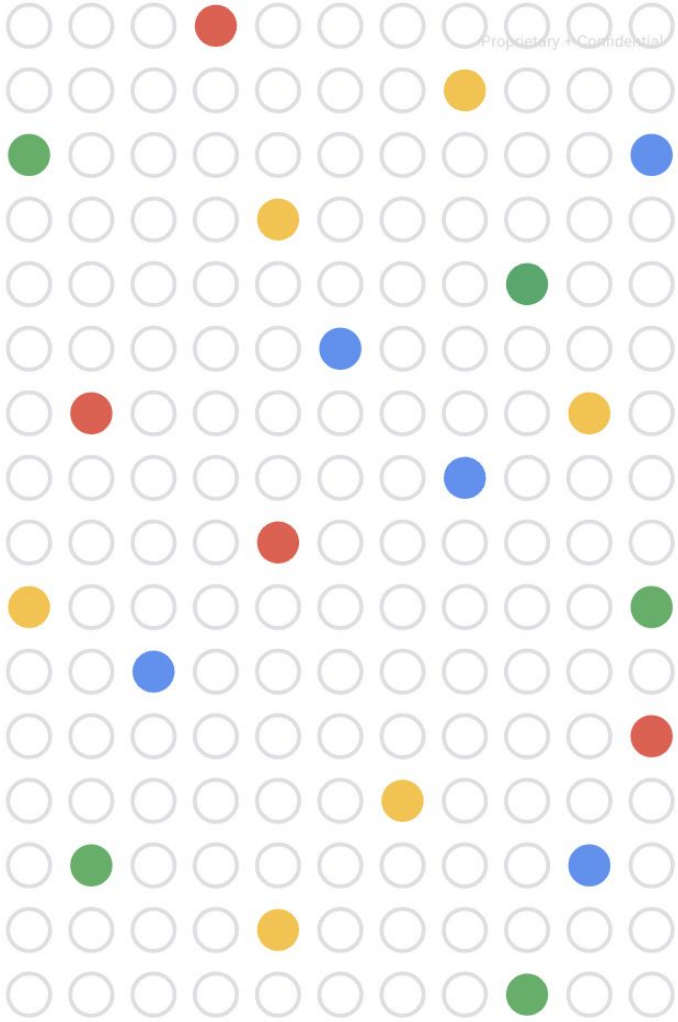
Muse

Text-to-image Generation via Masked Transformers

Dilip Krishnan

Research Scientist, Google DeepMind

http://muse.github.io



Muse: an efficient text-to-image generation and editing model, based on parallel decoding of tokens, developed in 2022-2023

Text-to-Image Generation

Text: A Welsh corgi holding a sign in its mouth that says 'Muse'.

How is Muse different from prior approaches?

Muse uses a new paradigm, different from both diffusion models and autoregressive models.

Fast — **Fewer iterations than both autoregressive models and diffusion**

High-Fidelity — **SOTA CLIP / FID**; Deep understanding of spatial relationships

Flexible — **Editing applications** without fine-tuning

Key Idea

- MaskGIT: Masked image modeling in the token space, inspired by MLM's
- Predict all the masked image tokens, just like BERT
- Crucial design for image generation and editing
  - Variable masking scheduling
  - Sampling approach

