and associated language descriptions or textual data.

- **Multimodal Learning**: VLMs are at the forefront of multimodal learning, where the goal is to build models that can process and relate information from more than one modality – in this case, vision (images, videos) and language (text).
- **Joint Embedding Space**: They typically work by mapping both visual and textual inputs into a common embedding space where the relationship between the two can be learned and understood.
- **Cross-Modal Understanding**: VLMs are trained to not only understand each modality independently but also to perform cross-modal tasks, such as image captioning, text-to-image generation, and visual question answering (VQA).



Vision-Language Model (VLM)

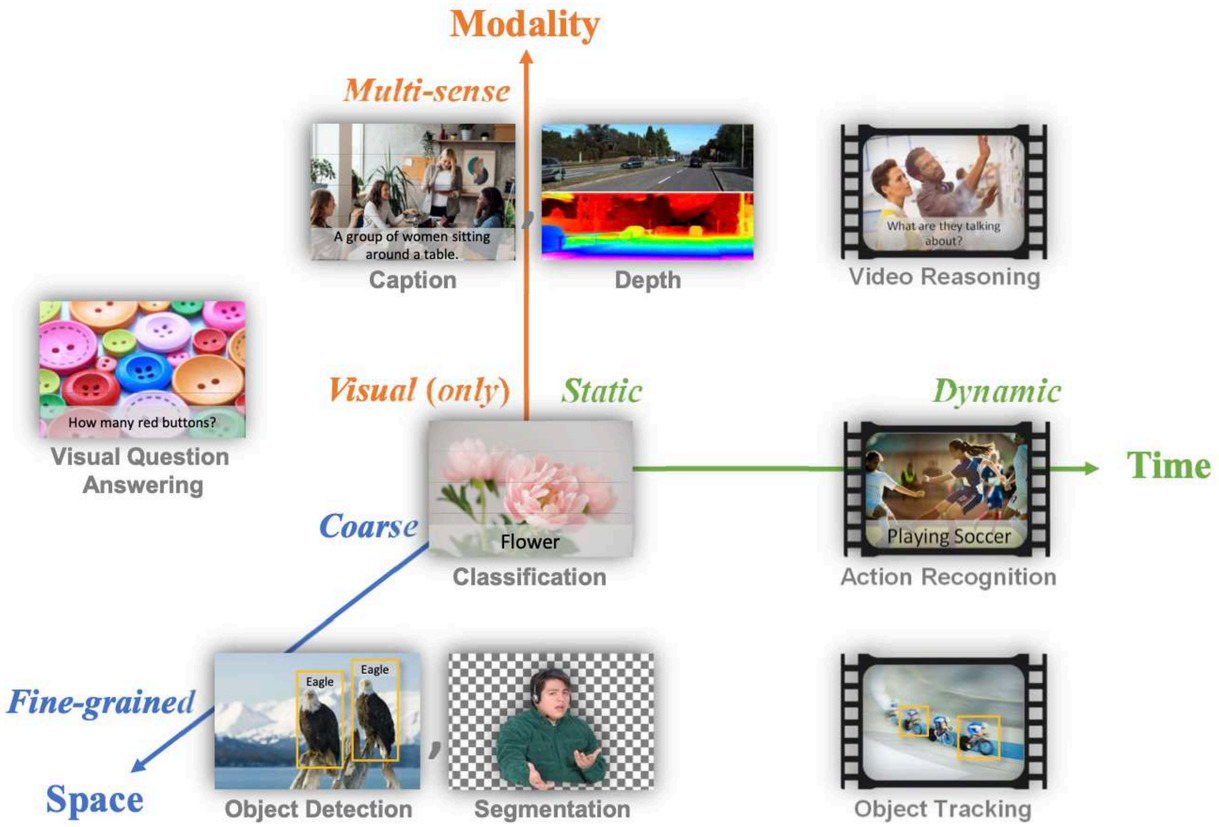
Flamingo: a Visual Language Model for Few-Shot Learning



Flamingo (Alayrac et al., 2022)

Multi-Modalities Vision-Language Model

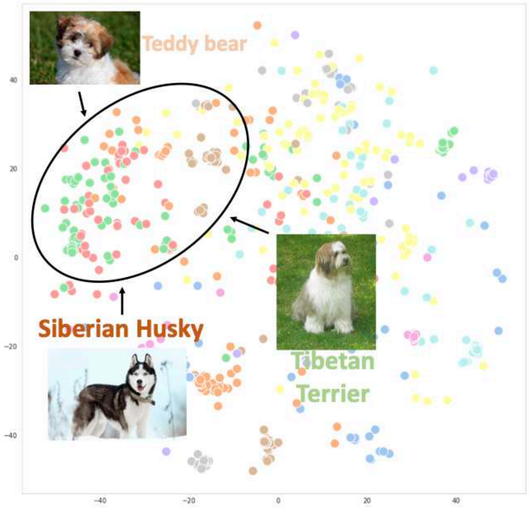
Florence: A New Foundation Model for Computer Vision (Lu et al., 2022)

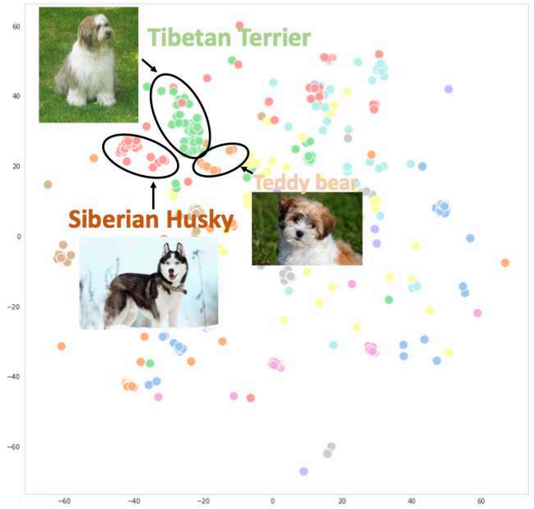


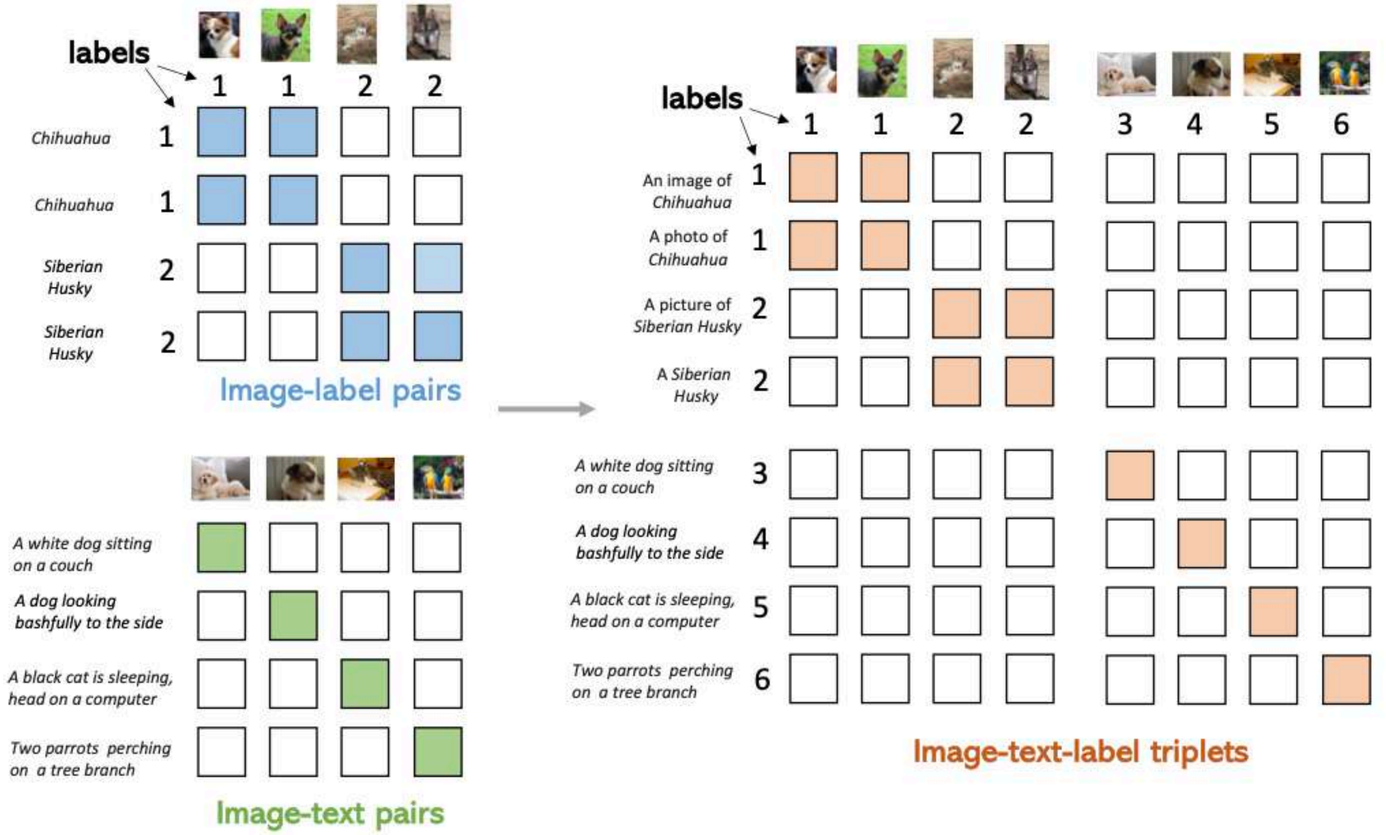
- STOA Multi-Model Vision-Language foundation model.
- Unified learning
- Transformer architectures and adaption
- ImageNet-1K zero-shot top-1 83.74%
- Release a dataset FLD-900M, containing 900M images and 900M texts.
- **Model size efficiency (only 647M)**
- Combine multiple previous works:
  - Model design: Dynamic DETR (Dai et al., 2021)
  - UniCL (Yang et al., 2022),
  - ZeRO optimizer, mix-precision training, etc.,

Multi-Modalities Vision-Language Model

Florence: A New Foundation Model for Computer Vision (Lu et al., 2022)







UniCL (Yang et al., 2022)

- Introduction
- **Video Action Anticipation**
- Introduction to Diffusion Models
- Synthesis and Future Directions
- Conclusion and Discussion

Challenge in Videos

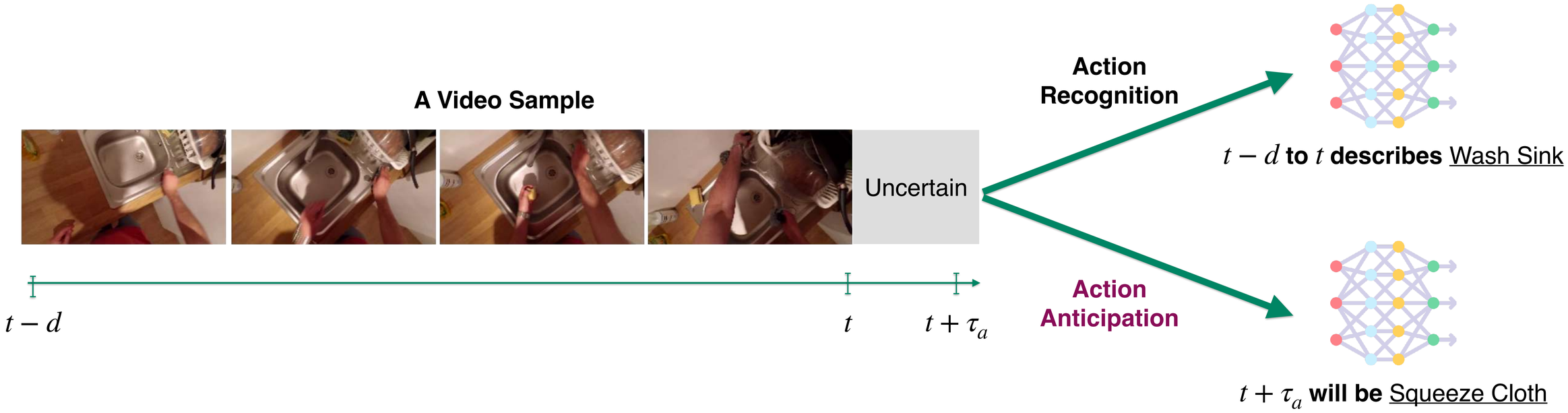
Curse of Dimensionality

- Video datasets present unique challenges for analysis and modeling, primarily due to their high-dimensional nature.
  - **High Dimensionality:** Managing large data volumes without losing crucial information.
  - **Temporal Dynamics:** Modeling frame sequences and changes over time.
  - **Computational Demands:** Requires more resources than text data processing.

