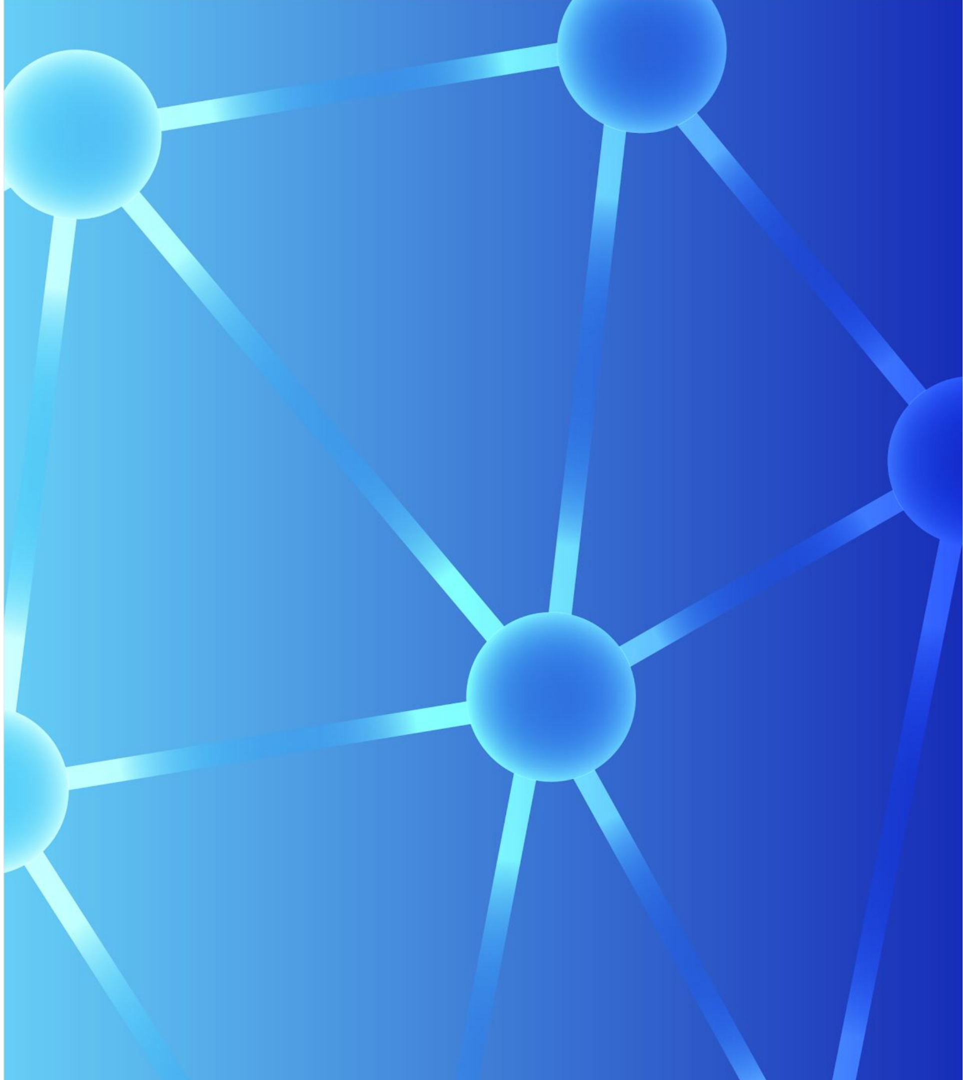
CVPR 2025 Tutorial: Efficient Text-to-Image/Video Modeling

Ameesh Makadia

12 June 2025

Google Research



Different perspectives on *efficiency*

Different perspectives on *efficiency*

Compression

More compact latent spaces → more efficient generation

Different perspectives on *efficiency*

Compression

More compact latent spaces → more efficient generation

Structured representations

Latent representation design that enables efficient modeling

Different perspectives on *efficiency*

Compression

More compact latent spaces → more efficient generation

Structured representations

Latent representation design that enables efficient modeling

Data sparsity

Generative models designed for data-sparse settings

Agenda

Part I - Compression (15 min)

Factorized latent representations for video

Part II - Structured representations (15 min)

Multiscale image generation with autoregressive models

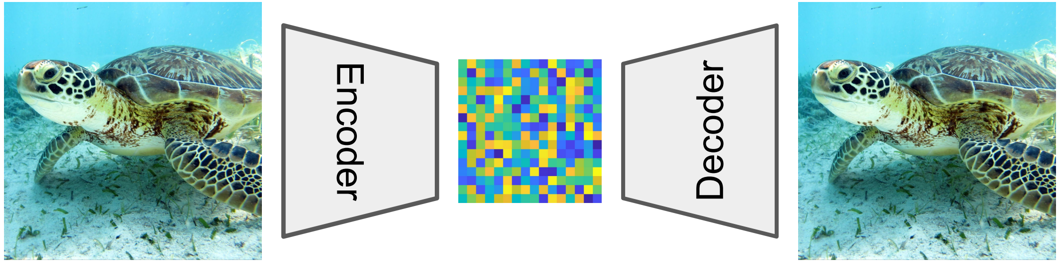
Part III - Data sparsity (< 10 min)

Diffusion models from a single 3D shape

Part I

Factorized latent representations for video

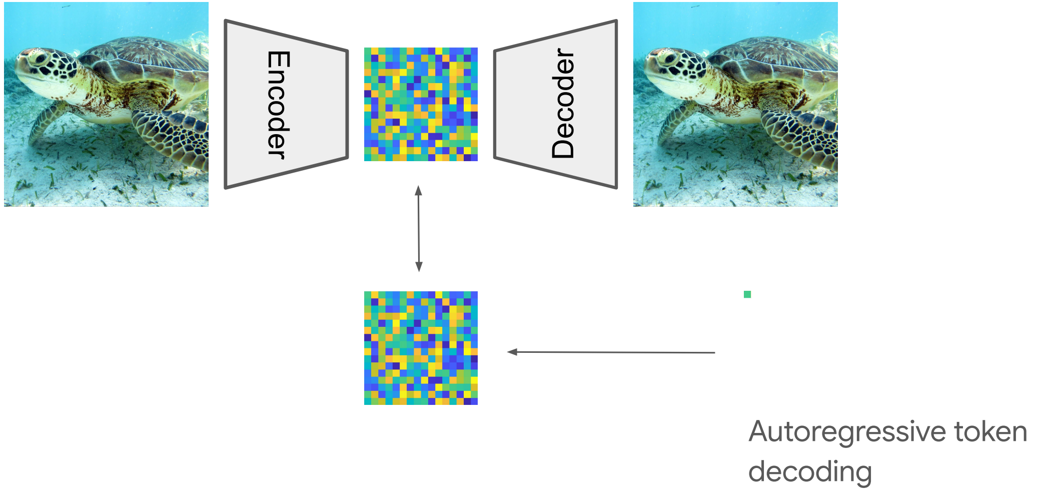
Latent generative models



- Reduce burden of generation in high dimension image/pixel space
- Reconstruction losses: pixel (MSE), perceptual (LPIPS), discriminator
- Latent representation is a heavily compressed, e.g. 512x512x3→64x64x4
- Individual tokens can be discrete (vector quantization) or continuous

van den Oord et al., *Neural Discrete Representation Learning*, 2017.
Razavi et al., *Generating Diverse High-Fidelity Images with VQ-VAE-2*, 2019.
Esser et al., *Taming Transformers for High-Resolution Image Synthesis*, 2020.
Rombach et al., *High-Resolution Image Synthesis with Latent Diffusion Models*, 2021.

Latent generative models



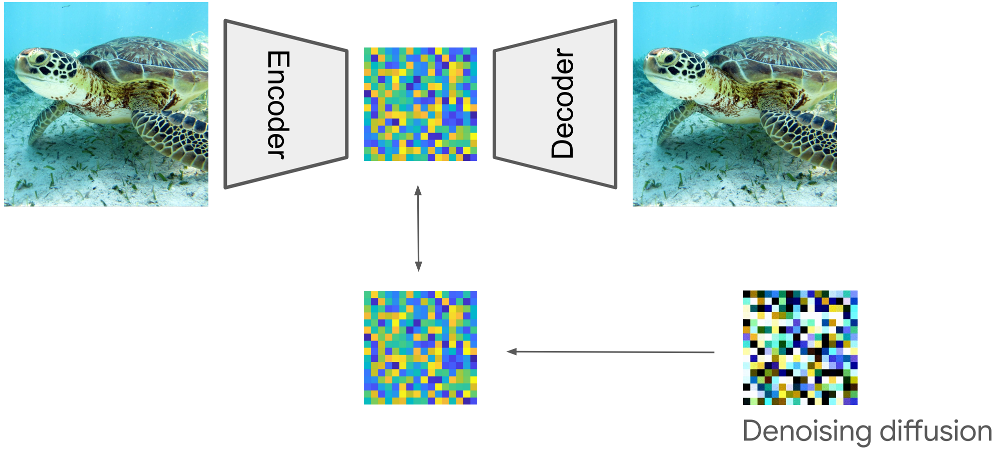
Stage 1: training autoencoder to learn latent feature space (image ⟶ visual tokens)

Stage 2: training a generative model for latent features

Autoregressive models (discrete tokens)

van den Oord et al., *Neural Discrete Representation Learning*, 2017.
Razavi et al., *Generating Diverse High-Fidelity Images with VQ-VAE-2*, 2019.
Esser et al., *Taming Transformers for High-Resolution Image Synthesis*, 2020.
Ramesh et al., *Zero-Shot Text-to-Image Generation*, 2021.
Yu et al., *Scaling Autoregressive Models for Content-Rich Text-to-Image Generation*, 2022.

