



Recent Advances in (Image) Generative Foundation Models

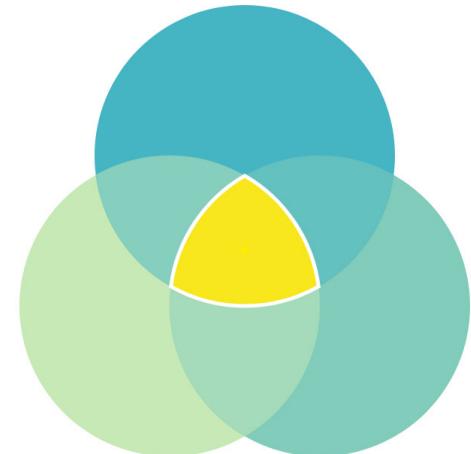
Zhengyuan Yang



Outline

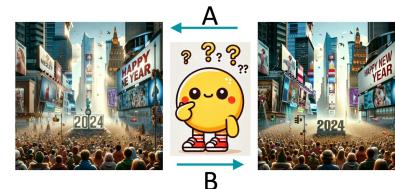
1. Training good generative foundation models
2. Image generation evaluation
3. Generative foundation models + LMMs

Training Good Generative Foundation Models



Evaluation

Usage beyond Generation

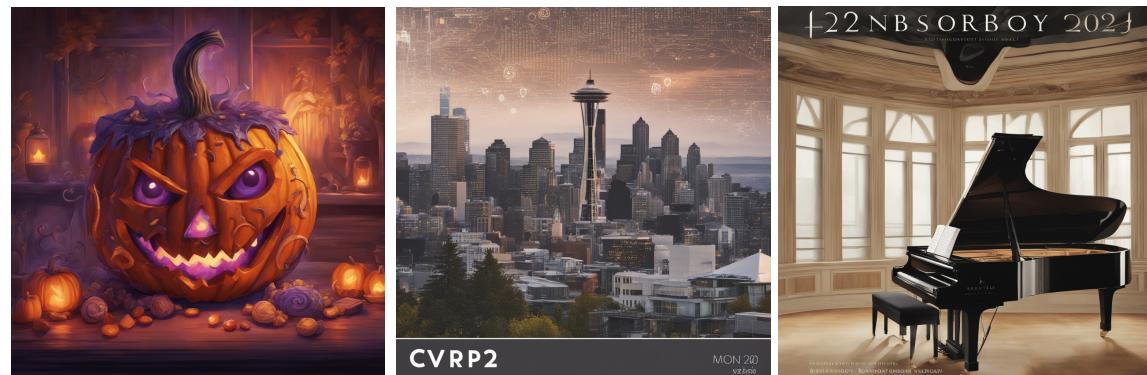


Training Better Image Generative Foundation Models

Compared with last year:

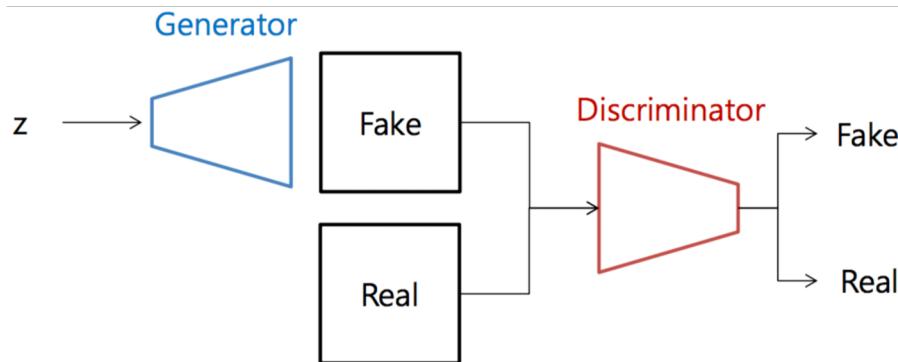
- Better quality
- Human alignment
- Faster inference (e.g. x50)

Last year, e.g. SDXL

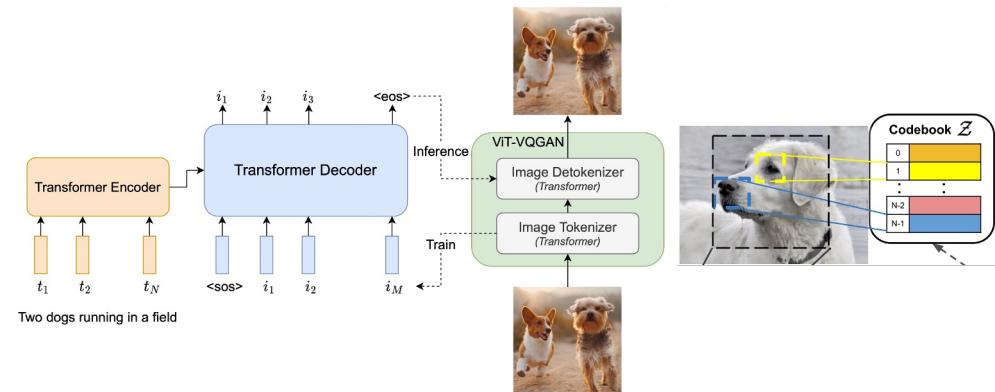


Text-to-Image Basics

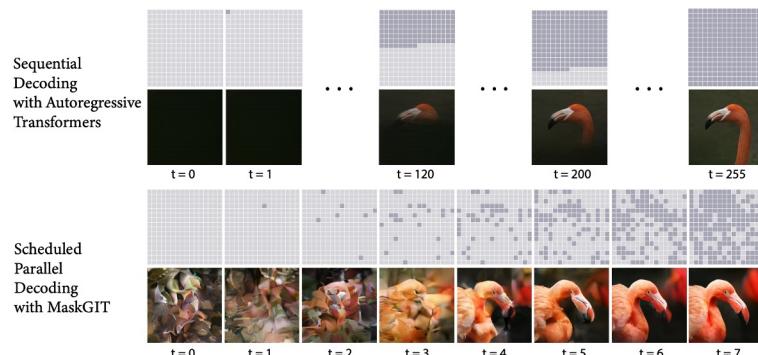
Generative Adversarial Networks (GAN)



Auto-regressive (AR)



Non-AR Transformer



Diffusion

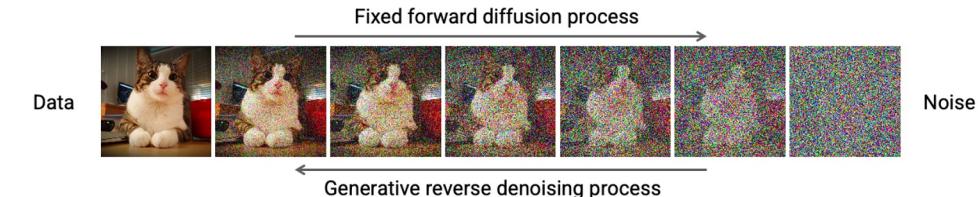


Image credit: Taming Transformer, Parti, MaskGIT, CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

Text-to-Image Basics

- Latent diffusion overview
 - Variational autoencoder (VAE)
 - Condition (text) encoder
 - Conditional denoising U-Net

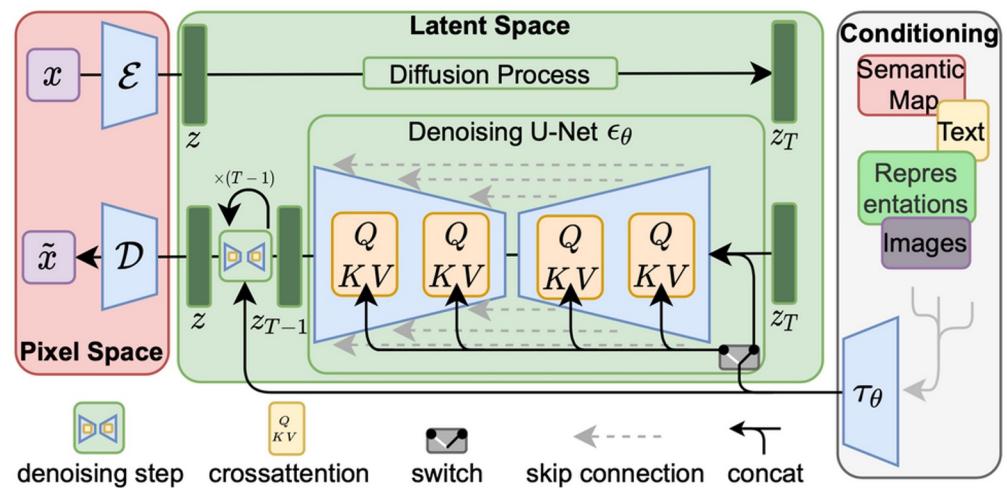
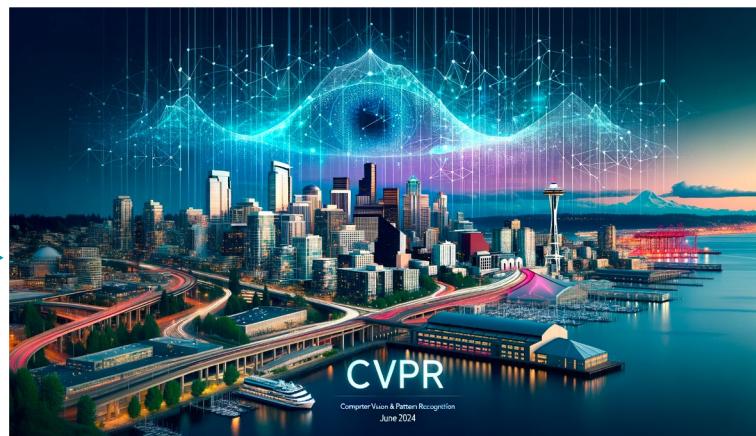
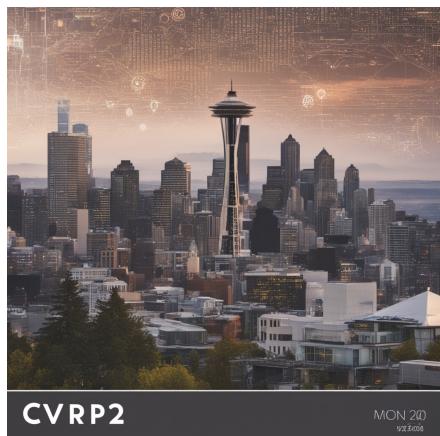
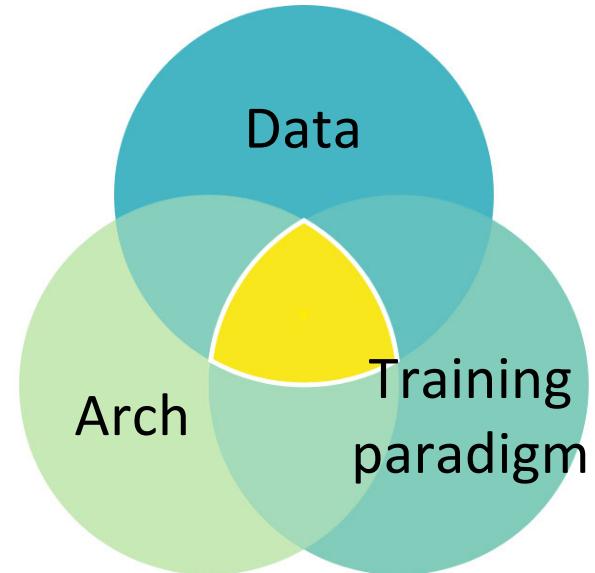


Image credit: "High-Resolution Image Synthesis with Latent Diffusion Models", VAE-Wikipedia

Training Good Generative Foundation Models

- ✓ Text-to-image basics
- 1. Data
- 2. Architecture
- 3. Training paradigm



Past year:
DALL-E 3
SD3
MJ6
PixArt
VAR
.....