- Comparing video data with text data:
  - **Dimensionality and Encoding**
    - **Text Data:** Encoded as vectors, typically ~20,000 elements for a word, capturing semantic properties.
    - **Video Data:** Far more complex; a 256 x 256 x 3 (RGB) frame over time T requires ~20,000 × T elements, as each frame is a high-resolution image.
  - **Discrete vs. Continuous Nature**
    - **Text Data:** Discrete in nature; words are distinct and ordered units.
    - **Video Data:** Continuous in both spatial and temporal aspects, where small changes can significantly alter content.
  - **Temporal Sensitivity and Sampling**
    - **Text Data:** Order is key, but lacks direct temporal encoding.
    - **Video Data:** Highly time-sensitive; frame rates and action durations within frames are critical for interpretation.

Video Action Anticipation

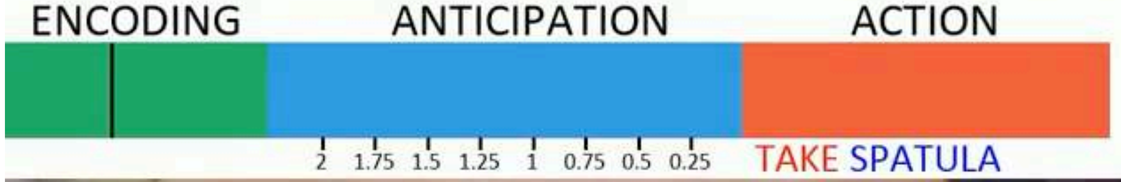
Action Recognition versus Action Anticipation



- The same video, when approached with different problem definitions (i.e., recognition vs. anticipation), can lead to varying objectives.
- This demonstrates that action anticipation requires a different solution than action recognition. Solutions directly adapted from action recognition are often sub-optimal.

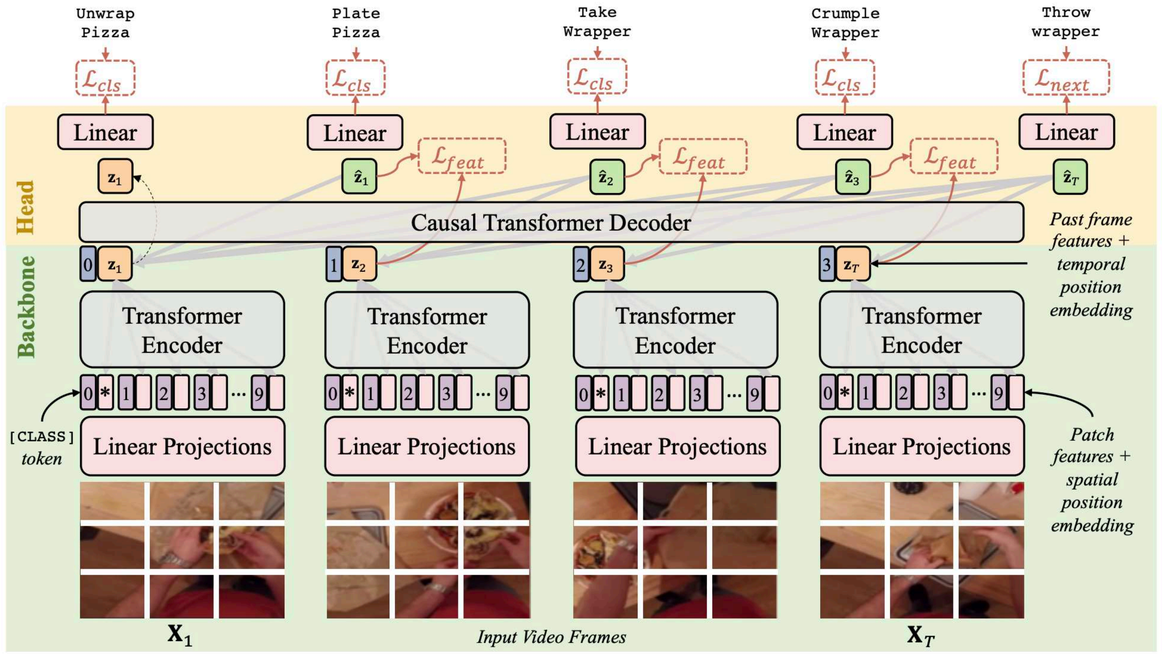
What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention

A. Furnari, G. M. Farinella *(egocentric action anticipation)*

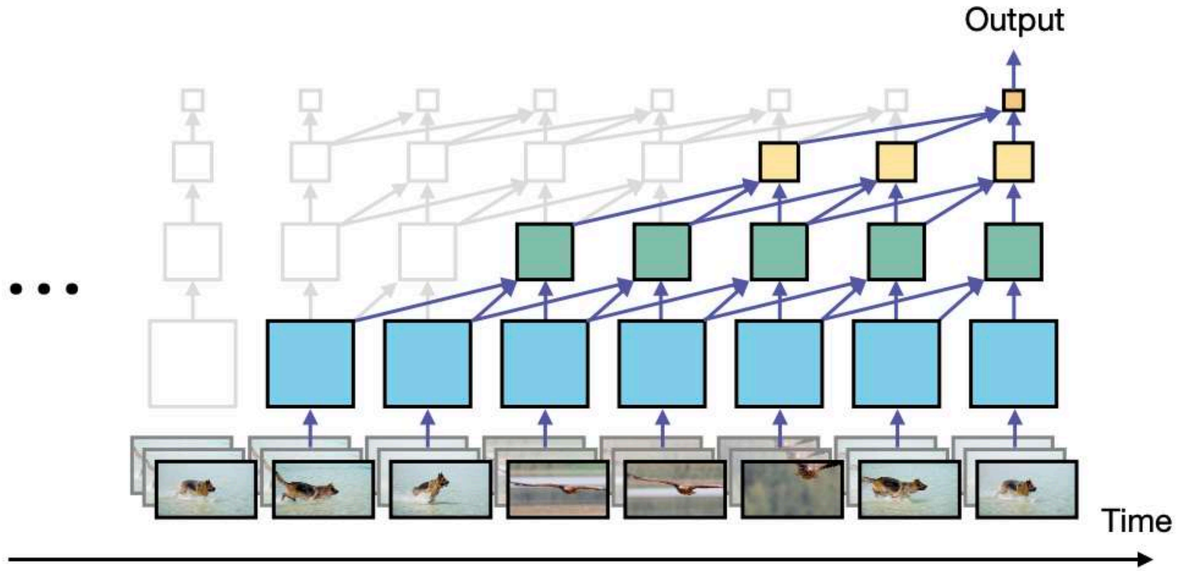


Encoding...

Baselines for Video Action Anticipation



AVT (Girdhar et al., 2021)

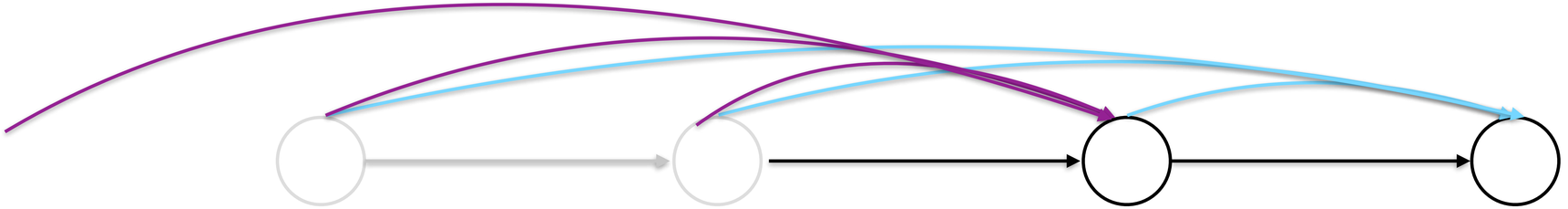


MeMViT (Wu et al., 2022)

- **Unlike the previous work, our general strategy for solving this involves modeling long-term dependencies and capturing subtle evidence that is conditional to the targets.**

Higher Order Recurrent Spatial-Temporal Transformer (HORST)

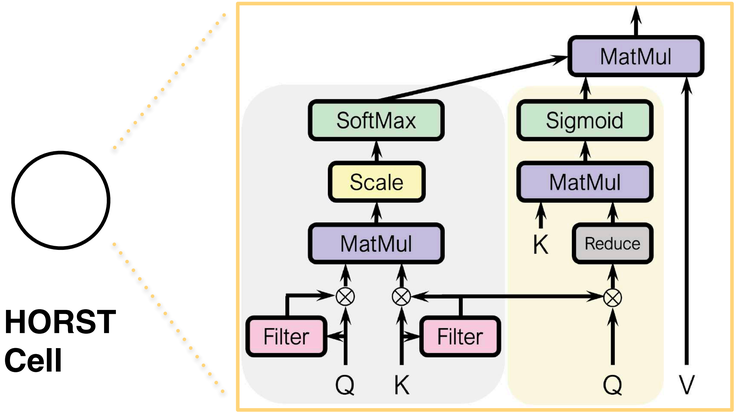
Tai, Tsung-Ming, et al. "Higher-Order Recurrent Network with Space-Time Attention for Video Early Action Recognition." *(ICIP 2022).*
Tai, Tsung-Ming, et al. "NVIDIA-UNIBZ Submission for EPIC-KITCHENS-100 Action Anticipation Challenge 2022." *arXiv preprint arXiv:2206.10869* (2022).



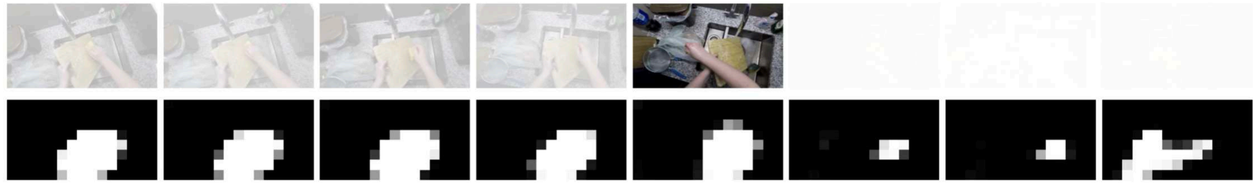
Higher-Order Recurrence

(last two states are demonstrated)

Spatial-Temporal Decomposition Attention



HORST Cell



Ground Truth: Wash Board; **Top-5 Predictions**: Wash Board, Put-Down Board, Turn-Off Tap, Put-Down Sponge, Stir Pan.