Latent generative models



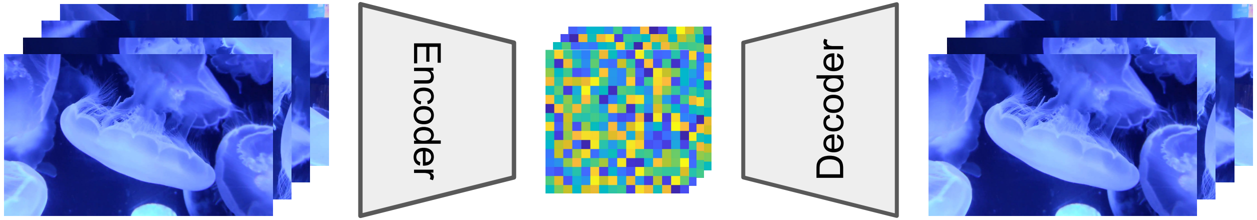
Stage 1: training autoencoder to learn latent feature space (image ⟶ visual tokens)

Stage 2: training a generative model for *latent features/tokens*

- Autoregressive models (discrete tokens)
- Diffusion models (continuous tokens)

Rombach et al., *High-Resolution Image Synthesis with Latent Diffusion Models*, 2021.
Peebles and Xie, *Scalable Diffusion Models with Transformers*, 2022.

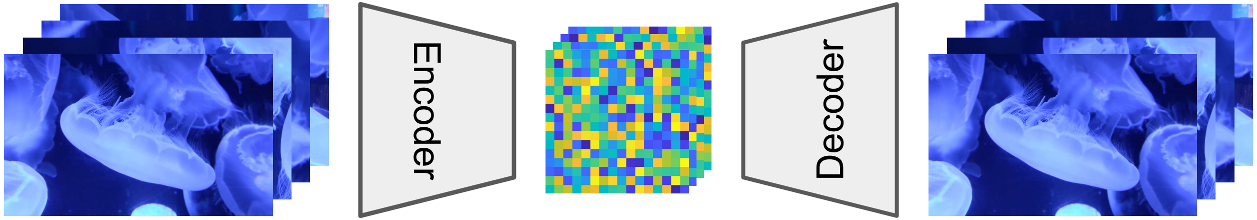
Video tokenization



Autoencoding spatiotemporal volumes

⟶ spatiotemporal latent features (H x W x T ⟶ H' x W' x T', O(HWT) storage)

Video tokenization



Autoencoding spatiotemporal volumes

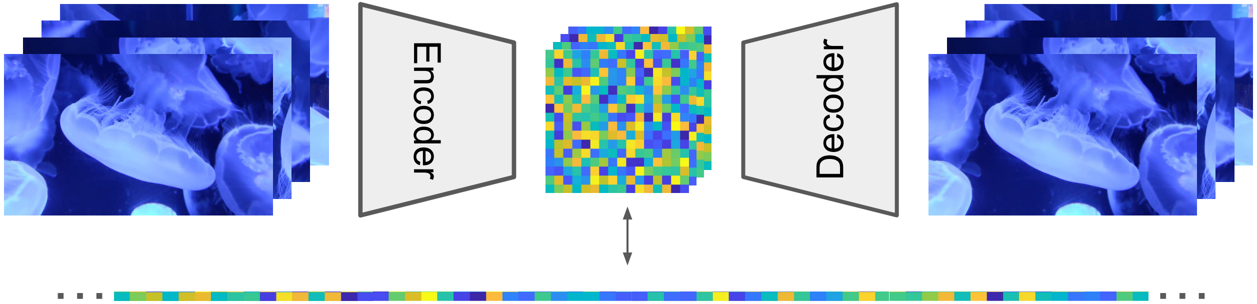
→ spatiotemporal latent features (H x W x T → H' x W' x T', O(HWT) storage)

Generative modeling w/spatiotemporal structure

3D U-Net (Video Diffusion Models, 2022)

Cicek et al., *3D U-Net: learning dense volumetric segmentation from sparse annotation*, 2016.
Ho et al., *Video Diffusion Models*, 2022.
Ho et al., *Imagen video: High definition video generation with diffusion models*, 2022.

Video tokenization



Autoencoding spatiotemporal volumes

→ spatiotemporal latent features (H x W x T → H' x W' x T', *O*(HWT) storage)

Generative modeling w/spatiotemporal structure

3D U-Net (Video Diffusion Models, 2022)

Sequence modeling (tokens unrolled into a 1D sequence)

Autoregressive transformers (TATS)

Masked transformers (Phenaki, Magvit, Magvit-v2)

Transformer diffusion (W.A.L.T.)

Ge et al., Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer, 2022.
Villegas et al., Phenaki: Variable length Video Generation From Open Domain Textual Descriptions, 2022.
Yu et al., MAGVIT: Masked Generative Video Transformer, 2022.
Yu et al., Language Model Beats Diffusion – Tokenizer is Key to Visual Generation, 2023.
Gupta et al., Photorealistic Video Generation with Diffusion Models, 2024.

Video tokenization

For sequence models (masked transformer, autoregressive, diffusion transformer), efficiency is directly tied to the latent size

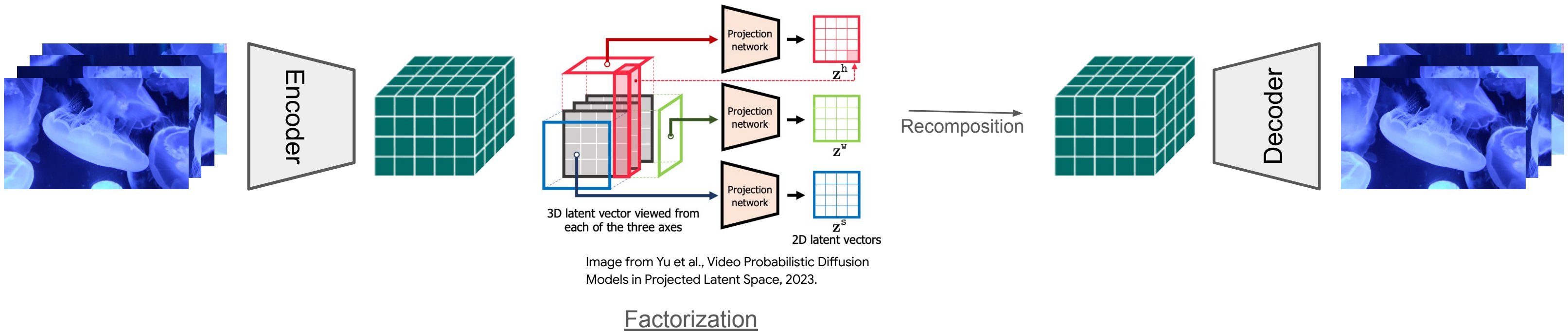
Can we further compress the latent space, without sacrificing reconstruction or generation quality?

Volumetric latent space – scales linearly with the input size

Plane-factorization (factorize volumetric data into orthogonal planes)

Size scales sublinearly with the input

Tri-plane factorization



Triplane representations commonly used for 3D generation tasks

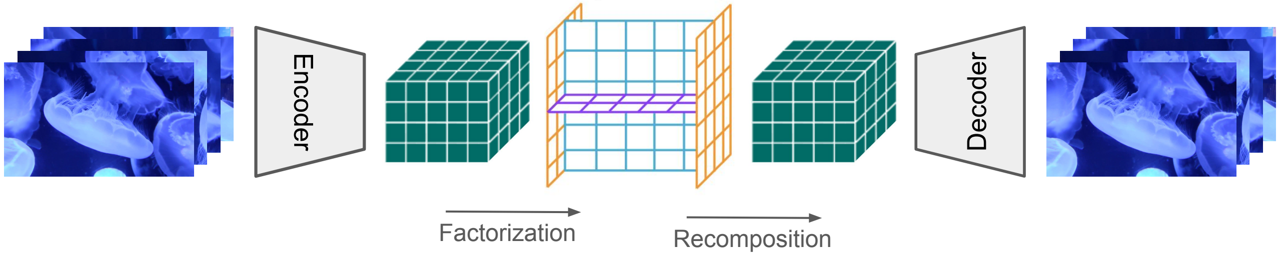
- 3D neural fields, 3D semantic scenes, 3D shapes

Recently applications to video tokenization: PVDM, HVDM, CMD

- Benefit from 2D diffusion models for image generation
- 2D conv UNets for each plane w/cross attention, fine-tuning DiT

Wu et al., *Sin3dm: Learning a diffusion model from a single 3d textured shape*, 2023.
Shue et al., *3D neural field generation using triplane diffusion*, 2022.
Yu et al., *Video Probabilistic Diffusion Models in Projected Latent Space*, 2023.
Kim et al., *Hybrid Video Diffusion Models with 2D Triplane and 3D Wavelet Representation*, 2024.
Yu et al., Efficient video diffusion models via content-frame motion-latent decomposition, 2024.

Four-plane factorization



Triplane tokenization

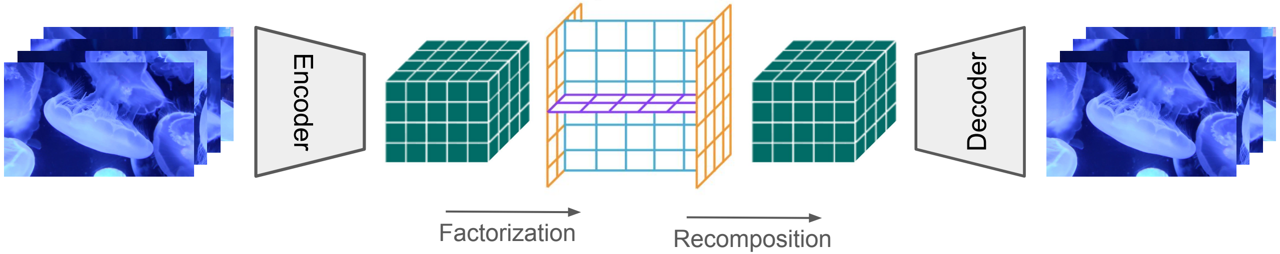
- Smaller latent sizes enable much faster generative model training and sampling
- Generation quality still lags behind volumetric latent generation
- Not easily adopted to all video generation tasks, e.g. frame extrapolation and interpolation

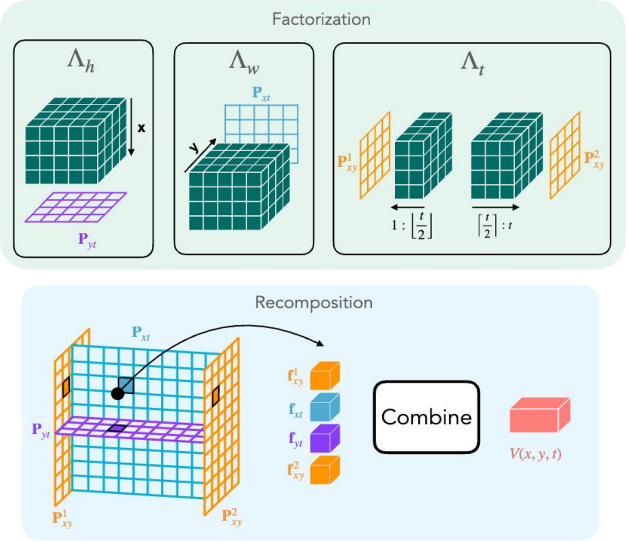
Four-plane factorization

- Two spatial planes (orange), two spatiotemporal planes (blue / purple)
- Structure allows flexibility for different image-conditioned video generation tasks
- Favorable efficiency vs quality tradeoff when introduced into volumetric architectures
  - 2x speedup in generative model training/sampling, comparable generation quality

Suhail et al., *Four-Plane Factorized Video Autoencoders*, 2024.