Data: Recaption (DALL-E 3)

- Data: the importance of re-caption and text encoder (T5)
- Less noisy; more detailed

			
Image	now at victorian plumbing.co.uk	is he finished...just about!	23 (19 of 30) 1200
SSC / Alt Text	a white modern bathtub sits on a wooden floor.	a quilt with an iron on it.	a jar of rhubarb liqueur sitting on a pebble background.
DSC	this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.	a quilt is laid out on a ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.	rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.

Image credit: "Improving Image Generation with Better Captions", DALL-E 3
 "DEsignBench: Exploring and Benchmarking DALL-E 3 for Imagining Visual Design"



User Input: A horse riding an astronaut

Expanded Prompt: Render of a humorous setting where a white horse, looking a bit puzzled, is sitting on top of an astronaut's back. The astronaut tries to balance the horse while surrounded by asteroids and space debris.

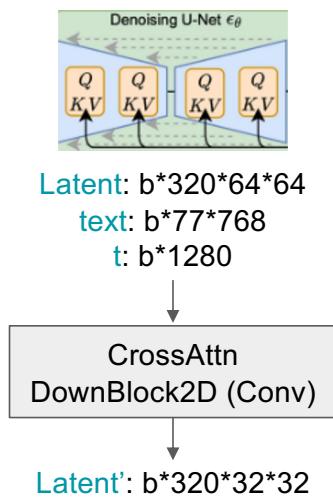
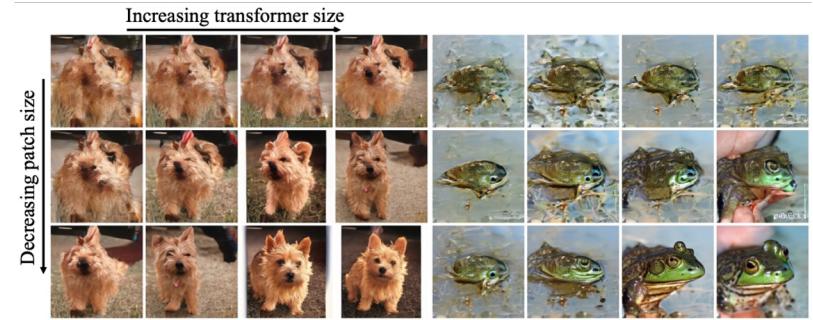


User Input: A woman stands wearing a black and white coat over a dark vest, orange shirt and pink tie.

Expanded Prompt: Photo of a confident woman with short blonde hair standing against a neutral background. She is wearing a black and white striped coat that reaches her knees. Underneath the coat, she has a dark vest and a bright orange shirt. Around her neck, she has tied a vibrant pink tie which stands out against the other colors.

Arch: Diffusion Models with Transformers

- Architecture: from U-Net to fully transformer
- Good scaling behavior



U-Net Block

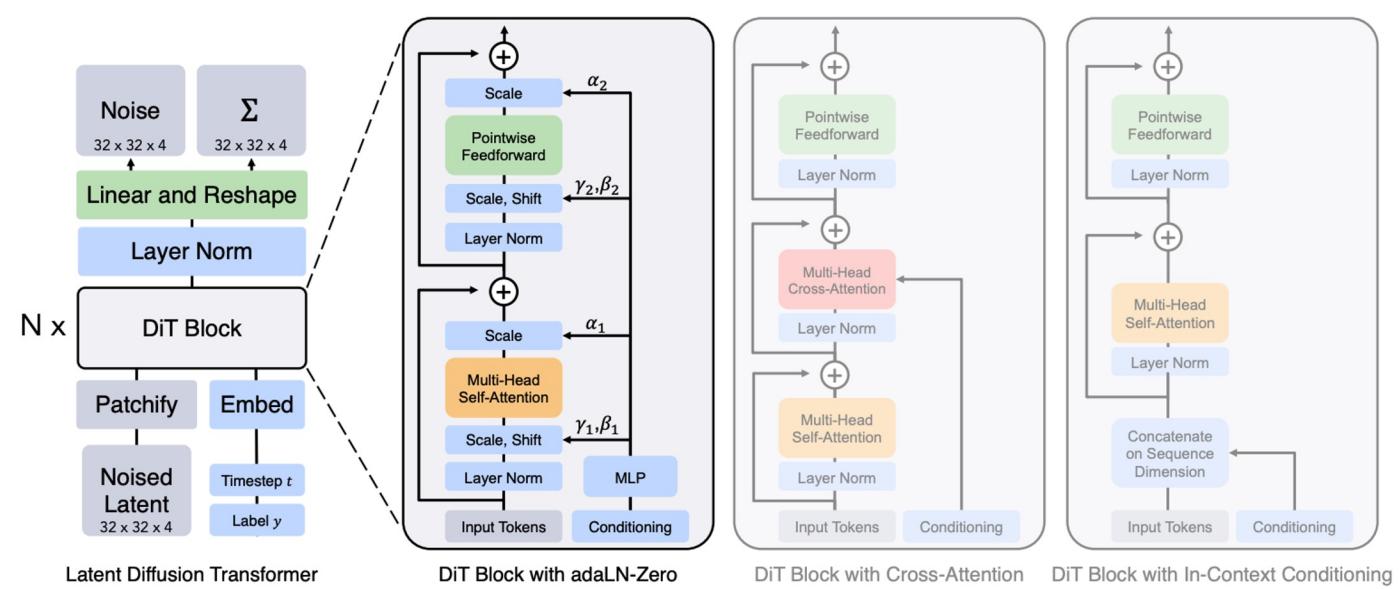


Image credit: “Scalable Diffusion Models with Transformers”, DiT

Training Paradigm: Rectified Flow

- Diffusion training:
 - Diffusion improvements
 - Rectified flow

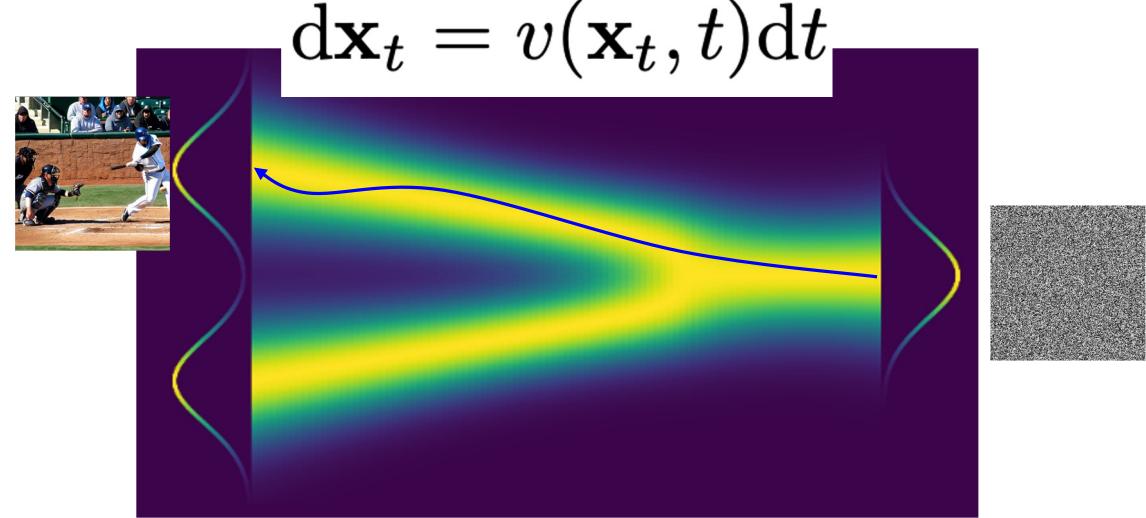


Image credit: "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow"
"An introduction to flow matching," Fjelde, Tor and Mathieu, Emile and Dutordoir, Vincent
"Flow Matching in Latent Space"

Case Study: Stable Diffusion 3

- A scaled-up open-sourced model with:
 - Data: recaption + T5 (and CLIP x2)
 - Architecture: diffusion transformer (MM-DiT)
 - Training Paradigm: rectified flow



an old rusted robot wearing pants and a jacket riding skis in a supermarket.



smiling cartoon dog sits at a table, coffee mug on hand, as a room goes up in flames. "This is fine," the dog assures himself.

Image credit: "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis", SD3

Training Paradigm: Autoregressive

- Scaling behavior of AR generation
- From next token prediction to next scale

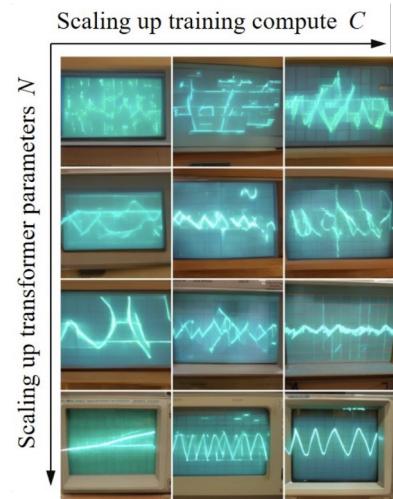
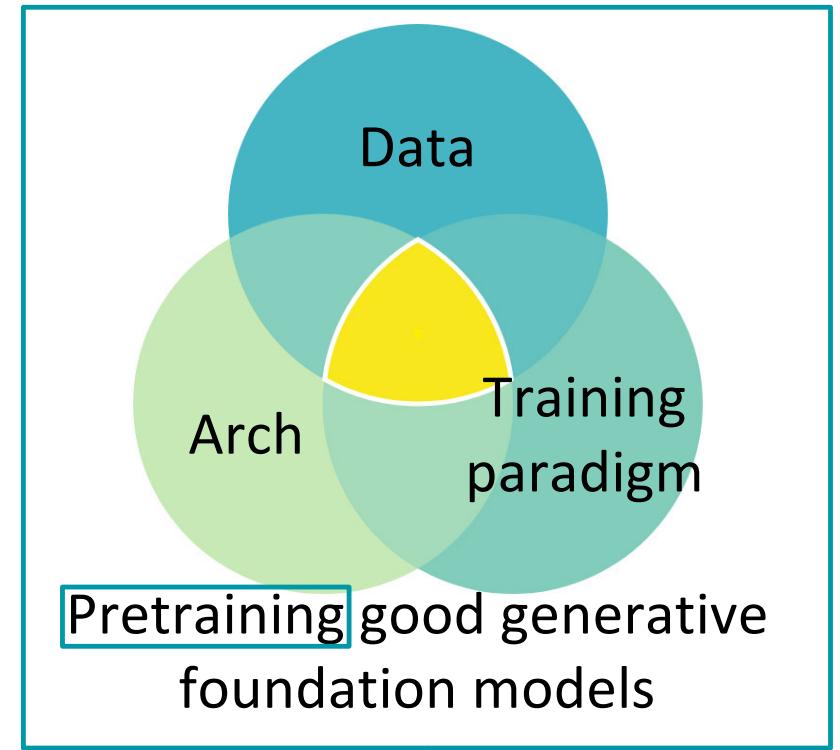


Image credit: “Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation”, LlamaGen
“Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction”, VAR