- As N-Gram modeling, HORST leverages higher-order recurrent design. Each timestep, the output is computed depending on multiple previous hidden states, instead of the last one.
- Spatial-Temporal attention learns to query the relevant information from higher-order queue with the decomposition attention of spatial and temporal branches

unibz NVIDIA

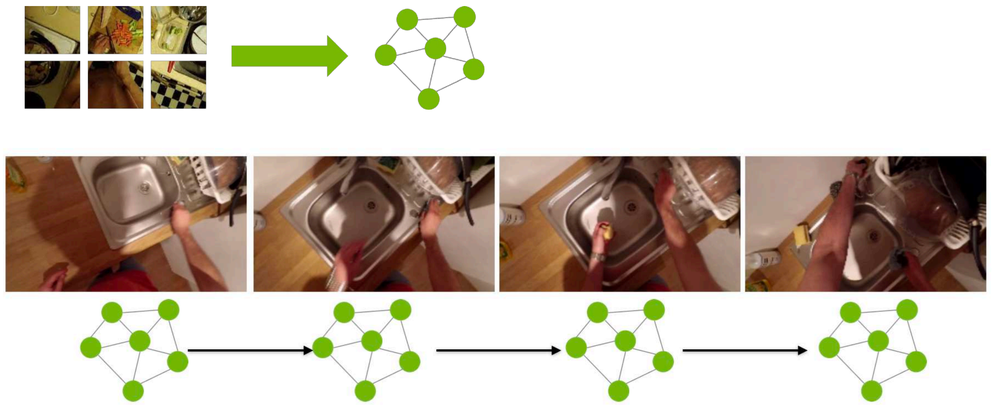
Unified Recurrent Modeling for Video Action Anticipation

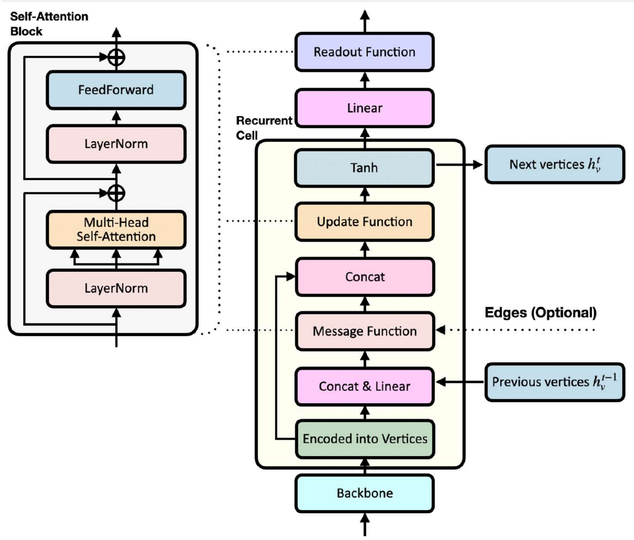
Tai, Tsung-Ming, et al. "Unified recurrence modeling for video action anticipation." *(ICPR 2022)*
Tai, Tsung-Ming, et al. "NVIDIA-UNIBZ Submission for EPIC-KITCHENS-100 Action Anticipation Challenge 2022." *arXiv preprint arXiv:2206.10869* (2022).

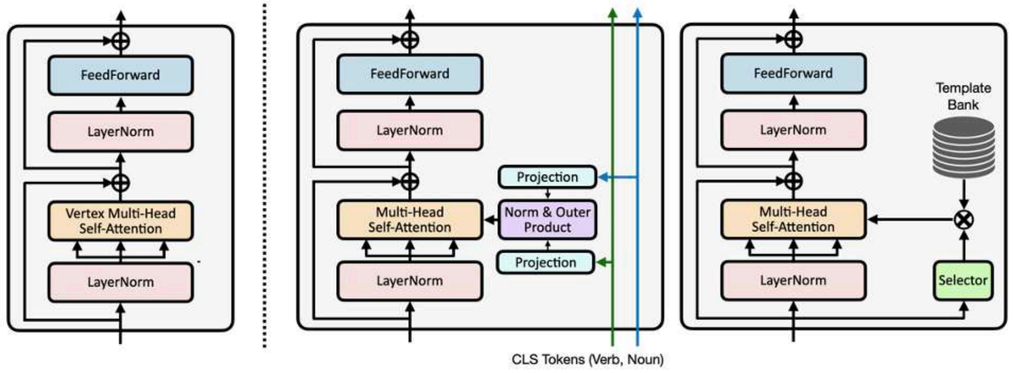
- We proposed MPNNEL, first encodes each frame onto a spatial graph representation, then leverages the message passing framework to learn the temporal propagation.
  - The message-passing framework is composed of **Message Function**, **Update Function**, and **Readout Function**.
  - We implement all of them by the multi-head self-attention.

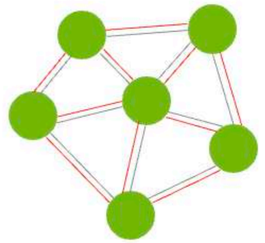
- The vertex of the spatial graph can be also augmented by our novel edge learning designs, which are
  - Class-Tokens Projection (CTP): which uses two learnable class tokens, supervised from verbs and nouns, and forms the estimated adjacency matrix to augment the edges.
  - Template Bank (TB): Select the learnable templates by the query and using them to augment the edges.

- Using the self-attention here can be treated as learning the information routing between vertices by attention weightings implicitly. Our edge learnings provide a flexible way to further augment the edges explicitly.



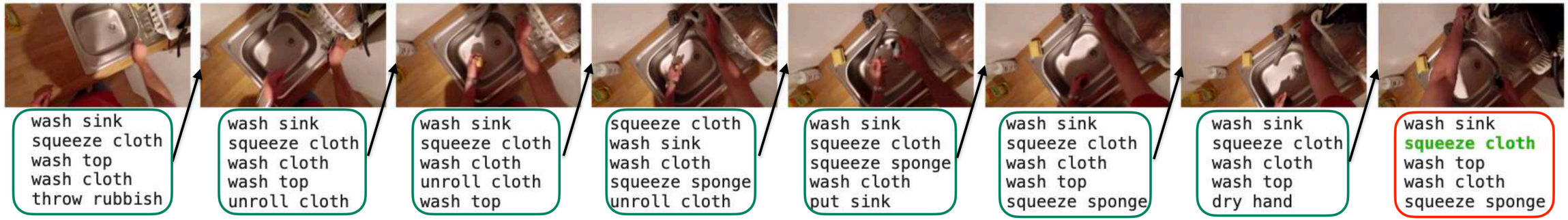




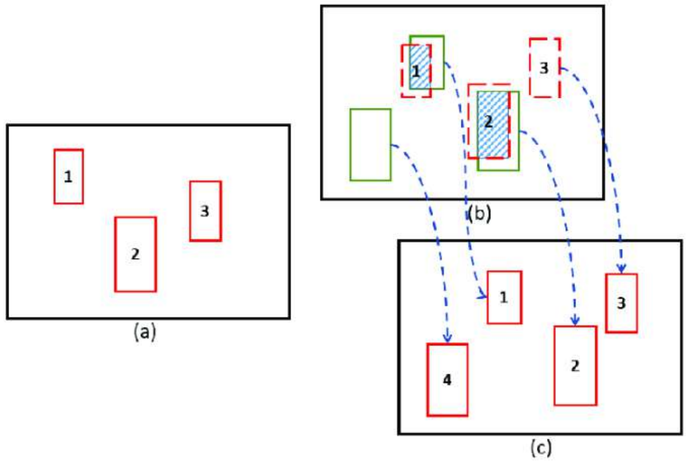


Inductive Attention for Video Action Anticipation

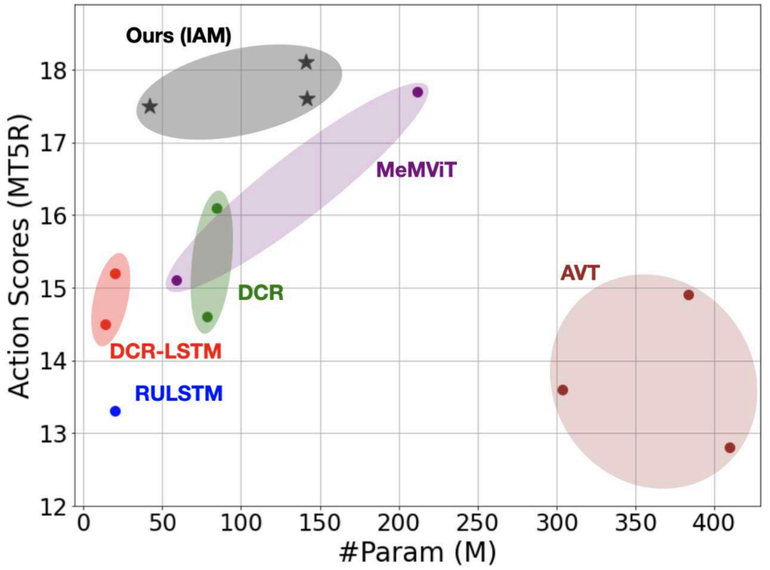
Tai, Tsung-Ming, et al. "Inductive Attention for Video Action Anticipation." *arXiv preprint arXiv:2212.08830* (2022).



- We further explore what should be the suitable "query" in attention to drive the higher-order recurrent model.
- In addition, we augment the higher-order recurrent model. Instead of modeling in the latent space, we explicitly utilize the previous predictions as reference (as attention "keys").
- From the experimental results, we found our design can lead to more efficient design in model size and generalize better on the unseen test set.



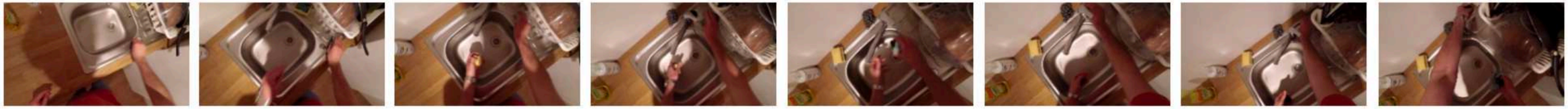
(For example, object tracking explicitly take the previous prediction as a reference)



Inductive Attention for Video Action Anticipation

(Under Review)

- The first model provides the transparent explainability of previous actions to the future prediction.
- Strong generalizability on the unseen tests.
- Evaluate on the large-scale EPIC-Kitchens dataset, which contains the 3806 daily kitchen activities.



Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive
squeeze cloth	wash cloth	wash cloth	squeeze cloth	wash cloth	squeeze cloth	squeeze cloth
wash sink	wash top	unroll cloth	unroll cloth	wash sink	wash plug	wash cloth
unroll cloth	unroll cloth	dry hand	wash plug	dry hand	wash cloth	hang hand
gather rubbish	dry hand	wash hand	pull cup	take cloth	pull cup	pull cup
put sink	wash plug	wash top	hang hand	put cloth	hang hand	squeeze sponge
Top-5 Negative	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**
apply bowl	let-go sponge	gather meat	sort broccoli	throw pear	roll omelette	roll lemon
gather broccoli	mix lettuce	gather broccoli	eat pork	put sheets	wrap box	pat omelette
insert bean:green	eat pork	eat broccoli	coat oil	pat omelette	eat broccoli	eat pork
check pie	check box	wash food	brush dough	unroll omelette	turn ring:onion	turn-on liquid:washing
close freezer	soak plate	insert pan:dust	eat broccoli	move omelette	pat noodle	roll omelette

wash sink
squeeze cloth
wash top
wash cloth
squeeze sponge

Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive
wash top	squeeze cloth	squeeze cloth	wash sink	squeeze cloth	wash sink	wash sink
wash cloth	unroll cloth	wash sink	wash cloth	squeeze sponge	dry hand	wash top
throw food	wash oven	wash floor	squeeze sponge	wash top	take cloth	dry hand
take cloth	shake hand	drink beer	wash top	wash oven	wash oven	wash oven
throw rubbish	gather rubbish	remove rubbish	wash sponge	put sponge	gather rubbish	wash hand
Top-5 Negative	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**
shake rubbish	peel squash	insert banana	shake peach	dry scissors	fill liquid	adjust plate
wash plug	insert nut	peel sausage	squeeze garlic	search salt	hang hand	dry sponge
shake straw	take nut	unroll omelette	carry bag	turn-on water	squeeze can	move funnel
dry cloth	cut apple	cut banana	wash straw	throw squash	shake liquid	empty container
hang hand	look bag	hang hand	squeeze caper	hang cup	shake courgette	filter plate

wash sink
squeeze cloth
wash top
wash cloth
squeeze sponge

Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive
wash top	wash top	wash top	wash sink	squeeze sponge	wash top	wash top
wash cloth	squeeze sponge	unroll cloth	unroll cloth	squeeze cloth	squeeze sponge	wash cloth
squeeze cloth	wash cloth	wash oven	pour liquid:washing	wash sponge	wash cloth	squeeze sponge
squeeze sponge	wash sponge	dry hand	wash hand	turn-off tap	dry hand	take cloth
take cloth	take cloth	squeeze sponge	dry hand	take cloth	take cloth	dry hand
Top-5 Negative	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**
shake straw	shake straw	shake straw	throw squash	shake straw	shake straw	hang hand
shake rubbish	shake rubbish	brush sink	smell candle	throw squash	shake rubbish	move coffee
throw kale	shake caper	adjust straw	let-go sponge	shake rubbish	brush sink	shake straw
hang hand	wash plug	shake caper	soak plate	wash straw	move filter	squeeze caper
move stand	pat napkin	take alarm	shake caper	drop sponge	move coffee	shake rubbish

wash sink
squeeze cloth
wash top
wash cloth
squeeze sponge

(GT: take paper)
Rank 28 of 3806

Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive
apply olive	take pizza	take pizza	put pizza	put pizza	put pizza	apply olive
put mushroom	apply olive	apply olive	flatten dough	take pizza	apply olive	cut pizza
put potato	apply basil	apply basil	apply olive	apply basil	open pizza	remove pizza
throw rubbish	take basil	check pizza	fold bag	apply olive	put fruit	put pin:rolling
throw skin	open oven	take basil	put potato	take basil	close bin	put food
Top-5 Negative	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**
look bottle	check board:chopping	look sauce	turn-on dishwasher	eat bread	pat spatula	look can
sort fork	search book	remove jar	sort fork	press coriander	insert ring:onion	take backpack
open kale	rip wrap:plastic	look can	check drawer	take candle	move omelette	take pear
put parsley	insert ring:onion	insert holder	scrub scissors	open corn	check pie	insert cap
insert oven	insert cloth	lift dishwasher	look box	look box	put pie	pull box

put pizza
take pizza
insert pizza
cut pizza
remove pizza

Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive
put pizza	put pizza	insert pizza	flatten dough	insert pizza	take pizza	put pizza
take pizza	insert pizza	remove pizza	cut pizza	flatten dough	insert pizza	take pizza
insert pizza	check pizza	cut pizza	apply dough	cut pizza	cut pizza	insert pizza
close bin	open pizza	flatten dough	remove pizza	apply olive	put sausage	drink beer
open pizza	wrap plate	open oven	put fruit	put fruit	remove pizza	open pizza
Top-5 Negative	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**
cut corn	cut corn	move pen	rip wrap:plastic	dry food	remove floor	cut corn
pat garlic	press onion	put pear	divide pizza	look bottle	throw oil	sprinkle tomato
press onion	remove stalk	take toaster	unroll clothes	sort broccoli	throw butter	close oregano
check ginger	apply onion	eat broccoli	move fridge	pour bean:green	take pear	put pear
check onion	sort onion	remove cucumber	dry food	brush dough	close paper	scrape skin

put pizza
take pizza
insert pizza
cut pizza
remove pizza

Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive	Top-5 Positive
apply olive	take pizza	put pizza	flatten dough	remove pizza	put pizza	apply olive
cut pizza	insert pizza	insert pizza	insert pizza	insert pizza	open pizza	take pizza
remove pizza	check pizza	check pizza	take pizza	take pizza	close bin	cut pizza
apply basil	open pizza	cut pizza	fold bag	cut pizza	drink beer	put fruit
take aubergine	open oven	open pizza	divide dough	apply basil	remove salad	put food
Top-5 Negative	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**	**Top-5 Negative**
squeeze garlic	cut corn	cut corn	search plate	cut corn	cut corn	check board:chopping
pat burger	press onion	press onion	lift foil	attach tray	close pepper	unroll clothes
squeeze sausage	apply onion	apply onion	cut pork	take toaster	put shell:egg	move fridge
mix sausage	sort onion	sort onion	sort pizza	rub sauce	insert shell:egg	scrape salmon
sort onion	squeeze garlic	remove stalk	hold tray	shake tomato	press onion	insert oven

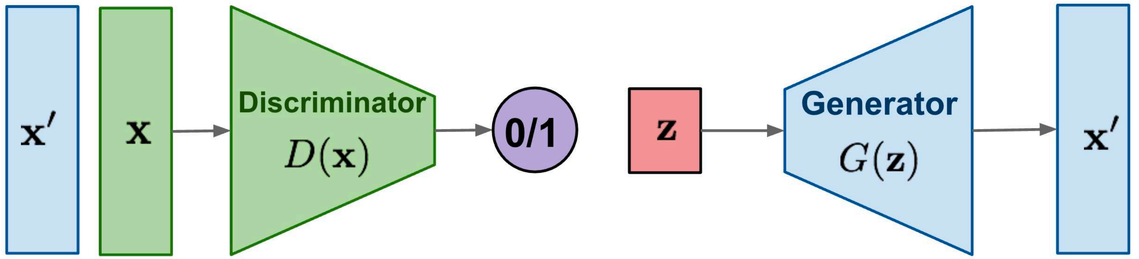
put pizza
take pizza
insert pizza
cut pizza
remove pizza

- Introduction
- Video Action Anticipation
- **Introduction to Diffusion Models**
- Synthesis and Future Directions
- Conclusion and Discussion

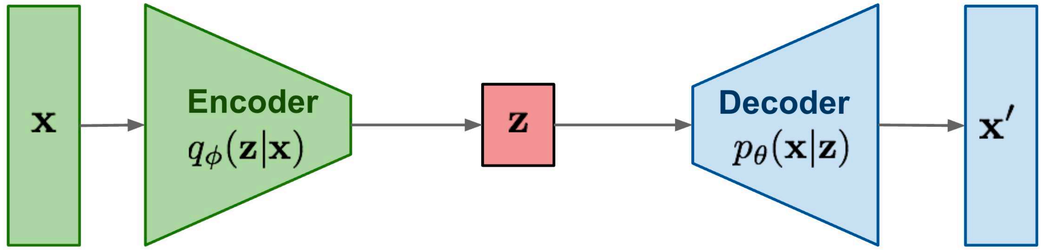
Generative Models

Overview

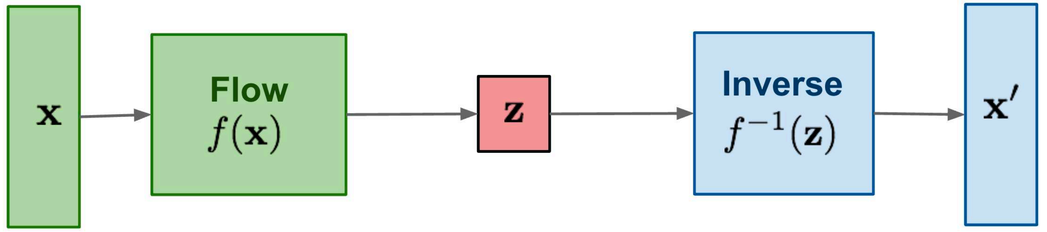
GAN: Adversarial training



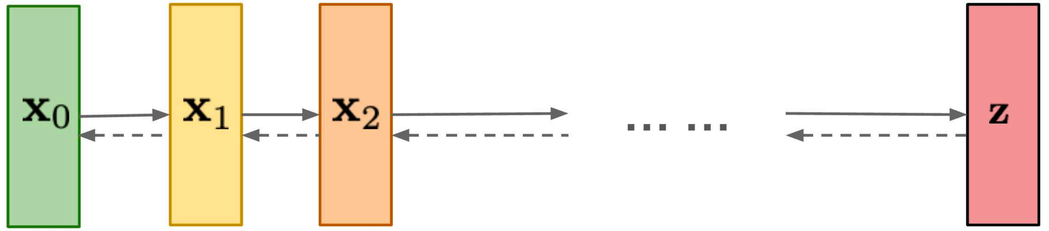
VAE: maximize variational lower bound



Flow-based models: Invertible transform of distributions

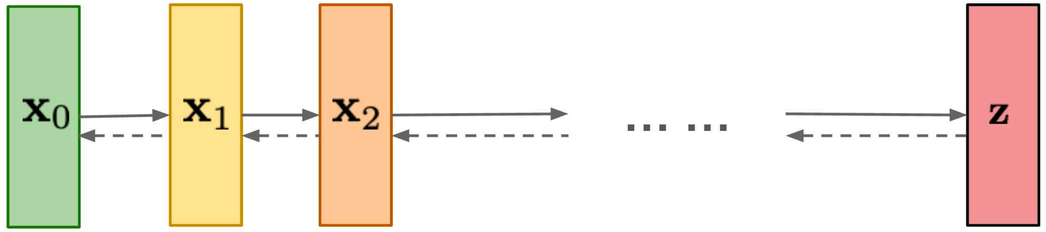


Diffusion models: Gradually add Gaussian noise and then reverse



Markov Chain and Langevin Dynamics

Diffusion models:
Gradually add Gaussian noise and then reverse



Markov chain assumption: $P(x_{t+1} \mid x_{0:t}) = P(x_{t+1} \mid x_t)$, which simplifies the way to model the event in continuous time.

- The diffusion model is based on Langevin (stochastic) Dynamics:
  - $dx_t = \nabla \log p(x_t)dt + \sqrt{2}dB_t$
    - Where B_t is the Brownian Motion with zero mean.
- Utilize the Langevin Dynamics to infer the next small step $t + dt$:
  - $x_{t+dt} = x_t + \nabla \log p(x)dt + \sqrt{2dt} \cdot N(0,1)$ // this $\sqrt{2dt}$ ensures the following Gaussian noise sampling follows the Brown Motion.
  - $x_{t+dt} = x_t + \nabla \log p(x)dt + N(0,2dt)$

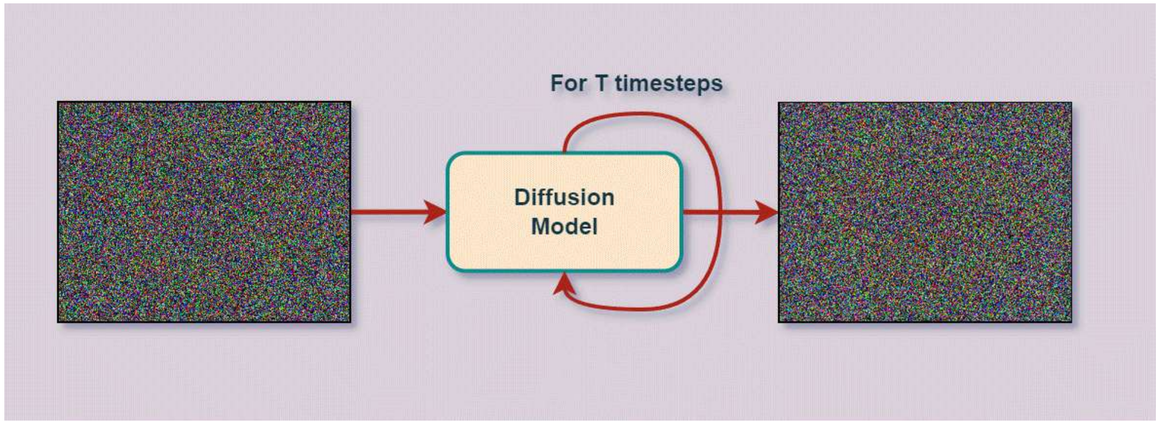
- Here, we can see two different perspectives to model this equation:
  - **Continuous Representation:** solving the SDE (and its reverse version) and using a neural network to approximate $\nabla \log p(x)$
  - **Discrete Representation:** solving the Markov chain as each timestep models the small change dt, which is $x_{t+1} = P(x_{t+1} \mid x_t)$

Today's focus!