Four-plane factorization





<u>Factorization</u>

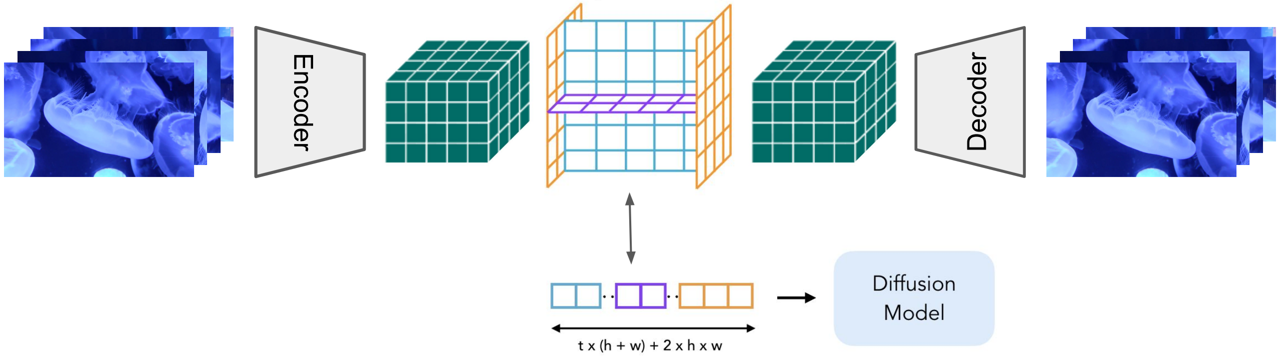
The simplest operator (mean pooling) generalizes best, compared to learned linear projection, or transformer (PVDM)

Spatial planes are obtained after splitting the volume into two non-overlapping segments along time

<u>Recomposition</u>

Features are combined through concatenation to reconstitute the volume

Four-plane factorization



Adopt the W.A.L.T. framework for analysis

- Encoder is Magvit-v2 causal 3D convolution architecture (also used by OpenSora, CogVideoX, ...)
- Continuous 8-dimensional tokens
- Generation is diffusion transformer model

W.A.L.T. + Four-plane tokenization

- Introduce volume factorization and recomposition steps at the latent bottleneck
- All other AE/Diffusion details mirror W.A.L.T.

Gupta et al., *Photorealistic Video Generation with Diffusion Models*, 2024.
Zheng et al., *Open-Sora: Democratizing Efficient Video Production for All*, 2024.
Yang et al., *CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer*, 2024.

Reconstruction

Kinetics-600 dataset, 17 frame videos

Res.	Method	PSNR↑	SSIM↑	LPIPS↓	Seq.Len
128x128	Volumetric	27.64	0.85	0.049	1280
	4Plane	27.11	0.82	0.051	672
256x256	W.A.L.T.	26.27	0.79	0.089	1280
	Four-plane	25.67	0.77	0.104	672
	WF-VAE	27.86	0.83	0.064	1280
	Four-plane-WF-VAE	26.98	0.81	0.073	672

Number of frames	PSNR↑	SSIM↑	LPIPS↓
17	27.11	0.82	0.051
21	26.95	0.82	0.051
25	26.51	0.81	0.052

Four-plane reconstruction for longer videos

256x256 tokenizers - extra layer to the encoder and decoder

Comparable reconstruction metrics despite half the sequence length

WF-VAE is the AE architecture for OpenSoraPlan

Carreira et al., *A Short Note About Kinetics-600*, 2018.

Li et al., *WF-VAE: Enhancing Video VAE by Wavelet-Driven Energy Flow for Latent Video Diffusion Model*, 2024.

Lin et al., *Open-Sora Plan: Open-Source Large Video Generation Model*, 2024.

Generation

Tokenizer: Kinetics-600 dataset, 17 frame videos
Diffusion model trained on UCF-101

	Class Conditional Generation (FVD ↓)		Params	Steps
	UCF-101 (128x128)	UCF-101 (256x256)		
MAGVIT	76	-	306M	48
MAGVIT-v2	58	-	307M	24
WALT	39	84.68	214M	50
Four-plane	38	58.27	214M	50

	Class Conditional Generation (FVD ↓)		Params	Steps
	UCF-101 (128x128)	UCF-101 (256x256)		
PVDM	-	399.4	-	400
HVDM	-	303.1	63M	100
CMD	73	-	-	-
Tri-plane	52	-	214M	50
Four-plane	38	58.27	214M	50

Generation cost (TPU-v5e-2x2, four 17-frame 128x128 videos):
0.71s Four-plane, 1.59s W.A.L.T. (> 2x faster)

Yu et al., *MAGVIT: Masked Generative Video Transformer*, 2022.
Yu et al., *Language Model Beats Diffusion – Tokenizer is Key to Visual Generation*, 2023.
Yu et al., *Video Probabilistic Diffusion Models in Projected Latent Space*, 2023.
Kim et al., *Hybrid Video Diffusion Models with 2D Triplane and 3D Wavelet Representation*, 2024.
Yu et al., Efficient video diffusion models via content-frame motion-latent decomposition, 2024.

Class-conditional generation

256x256 class-conditional generation

W.A.L.T. Four-plane

Wall pushups

W.A.L.T. Four-plane

Surfing

W.A.L.T. Four-plane

Pushups

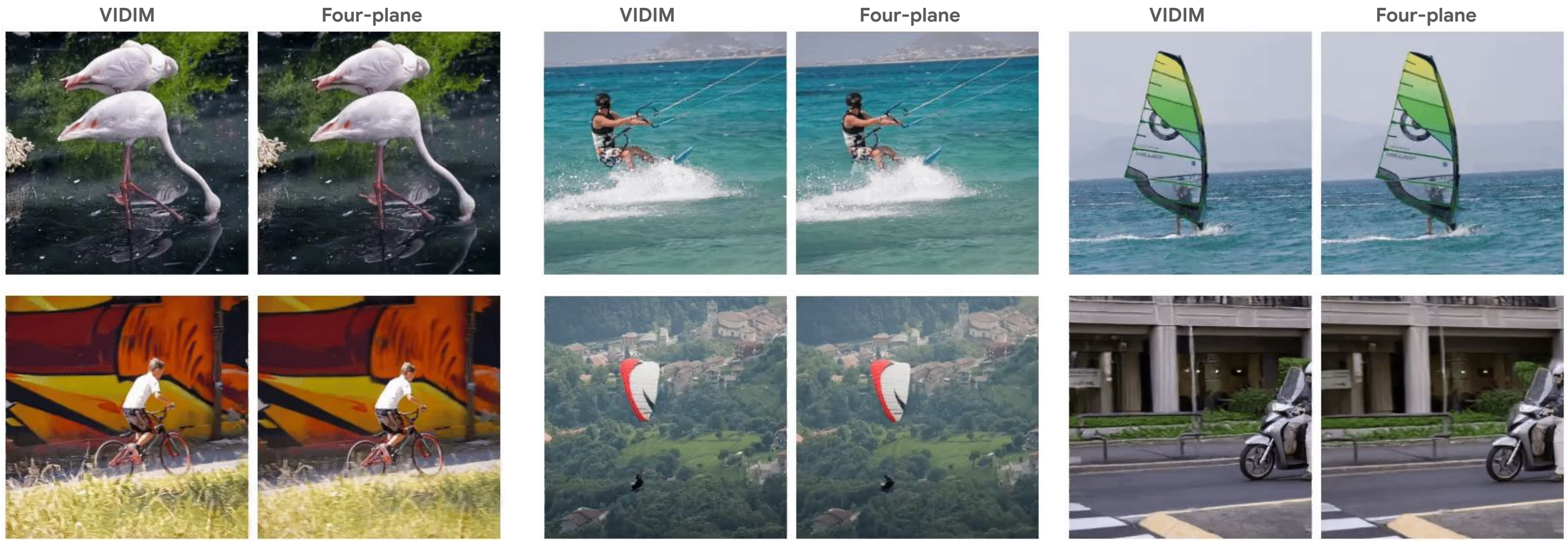
Lunges

Handstand walking

Billiards

Interpolation

256x256 resolution 9-frame interpolation
VIDIM: cascaded diffusion models



Jain et al., *Video Interpolation with Diffusion Models*, 2024.