Training Good Generative Foundation Models

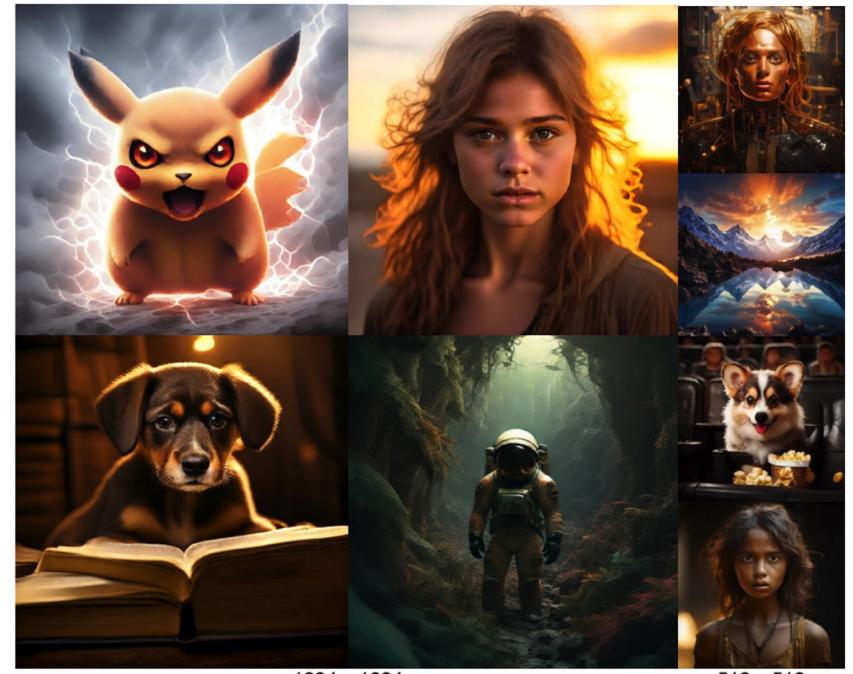
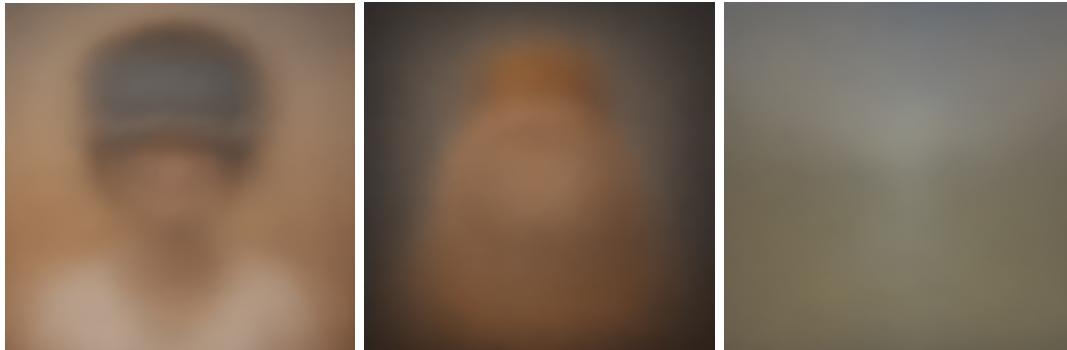
- ✓ Text-to-image basics
- 1. Data: better captions (DALL-E 3) ✓
- 2. Architecture: ✓ diffusion transformer (DiT)
- 3. Training paradigm: ✓
 - Flow matching (rectified flow)
 - Next-scale prediction (VAR)



Slow inference?
Human alignment?

Diffusion Inference Acceleration

- Diffusion inference speed
 - Advanced diffusion samplers
(50 -> ~10 steps)
 - Diffusion distillation
(50 -> 1-4 steps)



One-step with
Diffusion Distillation

Image credit: "Improving Adversarial Distillation of Diffusion Models"

Consistency Models

- Consistency distillation

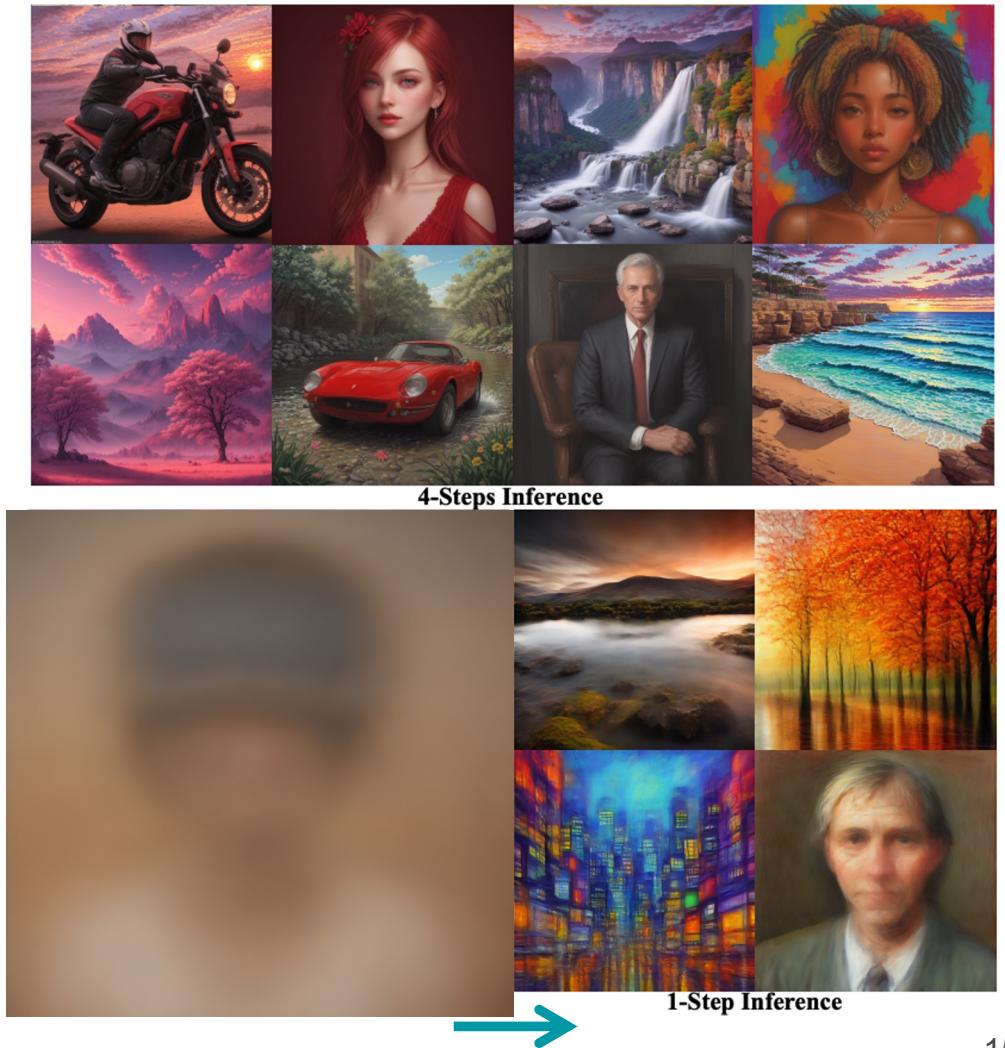
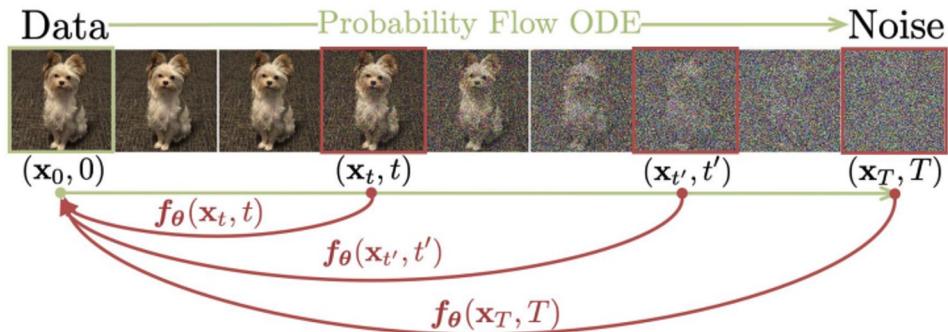


Image credit: "Consistency Models",
"Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference"

Adversarial Training

- SD initialized generator and discriminator

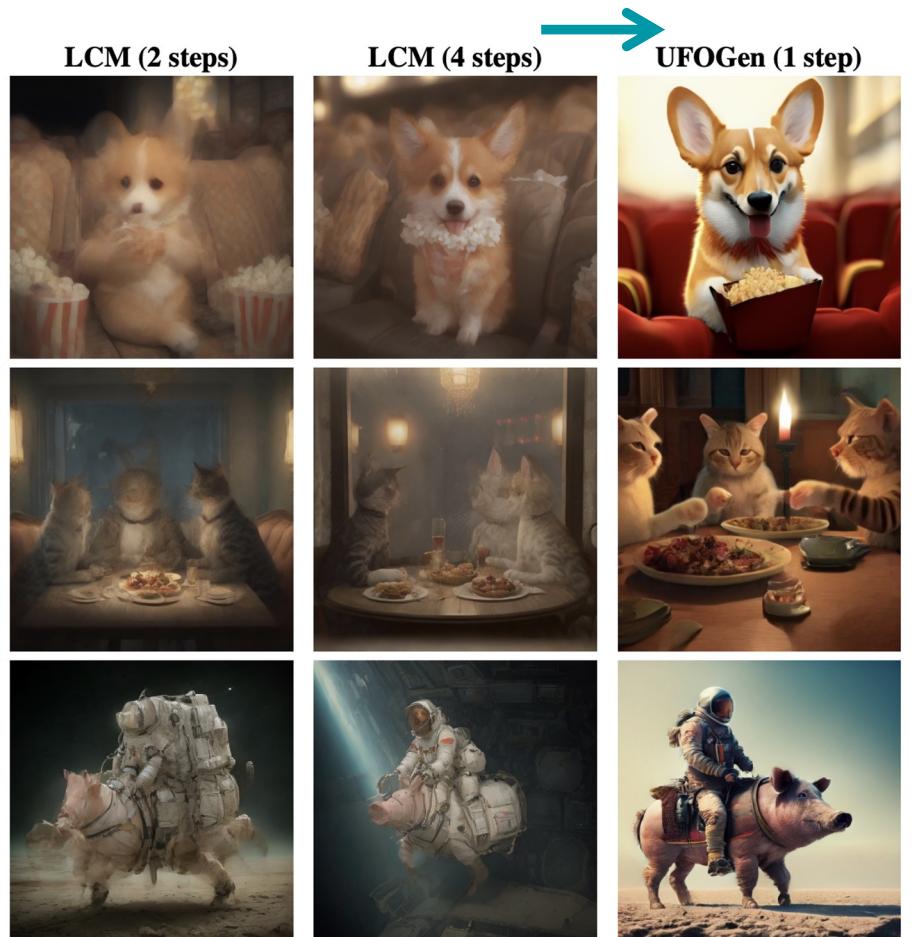
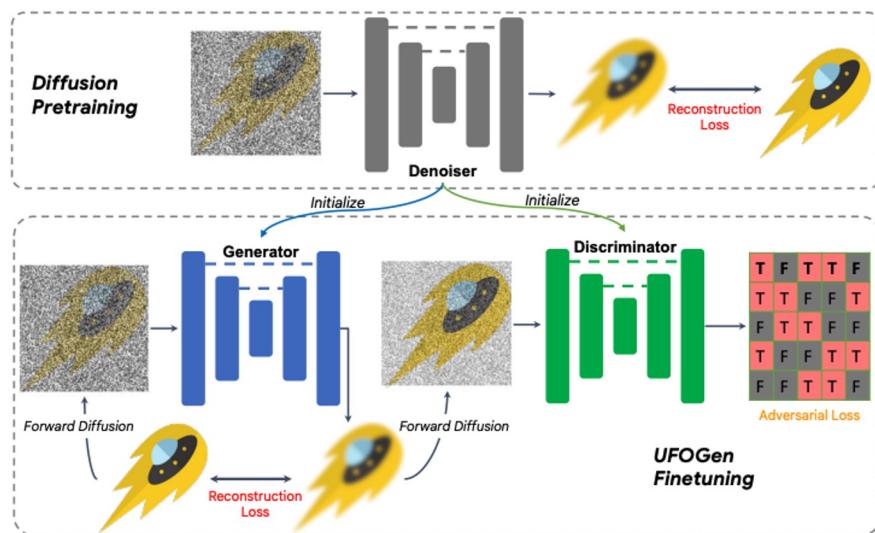