Denoising Diffusion Probabilistic Models

DDPM (Ho et al., 2020)

- DDPM is a generative model, modeling as a Markov chain, composed of **Diffusion** and **Denoising** phases.
  - **Diffusion** fused the images with noise.
  - **Denoising** remove the noise in the image.



- The Markov chain processes for T steps, where T controls the noise strengths in diffusion/denoising phase; T should be sufficiently large (in other words, noise step should be sufficiently small).

- For each step in T,
  - **Diffusion** utilizes a predefined noises in mixtures.
  - **Denoising** employs a parameterized neural network θ to restore to restore the noisy input.

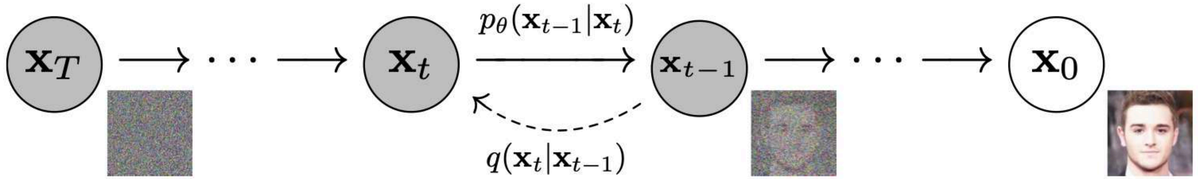
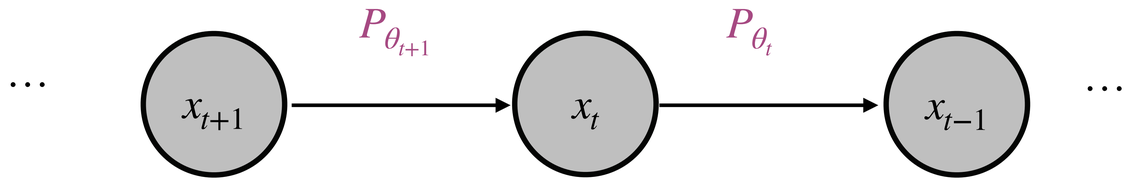


Figure 2: The directed graphical model considered in this work.

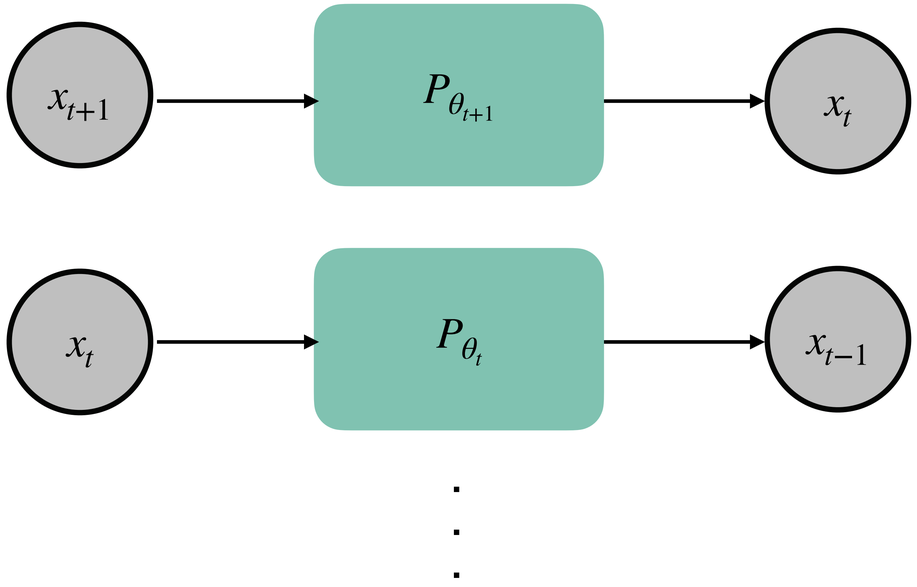
Denoising Diffusion Probabilistic Models

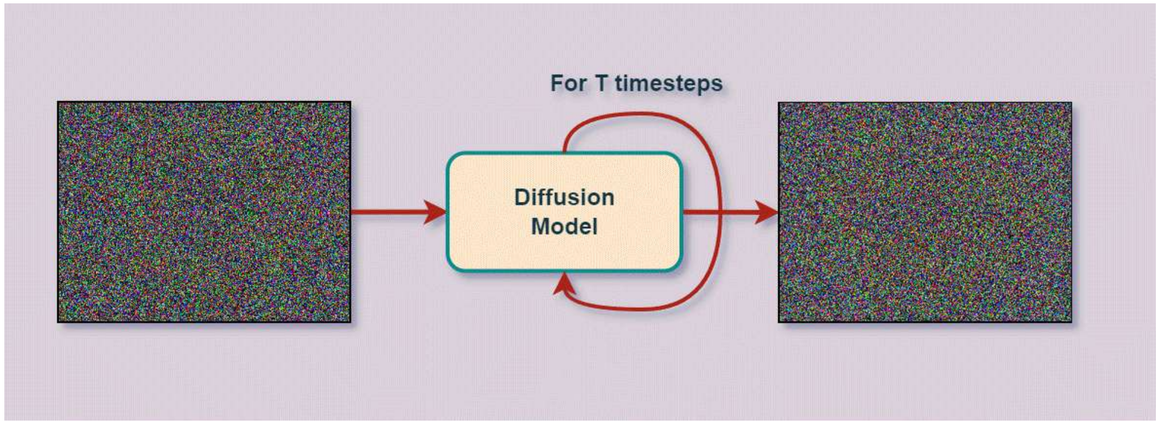
DDPM (Ho et al., 2020)



Note the noise mixtures differently in each diffusion step,
As a result, $P_{\theta_{t+1}}$ needs to be different than P_{θ_t}

It is unrealistic to define a total T models for DDPM 😳





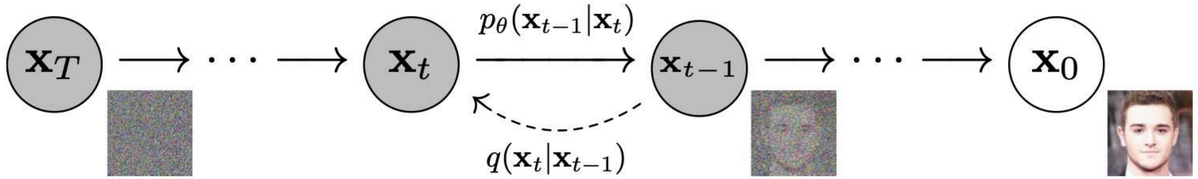
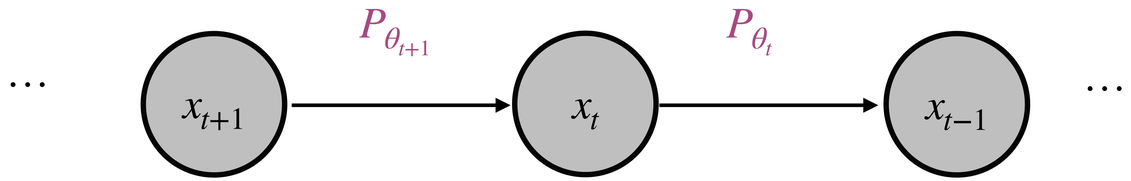
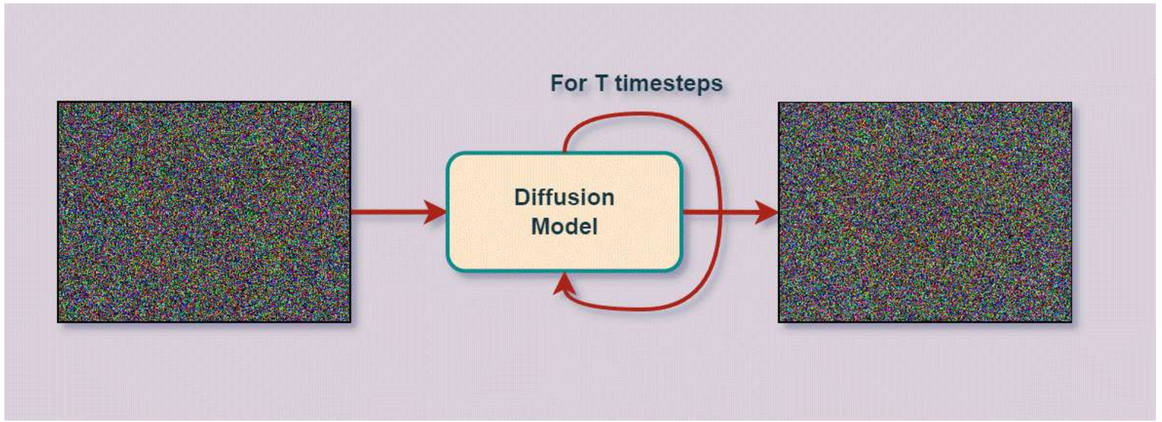


Figure 2: The directed graphical model considered in this work.

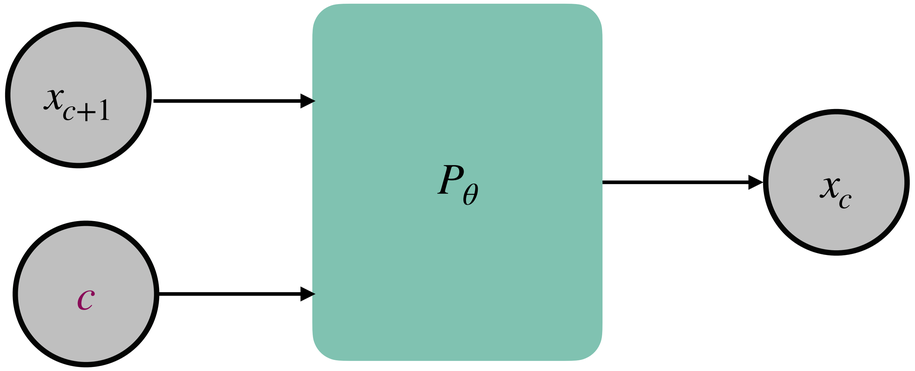
Denoising Diffusion Probabilistic Models

DDPM (Ho et al., 2020)





We parameterize the timestep as an auxiliary input (condition) to marginalize the model 😊



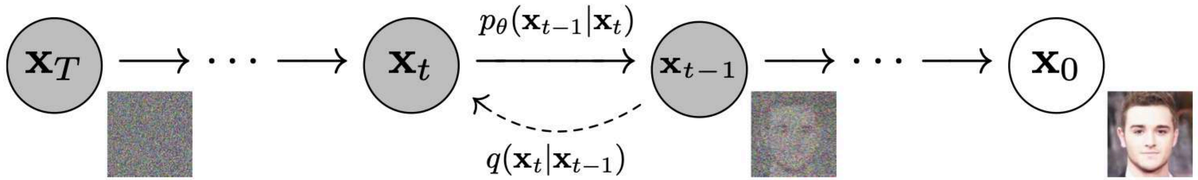


Figure 2: The directed graphical model considered in this work.