Text-to-Video

300M internet videos

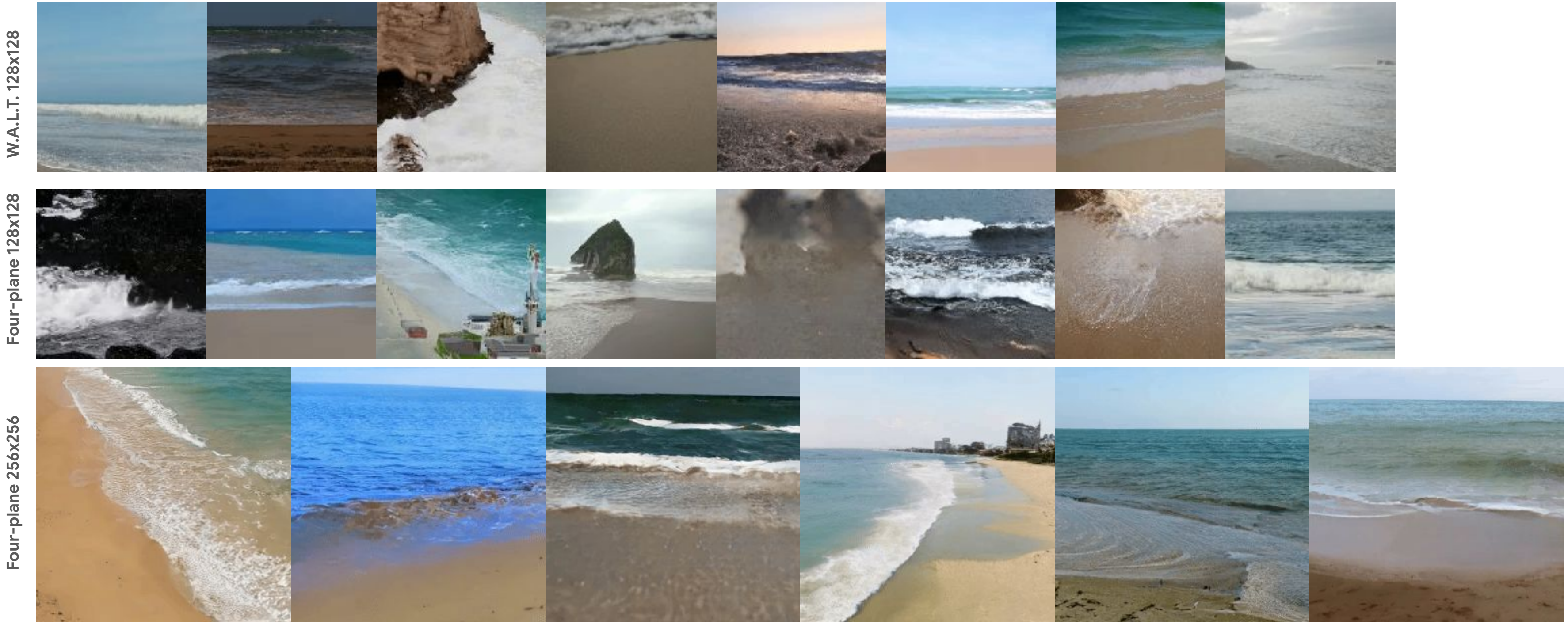
FVD (17 frame, 128x128): 18.22 for W.A.L.T., 20.24 for Four-plane

“Flying over the mountains with a river”

Text-to-Video

300M internet videos

FVD (17 frame, 128x128): 18.22 for W.A.L.T., 20.24 for Four-plane

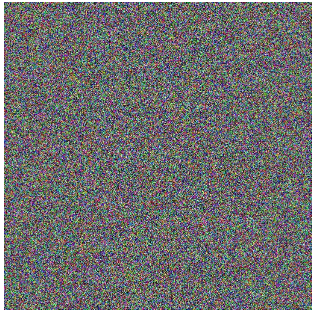


"A wave reaching the beach"

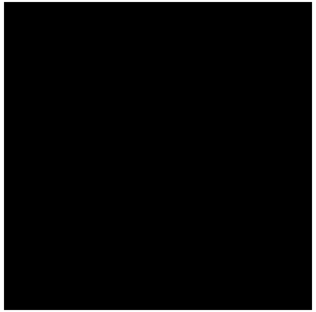
Part II

Multiscale image generation with autoregressive models

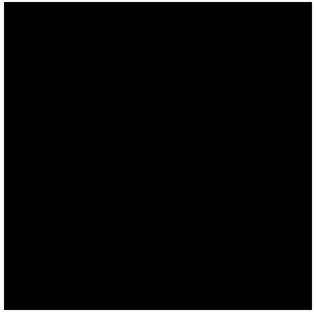
Latent Generative Models for Images



Diffusion models



Autoregressive models



Masked models

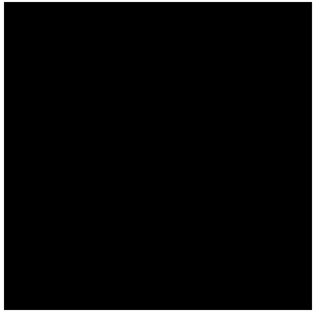
Language Models

Illustration is in pixels space, but latent models operate in the AE latent space (tokens)

Autoregressive models

Recent trend: more powerful autoregressive image generation models

- **LlamaGen** (Sun et al, "*Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation*", 2024)
- **VAR** (Tian et al, "Visual autoregressive modeling: Scalable image generation via next-scale prediction", 2024)
- **Open-MAGVIT2** (Luo et al., "Open-MAGVIT2 An Open-Source Project Toward Democratizing Auto-regressive Visual Generation", 2024)
- …


Autoregressive generation

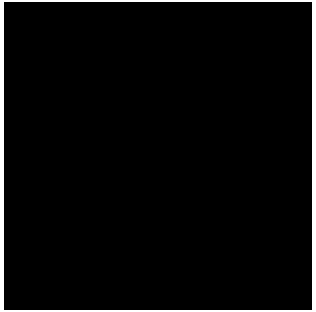
Strategic

- Borrow from widely successful LLM architectures
- Ideal for multimodal applications where all representations are discrete tokens

Autoregressive models

Recent trend: more powerful autoregressive image generation models

- **LlamaGen** (Sun et al, "Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation", 2024)
- **VAR** (Tian et al, "Visual autoregressive modeling: Scalable image generation via next-scale prediction", 2024)
- **Open-MAGVIT2** (Luo et al., "Open-MAGVIT2 An Open-Source Project Toward Democratizing Auto-regressive Visual Generation", 2024)
- …



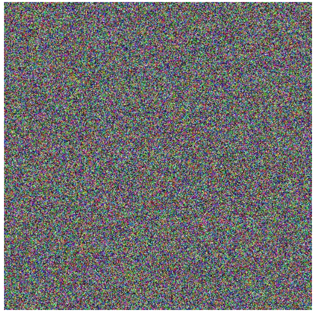
Autoregressive models

Strategic

- Borrow from widely successful LLM architectures
- Ideal for multimodal applications where all representations are discrete tokens

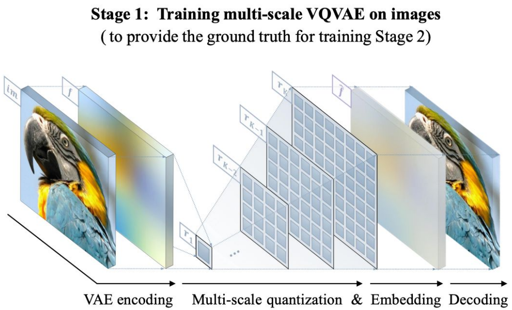
Drawback

- Conditioning for next-token prediction is not ideal (partial image)
- Prefer conditioning on a noisy version of the full image (diffusion models!)

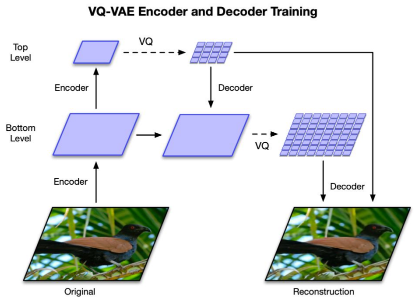


Diffusion models

Multiscale (coarse-to-fine) tokenization



VAR, Tian et al., 2024



VQ-VAE-2, Razavi et al., 2019

Multiscale tokenizers

VAR (Tian et al, "Visual autoregressive modeling: Scalable image generation via next-scale prediction", 2024)

VQ-VAE-2 (Razavi et al., Generating Diverse High-Fidelity Images with VQ-VAE-2, 2019)

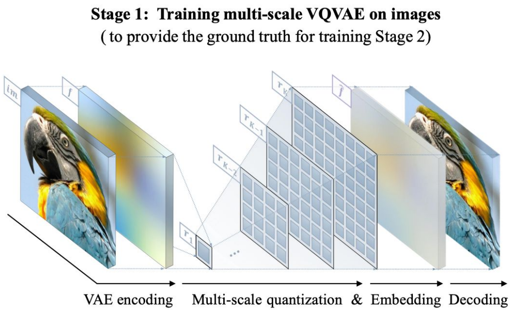
RQ-VAE (Lee et al., "Autoregressive Image Generation using Residual Quantization", 2022)

Multiscale quantization of the latent space

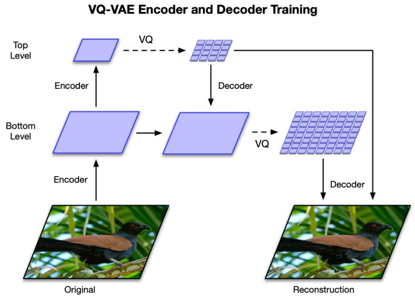
Residual design

Better conditioning – next-scale prediction depends on previous scales

Multiscale (coarse-to-fine) tokenization



VAR, Tian et al., 2024



VQ-VAE-2, Razavi et al., 2019

Multiscale tokenizers – coarse-to-fine quantization of the *latent* space

Example, the coarsest token map does not correspond to the coarse image

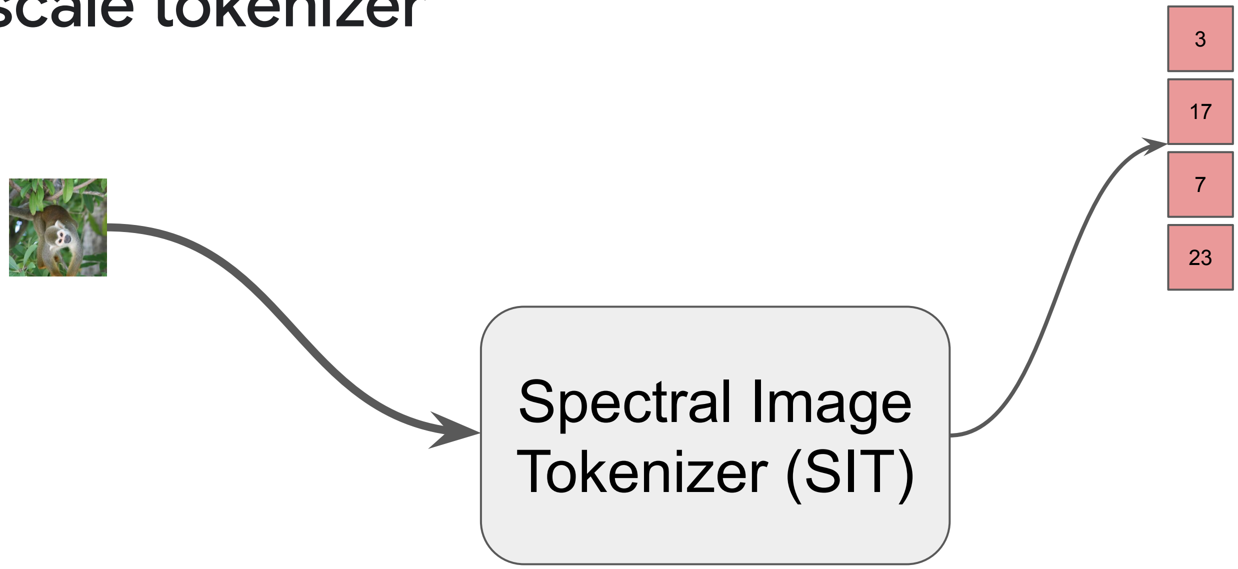
Tokenizing multiscale image representations