Image credit: "UFOGen: You Forward Once Large Scale Text-to-Image Generation via Diffusion GANs"

Adversarial Diffusion Distillation

- Other discriminator design
 - Full pretrained diffusion, DINOv2, diffusion bottleneck feature

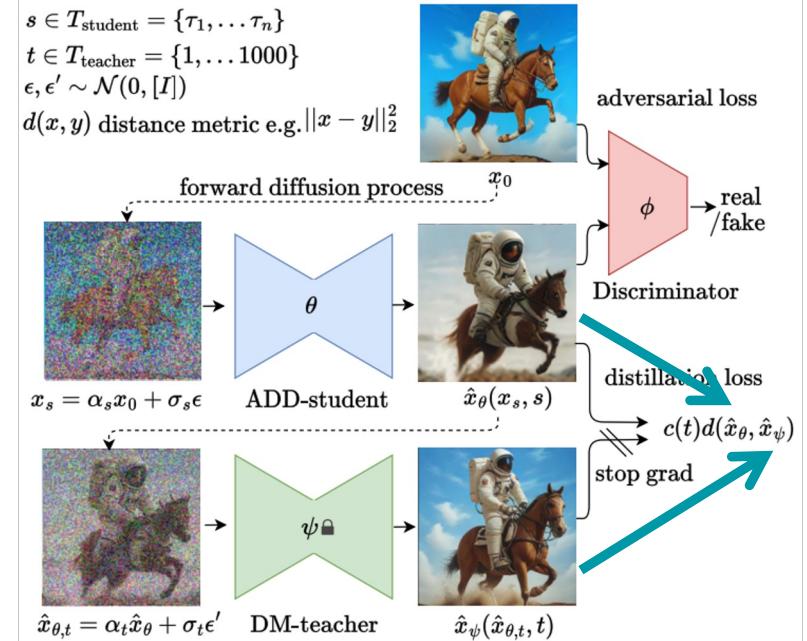


Image credit: “Adversarial Diffusion Distillation” (SDXL Turbo)

“Improved Distribution Matching Distillation for Fast Image Synthesis” (DMD/DMD2)

Adversarial Diffusion Distillation

- Other discriminator design
 - Full pretrained diffusion, DINOv2, diffusion bottleneck feature
- Distillation loss
- Distribution matching distillation

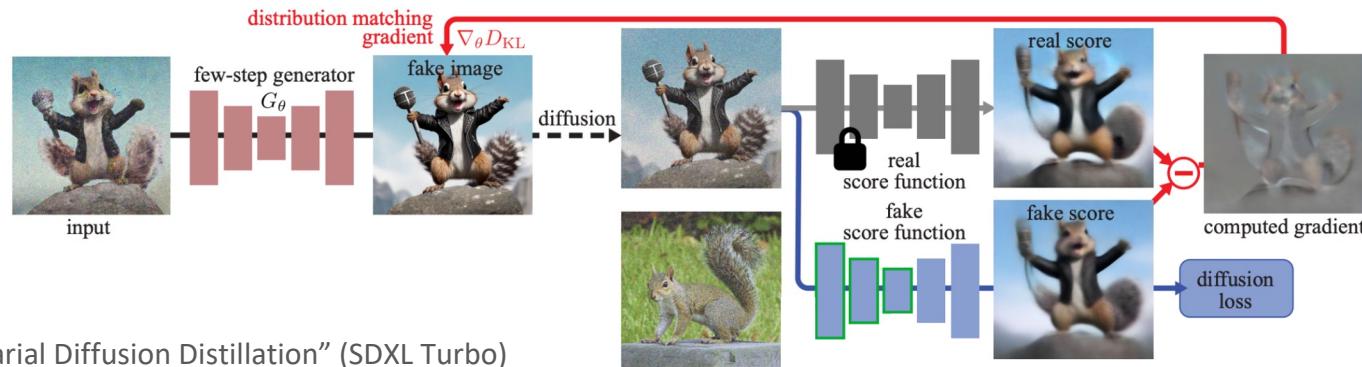
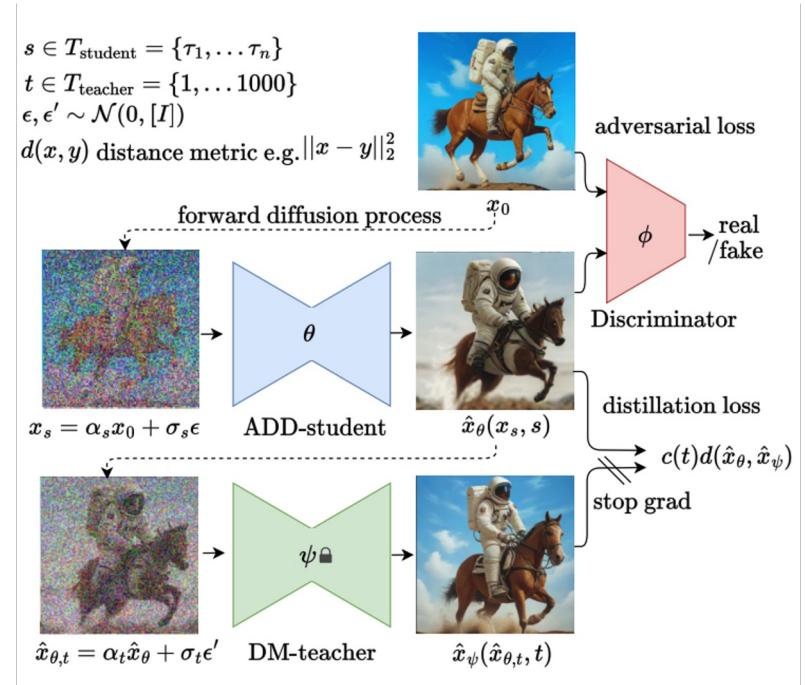


Image credit: “Adversarial Diffusion Distillation” (SDXL Turbo)

“Improved Distribution Matching Distillation for Fast Image Synthesis” (DMD/DMD2)

Diffusion Inference Acceleration

a shiba inu wearing a beret and black turtleneck



a young girl playing piano



A train ride in the monsoon rain in Kerala. With a Koala bear wearing a hat looking out of the window. There is a lot of coconut trees out of the window.



DMD2 (Ours)

LCM

Turbo

Lightning

Teacher

Image credit: "Improved Distribution Matching Distillation for Fast Image Synthesis" (DMD/DMD2)

From Image to Video



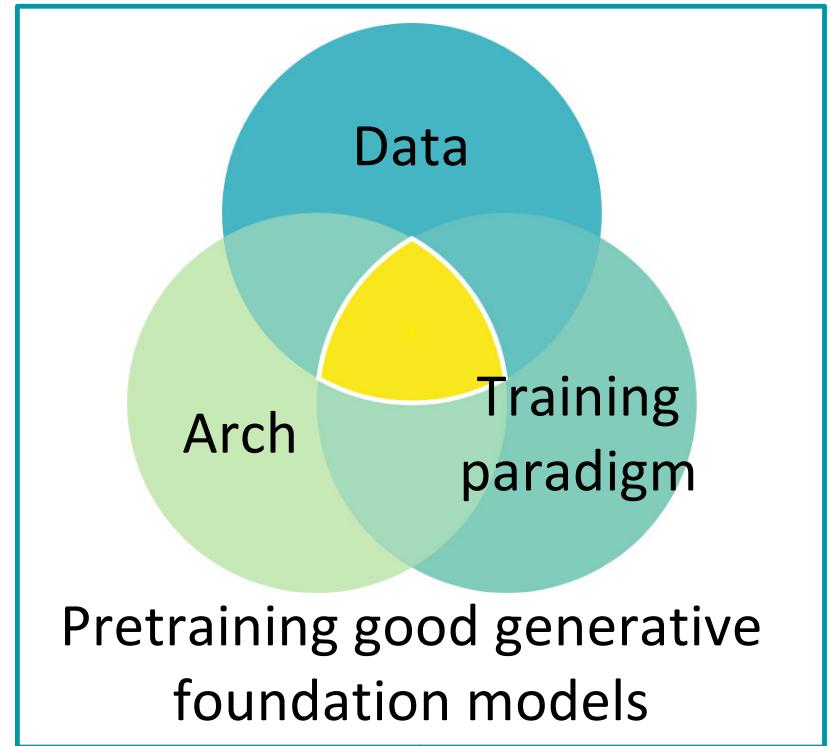
Teacher 50 steps



Motion Consistency Model 4 steps

Image credit: "Motion Consistency Model: Accelerating Video Diffusion with Disentangled Motion-Appearance Distillation"

Post Training, Human Alignments



✓ Slow inference

Human alignment?

Post Training, Human Alignments: Last Year's Tutorial

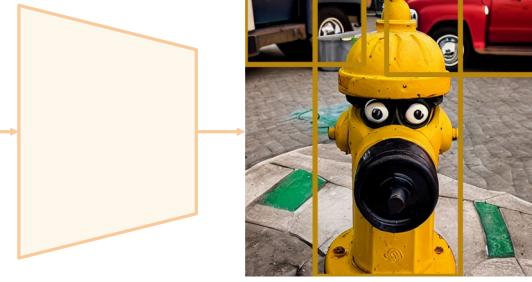
Controllable generation

Image-level: a yellow fire hydrant with a cartoon face drawn on it.

a truck is parked next to a trash can.

a yellow fire hydrant with a face on it and black eyes.

a red truck is parked in a parking lot.



Editing

"Swap sunflowers with roses"



"Add fireworks to the sky"



Better following prompts

"A horse and a dog"



"A painting of an elephant with glasses"



Stable Diffusion



+Attend-and-Excite



Concept customization



Input images



in the Acropolis



swimming



sleeping

in a doghouse



in a bucket

Image credit: ReCo, InstructPix2Pix, Attend-and-Excite, DreamBooth

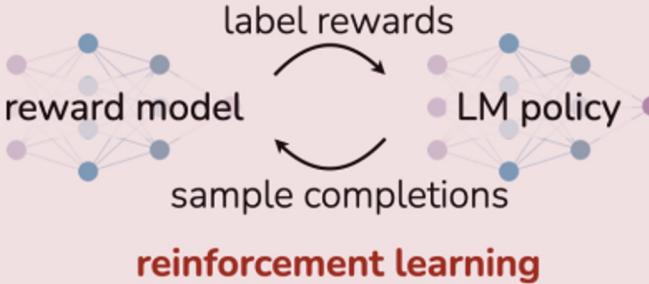
Human Feedback Learning in LLMs

Reinforcement Learning from Human Feedback (RLHF)

x: "write me a poem about
the history of jazz"



maximum likelihood



Direct Preference Optimization (DPO)

x: "write me a poem about
the history of jazz"



maximum likelihood



- With, w/o reward model
- LM policy, maximum likelihood

Image credit: "Training language models to follow instructions with human feedback"
"Direct Preference Optimization: Your Language Model is Secretly a Reward Model"

Fine-tuning Diffusion Models with Reward Model

Specialized model as reward model:
Aesthetic, I2T+BertScore, ImageReward, etc.