Denoising Diffusion Probabilistic Models -- Diffusion Phase

DDPM (Ho et al., 2020)

- The diffusion phase $q(x_{1:T}|x_0) := \Pi_{t=1}^{T} q(x_t|x_{t-1})$,

  where $q(x_t|x_{t-1}) := N(\sqrt{\frac{a_t}{a_{t-1}}}x_{t-1}, (1-\frac{a_t}{a_{t-1}})I)$.

  For each diffusion step, the noise mixtures in different strengths rather than the same gaussian noise.

  The scheduler is predefined by a decreasing sequence $a_{1:T} = [0,1)^T$. A popular choice is cosine decay (Nichol, Alexander Quinn, and Prafulla Dhariwal. 2021.)

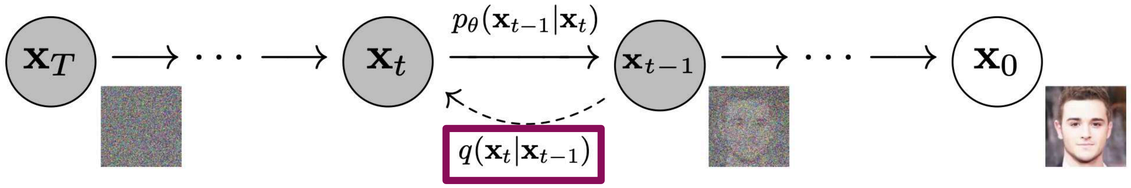


Figure 3. Latent samples from linear (top) and cosine (bottom) schedules respectively at linearly spaced values of t from 0 to T. The latents in the last quarter of the linear schedule are almost purely noise, whereas the cosine schedule adds noise more slowly

- Since the noise distribution is predefined and known, we can directly approximate $q(x_t|x_0) = N(\sqrt{\bar{a}_t}x_0, \sigma^2 I)$, where $\bar{a}_t = \Pi_{s=1}^{t} a_s$.

  - Which is, $x_t = \sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon, \epsilon \sim N(0,1)$

We can see when $t = T$, if and only if $a_T \to 0$, the x_T converges to $N(0,1)$ for all x_0

Denoising Diffusion Probabilistic Models -- Diffusion Phase

DDPM (Ho et al., 2020)

- The diffusion phase $q(x_{1:T}|x_0) := \Pi_{t=1}^{T} q(x_t|x_{t-1})$,

  where $q(x_t|x_{t-1}) := N(\sqrt{\frac{a_t}{a_{t-1}}}x_{t-1}, (1-\frac{a_t}{a_{t-1}})I)$.

  For each diffusion step, the noise mixtures in different strengths rather than the same gaussian noise.

  The scheduler is predefined by a decreasing sequence $a_{1:T} = (0,1]^T$. A popular choice is cosine decay (Nichol, Alexander Quinn, and Prafulla Dhariwal. 2021.)

- Since the noise distribution is predefined and known, we can directly approximate $q(x_t|x_0) = N(\sqrt{\bar{a}_t}x_0, \sigma^2 I)$, where $\bar{a}_t = \Pi_{s=1}^{t} a_s$.

  - Which is, $x_t = \sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon, \epsilon \sim N(0,1)$

We can see when $t = T$, if and only if $a_T \to 0$, the x_T converges to $N(0,1)$ for all x_0

Present the expected ratio of x_t to x_{t-1} is $\sqrt{\frac{a_t}{a_{t-1}}}$, and the variance, σ^2, is $(1-\frac{a_t}{a_{t-1}})$.

$$
\begin{aligned}
x_t &= \sqrt{a_t}x_{t-1} + \sqrt{1-a_t}\epsilon \\
&= \sqrt{a_t a_{t-1}}x_{t-2} + \sqrt{(1-a_t) + a_t(1-a_{t-1})}\epsilon \\
&\cdots \\
&= \sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon
\end{aligned}
$$

where $\bar{a}_t = \Pi_{s=1}^{t} a_s$

Given that,

$A = N(0,\sigma_a^2 I)$

$B = N(0,\sigma_b^2 I)$

$Z = A + B$

$= N(0,(\sigma_a^2 + \sigma_b^2)I)$

And,

$N(0,\sigma^2 I) = \sigma N(0,I)$

Denoising Diffusion Probabilistic Models -- Denoising Phase

DDPM (Ho et al., 2020)

- Reverse process iteratively inference the $p_\theta(x_{t-1}\,|\,x_t)$ for $t = T, \ldots, 1$, parameterized by trainable weights θ.

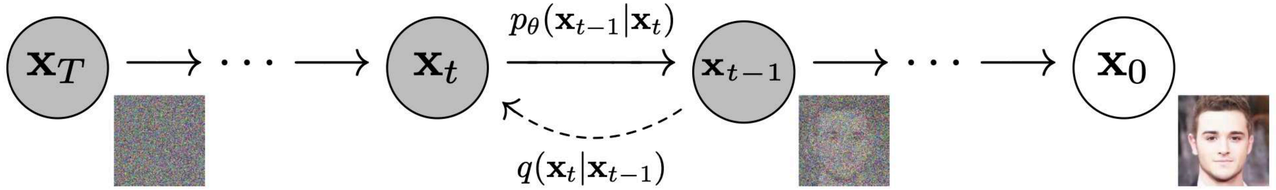
- In DDPM, the reverse chain is defined by $x_{t-1} = \frac{1}{\sqrt{a_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{a}_t}}\epsilon)$

  - Since noise is intractable by only the noisy observation x_t, a neural network $f_\theta(x_t, t)$ is deployed to predict the noise ϵ with the noisy observation x_t input.

- Alternatively, observing $q(x_t\,|\,x_0) = N(\sqrt{\bar{a}_t}x_0, \sigma^2 I)$

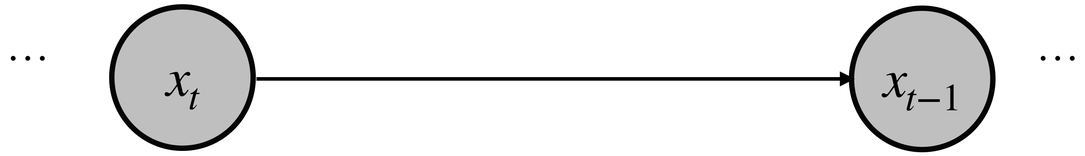
  - We can directly predict $\hat{x}_0 = f'_\theta(x_t, t)$, and obtain x_t by $x_t = q(x_{t-1}\,|\,f'_\theta(\hat{x}_0\,|\,x_t, t) = N(\sqrt{\bar{a}_t}\hat{x}_0, \sigma^2 I)$

  - where $\sigma^2 = \frac{1-\bar{a}_{t-1}}{1-\bar{a}_t}(1-a_t)$ (Sec 3.2 in DDPM paper)



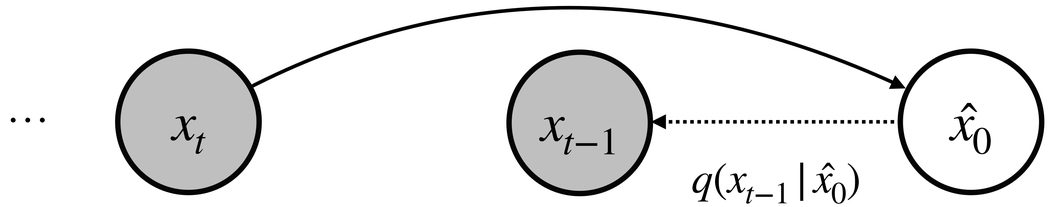
1.

$$x_{t-1} = \frac{1}{\sqrt{a_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{a}_t}} f_\theta(x_t, t))$$



2.

$$\hat{x}_0 = f_\theta(x_t, t)$$



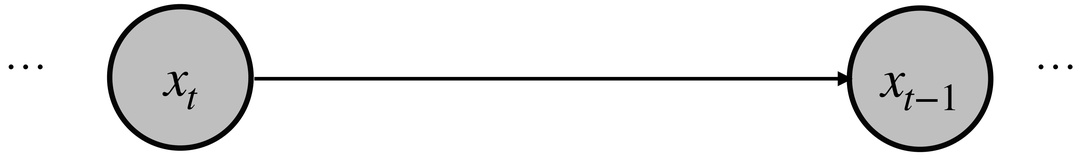
Denoising Diffusion Probabilistic Models -- Train Stage

DDPM (Ho et al., 2020)

- Method 1, $f_\theta(x_t, t)$ is the neural network aims to predict the gaussian noises ϵ by the noisy observation x_t,
  - Since we can access the ϵ during training, we can supervised f_θ by minimizing $||f_\theta(\sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon, t) - \epsilon||_2^2$
- Method 2, $f_\theta(x_t, t)$ directly predicts the x_0, the supervised target, in this case, is straightforward, minimizing $||f'_\theta(x_t, t) - x_0||_2^2$

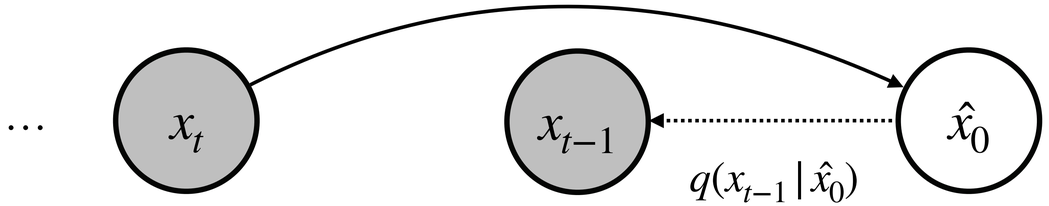
1.

$$x_{t-1} = \frac{1}{\sqrt{a_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{a}_t}} f_\theta(x_t, t))$$



2.

$$\hat{x}_0 = f_\theta(x_t, t)$$



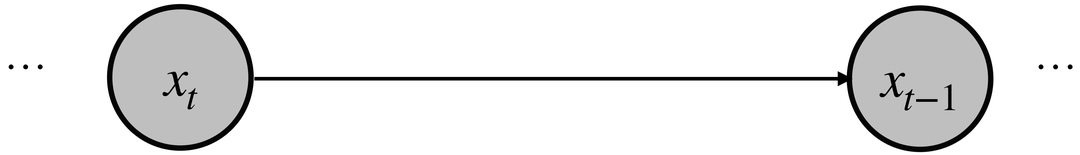
Denoising Diffusion Probabilistic Models -- Train Stage

DDPM (Ho et al., 2020)

- Method 1, $f_\theta(x_t, t)$ is the neural network aims to predict the gaussian noises ϵ by the noisy observation x_t,
  - Since we can access the ϵ during training, we can supervised f_θ by minimizing $||f_\theta(\sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon, t) - \epsilon||_2^2$
- Method 2, $f_\theta(x_t, t)$ directly predicts the x_0, the supervised target, in this case, is straightforward, minimizing $||f'_\theta(x_t, t) - x_0||_2^2$
  - x_0 (original image) usually is more stable in numeric computation than ϵ (noise). **<u>Method 2 could result in more stable training.</u>**

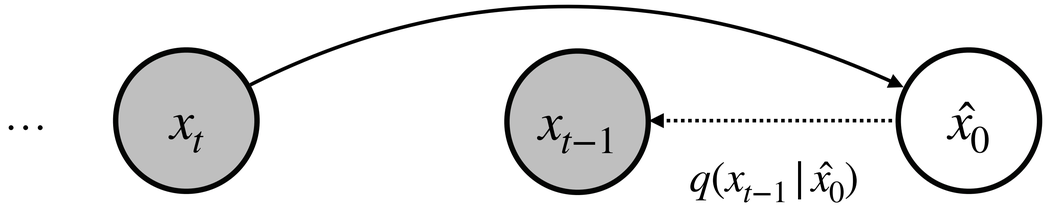
1.

$$x_{t-1} = \frac{1}{\sqrt{a_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{a}_t}} f_\theta(x_t, t))$$



2.

$$\hat{x}_0 = f_\theta(x_t, t)$$



High-Level Interpretation of DDPM

"Denoising" Perspective

- From an image with pure noises, DDPM gradually removes the noise bit by bit in the image and reveals the underneath content.