Input is a coarse-to-fine *image* representation (Discrete Wavelet Transform)

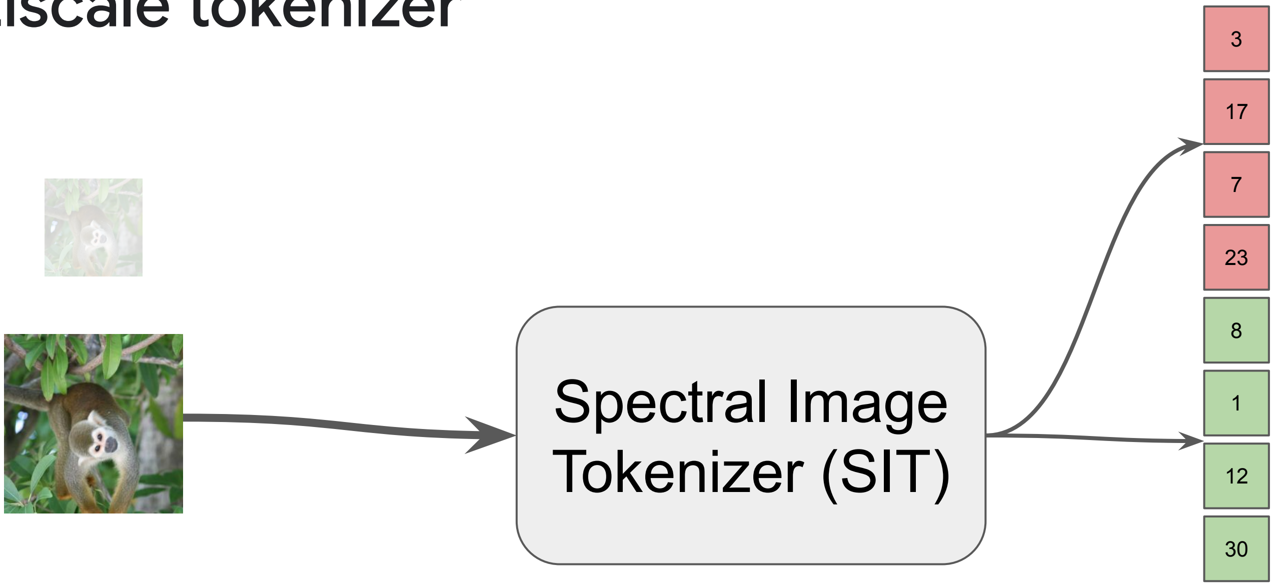
How do you tokenize the DWT?

SIT (Esteves et al., “Spectral Image Tokenizer,” 2024)