Reinforcement Learning

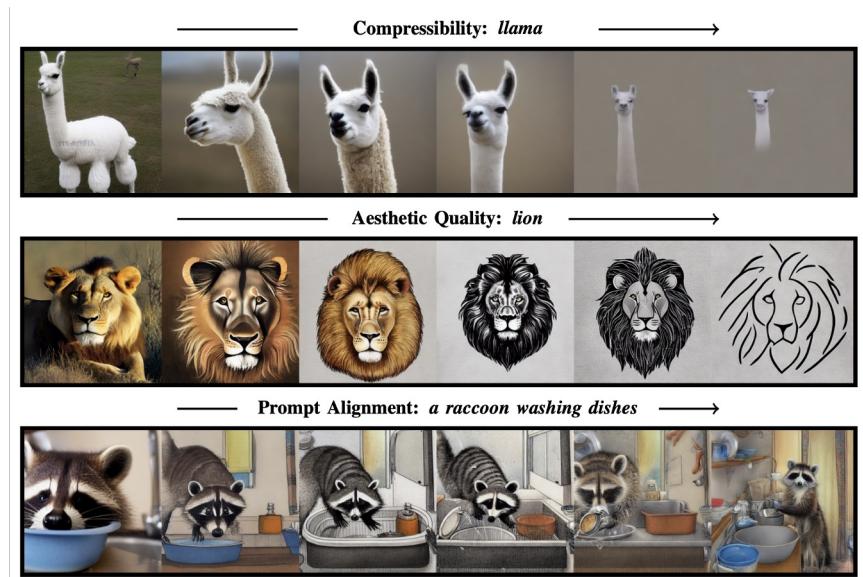
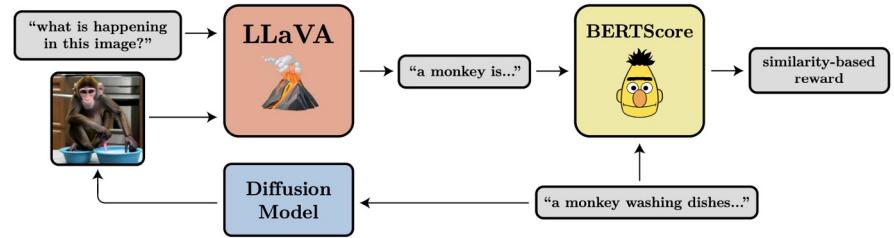
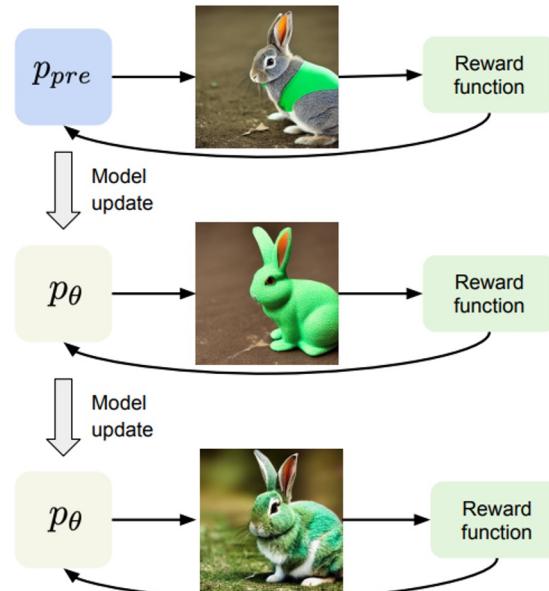


Image credit: "Training Diffusion Models with Reinforcement Learning" (DDPO)
"DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models"

Fine-tuning Diffusion Models towards Diverse Generation

- **Diversity Rewards:**
Customize with a set of diverse reference images

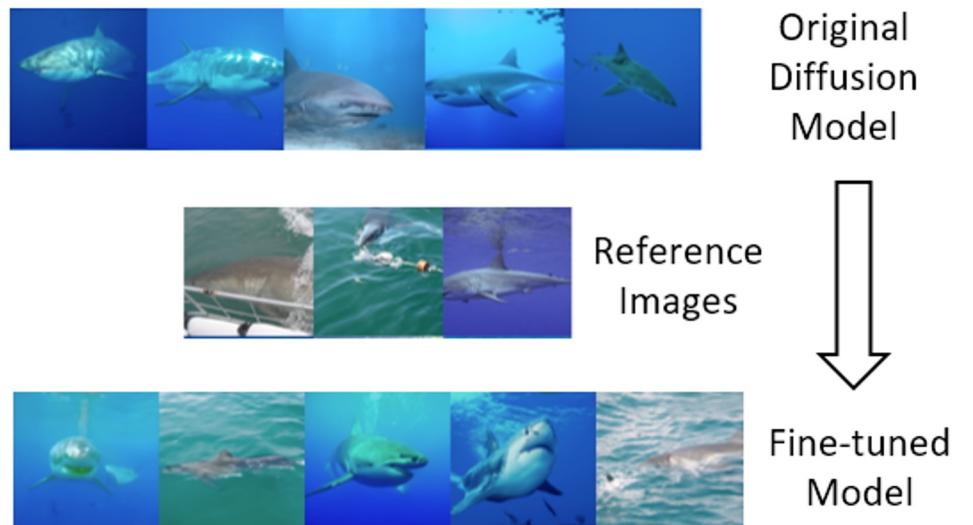
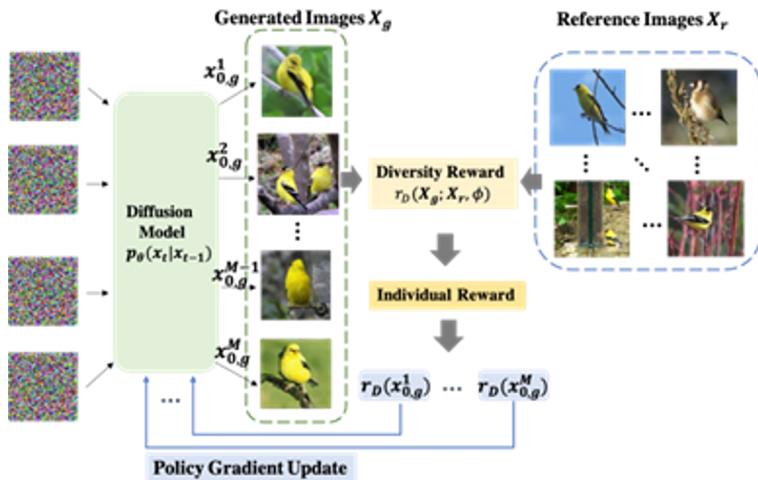
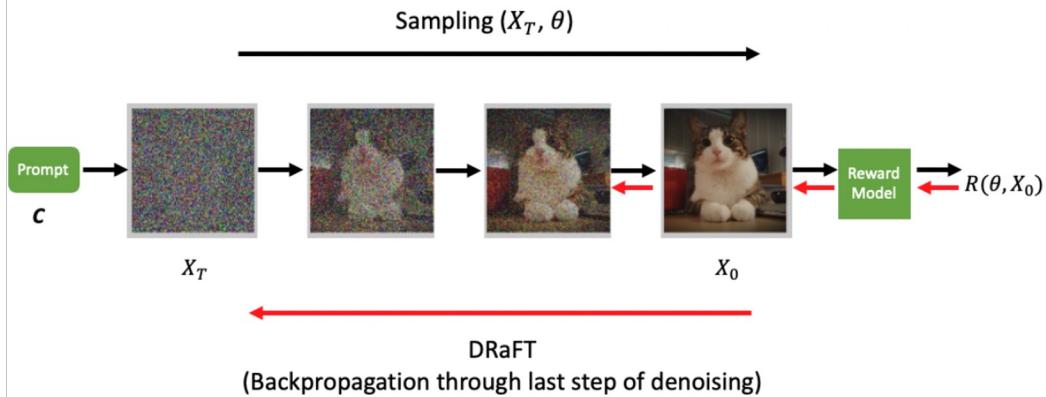


Image credit: "Training Diffusion Models Towards Diverse Image Generation with Reinforcement Learning"

Fine-tuning Diffusion Models with Reward Model

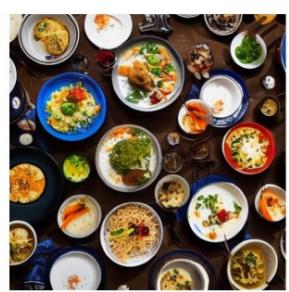
Direct Reward Back Propagation



A stunning beautiful oil painting of a lion, cinematic lighting, golden hour light.



Highly detailed photograph of a meal with many dishes.



A racoon washing dishes.



Stable Diffusion 1.4

After Reward
Fine-Tuning



Image credit: "Directly Fine-Tuning Diffusion Models on Differentiable Rewards"

Directly Fine-tuning Diffusion Models with Preference

- No need of reward model
- Requires datasets of {(prompt, prefer image, dislike image)}

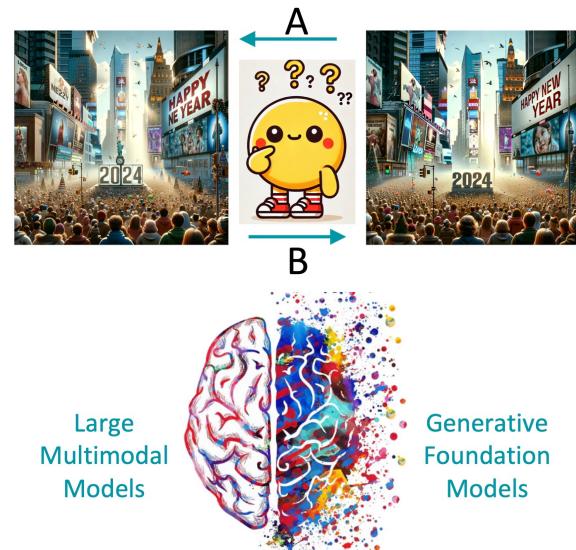
$$L_{\text{DPO-Diffusion}}(\theta) = -\mathbb{E}_{(\mathbf{x}_0^w, \mathbf{x}_0^l) \sim \mathcal{D}} \log \sigma \left(\beta \mathbb{E}_{\substack{\mathbf{x}_{1:T}^w \sim p_\theta(\mathbf{x}_{1:T}^w | \mathbf{x}_0^w) \\ \mathbf{x}_{1:T}^l \sim p_\theta(\mathbf{x}_{1:T}^l | \mathbf{x}_0^l)}} \left[\log \frac{p_\theta(\mathbf{x}_{0:T}^w)}{p_{\text{ref}}(\mathbf{x}_{0:T}^w)} - \log \frac{p_\theta(\mathbf{x}_{0:T}^l)}{p_{\text{ref}}(\mathbf{x}_{0:T}^l)} \right] \right)$$



