UCF Center for Research in Computer Vision

CAP6412
Advanced Computer Vision

Mubarak Shah

shah@crcv.ucf.edu

HEC-245

Lecture-12: Diffusion Models-Introduction Part-III

Contents

- Stable Diffusion
- Dream Booth
- Control Net

Contents

- **Stable Diffusion**
- Dream Booth
- Control Net

This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Stable Diffusion CEO Emad Mostaque

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach[1] * Andreas Blattmann[1] * Dominik Lorenz[1] Patrick Esser[R] Björn Omm
[1]Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany [R]Runway ML

https://github.com/CompVis/latent-diffusion

Abstract

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text or bounding boxes and high-resolution synthesis becomes possible in a convolutional manner. Our latent diffusion models (LDMs) achieve new state of the art scores for image inpainting and class-conditional image synthesis and highly competitive performance on various tasks, including unconditional image generation, text-to-image synthesis, and super-resolution, while significantly reducing computational requirements compared to pixel-based DMs.

1. Introduction

Image synthesis is one of the computer vision fields with the most spectacular recent development, but also among



Figure 1. Boosting the upper bound on achievable quality with less agressive downsampling. Since diffusion models offer excellent inductive biases for spatial data, we do not need the heavy spatial downsampling of related generative models in latent space, but can still greatly reduce the dimensionality of the data via suitable autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at 512^2 px. We denote the spatial downsampling factor by f. Reconstruction FIDs [28] and PSNR are calculated on ImageNet-val. [12]; see also Tab. 8.

results in image synthesis [29,82] and beyond [7,44,47,56], and define the state-of-the-art in class-conditional image synthesis [15,30] and super-resolution [70]. Moreover, even unconditional DMs can readily be applied to tasks such as inpainting and colorization [82] or stroke-based synthesis [52], in contrast to other types of generative models [19,45,67]. Being likelihood-based models, they do not exhibit mode-collapse and training instabilities as GANs and, by heavily exploiting parameter sharing, they can model highly complex distributions of natural images without involving billions of parameters as in AR models [65].

Democratizing High-Resolution Image Synthesis DMs

High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion)

Robin Rombach, Andreas Blattmann, Dominik Lorenz,
Patrick Esser, Bjorn Ommer

Presentation by:

Ronald Campos, Muhammad Asad Haider,
Suneet Tipirneni, Stefan Werleman
https://www.youtube.com/watch?v=3MI7MBOhnVs

Paper Details

Presented: IEEE/CVF Conference on Computer Vision and Pattern Recognition

Year: 2022

Citations: 2023: 372

2025: 15,417

Paper: https://arxiv.org/pdf/2112.10752.pdf

Code: https://github.com/CompVis/latent-diffusion

Issues with standard diffusion models

- Standard diffusion models calculate probabilities for fine-grain details that may not be perceptible by humans.
- Instead we need to way to reduce dimensionality while also preserving as much information as possible.

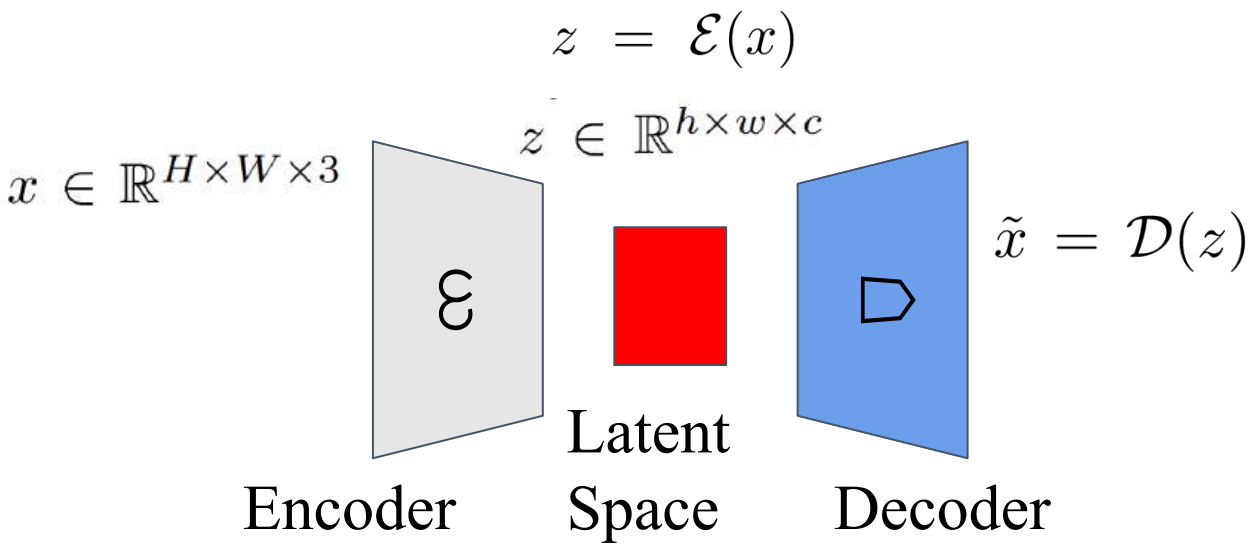
- Training a standard diffusion model can take 150-1000 **V100-days**
  - A V100-days measures how many days it takes to train a model using a single Nvidia V100 GPU
  - 1 V100-day = 2 google TPUv3's-days

How to reduce Dimensionality?

- Pixel space
  - Computational overhead
  - Not be able to generate high resolution images

- Solution?
  - Pixel Space -> Learned Latent Space
  - Auto Econder

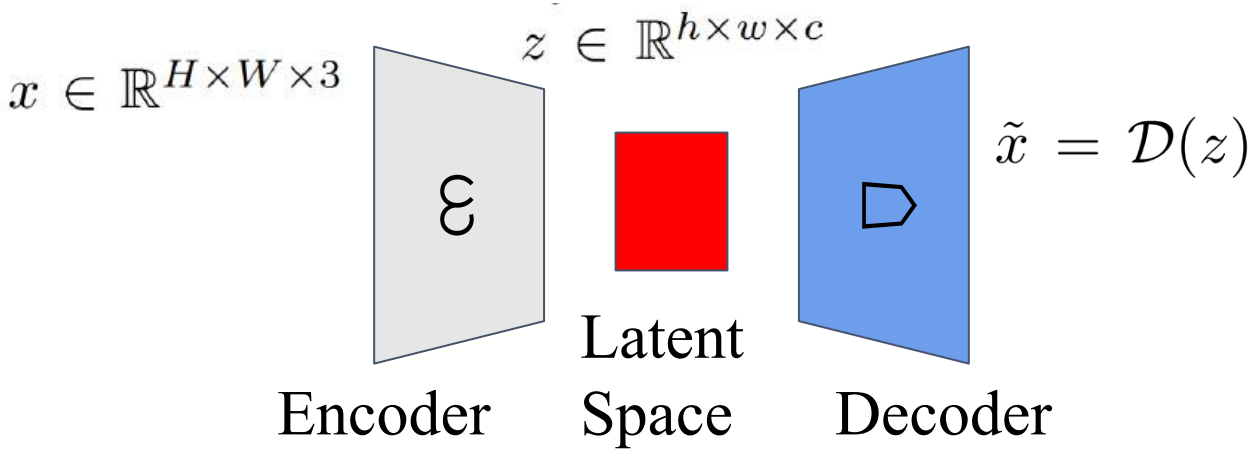
Auto Econder



$$f = \frac{H}{h} = \frac{W}{w}$$

Auto Econder

$$z = \mathcal{E}(x)$$



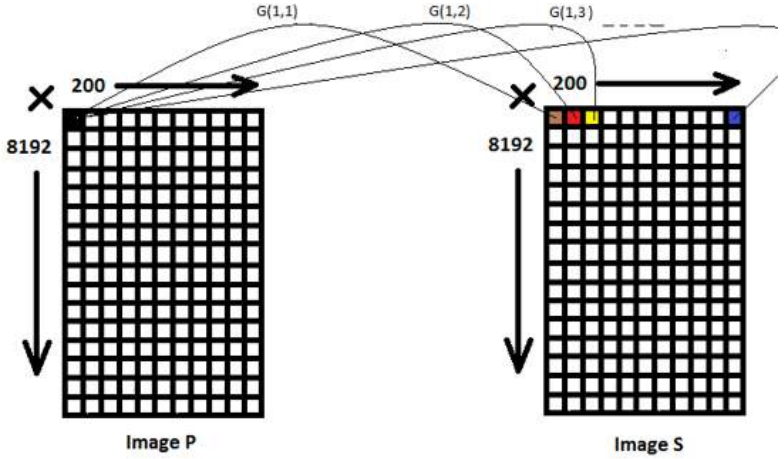
$$f = \frac{H}{h} = \frac{W}{w}$$

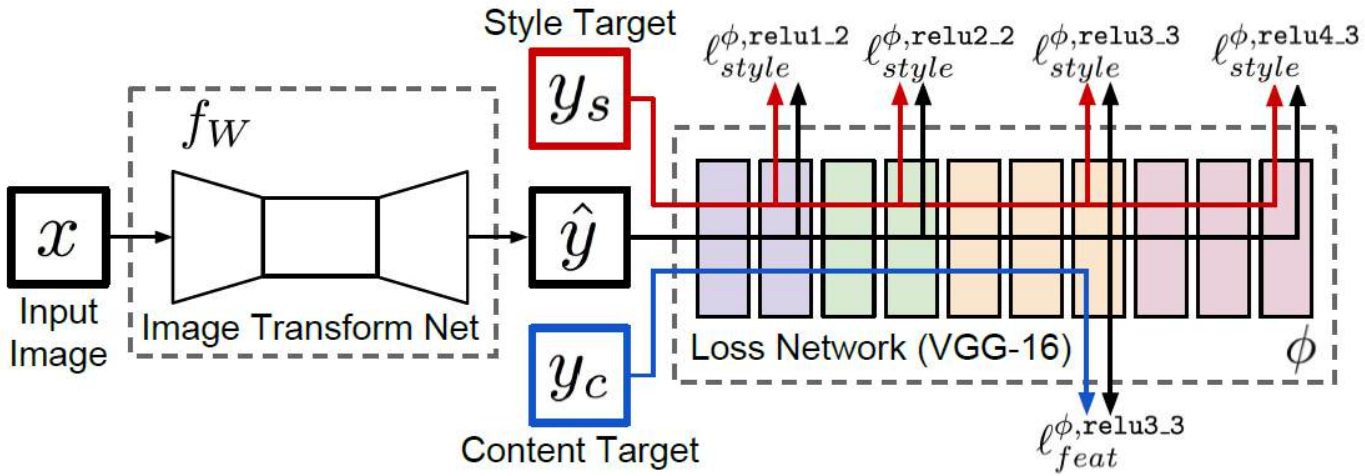
$$L_{\text{Autoencoder}} = \min_{\mathcal{E},\mathcal{D}} \max_{\psi} \Big(L_{rec}(x, \mathcal{D}(\mathcal{E}(x))) - L_{adv}(\mathcal{D}(\mathcal{E}(x))) + \log D_{\psi}(x) + L_{reg}(x; \mathcal{E}, \mathcal{D}) \Big)$$

Reconstruction Loss Adversarial Loss Regularization

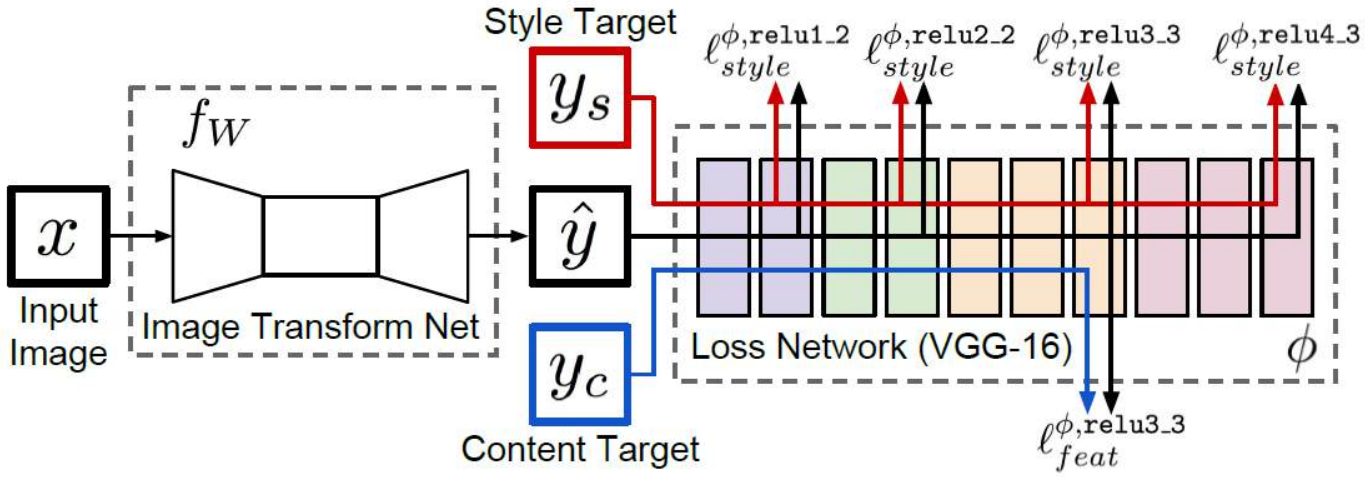
Reconstruction Loss

- L1/L2 + Perceptual
- Why Perceptual Loss
  - Style
  - Features
- How Perceptual Loss Works?