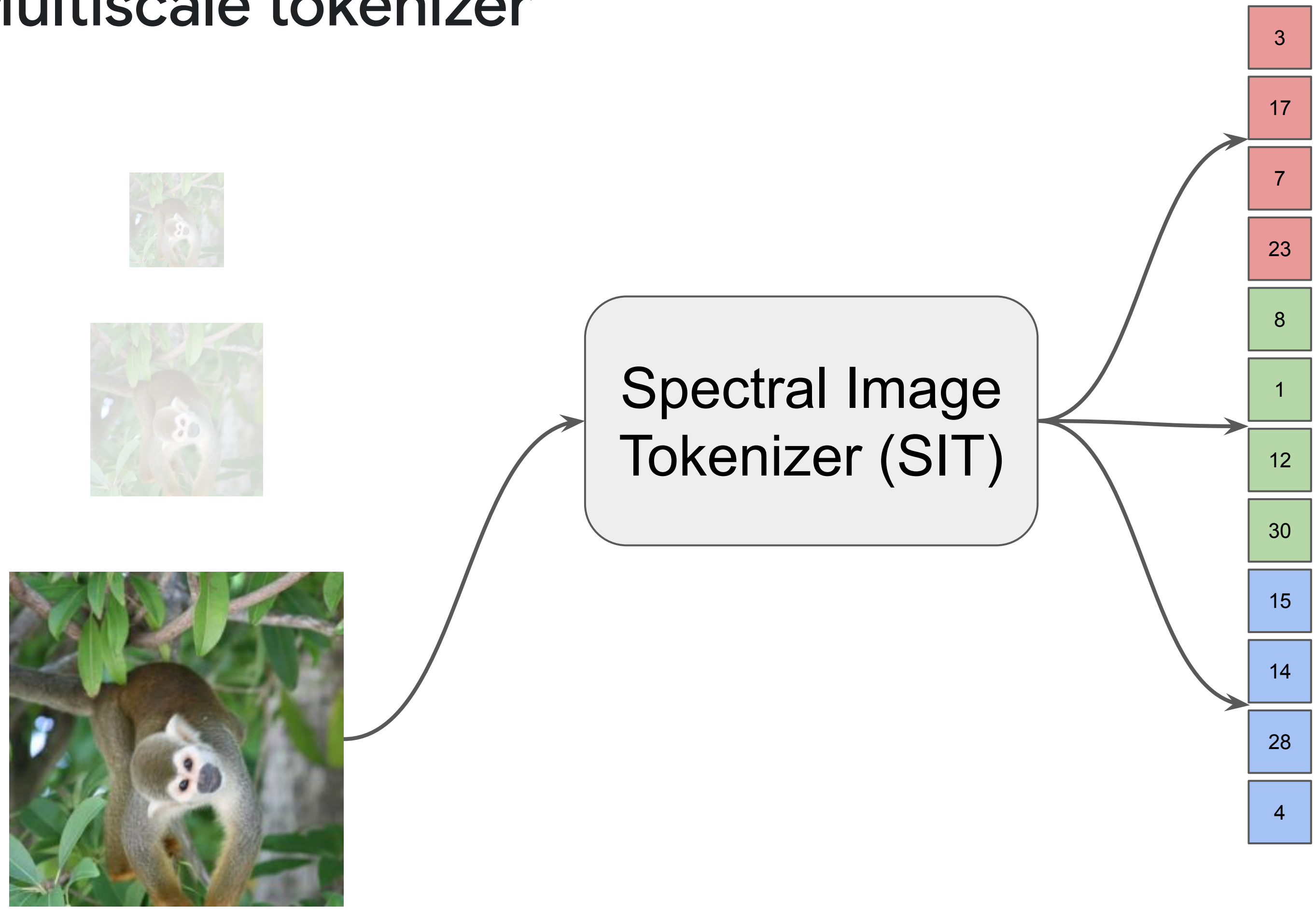
Multiscale tokenizer



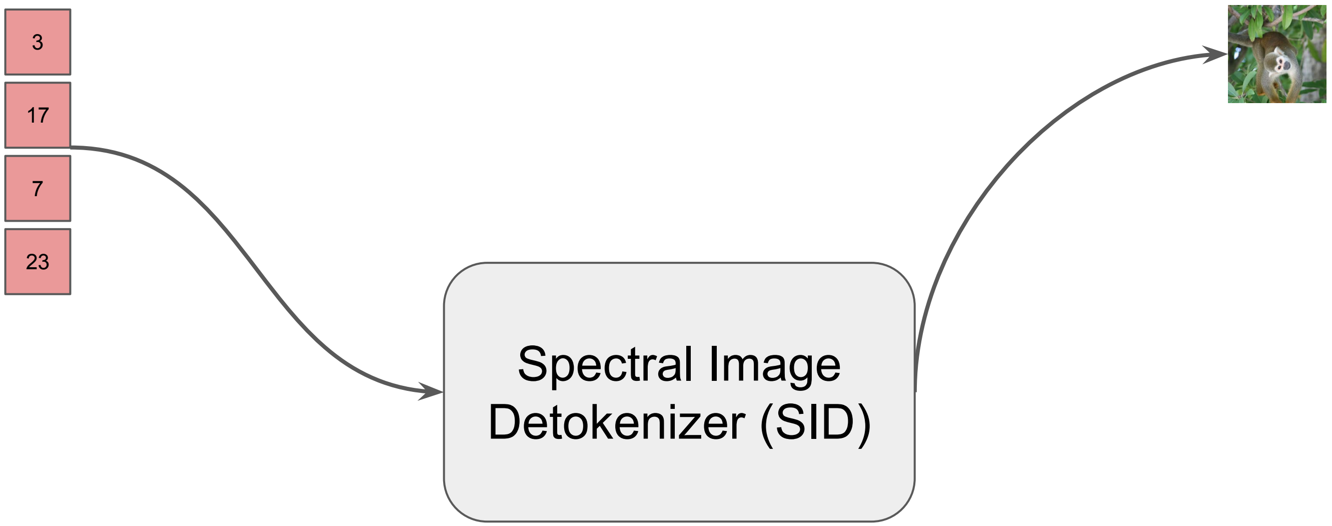
Multiscale tokenizer



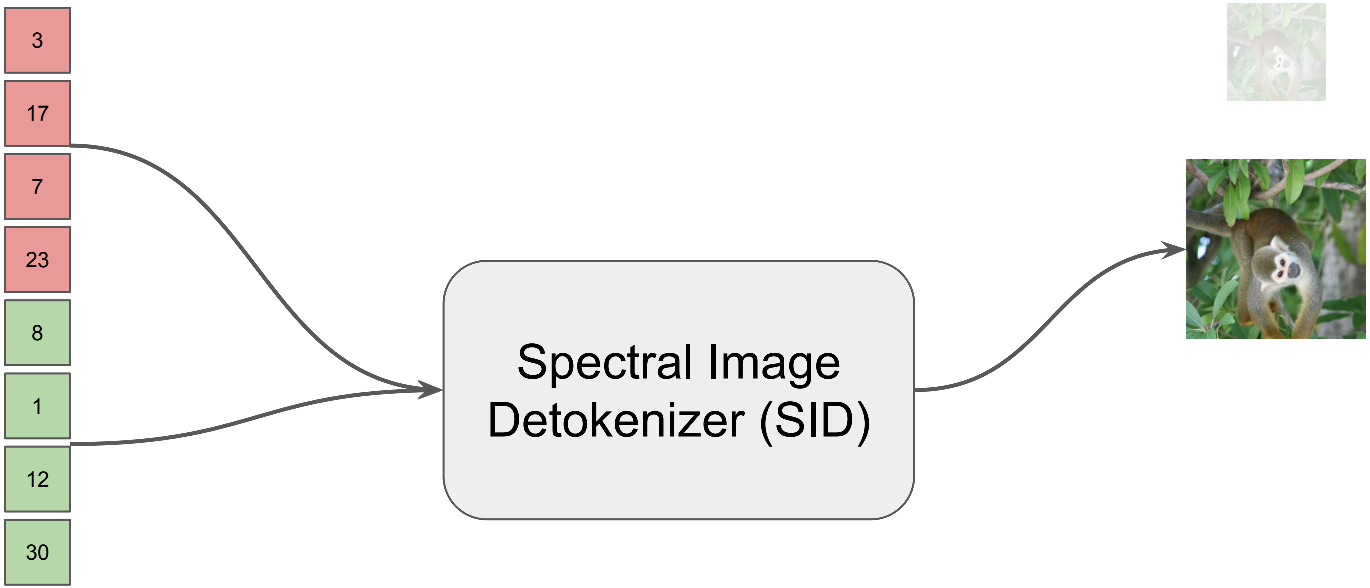
Multiscale tokenizer



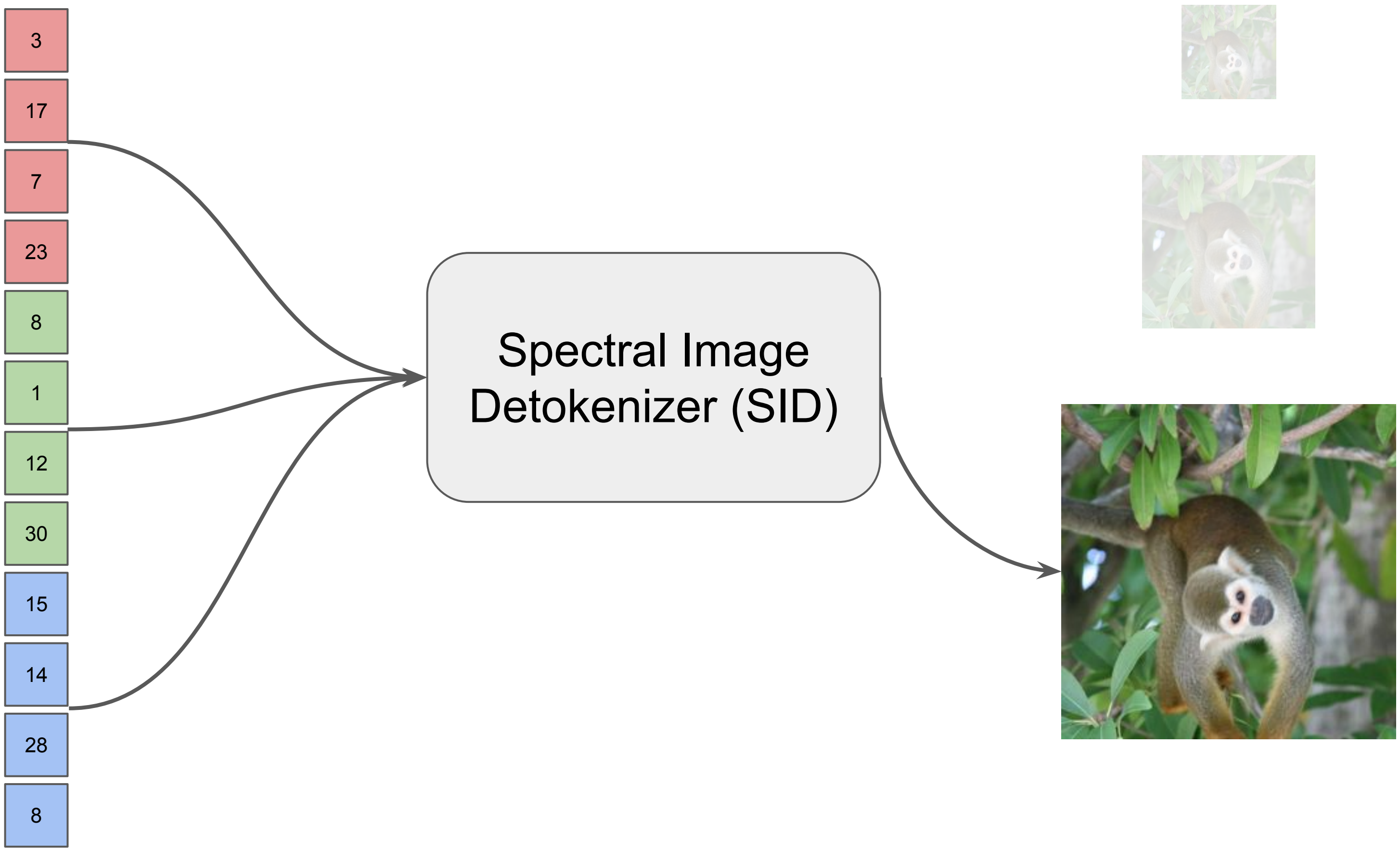
Multiscale detokenizer



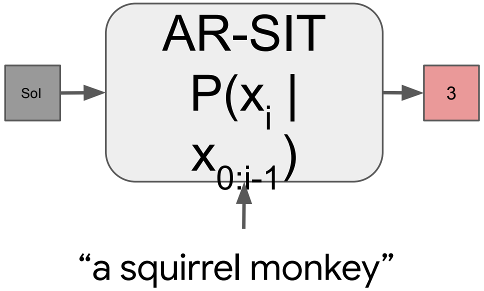
Multiscale detokenizer



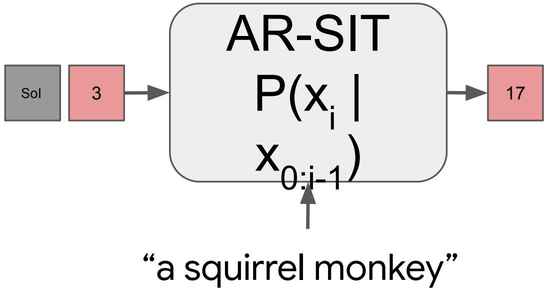
Multiscale detokenizer



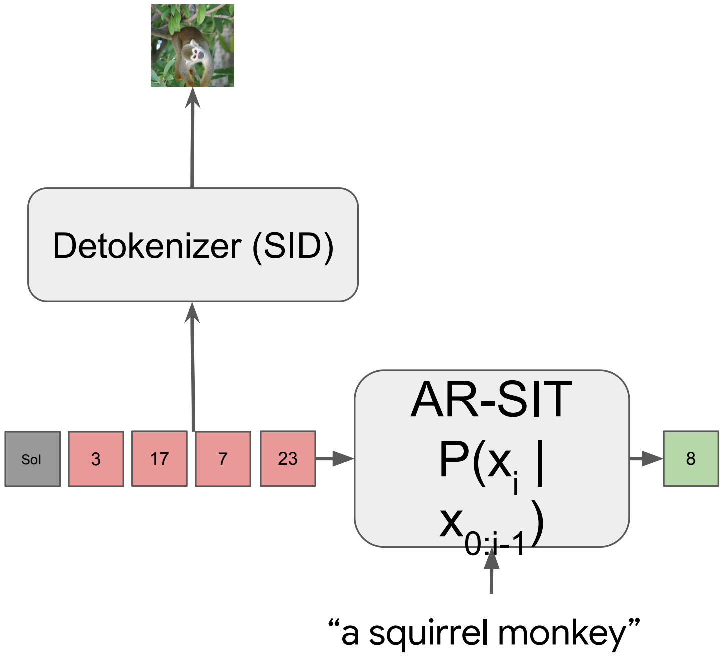
(Coarse-to-fine) autoregressive generation



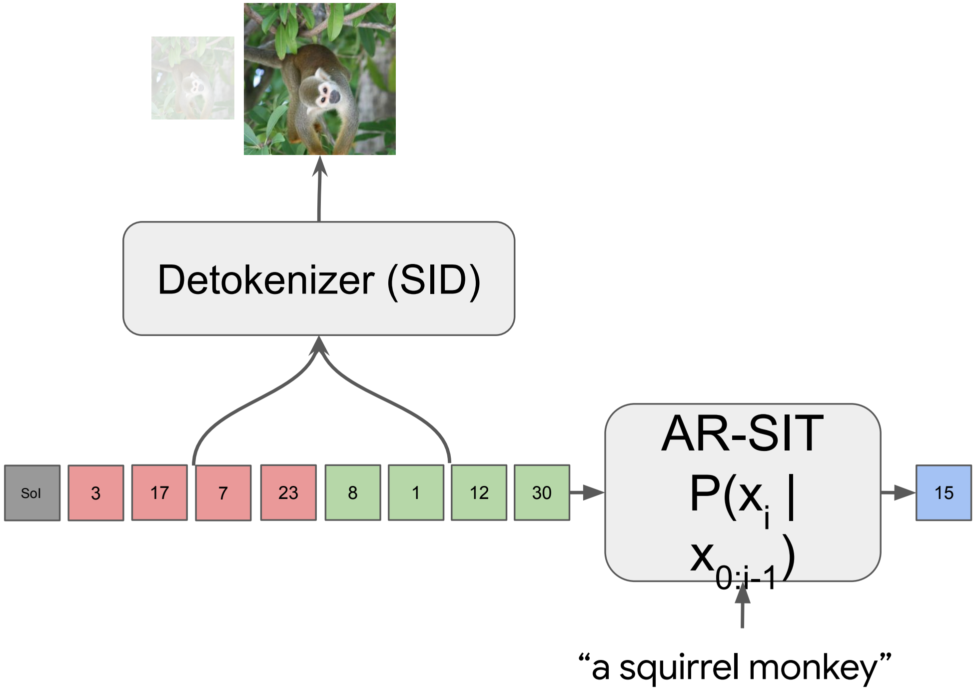
(Coarse-to-fine) autoregressive generation