- In this perspective, the neural network θ serves as a Gaussian denoiser.

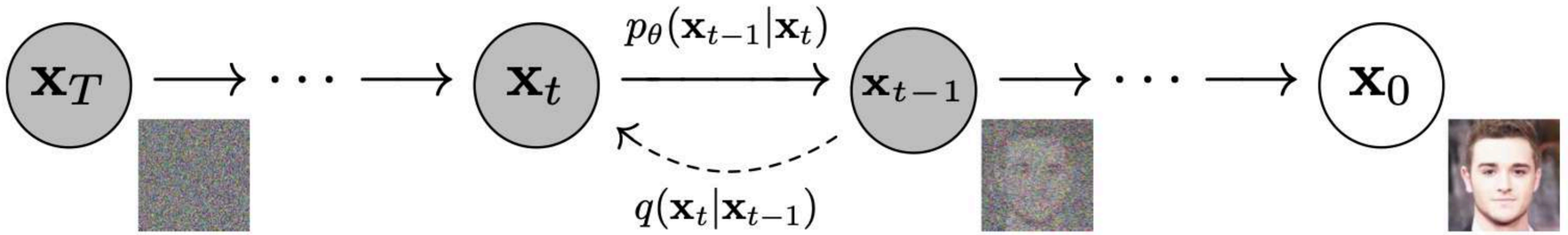


Figure 2: The directed graphical model considered in this work.

High-Level Interpretation of DDPM

"Entropy" Perspective

- Starting from an uncertain state, DDPM gradually excludes the least possible candidate until the only possible state remains.

- In this perspective, the neural network θ learns to reduce the entropy of the corresponding input.
  - (ref: Shannon entropy of a random variable x is defined as $E[-logp(x)]$)

High entropy state
(possible to be anything)

Low entropy state
(as it is)

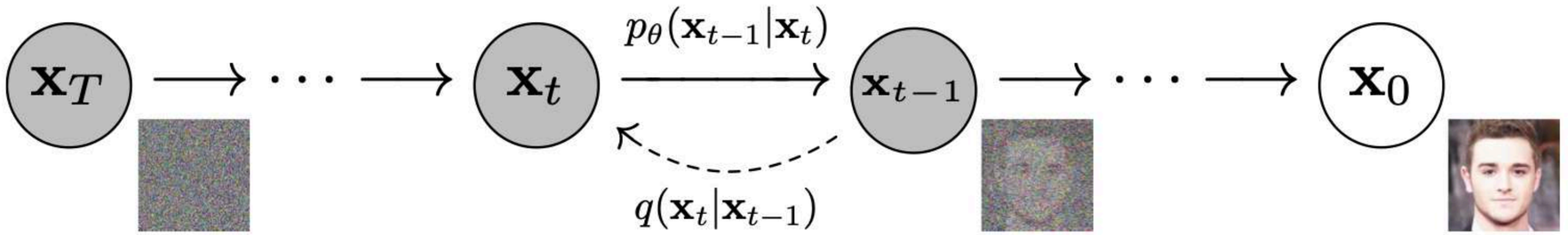
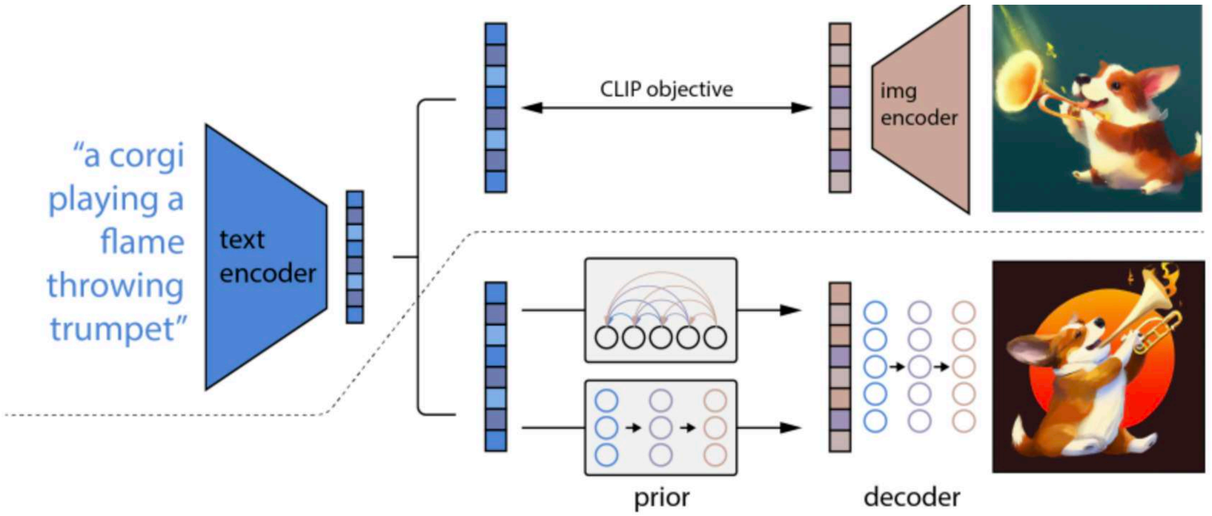
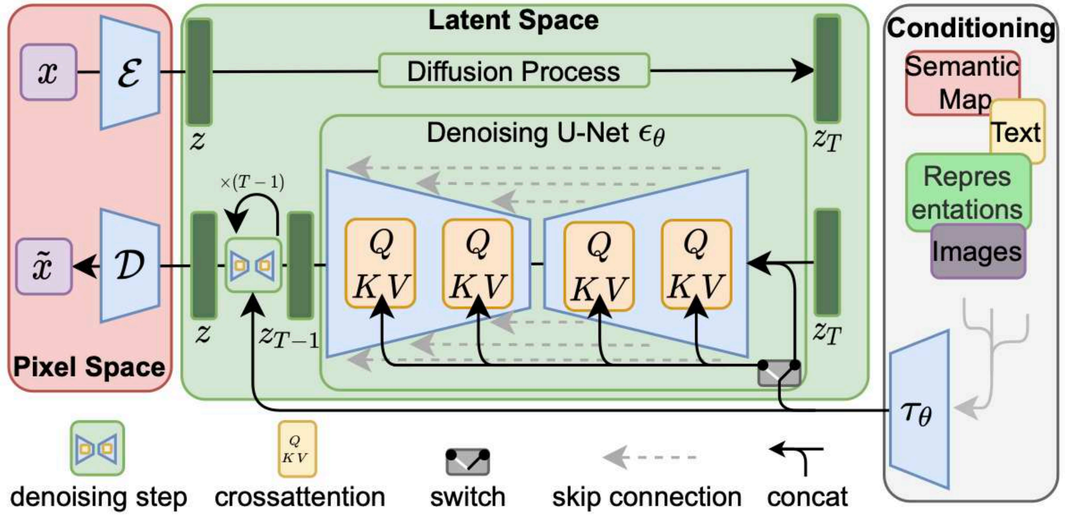


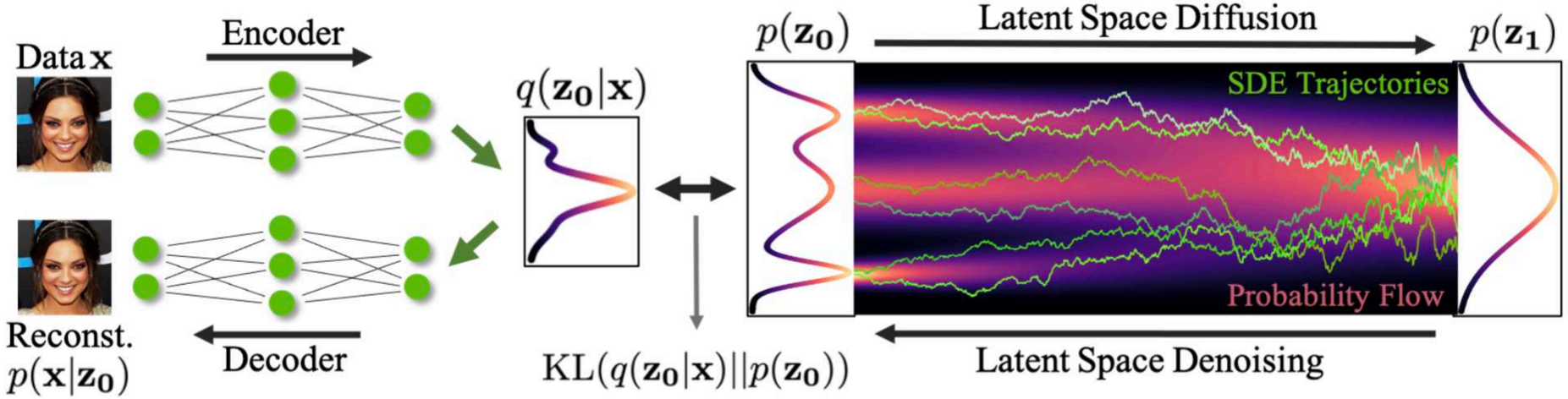
Figure 2: The directed graphical model considered in this work.