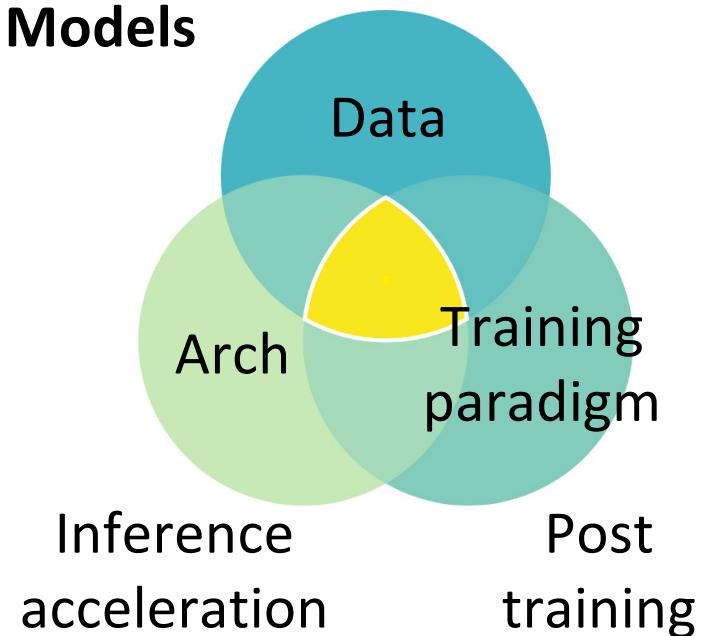
Image credit: "Diffusion Model Alignment Using Direct Preference Optimization"

Training Good Generative Foundation Models

- ✓ Advances in data, architecture, and training paradigm
- ✓ Diffusion inference acceleration
- ✓ Post training for human alignment



Training Good Generation Models



Inference
acceleration

Post
training

Inference
acceleration

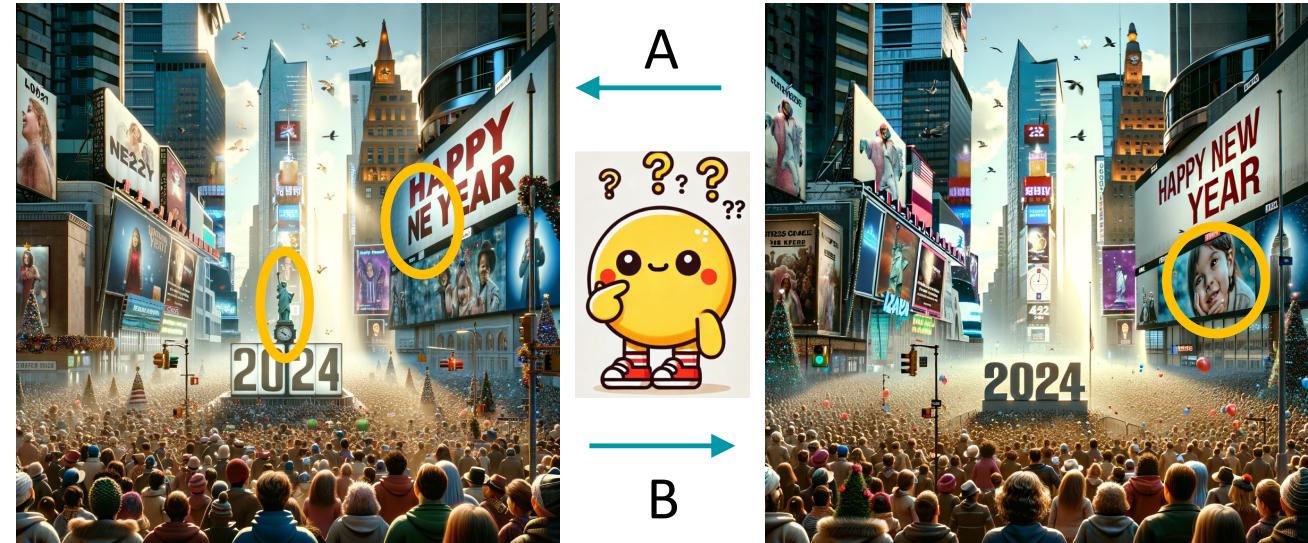
Evaluation

Usage beyond
generation

Generation Evaluation

- Beyond FID and CLIP score
 - Checking details
 - Diverse evaluation aspects
 - Emergent scenarios

Prompt: Photorealistic scene capturing the heart of Times Square during the exhilarating New Year's Eve countdown ushering in 2024. The area is densely packed with jubilant individuals, their faces reflecting the joy and optimism of welcoming a new year. Skyscrapers adorned with brilliant neon signs and screens add to the ambiance, painting the night with a myriad of colors. Central to the festivities is the New Year's Eve ball, steadily descending to mark the transition. Dominating the visual landscape, a grand digital screen prominently displays the messages 'Happy New Year' and '2024', symbolizing the collective celebration and the dawn of new possibilities.



Checking Details

- T2I-CompBench
 - VQA: attribute
 - UniDet: spatial
 - MiniGPT4
- GenEval:
object presence,
count, position,
attribute

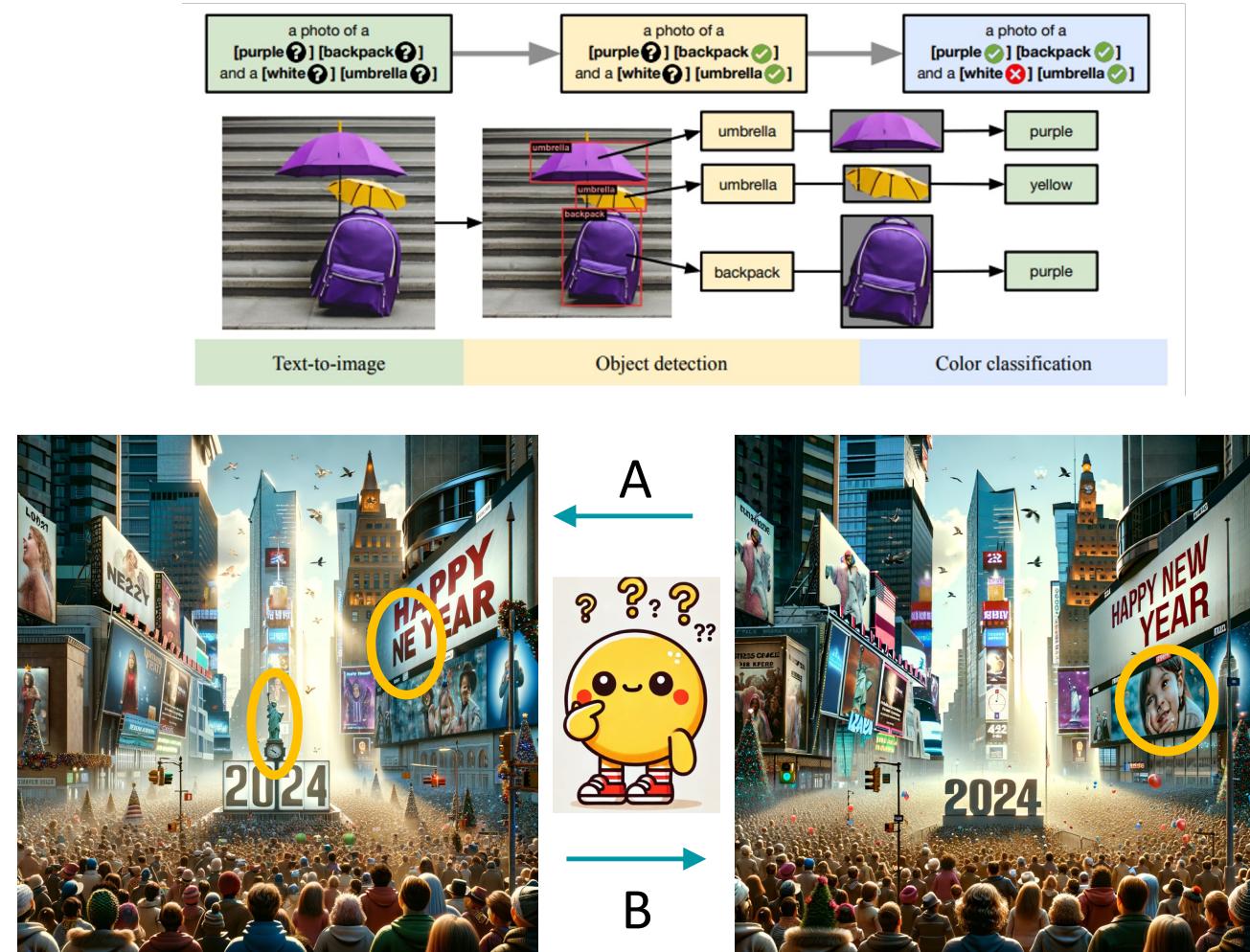
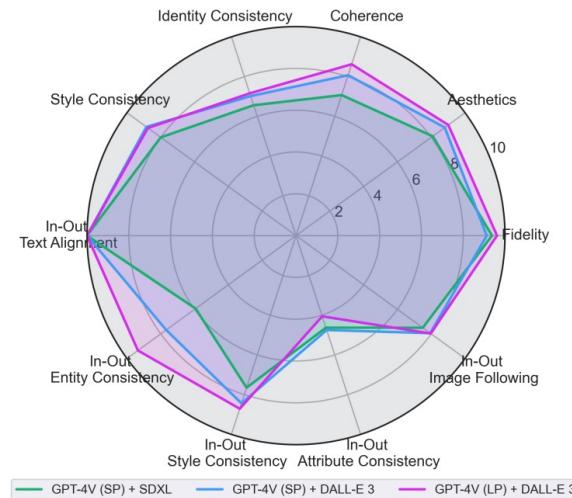


Image credit: "GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment"

Diverse Evaluation Aspects

- LMM as unified metric
- Extend to multimodal inputs and outputs



Following is the prompt we give to the vision-enabled GPT-4 model to perform our automated drawbench evaluation:

You are responsible for judging the faithfulness of images generated by a computer program to the caption used to generate them. You will be presented with an image and given the caption that was used to produce the image. The captions you are judging are designed to stress-test image generation programs, and may include things such as:

1. Scrambled or mis-spelled words (the image generator should draw an image associated with the probably meaning)
2. Color assignment (the image generator should apply the correct color to the correct object)
3. Counting (the correct number of objects should be present)
4. Abnormal associations, for example 'elephant under a sea', where the image should depict what is requested.
5. Descriptions of objects, the image generator should draw the most commonly associated object.
6. Rare single words, where the image generator should create an image somewhat associative with the specified image.
7. Images with text in them, where the image generator should create an image with the specified text in it. You need to make a decision as to whether or not the image is correct, given the caption. You will first think out loud about your eventual conclusion, enumerating reasons why the image does or does not match the given caption. After thinking out loud, you should output either 'Correct' or 'Incorrect' depending on whether you think the image is faithful to the caption.