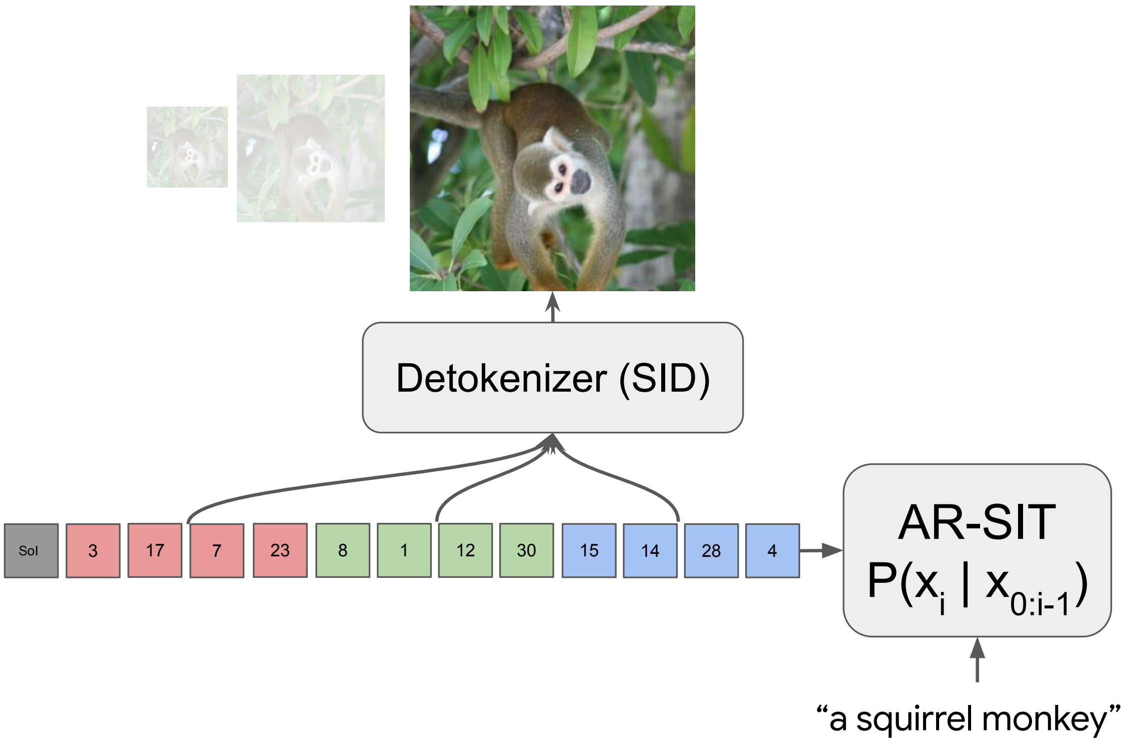
(Coarse-to-fine) autoregressive generation



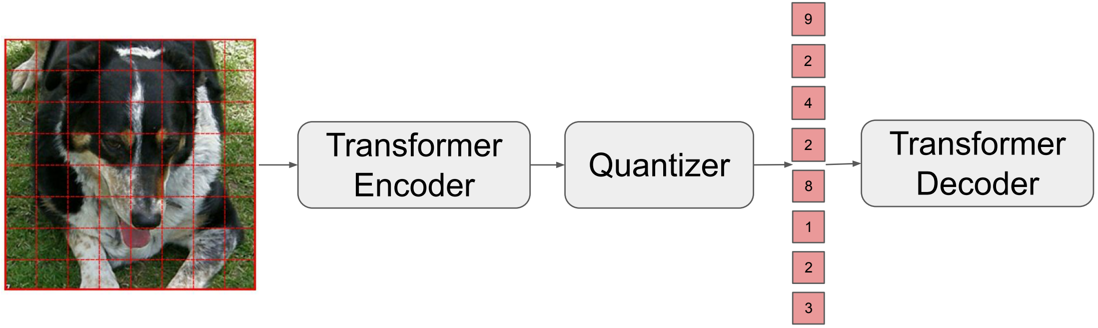
(Coarse-to-fine) autoregressive generation



(Coarse-to-fine) autoregressive generation



Spectral image tokenizer

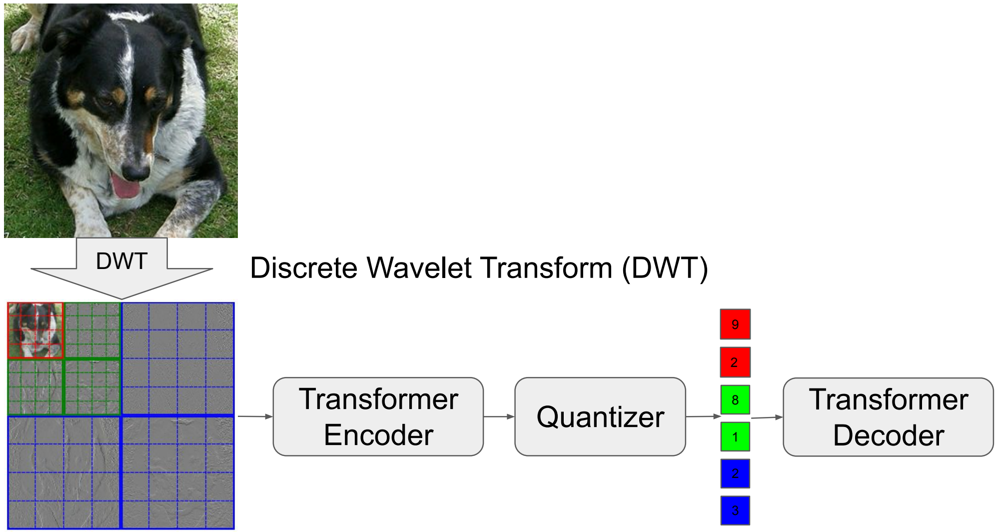


Prior work (e.g. ViT-VQGAN)

ViT-VQGAN: Transformer encodes patches of the input image

Yu et al., *Vector-quantized Image Modeling with Improved VQGAN*, 2021.

Spectral image tokenizer



SIT: Transformer encodes patches of the Discrete Wavelet Transform (Haar wavelets)

Haar Wavelet Transform

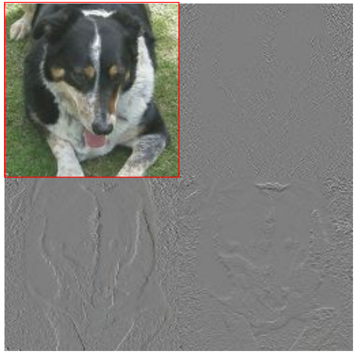
1	1
1	1

1	-1
1	-1

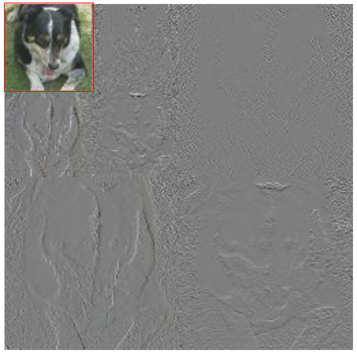
1	1
-1	-1

1	-1
-1	1

Haar filters

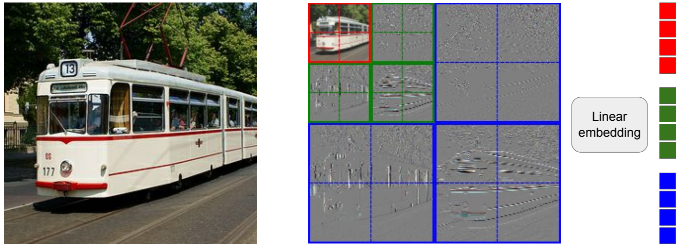


One scale



Two scales

Spectral image tokenizer - details

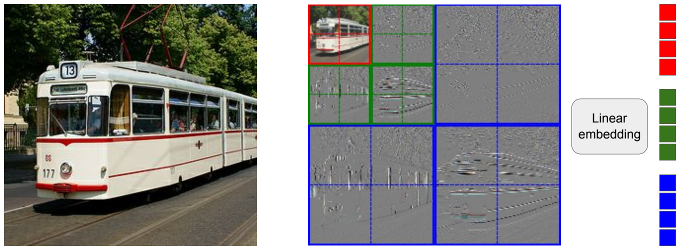


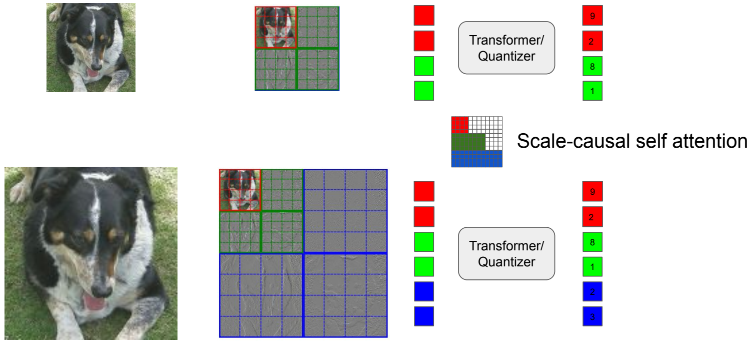
Larger patches for higher frequencies ⟶ Same number of tokens per scale

- Limits sequence length
- Higher frequencies are compressed more (desirable since they are sparser)

Different quantizer codebooks per scale (content distribution changes across scales)