DALL-E 2



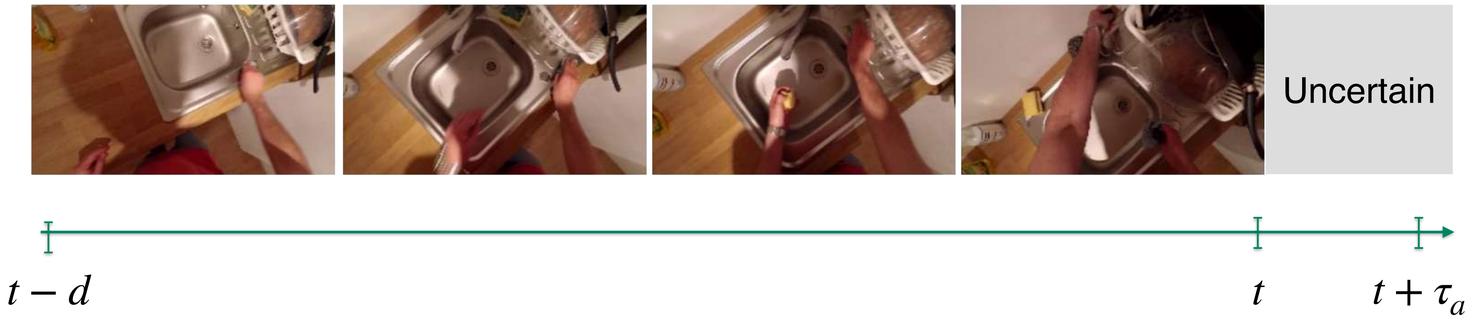
Stable-Diffusion



LSGM: Score-based Generative Modeling in Latent Space

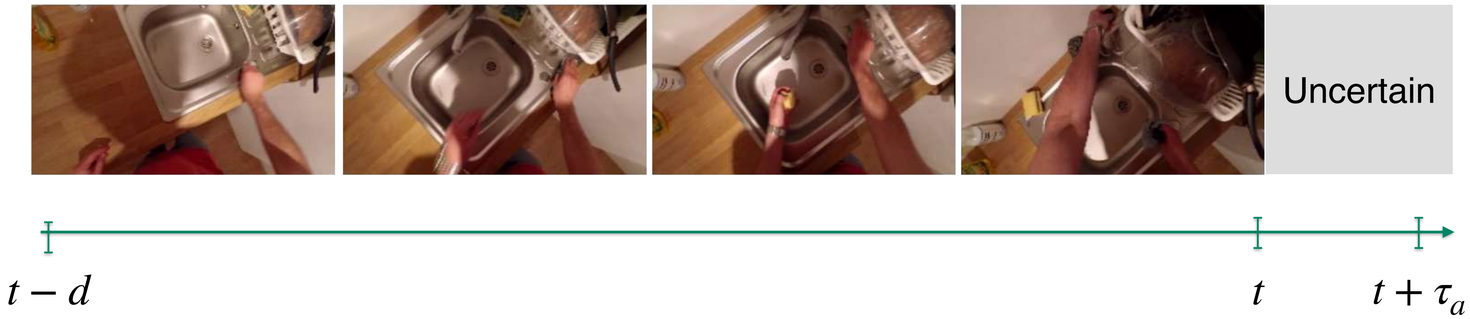
- Introduction
- Video Action Anticipation
- Introduction to Diffusion Models
- **Synthesis and Future Directions**
- Conclusion and Discussion

Integrating Video Action Anticipation with Diffusion Models



- On Frame
- On Action

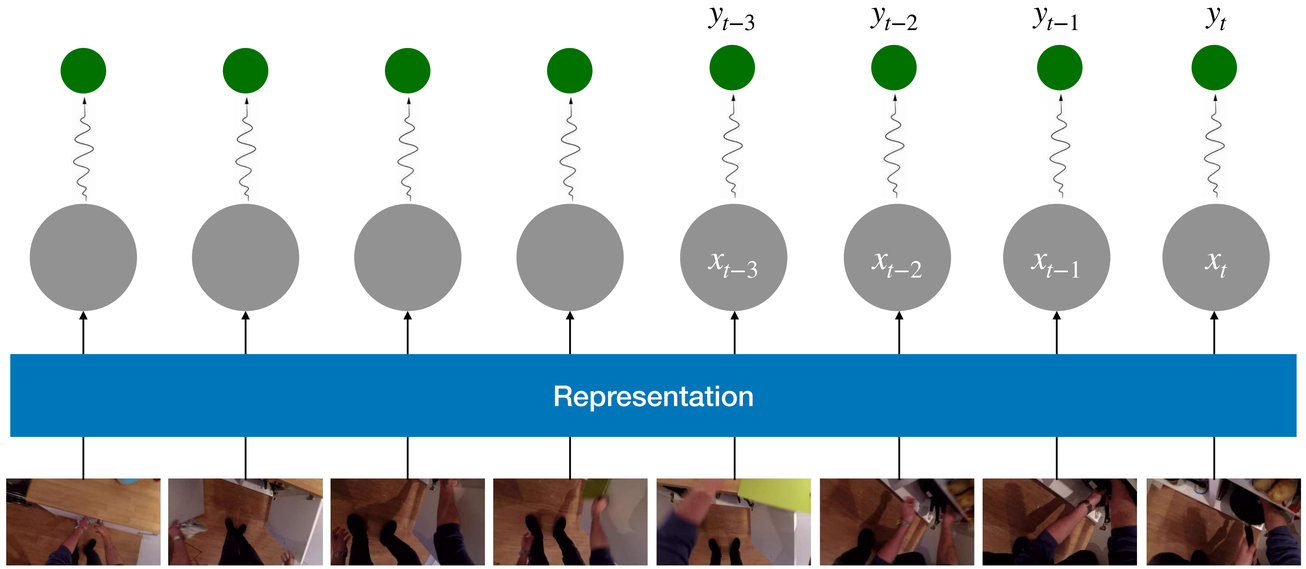
Integrating Video Action Anticipation with Diffusion Models



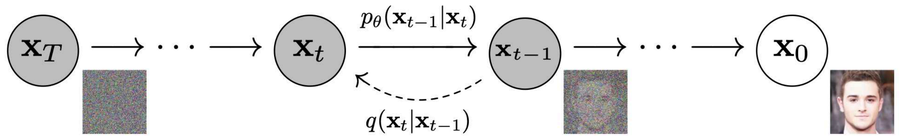
- On Frame
- **On Action**

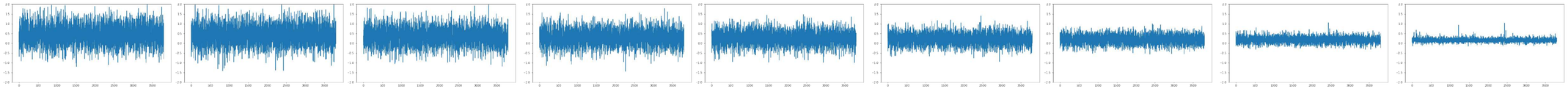
- Consider a differential equations:
  - $\frac{\partial x}{\partial t} = f(x)$
  - $x(t_0) = x_0$

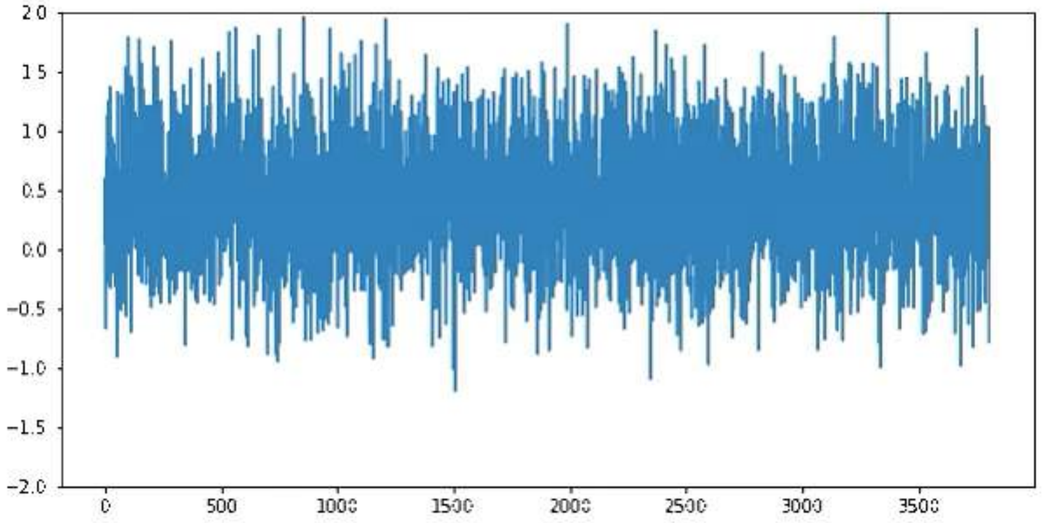
- The Picard-Lindelof theorem states that at least locally, if f is Lipschitz, there exists a unique solution $x(t)$. Which implies:
  - $x(t + \partial t) = x(t) + \partial t \cdot f(x(t))$.



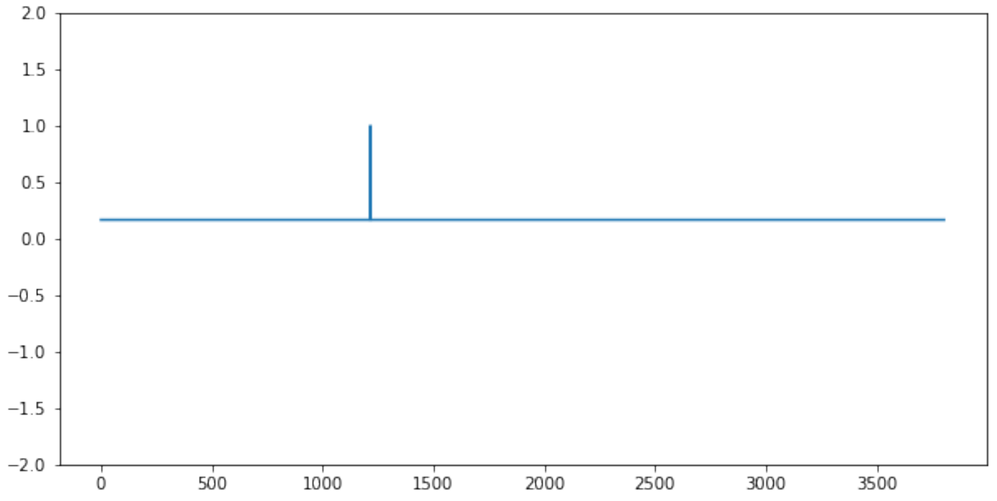
Integrating Video Action Anticipation with Diffusion Models







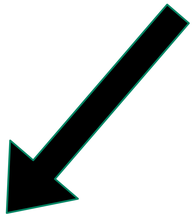
Prediction

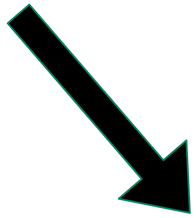


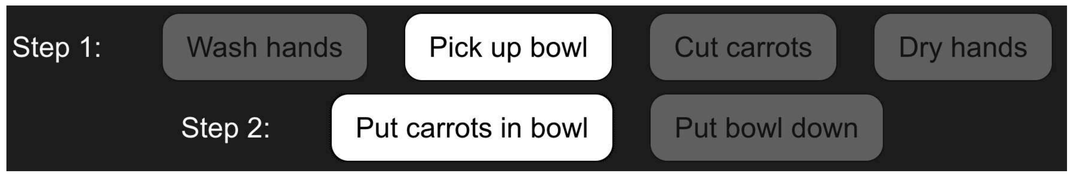
Ground Truth

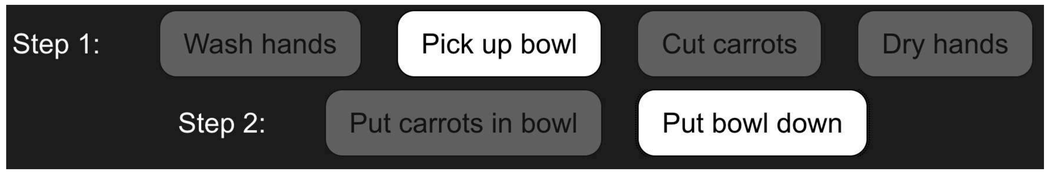
Manipulate the Environments (GAIA-1, Wayve)

Analysis the Video Action Causality (UniSim, Yang et al. 2023)









- Introduction
- Video Action Anticipation
- Introduction to Diffusion Models
- Synthesis and Future Directions
- **Conclusion and Discussion**

Conclusion and Discussion

- We've discussed the evolution of generative models, noting the progression from large language models (LLMs) to the current research focus on vision-language models (VLMs).
- The next frontier in the vision domain is video, which presents a significant challenge due to its inherent spatial and temporal complexities.
- The diffusion model operates on a discrete interpretation of Langevin Dynamics.
- The diffusion model is capable of synthesizing high-fidelity modalities, potentially serving the substantial data requirements of foundational models.
- **Can Generative AI and Video Analysis Redefine the Future Vision?**