Perceptual Loss



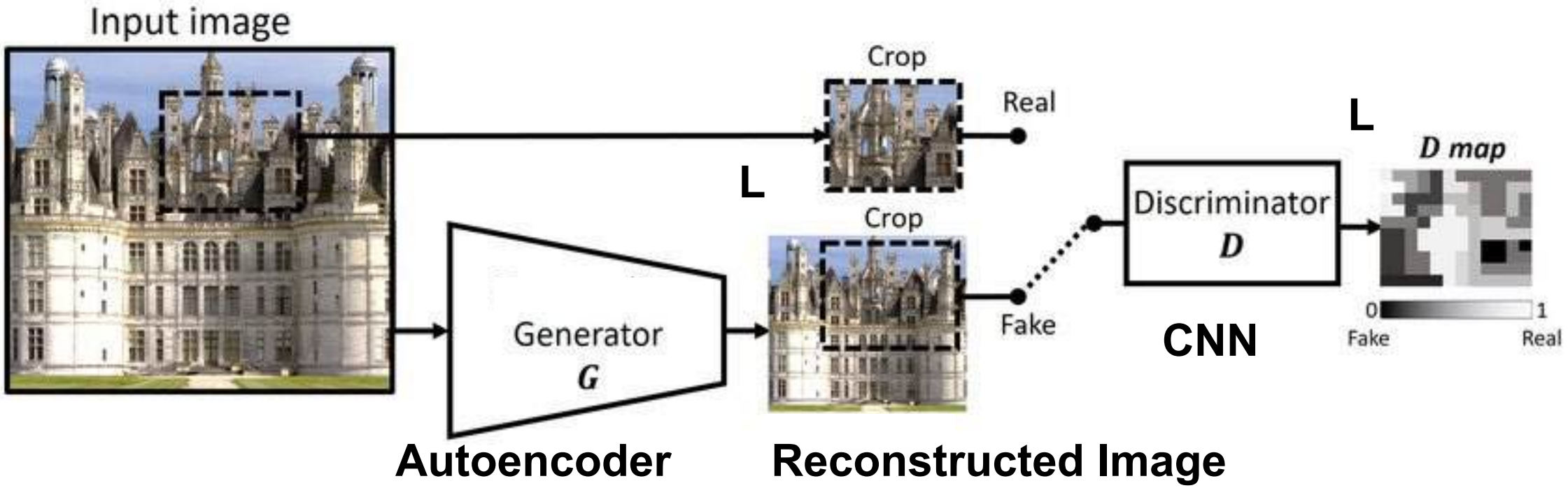
let $\phi_j(x)$ be the activations at the jth layer of the network ϕ

$$\ell^{\phi,j}_{feat}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \|\phi_j(\hat{y}) - \phi_j(y)\|_2^2$$

$$G^{\phi}_j(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'}.$$

$$\ell^{\phi,j}_{style}(\hat{y}, y) = \|G^{\phi}_j(\hat{y}) - G^{\phi}_j(y)\|_F^2.$$

Adversarial Loss



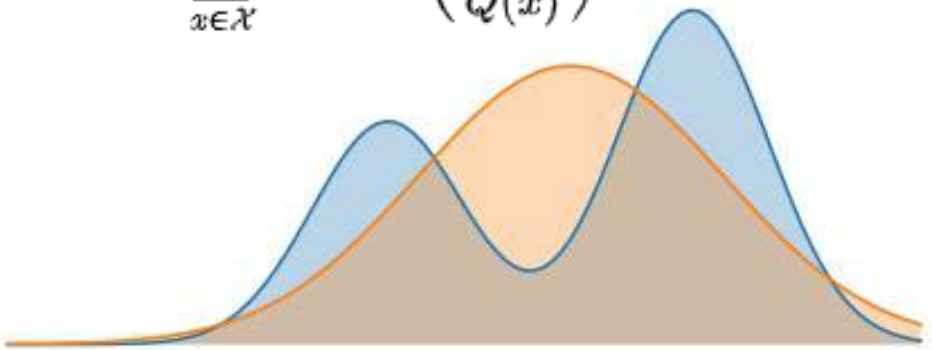
$$-L_{adv}(\mathcal{D}(\mathcal{E}(x))) + \log D_\psi(x)$$

$$-L_{adv}(\mathcal{D}(\mathcal{E}(x))) = -log[(D_\psi(D(\mathcal{E}(x))))]$$

Method - Regularization

- KL-Reg (Kullback Leibler)
  - Distribution Similarity

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$



To avoid arbitrarily scaled latent spaces, regularize the latent z

$$\mathcal{N}(z; 0, 1) \qquad \mathcal{N}(z; \mathcal{E}_{\mu}, \mathcal{E}_{\sigma^2})$$

Method

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]$$

Benefits

- Efficient
- Better semantic representation
- Train (Autoencoder) once!

Method - Conditioning

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t)$$

$$\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$$

$$W_Q^{(i)} \in \mathbb{R}^{d \times d_\tau}$$

$$K = W_K^{(i)} \cdot \tau_\theta(y)$$

$$W_K^{(i)} \in \mathbb{R}^{d \times d_\tau}$$

$$\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$$

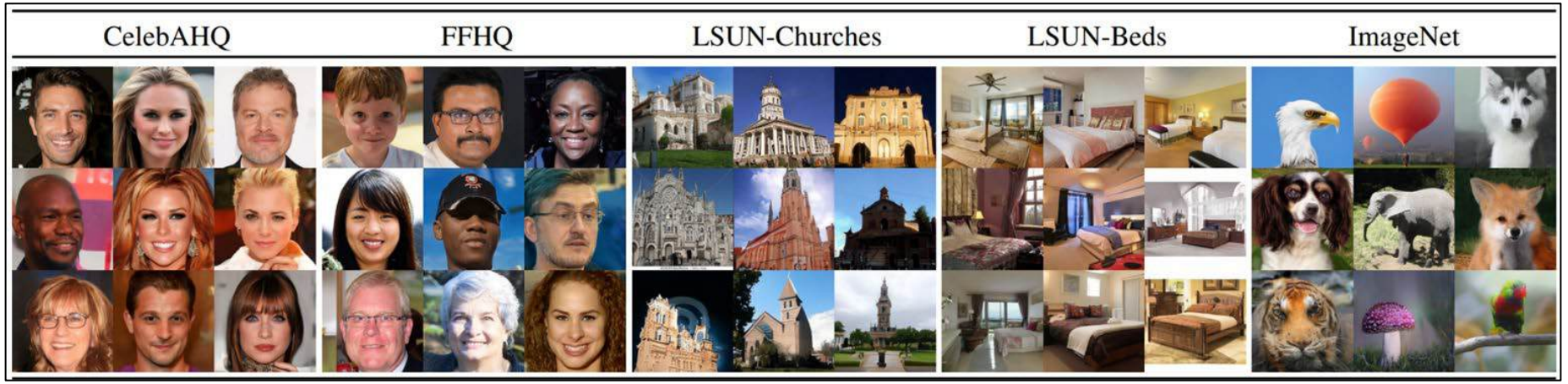
$$V = W_V^{(i)} \cdot \tau_\theta(y)$$

$$W_V^{(i)} \in \mathbb{R}^{d \times d_\epsilon^i}$$

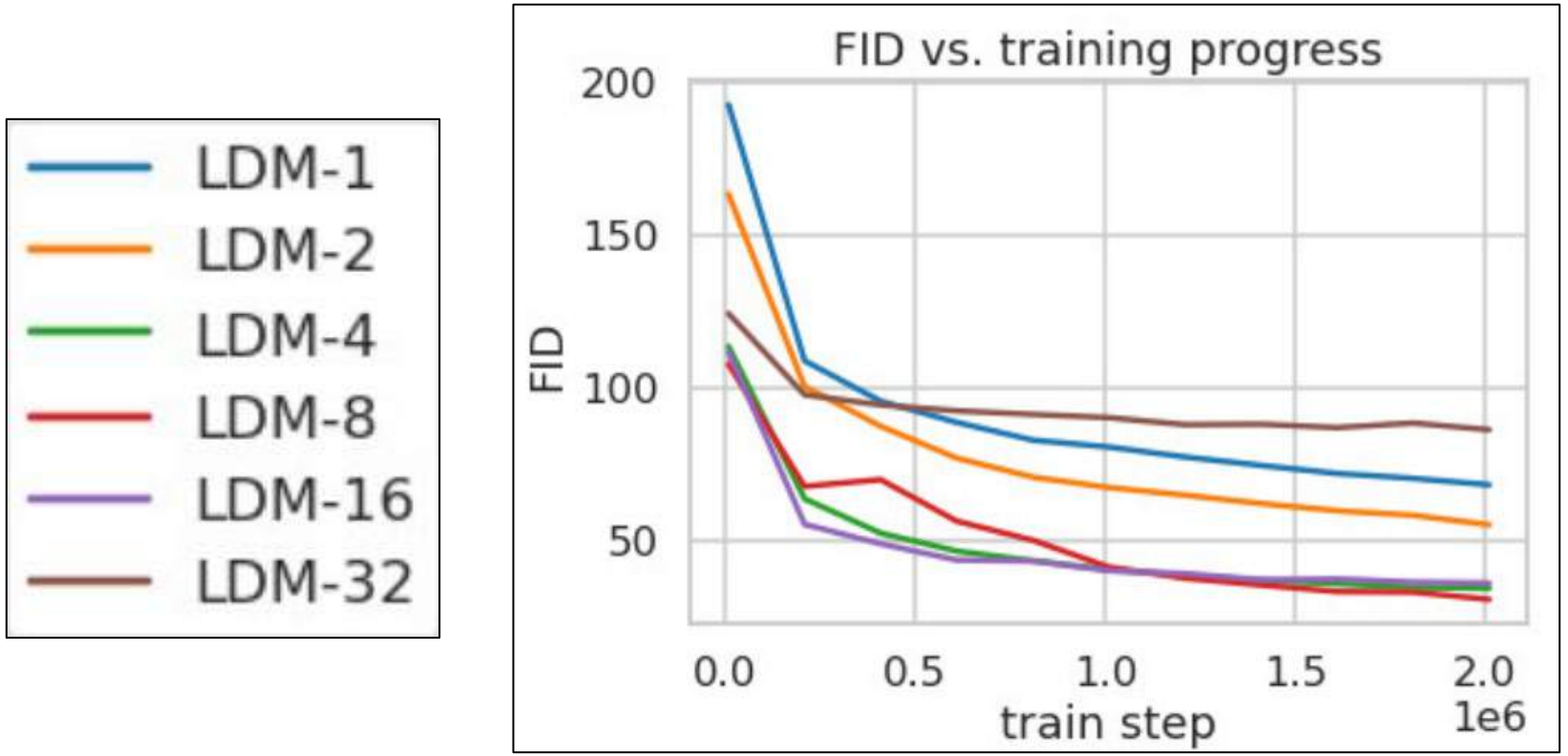
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V.$$