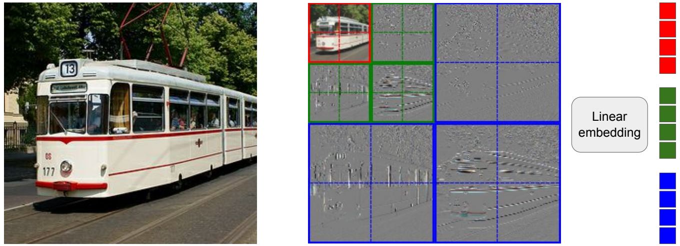
Spectral image tokenizer - details

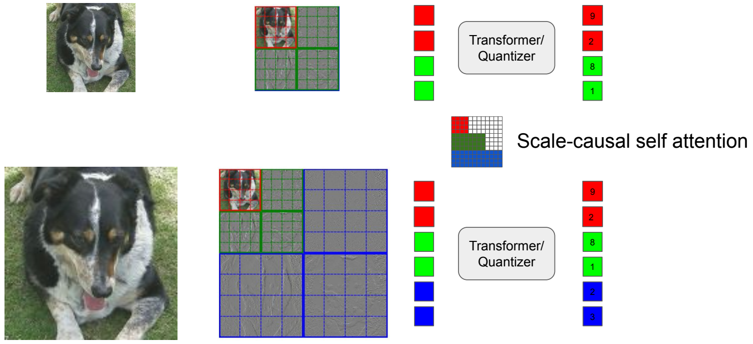




- Larger patches for higher frequencies ⟶ Same number of tokens per scale
  - Limits sequence length
  - Higher frequencies are compressed more (desirable since they are sparser)
- Different quantizer codebooks per scale (content distribution changes across scales)
- Scale-causal self-attention
  - Ensures different inputs with same lower frequency coefficients have identical tokens at those scales

Spectral image tokenizer - details





Larger patches for higher frequencies → Same number of tokens per scale

- Limits sequence length
- Higher frequencies are compressed more (desirable since they are sparser)

Different quantizer codebooks per scale (content distribution changes across scales)

Scale-causal self-attention

- Ensures different inputs with same lower frequency coefficients have identical tokens at those scales

Autoregressive transformer based on Parti (Yu et al., "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation", 2022)

- Different token embeddings per scale

Multiscale reconstruction (ImageNet)

	LPIPS ↓	PSNR ↑	L1 ↓	FID ↓	IS ↑	images/s ↑
Resolution: 512 × 512						
ViT-VQGAN	0.320	22.4	0.042	6.92	151.5	593
SIT-5 (Ours)	0.260	22.0	0.046	2.65	192.0	410
SIT-6 (Ours)	0.239	23.1	0.040	1.74	203.7	320
Resolution: 256 × 256						
ViT-VQGAN (reported)	-	24.8	0.032	1.99	184.4	-
ViT-VQGAN (reproduced)	0.167	25.0	0.031	2.33	184.0	-
ViT-VQGAN (no LL)	0.163	23.8	0.038	1.20	194.6	626
SIT-4 (Ours)	0.144	24.0	0.037	1.20	199.5	596
SIT-5 (Ours)	0.135	24.5	0.035	0.97	202.3	411
SIT-SC-5 (Ours)	0.161	24.1	0.037	1.33	193.7	411
Resolution: 128 × 128						
ViT-VQGAN	0.185	26.3	0.030	3.77	117.3	626
SIT-SC-5 (ours)	0.159	27.1	0.027	2.13	129.3	582
Resolution: 64 × 64						
ViT-VQGAN	0.129	28.8	0.023	3.53	21.0	627
SIT-SC-5 (ours)	0.111	31.3	0.017	1.39	30.1	847
Resolution: 32 × 32						
ViT-VQGAN	0.214	23.3	0.045	-	3.7	627
SIT-SC-5 (ours)	0.029	36.8	0.010	0.31	3.5	825
Resolution: 16 × 16						
ViT-VQGAN	0.127	24.9	0.039	-	1.7	627
SIT-SC-5 (ours)	0.013	41.3	0.006	0.09	1.8	2620

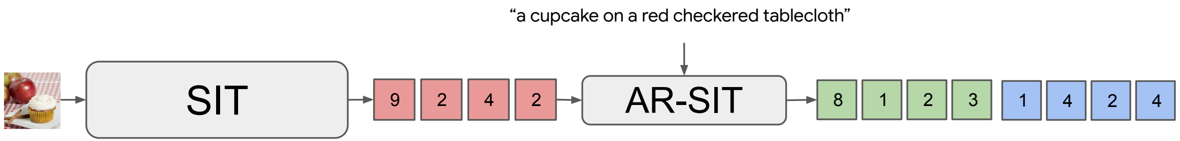
Multiscale generation (text-to-image on MSCOCO)

	FID ↓	IS ↑	images/s ↑	images/Gb ↑
Resolution: 256 × 256				
Parti350M (reported)	14.1	-	-	-
Parti350M	12.4	36.5	7.8	12.0
AR-SIT-SCD-4	12.6	37.3	6.5	8.0
Resolution: 128 × 128				
Parti350M	11.2	33.5	7.6	12.0
AR-SIT-SCD-4	11.4	33.2	12.6	12.0
Resolution: 64 × 64				
Parti350M	10.5	16.9	7.6	12.0
AR-SIT-SCD-4	11.4	18.6	24.5	16.0
Resolution: 32 × 32				
Parti350M	5.8	2.9	7.7	7.7
AR-SIT-SCD-4	7.6	3.2	74.7	28.0

Lin et al., "Microsoft COCO: Common Objects in Context", 2014

Class-conditional generation @ 512x512

Text-guided upsampling 16x16⟶256x256



"a motorcycle parked outside in a parking lot near the beach."

"a white table with glasses and red white and orange flowers"

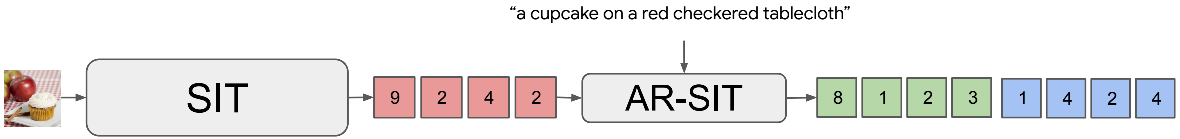
"a cupcake on a plate on a red checkered tablecloth"

"a piece of chocolate cake sitting on top of a wooden cutting board."

"a train is making its way down the tracks."

"crust topped with meat and vegetables on a stone pizza board."

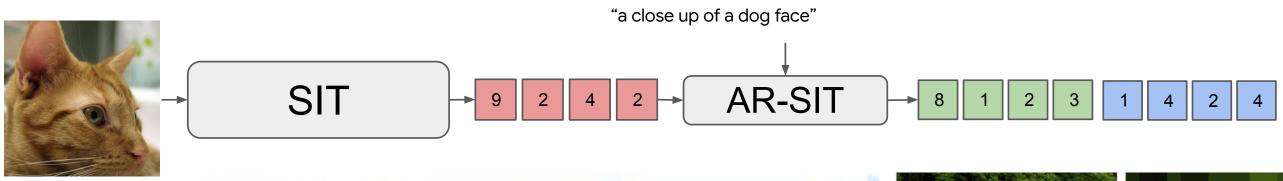
Text-guided upsampling 16x16⟶256x256

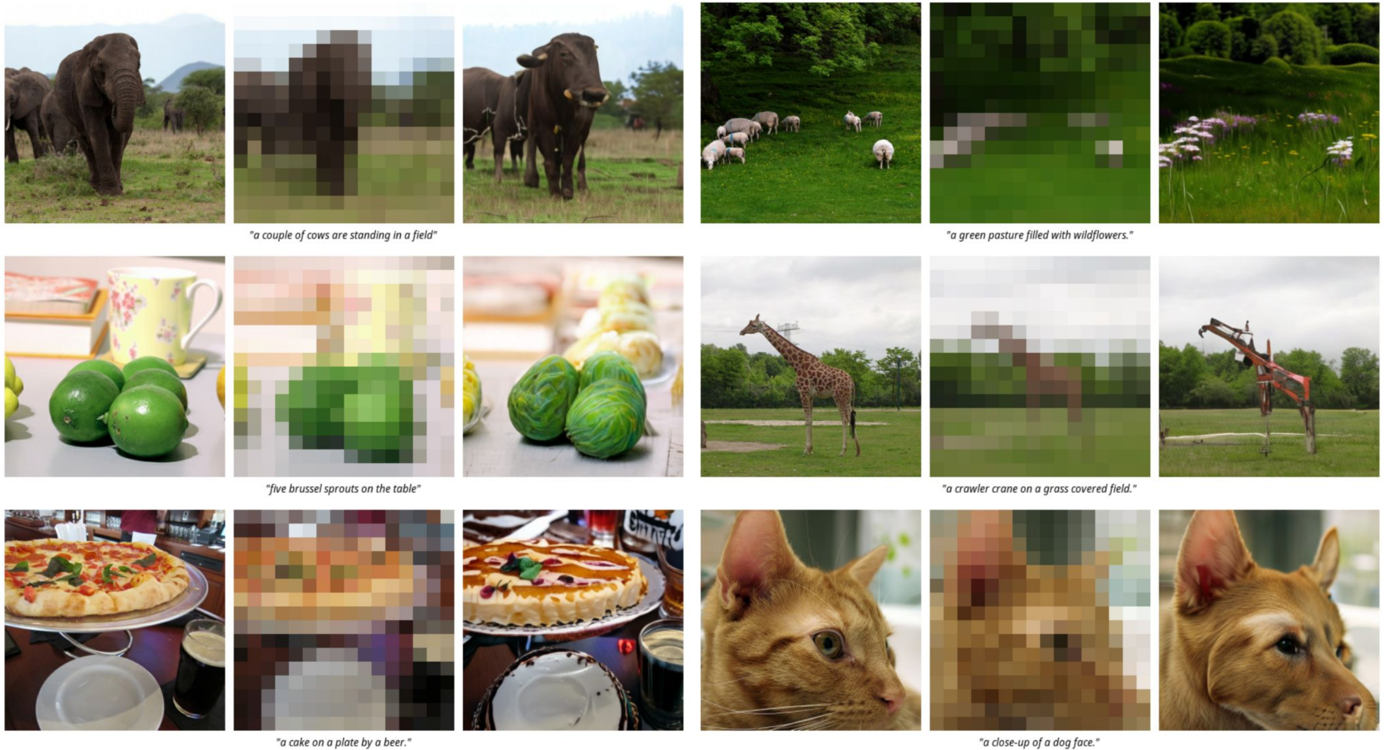


“an assortment of some colorful vases on display on a table”

“a cupcake on a red checkered tablecloth”

Text-guided editing





Part III

Diffusion models from a single 3D shape