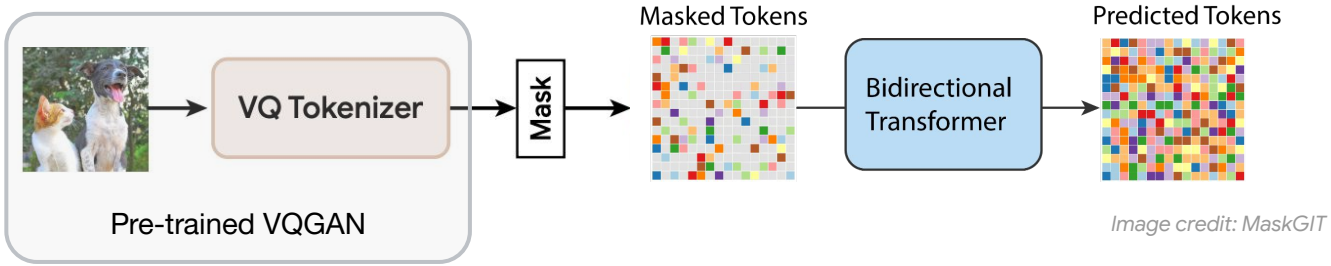
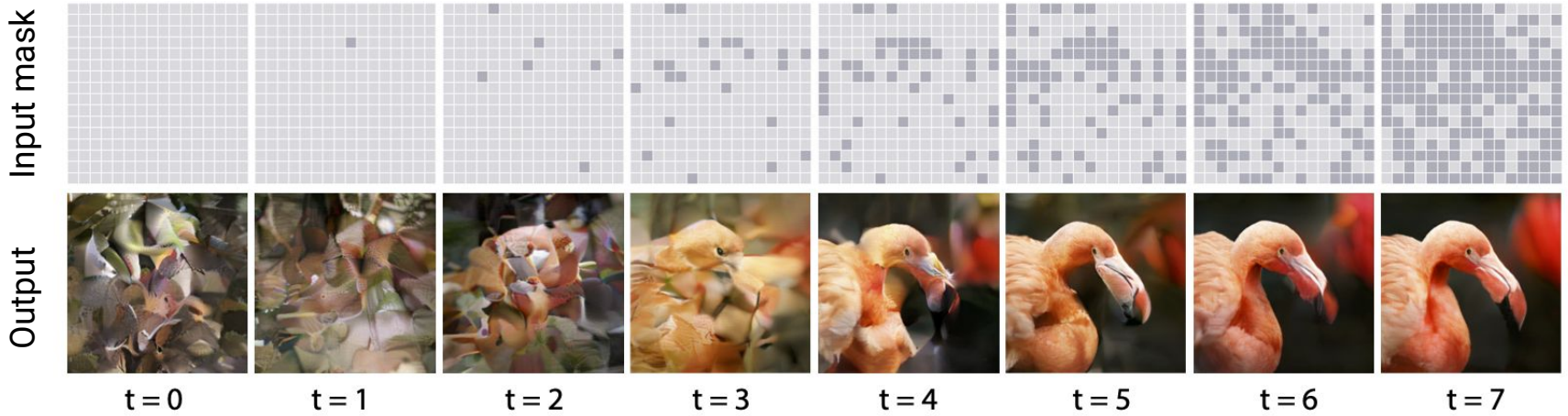


Image credit: MaskGIT

MaskGIT: Masked Generative Image Transformer, Chang et. al., 2022
VQGAN: Taming Transformers for High-Resolution Image Synthesis, Esser et. al., 2021

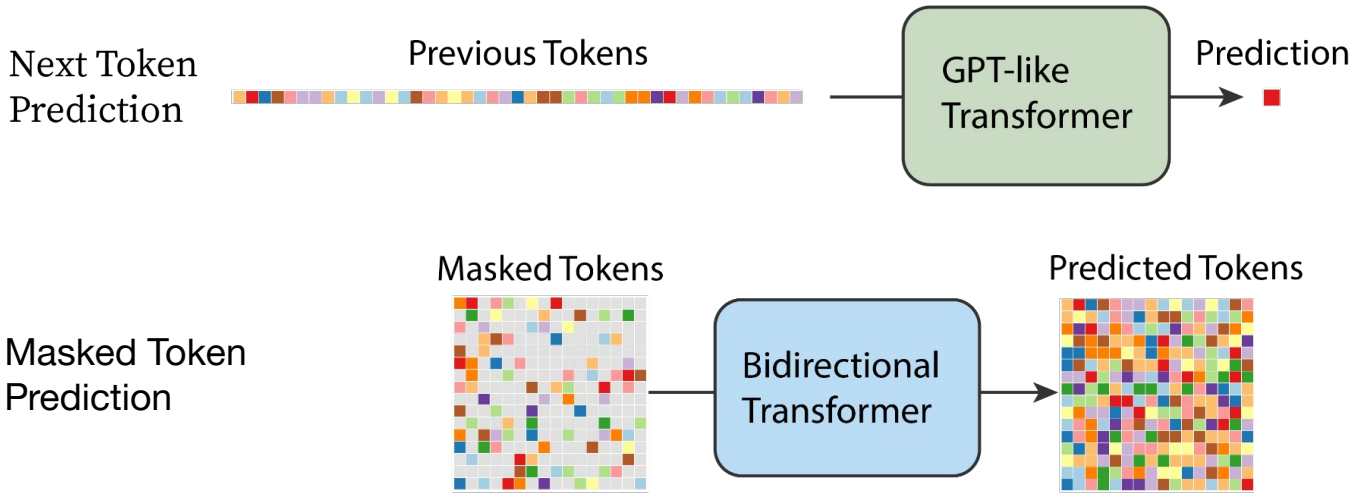
Sampling of masked models

- Parallel decoding in constant numbers of steps
  Predict → Mask out → Re-predict…