Experiments

- Conduct four experiments evaluate the performance of the LDMs

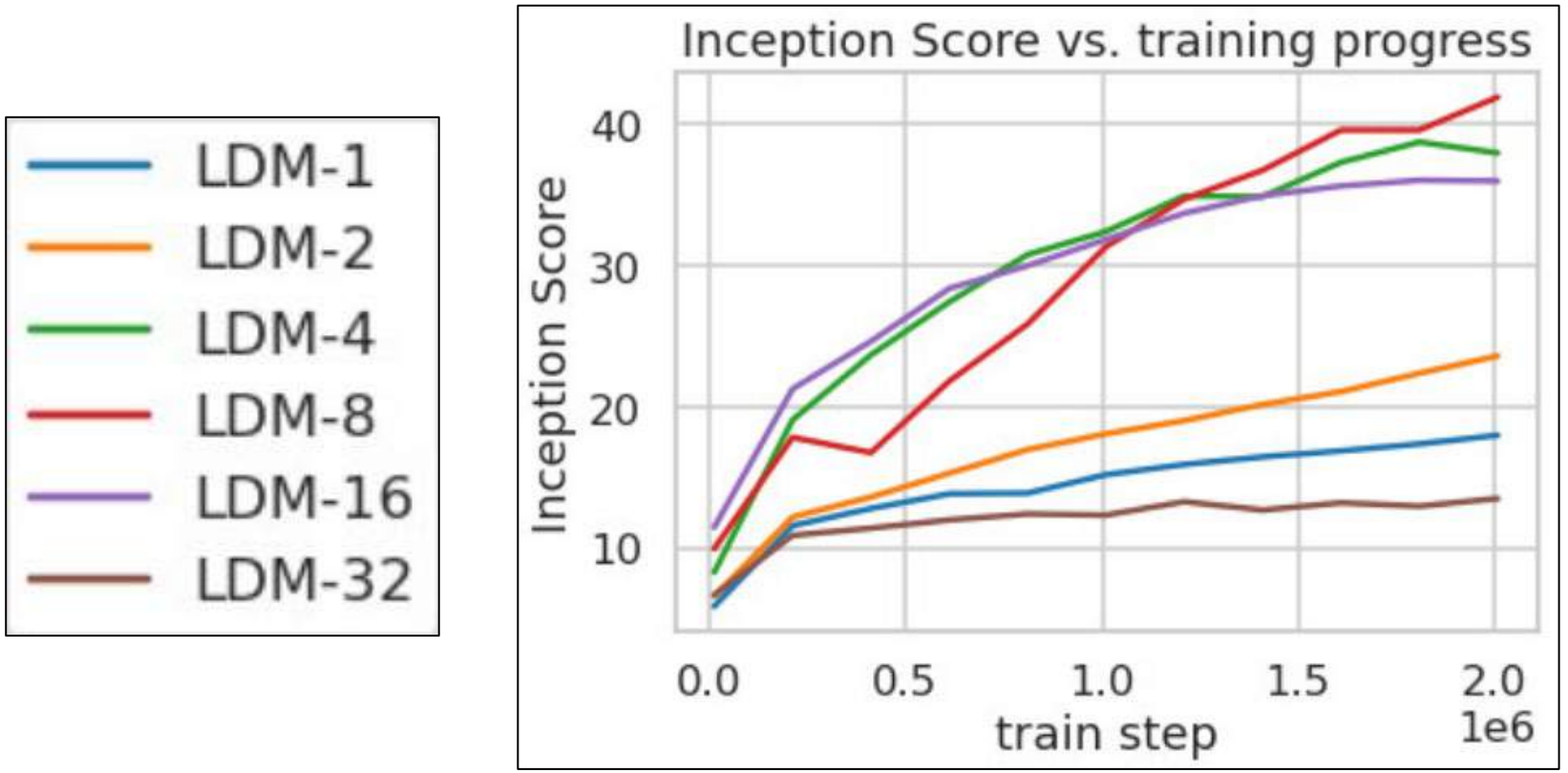


Experiments



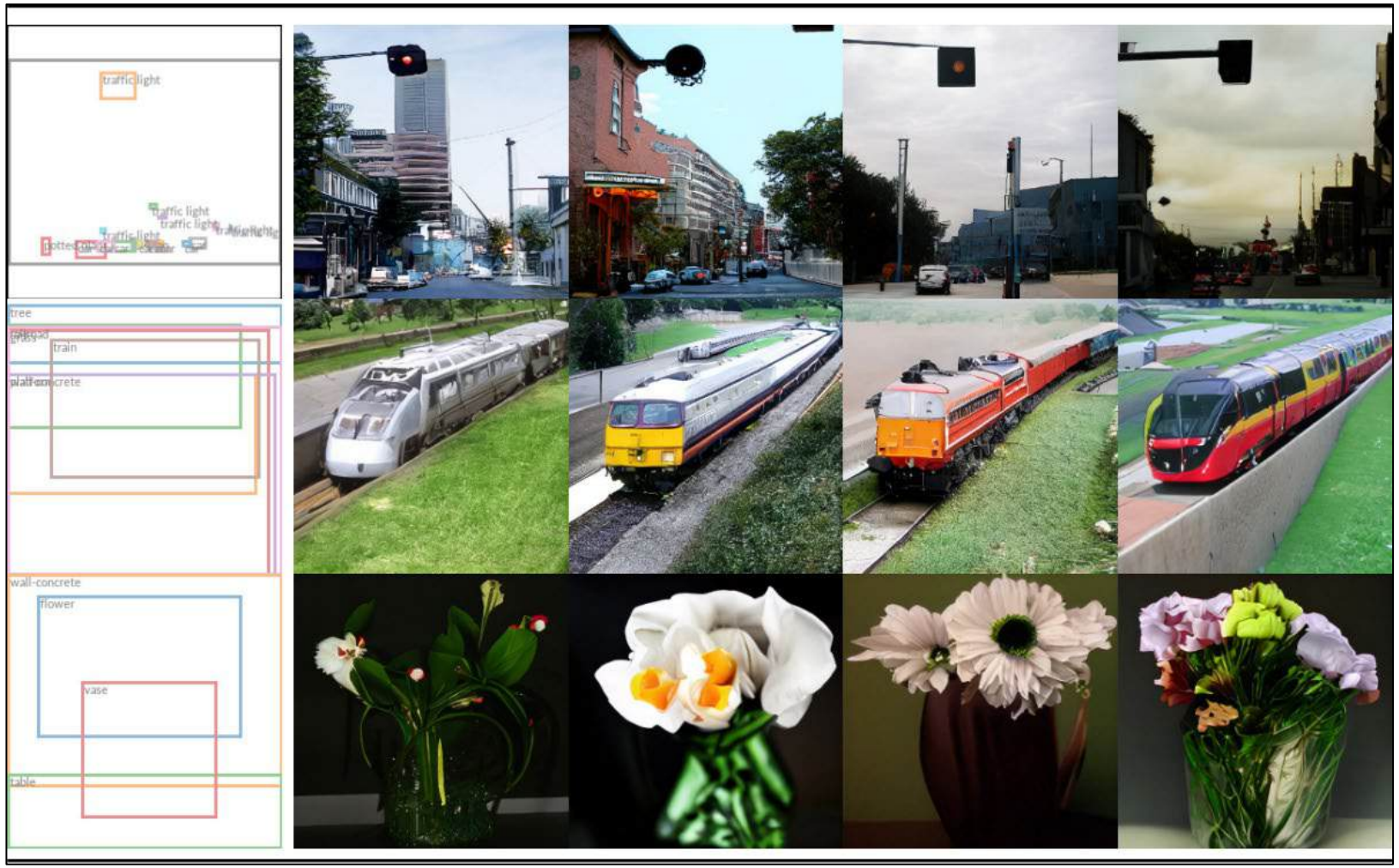
LDMs 4 and 8 produced the best FIDs

Experiments



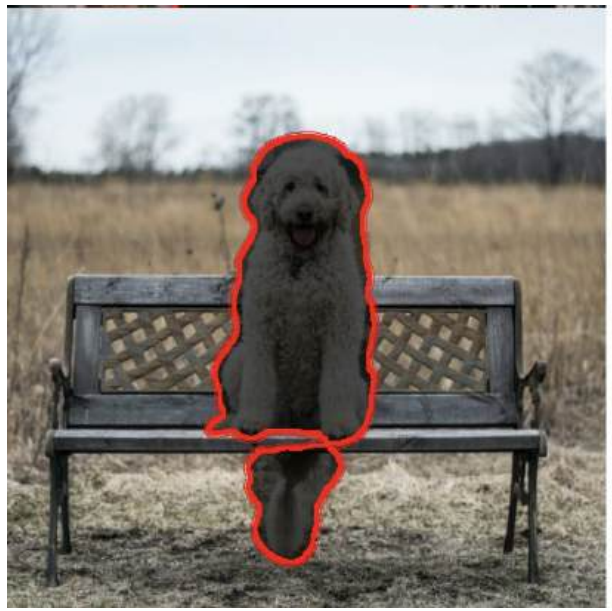
LDMs 4 and 8 produced the best Inception scores

Conditional Latent Diffusion



Inpainting Results















Data expansion via Diffusion for Clothes-Changing Person Re-ID

Nyle Siddiqui, Florinel Alin Croitoru, Gaurav Kumar Nayak,
Radu Tudor Ionescu, Mubarak Shah
(WACV-2025)

FEB 28 – MAR 4 TUCSON, ARIZONA, WACV 2025

Person Re-Identification

- To match an individual across images captured by cameras at different times and locations
- Re-ID is similar to geo-localization, here also aim is to learn robust embedding
- Cloth-Changing Re-ID is challenging



Clothes-Changing Re-Identification



Celeb Re-ID dataset (approved)

Cloth-Changing Re-ID Dataset



Cloth-changing person re-identification with self-attention, V Bansal, GL Foresti, N Martinel
Proceedings of the IEEE/CVF Winter Conference on Applications …, 2022

Approach: Enrich Training Dataset by Generating samples





Batman shirt, Jeans, Gray shoes



White jacket, Black pants, White shoes



Brown shirt, Blue pants, Brown shoes, Belt



Generating Clothes-Changed Images

G00075 - "yes" to consent

Qualitative Results

original



controlled

generated











G00160 - "yes" to consent

Qualitative Results

original



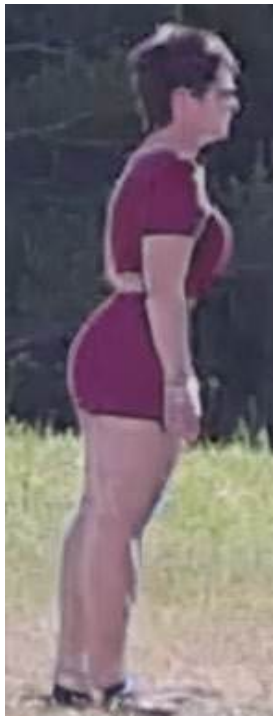
100m

generated











G00160 - "yes" to consent

Qualitative Results





PRCC Results

Model	Top-1	mAP
GI-ReID (CVPR '22)	37.5	-
AFD-Net (IJCAI '21)	42.8	-
RCSANet (ICCV '21)	50.2	49.6
3DSL (CVPR '21)	51.3	-
FSAM (CVPR '21)	54.5	-
CAL (baseline) (CVPR '22)	55.2	55.8
AIM (CVPR '23)	57.9	58.3
Ours	**66.5 (+8.6)**	**63.0 (+4.7)**

Our Method as a General Add-On (Data Expansion)

Standard Re-ID Models (Non-Clothes-Changing)		
Model	Top-1	mAP
PCB	41.8	38.7
PCB + Ours	**53.3 (+11.5)**	**50.7 (+12.0)**

Clothes-Changing Re-ID Models		
Model	Top-1	mAP

Contents

- **Stable Diffusion**
- Dream Booth
- Control Net

Contents

- Stable Diffusion
- **Dream Booth**
- Control Net

This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Nataniel Ruiz*,1,2 Yuanzhen Li[1] Varun Jampani[1]
Yael Pritch[1] Michael Rubinstein[1] Kfir Aberman[1]
[1] Google Research [2] Boston University



Figure 1. With just a few images (typically 3-5) of a subject (left), *DreamBooth*—our AI-powered photo booth—can generate a myriad of images of the subject in different contexts (right), using the guidance of a text prompt. The results exhibit natural interactions with the environment, as well as novel articulations and variation in lighting conditions, all while maintaining high fidelity to the key visual features of the subject.