A few rules:

1. Do not nitpick. If the caption requests an object and the object is generally depicted correctly, then you should answer 'Correct'.
2. Ignore other objects in the image that are not explicitly mentioned by the caption; it is fine for these to be shown.
3. It is also OK if the object being depicted is slightly deformed, as long as a human would recognize it and it does not violate the caption.
4. Your response must always end with either 'incorrect' or 'correct'
5. 'Incorrect' should be reserved for instances where a specific aspect of the caption is not followed correctly, such as a wrong object, color or count.
6. You must keep your thinking out loud short, less than 50 words.

```
image(<image_path>
<prompt>)
```

Image credit: "Improving Image Generation with Better Captions", DALL-E 3
"Openleaf: Open-domain interleaved image-text generation and evaluation"

Emergent Scenarios Evaluation

- Visual design

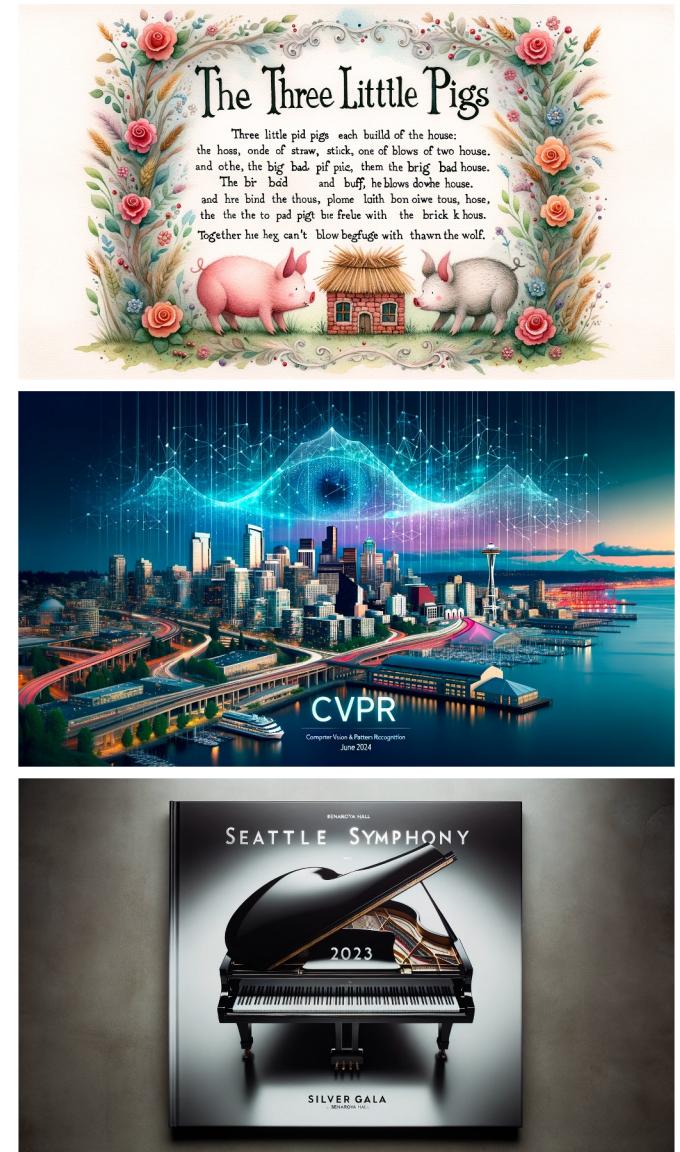
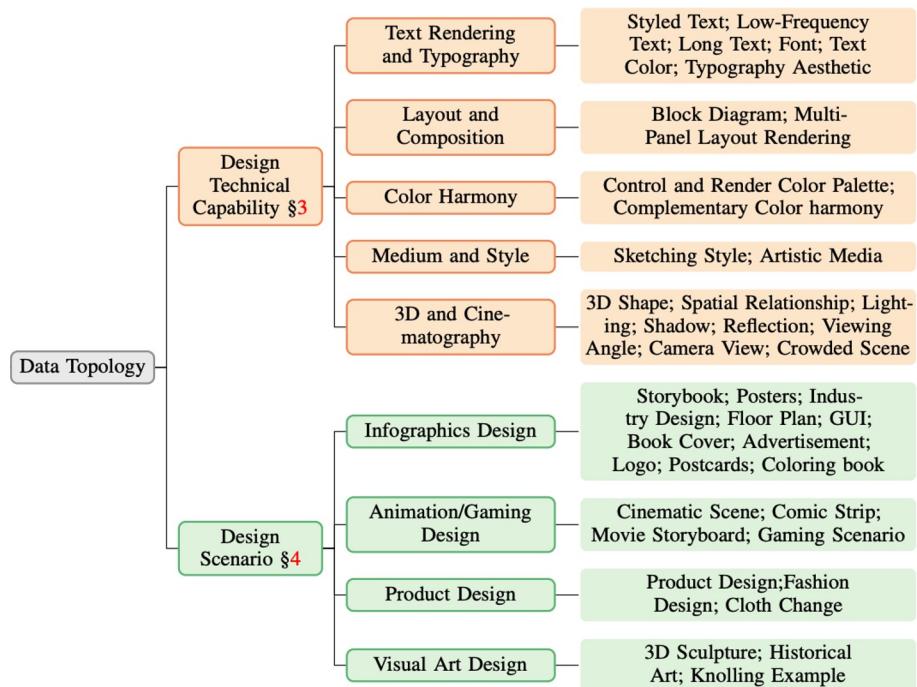


Image credit: "DEsignBench: Exploring and Benchmarking DALL-E 3 for Imagining Visual Design"

Emergent Scenarios Evaluation

- Interleaved generation:
multimodal inputs and outputs



Image credit: "Openleaf: Open-domain interleaved image-text generation and evaluation"

User Input

User Query: Can you spin a tale about a young witch and her talking raven companion?

Please follow the identity of the following two images.









In the heart of an ancient forest, there resided a young witch with hair as golden as the morning sun, cascading in voluminous waves down her back. Rosalind, dressed in her favourite dark green attire, was determined and swift; her eyes sparkling with the unwavering spirit of youth. Each dusk, with a wand that twinkled like the starlit sky, she practiced her spells and incantations, aspiring to be the most potent witch the land had ever seen.

Accompanying Rosalind was Alaric, her trusted raven companion with dark, enigmatic feathers that harbored a spectral array of colors shifting in the light. Alaric wasn't like other ravens; he had the gift of speech, a voice as serene and mystic as the moonlit night. Together, they explored the clandestine corridors of magic that lay embedded in the mystical woods, unraveling secrets that were forbidden and arcane.

One ominous night, whilst the moon was concealed behind the sinister clouds, Rosalind and Alaric stumbled upon an ancient spell book buried in the enigmatic depths of the forest. Names of witches and wizards long passed were scripted in the elegant, yet haunting calligraphy. A spell to unlock the unknown, a doorway to a world that existed parallel to theirs, stirred a tumultuous whirlwind of excitement and fear in Rosalind's heart.

Under the eerie silence of the looming trees encapsulating them, Rosalind, with Alaric perched on her shoulder, invoked the spell. The ominous verses, in synchrony with the enigmatic aura of the woods and Alaric's spectral voice echoing her chants, created an ethereal passage unveiling a world of marvel and terror. They were flung into an adventure that would test the potency of the Rosalind's spells and the wisdom of Alaric's voice, an escapade that was destined to either immortalize them in the annals of magic or obliterate their existence into the forbidden.

Generated Interleaved Content

Generative Foundation Models + LMMs

- Generative foundation models + LMMs
 - LMMs helps generation →
 - Unified text+image generation ←→
 - Diffusion for understanding ←
 - Synthesized data

Large
Multimodal
Models



Generative
Foundation
Models

LMMs Helps Generation Reasoning

- Reasoning and improving text prompts
 - Improve the input brief instructions
 - World knowledge

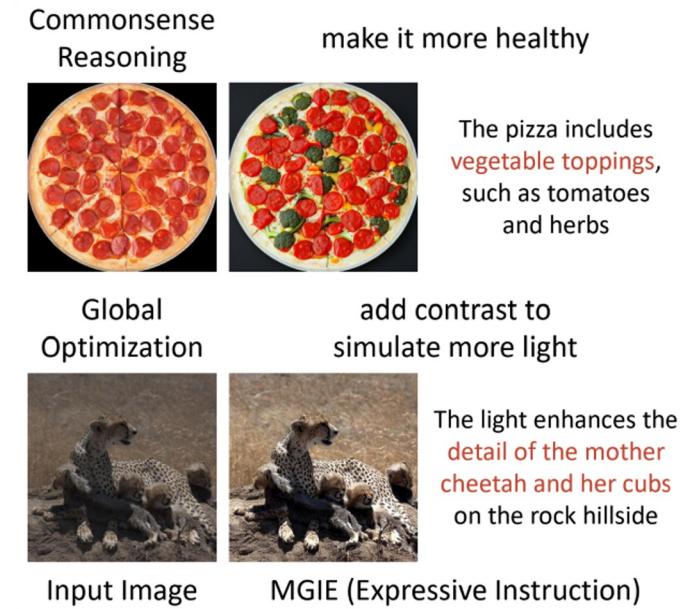


Image credit: "Guiding Instruction-based Image Editing via Multimodal Large Language Models"
"EditWorld: Simulating World Dynamics for Instruction-Following Image Editing"

LMMs Helps Generation Reasoning



- Idea2Img: GPT-4V + T2I:
 1. Revised prompt generation (Improving)
 2. Draft image selection (Assessing)
 3. Feedback reflection (Verifying)

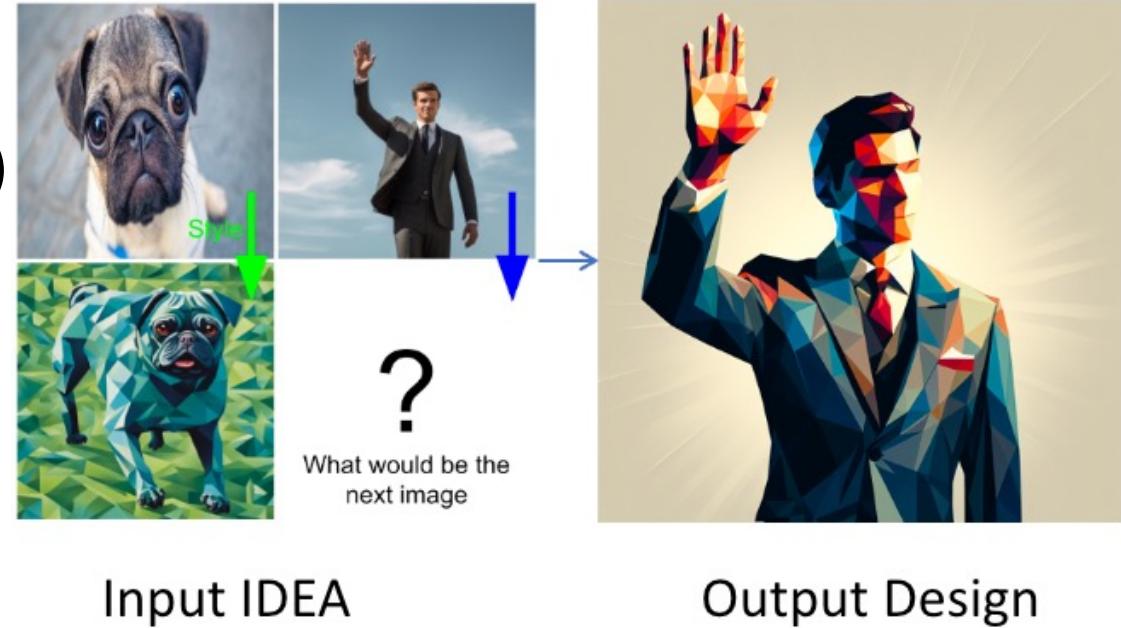
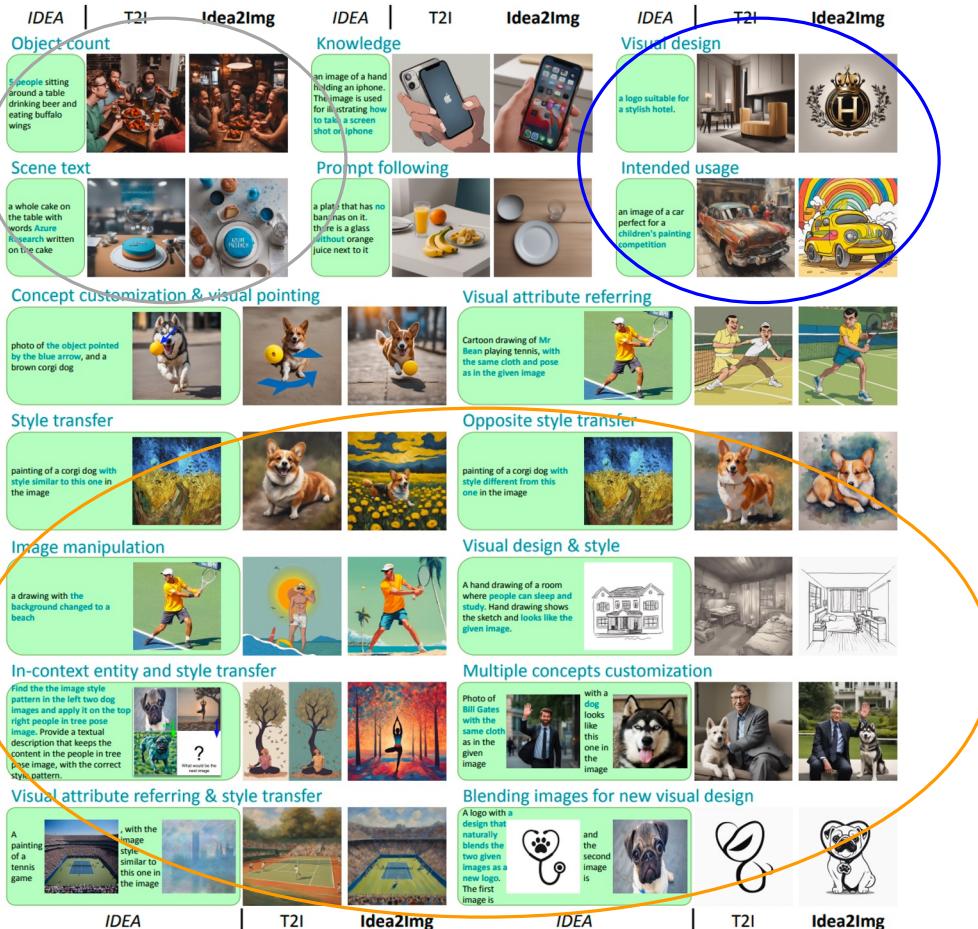