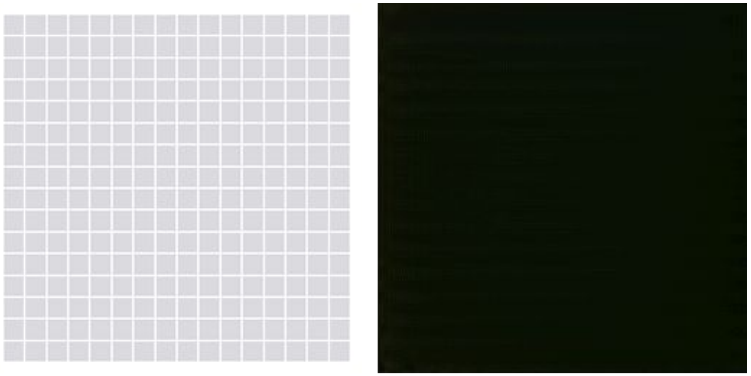
Three differences from autoregressive models

1. Bi-directional attention mechanism; all tokens attend to all others

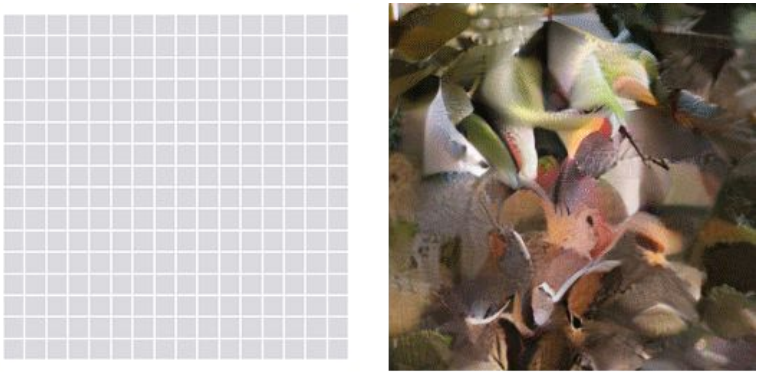


Three differences from autoregressive models

1. Bi-directional attention mechanism; all tokens attend to all others
2. Parallel decoding enables much faster sampling



Autoregressive Models (e.g. Parti, DALL·E) decodes in a raster scan order



Our parallel decoding is >20x faster

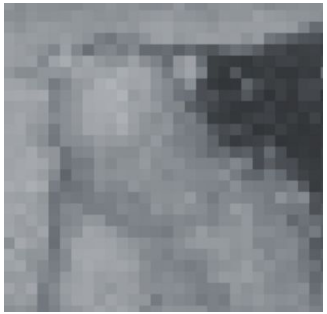
Three differences from autoregressive models

1. Bi-directional attention mechanism; all tokens attend to all others.
2. Parallel decoding enables much faster sampling speed.
3. Random masking during training enables zero-shot editing ability.

Inpainting

Extrapolation

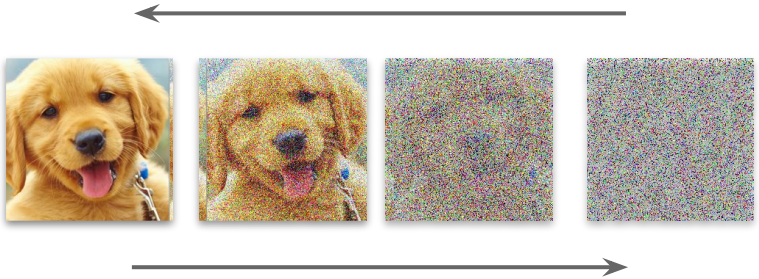
in any directions



Editing based on attention

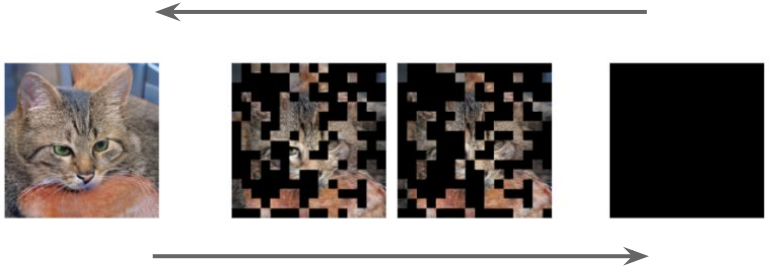
Connection to Discrete Diffusion

Continuous Diffusion
(e.g. Imagen-3, Dall-E 3)