Abstract

Large text-to-image models achieved a remarkable leap in the evolution of AI, enabling high-quality and diverse synthesis of images from a given text prompt. However, these models lack the ability to mimic the appearance of subjects in a given reference set and synthesize novel renditions of them in different contexts. In this work, we present a new approach for "personalization" of text-to-image diffusion models. Given as input just a few images of a subject, we fine-tune a pretrained text-to-image model such that it learns to bind a unique identifier with that specific subject. Once the subject is embedded in the output domain of the model, the unique identifier can be used to synthesize novel photorealistic images of the subject contextualized in different scenes. By leveraging the semantic prior embedded in the model with a new autogenous class-specific prior preservation loss, our technique enables synthesizing the subject in diverse scenes, poses, views and lighting conditions that do not appear in the reference images. We apply our technique to several previously-unassailable tasks, including subject recontextualization, text-guided view synthesis, and artistic rendering, all while preserving the subject's key features. We also provide a new dataset and evaluation protocol for this new task of subject-driven generation. Project page: https://dreambooth.github.io/

1. Introduction

Can you imagine your own dog traveling around the world, or your favorite bag displayed in the most exclusive showroom in Paris? What about your parrot being the main character of an illustrated storybook? Rendering such imaginary scenes is a challenging task that requires synthesizing instances of specific subjects (e.g., objects, animals) in new contexts such that they naturally and seamlessly blend into the scene.

Recently developed large text-to-image models have shown unprecedented capabilities, by enabling high-quality and diverse synthesis of images based on a text prompt written in natural language [51,58]. One of the main advantages of such models is the strong semantic prior learned from a large collection of image-caption pairs. Such a prior learns, for instance, to bind the word "dog" with various instances of dogs that can appear in different poses and contexts in an image. While the synthesis capabilities of these models are unprecedented, they lack the ability to mimic the appearance of subjects in a given reference set, and synthesize novel renditions of the *same subjects* in different contexts. The main reason is that the expressiveness of their output

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Nataniel Ruiz[1,2], Yuanzhen Li[1], Varun Jampani[1], Yael Pritch[1], Michael Rubinstein[1], and Kfir Aberman[1]

[1]*Google Research* [2]*Boston University*

CVPR-2023 Best Paper Award; Citations 2548

Presented By: Shane Davis, Joe Fioresi, Mitchell Klingler, & Nyle Siddiqui
https://www.youtube.com/watch?v=D641lhioXMc

Subject-Driven Generation – Personalization

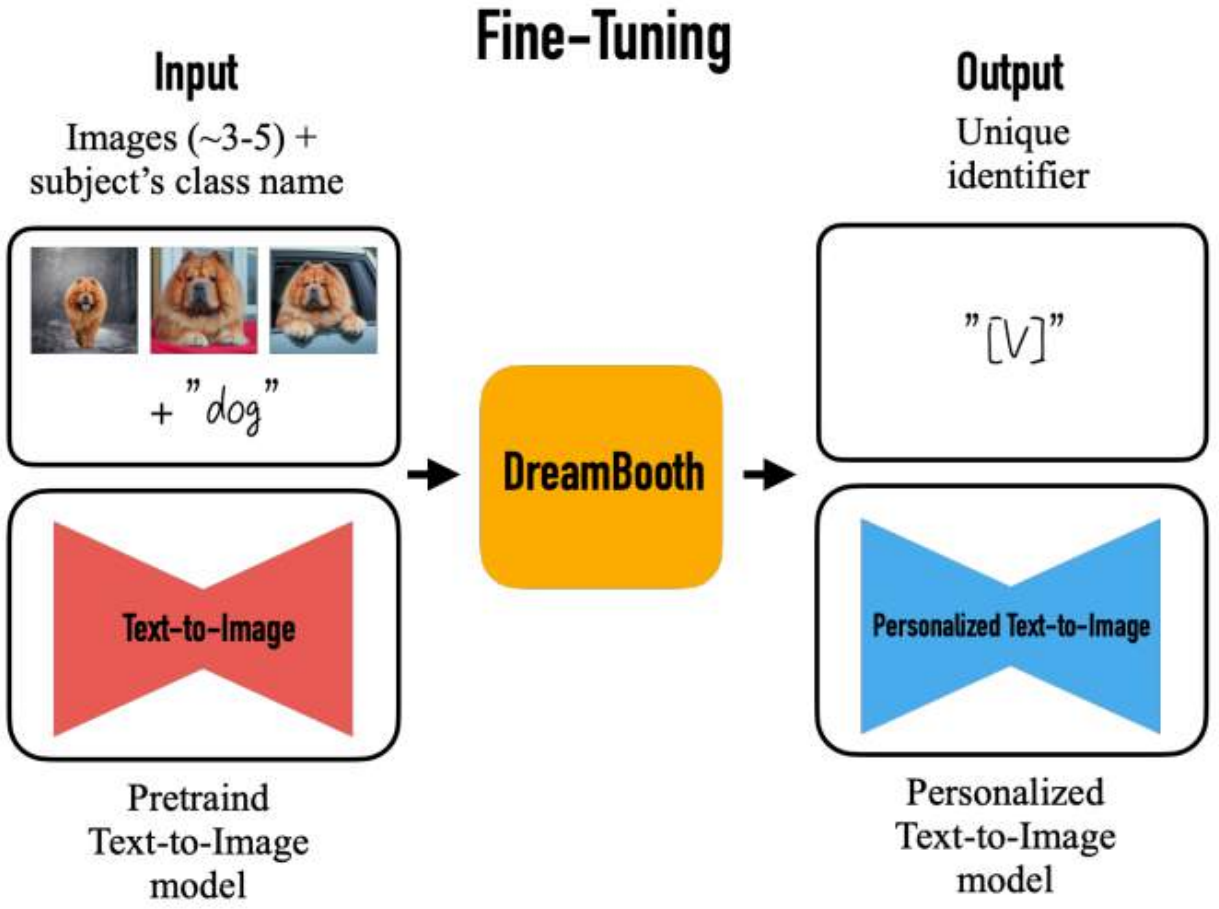


Input images

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2022). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242.

Methodology Overview

- Few-Shot text-guided diffusion
  - Fine-tune existing diffusion mode



Identified Gaps

- Most text-to-image generation models are unable to perform few-shot learning
  - **There is no efficient method to “inject” subjects into the model training**
    - Overfitting
    - Language Drift
    - Best ex.: GAN reproducing the same face given ~100 images [1]

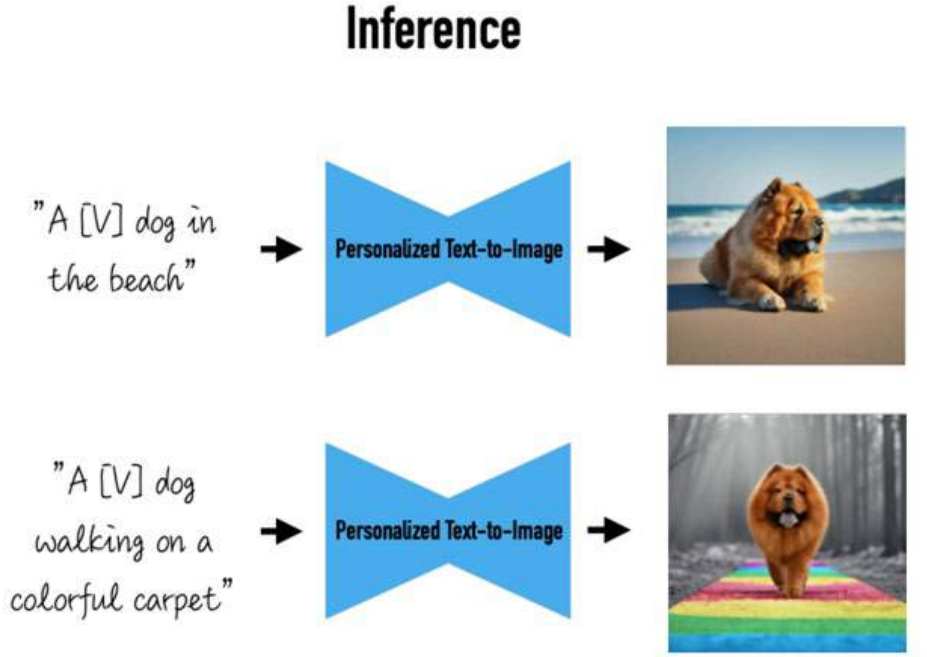
[1] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Rare-Token Identifier

- How to refer to a new subject?

“a [identifier] [class noun]”

- Diffusion model already has a ‘vocabulary’
  - Goal: Implant the subject into this vocabulary



Rare-Token Identifier: Naive Approaches

- Use English words to describe the new class
  - English lexicon has strong priors

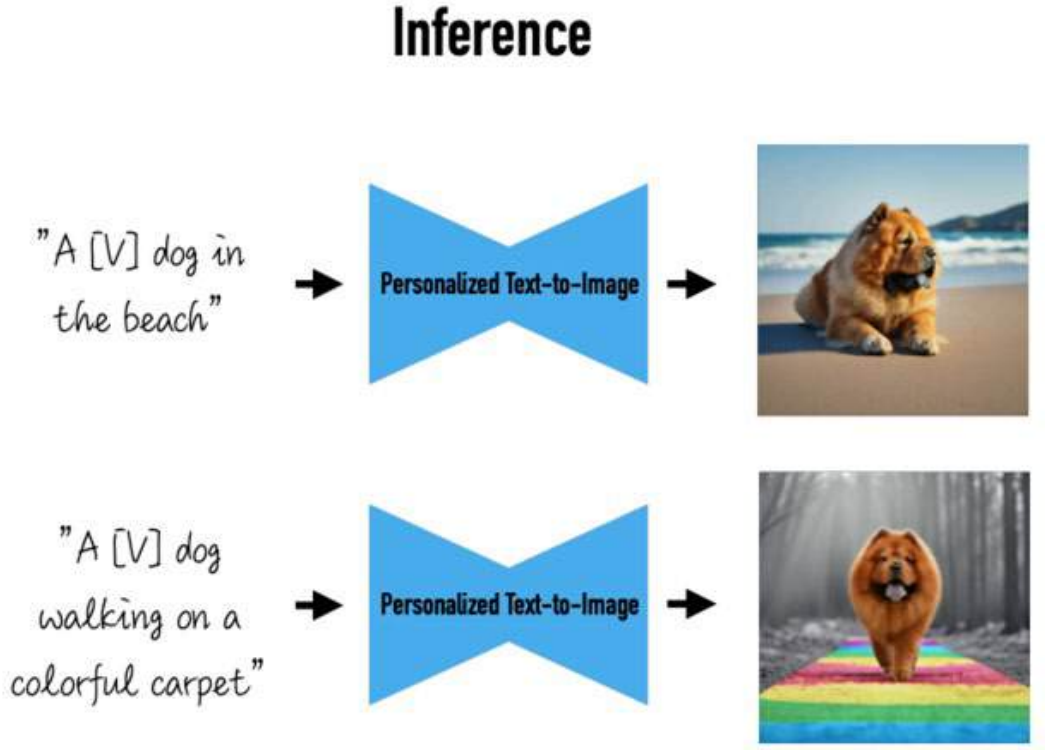
“unique” or “special”

- Generate a string of random characters
  - Nonsense identifier leads to literal artifacts

(e.g. “xxy5syt00”)

Rare-Token Identifier: Solution

1. Perform a rare-token lookup in the vocabulary
2. Invert the rare-token, resulting in plain text (1-3 characters)
3. Use this plain text as the **unique identifier (SKS)**