Image credit: "Idea2Img: Iterative Self-Refinement with GPT-4Vision for Automatic Image Design and Generation"

LMMs Helps Generation Reasoning



- Compared with T2I:
 - Image description -> arbitrary instruction
 - Text-only input -> multimodal idea input
 - Better visual and semantic quality

Image credit: "Idea2Img: Iterative Self-Refinement with GPT-4V(ision) for Automatic Image Design and Generation"

Unified Text+Image Generation

- Unified modeling:
 - Text, image autoregressive
 - Autoregressive text + diffusion decoder
 - Text, image diffusion

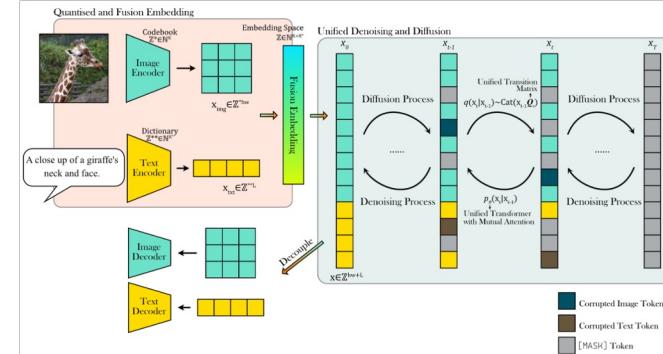
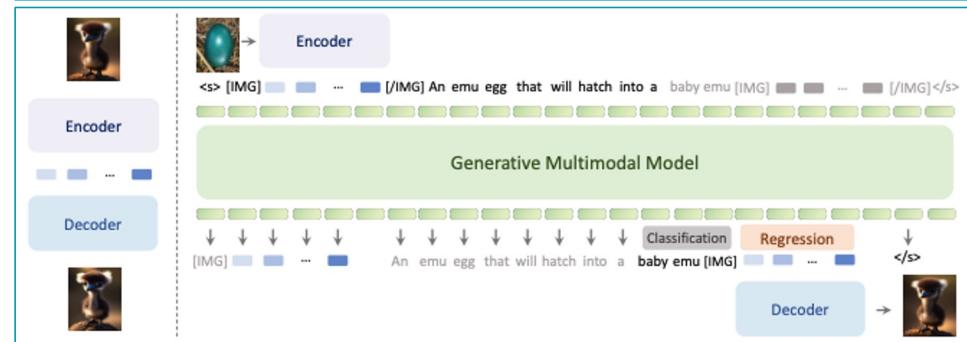
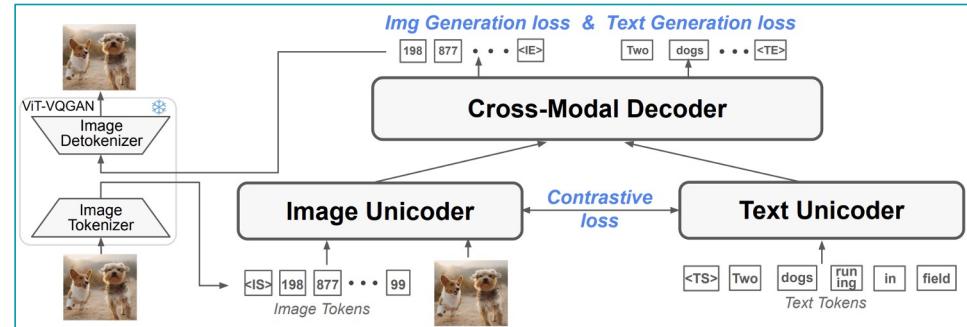


Image credit: “CoBIT: A Contrastive Bi-directional Image-Text Generation Model”
 “Generative Multimodal Models are In-Context Learners”
 “Unified Discrete Diffusion for Simultaneous Vision-Language Generation”

Discussion

- Beyond unified format, the grand vision of mutual beneficial
- Do we need perfect image (pixel) to benefit LMM learning?



Generative Foundation Models for Understanding

- Extract representation: extract UNet feature on a noisy image denoising process
- Classification
- Depth and dense prediction tasks

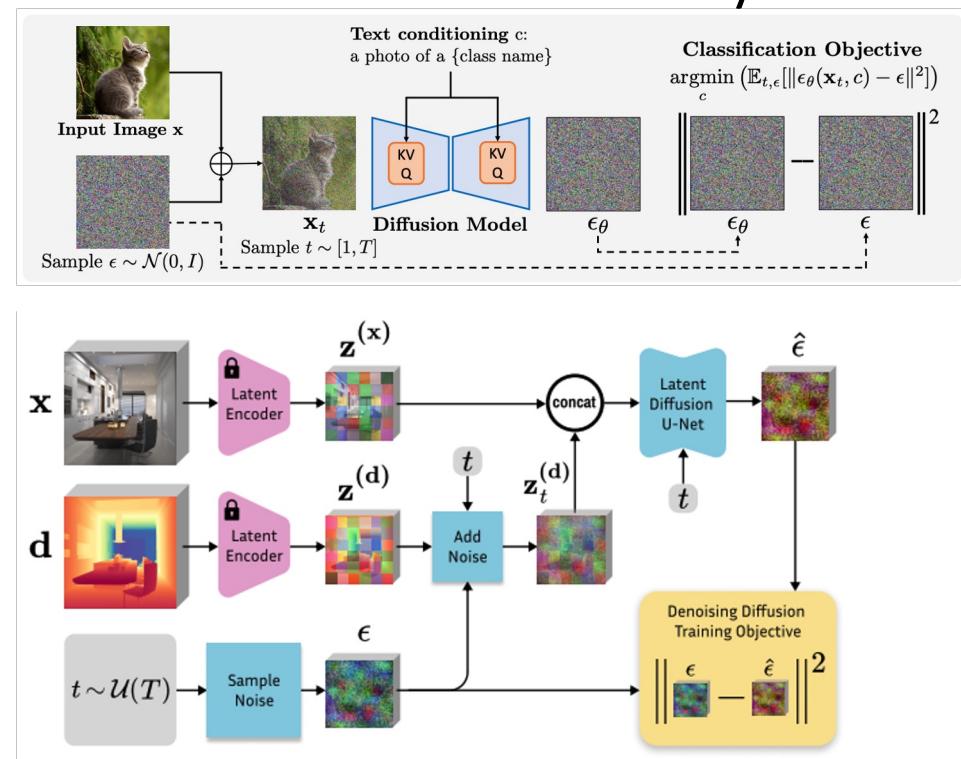


Image credit: "Your Diffusion Model is Secretly a Zero-Shot Classifier"
"Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation"

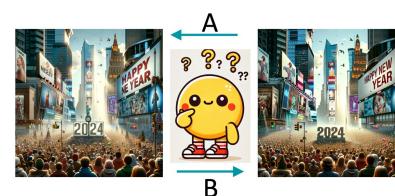
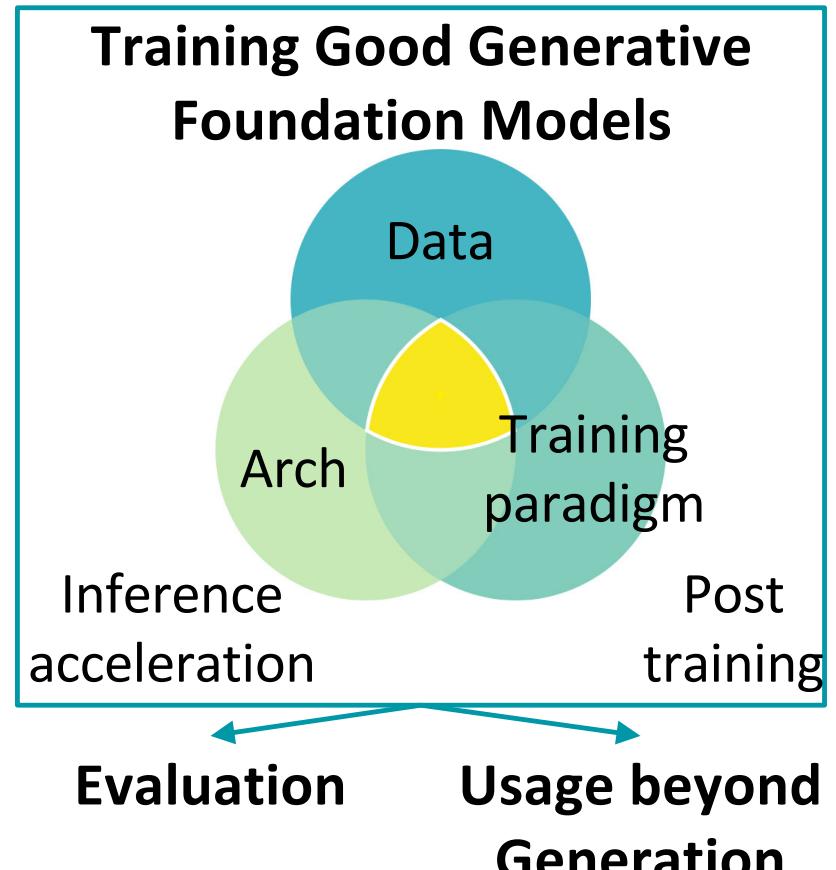
Discussion

- Promising results on dense output tasks like depth estimation, semantic correspondence, etc.
- Some discussions on insufficient as global visual representation
- Close the remaining gap and use diffusion as a vision encoder?



Summary

- Training
 - Data, arch, and training paradigm
 - Diffusion inference acceleration
 - Post training for human alignment
- Success and open problems in generation + LMMs
- Video and other modalities



Thank you!