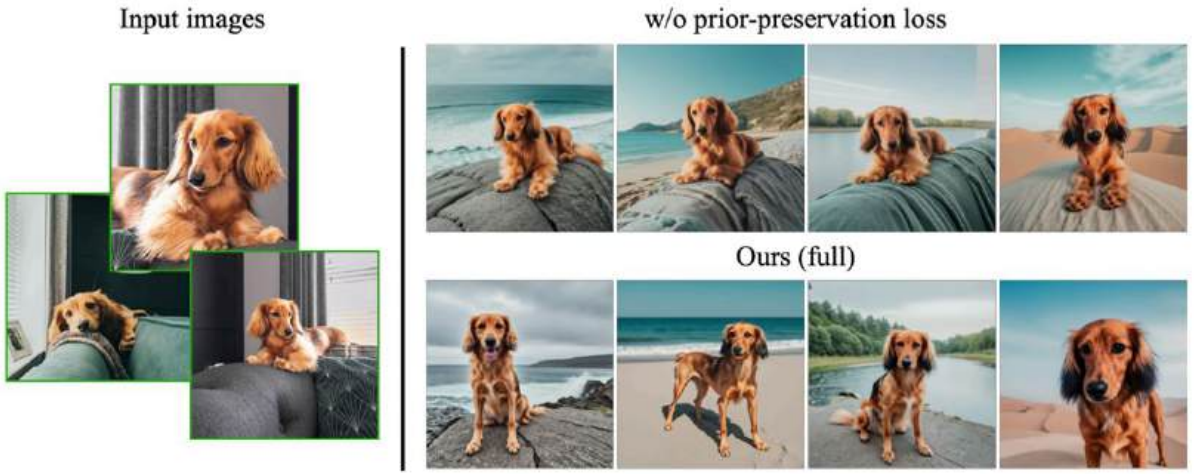
Identified Gaps

- Most text-to-image generation models are unable to perform few-shot learning
  - **There is no efficient method to “inject” subjects into the model training**
    - **Overfitting**
    - **Language Drift**
    - Best ex.: GAN reproducing the same face given ~100 images [1]

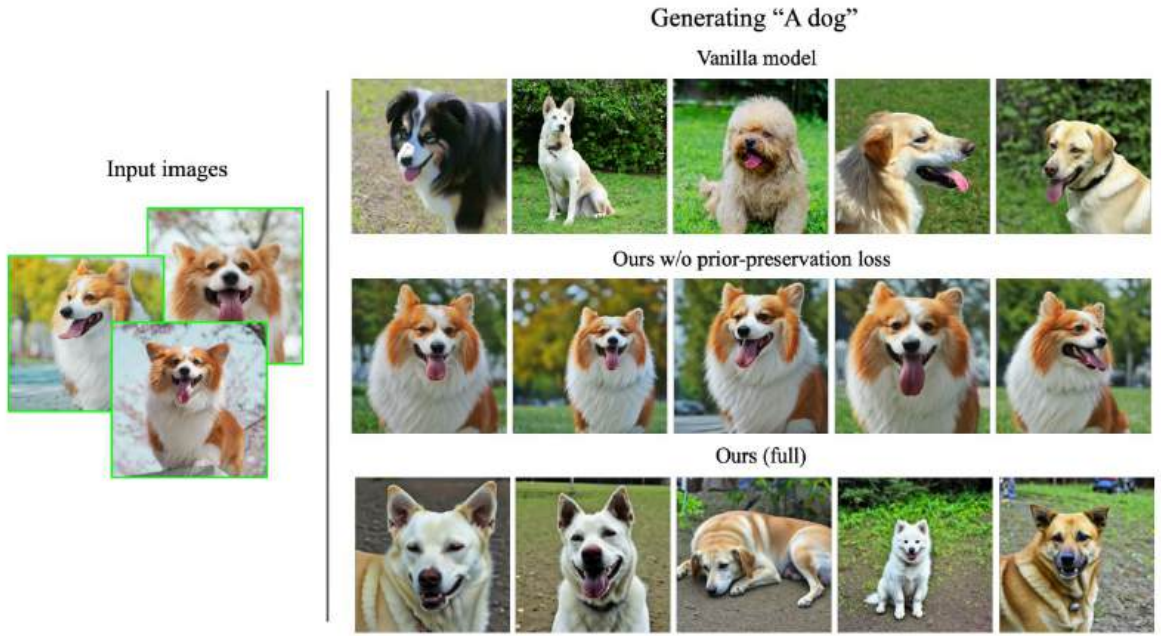
[1] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Autogenous Class-Specific Prior Preservation Loss

1. Overfitting

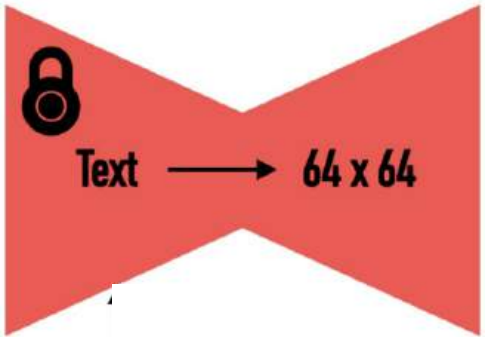


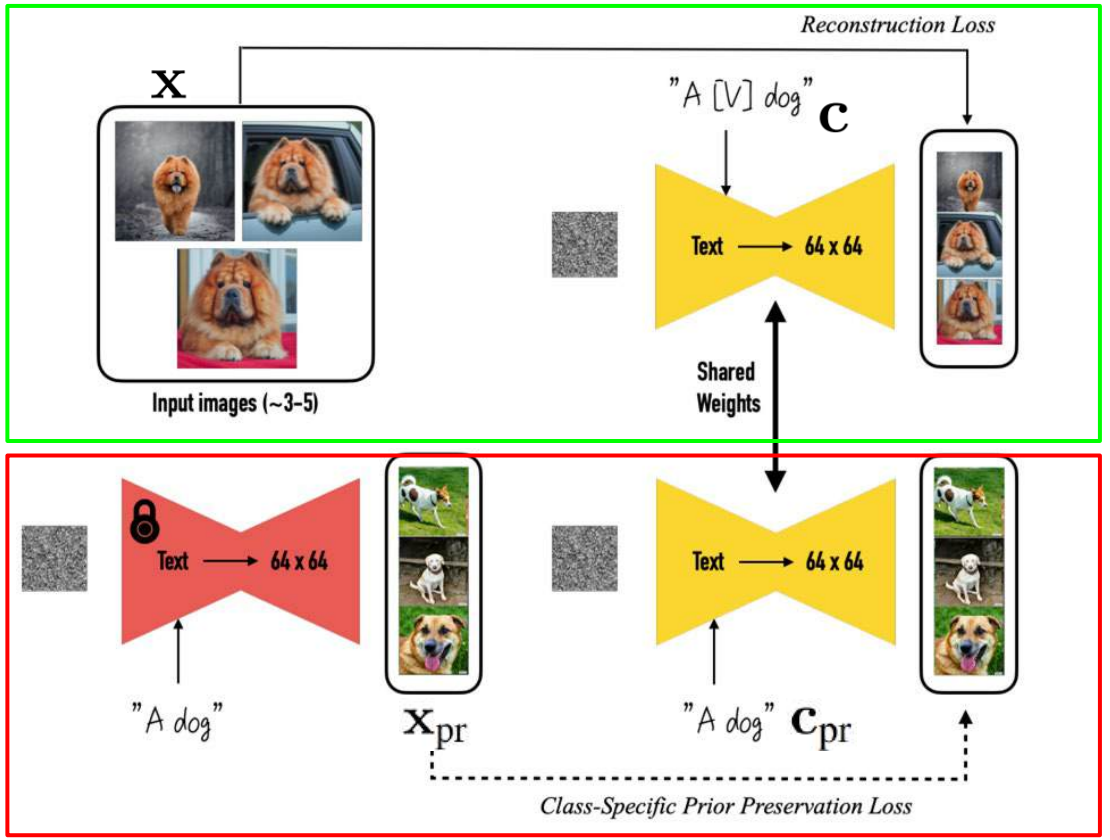
1. Language Drift
2. Autogenous: "self-produced”



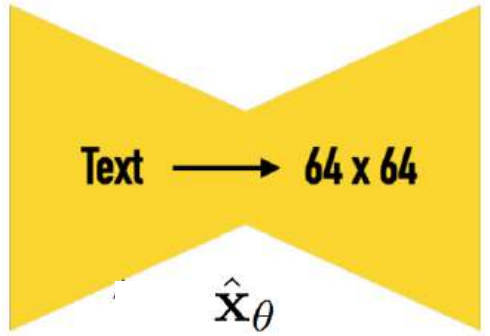
Autogenous Class-Specific Prior Preservation Loss

Frozen Imagen Model





Finetuned



Standard Diffusion Loss

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}\left[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2\right.$$

Autogenous Class-Specific Prior Preservation

$$\left.+\,\lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2\right]$$

Results

"A [V] dog in the beach"





Results: Recontextualization



A [V] backpack in the Grand Canyon





A [V] teapot pouring tea



A [V] vase in the ocean

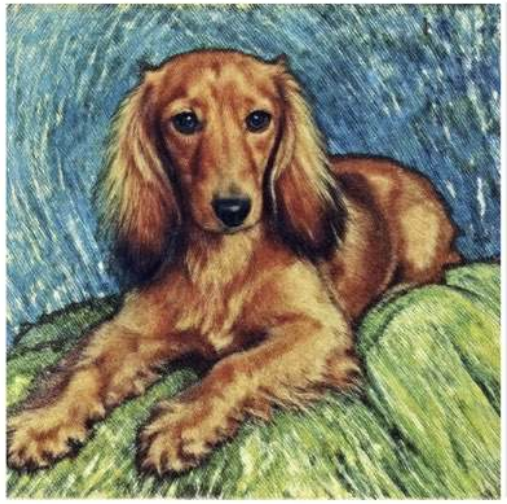




Results: Art Renditions



"A painting of a [V] dog in the style of Vincent Van Gogh"



"A painting of a [V] dog in the style of Leonardo Da Vinci"



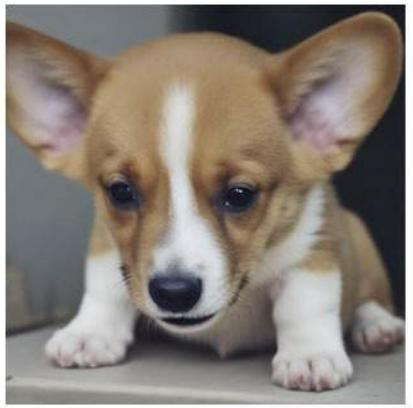
"A sculpture of a [V] dog in the style of Michelangelo



Results: Expression Manipulation



"A [state] [V] dog"



depressed



screaming



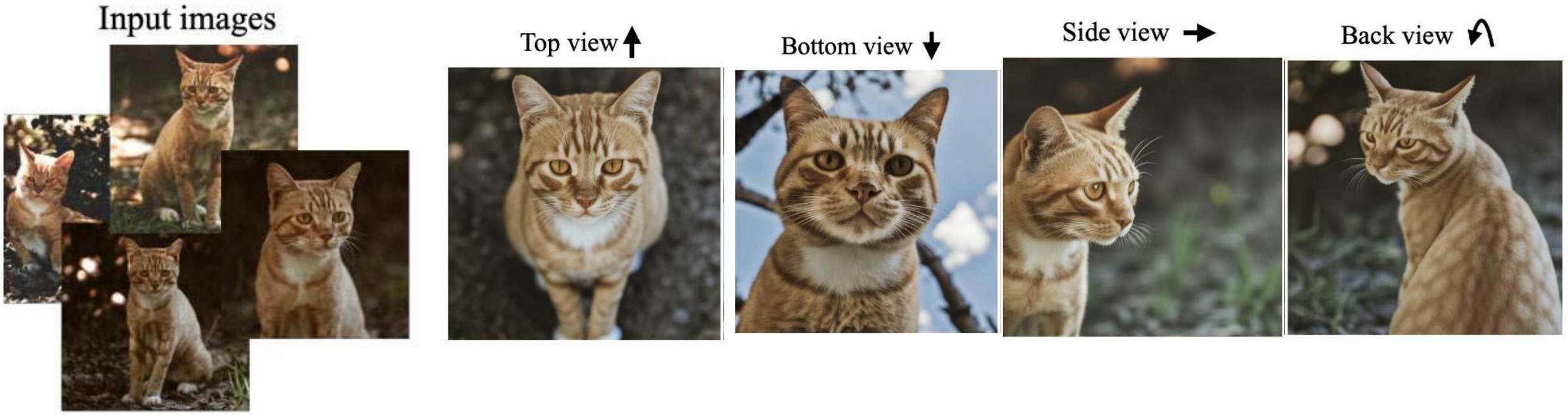
barking



sleeping

Results: View Synthesis

"[V] cat seen from the [direction]"



Results: Accessorization

Input images



"A [V] dog wearing a [accessory]"



Chef Outfit

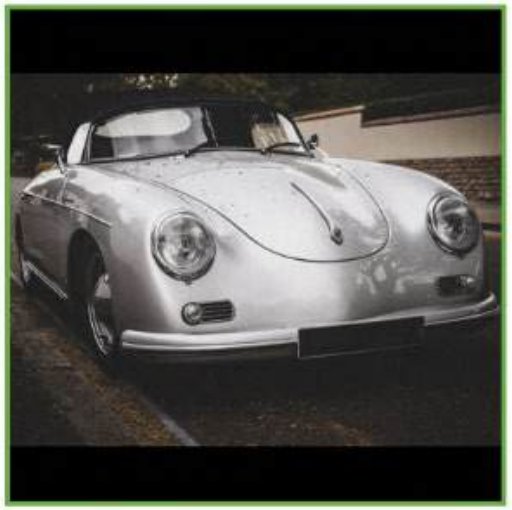


Purple Wizard Outfit



Angel Wings

Results: Property Modification



"A [color] [V] car"



purple



"A cross of a [V] dog and a [target species]"



hippo

Ablations - class noun prior

Input images



Fine-tuning

Inference

No class noun:
"A [V]"



A [V] | A [V] on top of blue fabric | A [V] with a river in the background

Incorrect class noun:
"A [V] dog"



A [V] dog | A [V] dog on top of blue fabric | A [V] dog with a river in the background

Correct class noun:
"A [V] sunglasses"



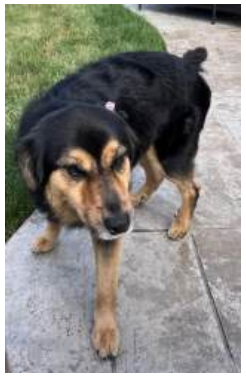
A [V] sunglasses | A [V] sunglasses on top of blue fabric | A [V] sunglasses with a river in the background

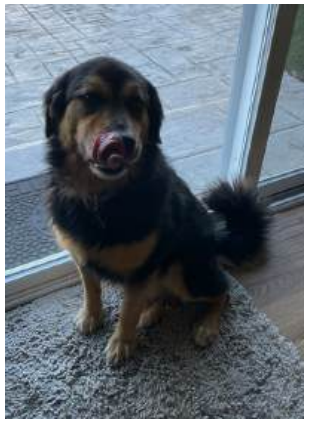
Experimentation

- Used Hugging Face's diffusers implementation of Dreambooth
  - Stable Diffusion v1.5
  - Rare Word Token Chosen: "sks" (List)
- Followed recommended hyperparameters

Results - Joe's Dog

Input Images:



Regularization Images (96 total):
"A photo of a dog"





"A photo of a sks dog at disney world":



"A photo of a sks dog running in the forest":



Results - Knightro

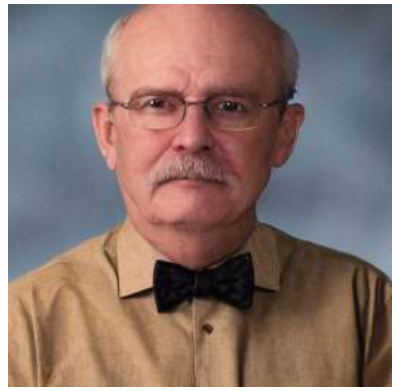
Input Images:







Regularization Images (200 total):
"A photo of a knight"


"A photo of a sks knight at the beach":



"A photo of a sks knight riding a bicycle":



Contents

- Stable Diffusion
- **Dreambooth**
- Control Net

Contents

- Stable Diffusion
- Dreambooth
- **Control Net**

David Mar Prize

[Cited by 3633]

Adding Conditional Control to Text-to-Image Diffusion Models

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala
Stanford University
{lvmin, anyirao, maneesh}@cs.stanford.edu

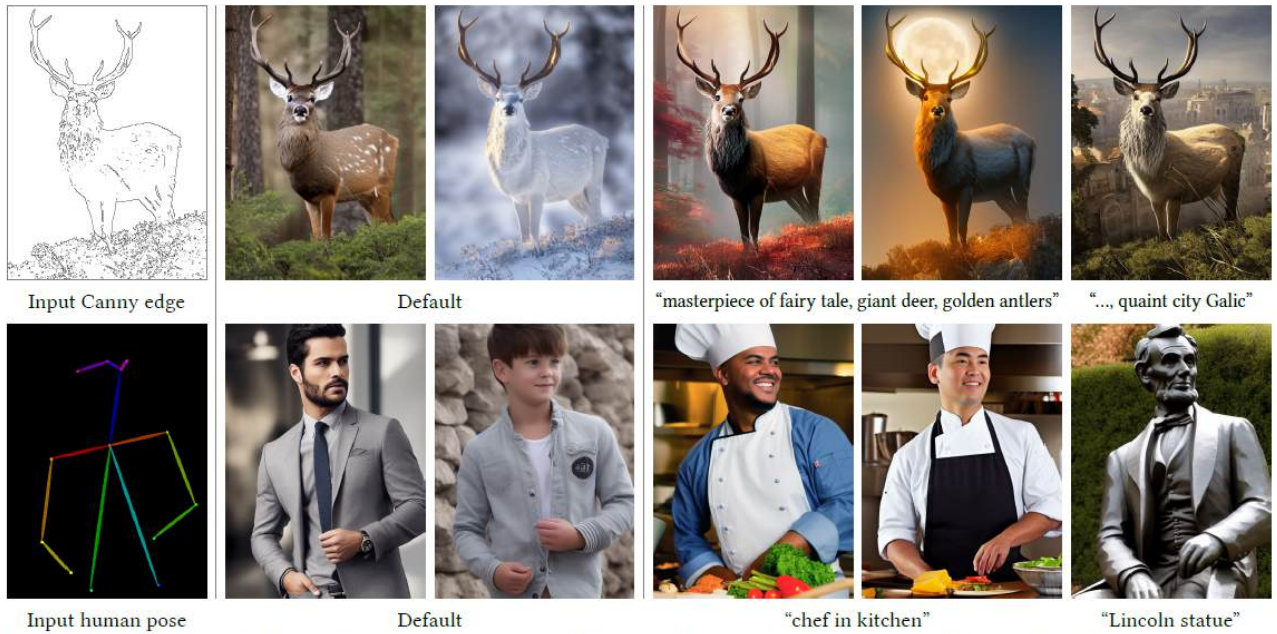


Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), *etc*., to control the image generation of large pretrained diffusion models. The default results use the prompt "a high-quality, detailed, and professional image". Users can optionally give prompts like the "chef in kitchen".

Abstract

We present ControlNet, a neural network architecture to add spatial conditioning controls to large, pretrained text-to-image diffusion models. ControlNet locks the production-ready large diffusion models, and reuses their deep and robust encoding layers pretrained with billions of images as a strong backbone to learn a diverse set of conditional controls. The neural architecture is connected with "zero convolutions" (zero-initialized convolution layers) that progressively grow the parameters from zero and ensure that no harmful noise could affect the finetuning. We test various conditioning controls, e.g., edges, depth, segmentation, human pose, etc., with

1. Introduction

Many of us have experienced flashes of visual inspiration that we wish to capture in a unique image. With the advent of text-to-image diffusion models [54, 62, 72], we can now create visually stunning images by typing in a text prompt. Yet, text-to-image models are limited in the control they provide over the spatial composition of the image; precisely expressing complex layouts, poses, shapes and forms can be difficult via text prompts alone. Generating an image that accurately matches our mental imagery often requires numerous trial-and-error cycles of editing a prompt, inspecting the resulting images and then re-editing the prompt.

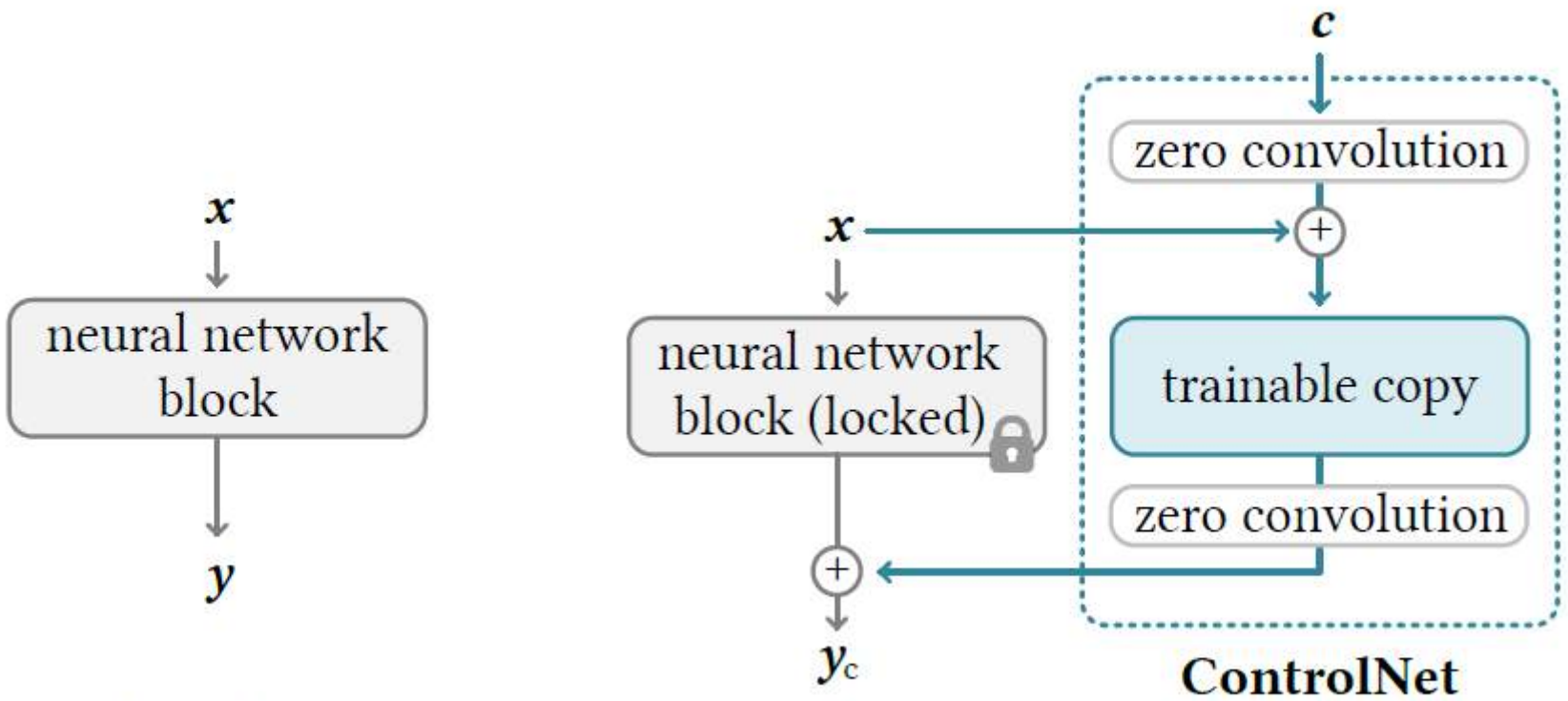
Can we enable finer-grained spatial control by letting

Control Net

- Text-to-image models are limited in the control they provide over the spatial composition of the image
- Enable finer grained spatial control by:
  - Let users provide additional images that directly specify their desired image composition:
    - Edge maps,
    - human pose skeletons,
    - segmentation maps,
    - depth,
    - normals

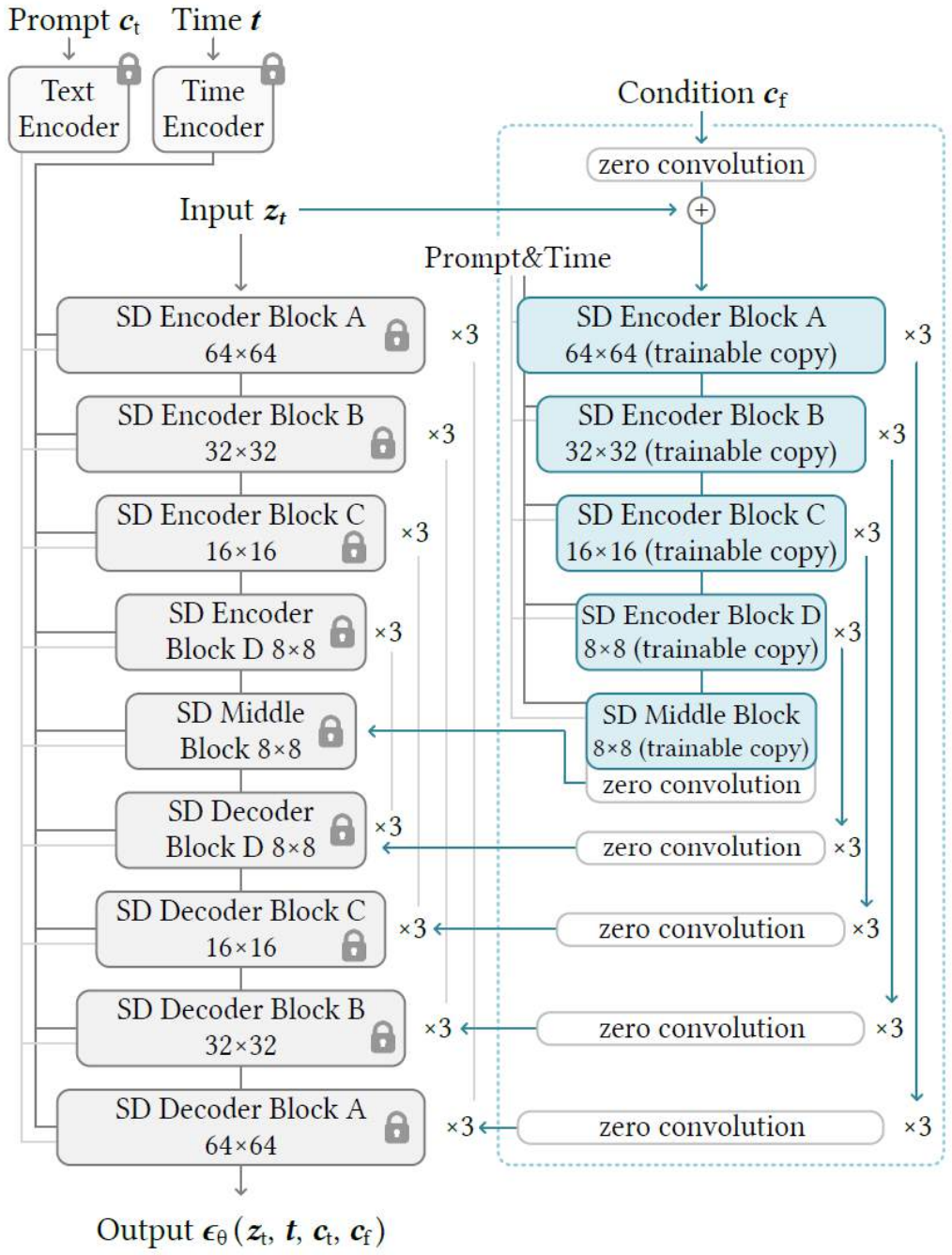
ControlNet

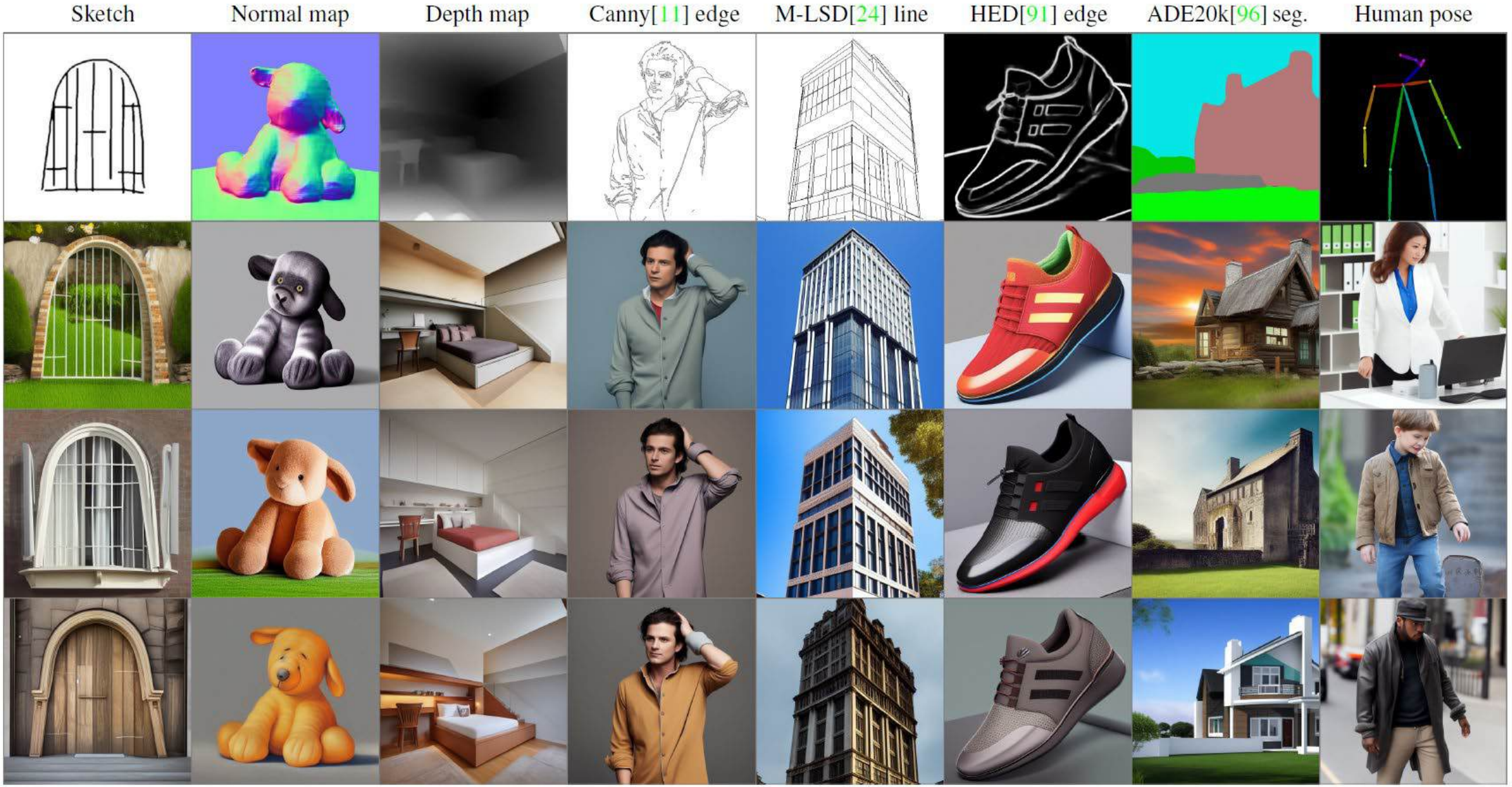
- An end-to-end neural network architecture that learns conditional controls for large pretrained text-to-image diffusion models:
    - Stable Diffusion



$$\boldsymbol{y} = \mathcal{F}(\boldsymbol{x}; \Theta).$$

$$\boldsymbol{y}_c = \mathcal{F}(\boldsymbol{x}; \Theta) + \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}; \Theta_{z1}); \Theta_c); \Theta_{z2}),$$


Prompt c_t
Time t
Text Encoder
Time Encoder
Condition c_f
zero convolution
Input z_t
Prompt&Time
SD Encoder Block A 64×64
SD Encoder Block A 64×64 (trainable copy)
SD Encoder Block B 32×32
SD Encoder Block B 32×32 (trainable copy)
SD Encoder Block C 16×16
SD Encoder Block C 16×16 (trainable copy)
SD Encoder Block D 8×8
SD Encoder Block D 8×8 (trainable copy)
SD Middle Block 8×8
SD Middle Block 8×8 (trainable copy)
zero convolution
SD Decoder Block D 8×8
SD Decoder Block C 16×16
SD Decoder Block B 32×32
SD Decoder Block A 64×64
×3
Output $\epsilon_\theta(z_t, t, c_t, c_f)$
(a) Stable Diffusion
(b) ControlNet


Sketch
Normal map
Depth map
Canny[11] edge
M-LSD[24] line
HED[91] edge
ADE20k[96] seg.
Human pose



Input canny map
Same CFG scale (9.0)
Same DDIM sampler
Same default prompt setting
("detailed high-quality professional image" without mentioning image contents)












without ControlNet
(using Stability's "official" method to add the channels to input layer, same as their depth-to-image structure)

SD + ControlNet

Generative AI (Computer Vision)

- **Text**
  - Large Language Models (LLM)
    - Generate/Predict the next word
    - Auto-regressive
  - **Visual Language Models (VLM)/Large Multimodal Models (LMM)**

- **Images, Videos, Audio,…**
  - **Diffusion Models**
    - **Generate Images**, Videos, audio, …

Can we do both Text and Image Generation?

NExT-GPT: Any-to-Any Multimodal LLM

Shengqiong Wu Hao Fei* Leigang Qu Wei Ji Tat-Seng Chua
NExT++, School of Computing, National University of Singapore

Project: https://next-gpt.github.io/

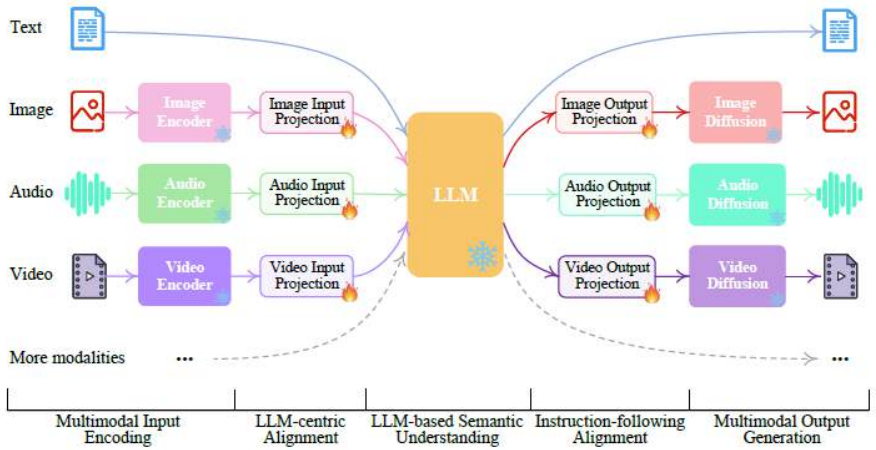
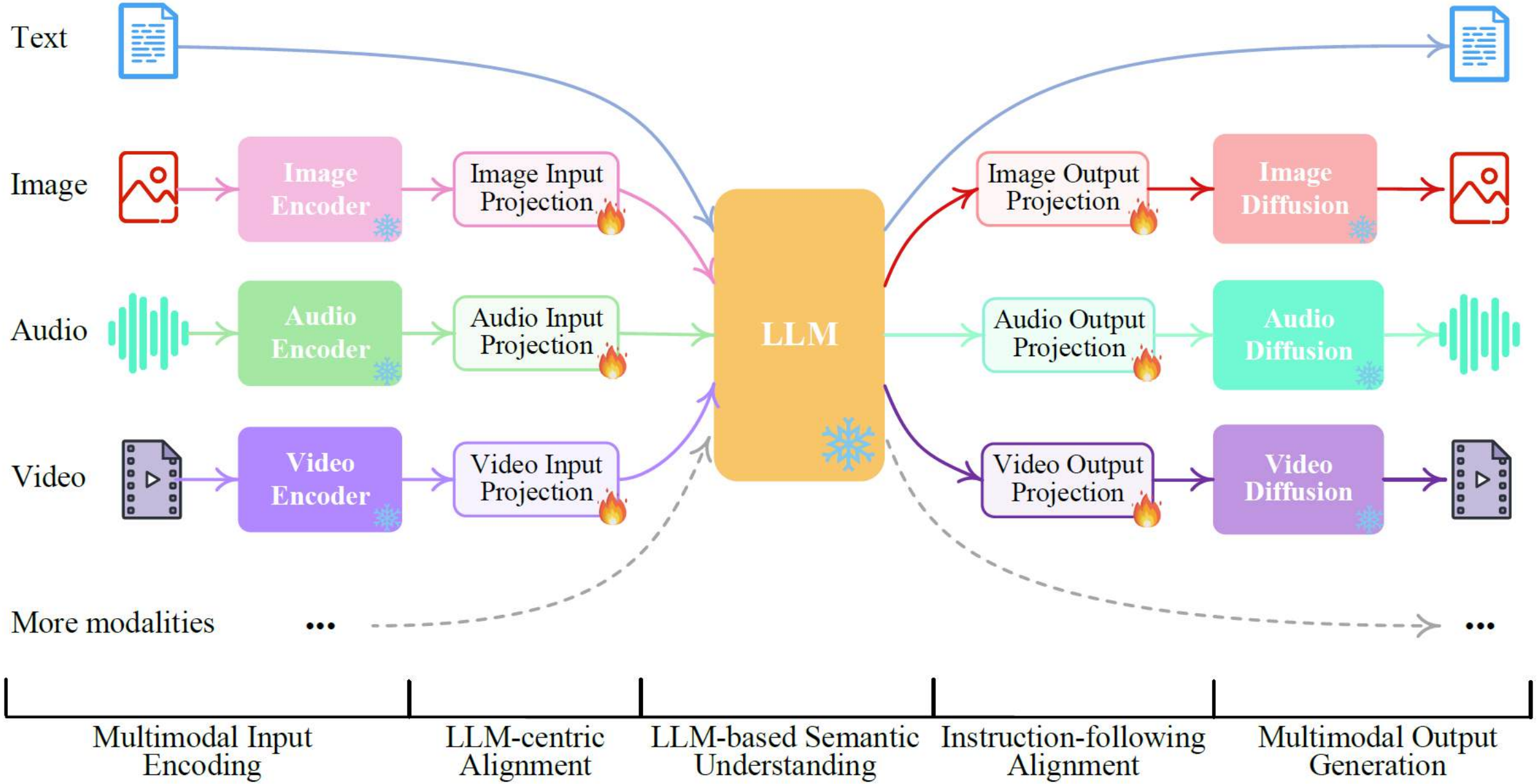


Figure 1: By connecting LLM with multimodal adaptors and diffusion decoders, NExT-GPT achieves universal multimodal understanding and any-to-any modality input and output.

Abstract

While recently Multimodal Large Language Models (MM-LLMs) have made exciting strides, they mostly fall prey to the limitation of only input-side multimodal understanding, without the ability to produce content in multiple modalities. As we humans always perceive the world and communicate with people through various modalities, developing any-to-any MM-LLMs capable of accepting and delivering content in any modality becomes essential to human-level AI. To fill the gap, we present an end-to-end general-purpose any-to-any MM-LLM system, **NExT-GPT**. We connect an LLM with multimodal adaptors and different diffusion decoders, enabling NExT-GPT to perceive inputs and generate outputs in arbitrary combinations of text, images, videos, and audio. By leveraging the existing well-trained highly-performing encoders and decoders, NExT-GPT is tuned with only a small amount of parameter (1%) of certain projection layers, which not only benefits low-cost training and also facilitates convenient expansion to more potential modalities. Moreover, we introduce a modality-switching instruction tuning (MosIT) and manually curate a high-quality dataset for MosIT, based on which NExT-GPT is empowered with complex cross-modal semantic understanding and content generation. Overall, our research showcases the promising possibility of building a unified AI agent capable of modeling universal modalities, paving the way for more human-like AI research in the community.

*Hao Fei is the corresponding author: haofei37@nus.edu.sg


Text
Image
Audio
Video
More modalities
Image Encoder
Audio Encoder
Video Encoder
Image Input Projection
Audio Input Projection
Video Input Projection
LLM
Image Output Projection
Audio Output Projection
Video Output Projection
Image Diffusion
Audio Diffusion
Video Diffusion
Multimodal Input Encoding
LLM-centric Alignment
LLM-based Semantic Understanding
Instruction-following Alignment
Multimodal Output Generation

Can we do both using one transformer?

SHOW-O: ONE SINGLE TRANSFORMER TO UNIFY MULTIMODAL UNDERSTANDING AND GENERATION

Jinheng Xie[1†] **Weijia Mao**[1†] **Zechen Bai**[1†] **David Junhao Zhang**[1†] **Weihao Wang**[2]
Kevin Qinghong Lin[1] **Yuchao Gu**[1] **Zhijie Chen**[2] **Zhenheng Yang**[2] **Mike Zheng Shou**[1*]

[1] Show Lab, National University of Singapore [2] ByteDance

ABSTRACT

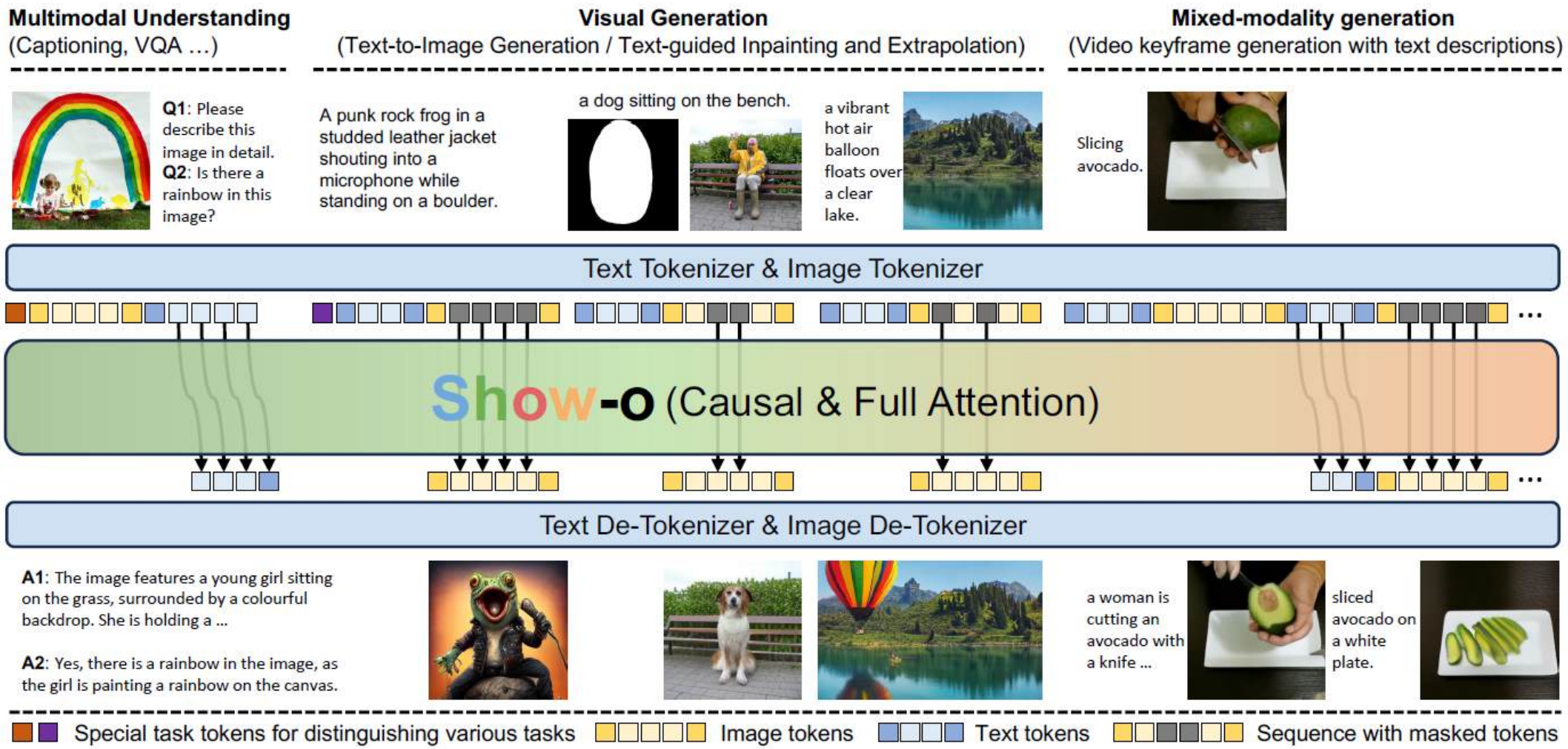
We present a unified transformer, *i.e.,* Show-o, that unifies multimodal understanding and generation. Unlike fully autoregressive models, Show-o unifies autoregressive and (discrete) diffusion modeling to adaptively handle inputs and outputs of various and mixed modalities. The unified model flexibly supports a wide range of vision-language tasks including visual question-answering, text-to-image generation, text-guided inpainting/extrapolation, and mixed-modality generation. Across various benchmarks, it demonstrates comparable or superior performance to existing individual models with an equivalent or larger number of parameters tailored for understanding or generation. This significantly highlights its potential as a next-generation foundation model. Code and models are released at https://github.com/showlab/Show-o.

1 INTRODUCTION

"Alone we can do so little; together we can do so much." – Helen Keller

Over the past few years, significant advancements have blossomed in the two key pillars of multimodal intelligence: understanding and generation (as depicted in Fig. 1(a) and (b)). For multimodal understanding, Multimodal Large Language Models (MLLMs) like LLaVA (Liu et al., 2024c) have demonstrated exceptional capabilities in vision-language tasks such as visual question-answering (VQA). For the other pillar of visual generation, denoising diffusion probabilistic models (DDPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020) have revolutionized the traditional generative paradigms (Kingma & Welling, 2013; Goodfellow et al., 2014), achieving unprecedented performance in text-to-image/video generation (Podell et al., 2023; Esser et al., 2024; Ho et al., 2022; Wu et al., 2023a).

Given these achievements in individual fields, it is only natural to explore the potential of connecting them. Recent works (Wu et al., 2023b; Ge et al., 2024; Ye et al., 2024a; Tang et al., 2024) have tried to assemble expert models from these two different domains to form a unified system that can handle both multimodal understanding and generation. However, existing attempts mainly treat each domain independently and often involve individual models responsible for understanding and generation separately (as shown on the left of Fig. 1(c)). For instance, NExT-GPT (Wu et al., 2023b) employs a base language model for multimodal understanding but requires an additional pre-


Multimodal Understanding
(Captioning, VQA …)
Visual Generation
(Text-to-Image Generation / Text-guided Inpainting and Extrapolation)
Mixed-modality generation
(Video keyframe generation with text descriptions)
Q1: Please describe this image in detail.
Q2: Is there a rainbow in this image?
A punk rock frog in a studded leather jacket shouting into a microphone while standing on a boulder.
a dog sitting on the bench.
a vibrant hot air balloon floats over a clear lake.
Slicing avocado.
Text Tokenizer & Image Tokenizer
Show-o (Causal & Full Attention)
Text De-Tokenizer & Image De-Tokenizer
A1: The image features a young girl sitting on the grass, surrounded by a colourful backdrop. She is holding a ...
A2: Yes, there is a rainbow in the image, as the girl is painting a rainbow on the canvas.
a woman is cutting an avocado with a knife ...
sliced avocado on a white plate.
Special task tokens for distinguishing various tasks
Image tokens
Text tokens
Sequence with masked tokens