Forward process is a destructive process by adding Gaussian noise

Discrete Diffusion



Forward pass is the absorbing state diffusion process in **image token space**

[Discrete Predictor-Corrector Diffusion Models for Image Synthesis](), Lezama et. al. 2023

Connection to Flow Models

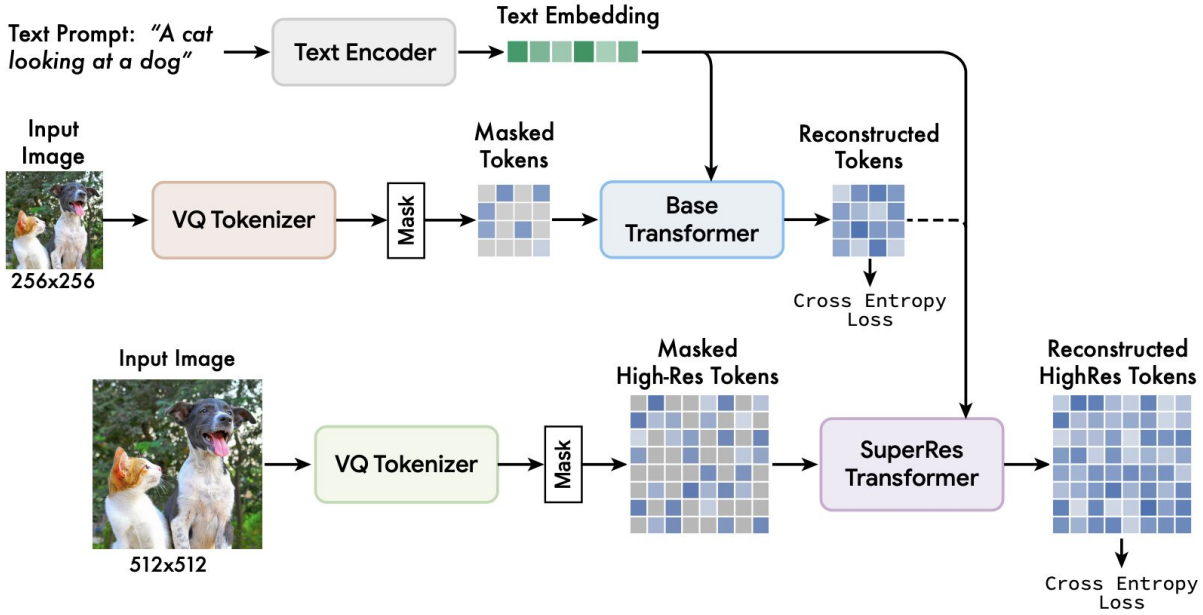
- Continuous Flow: transforms continuous distribution (e.g. Gaussian) to data distribution
- Discrete Flow: transforms discrete distribution (e.g. all masks) to data distribution
- Muse can be considered to be **a *time independent discrete flow model***

Discrete Flow Matching, Gat et. al. 2024

Muse Model

Four components:

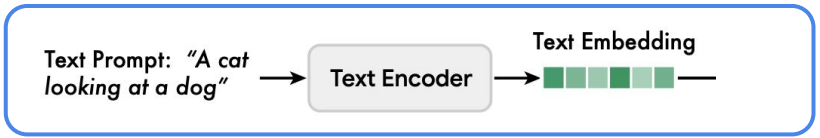
1. Pre-trained Text Encoder
2. Image Tokenizer
3. Base text-to-image generative transformer
4. SuperRes text-to-image generative transformer



Muse Model

Four components:

1. **Pre-trained Text Encoder**
2. Image Tokenizer
3. Base text-to-image generative transformer
4. SuperRes text-to-image generative transformer



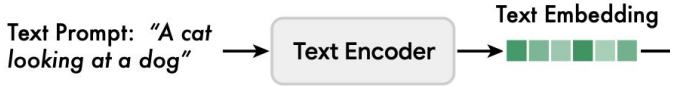
- T5-XXL pre-trained text model (4.6B parameters)
- Encodes text prompt into a sequence of 4096-D text embedding
- Chosen since it was used by the prior Imagen diffusion model

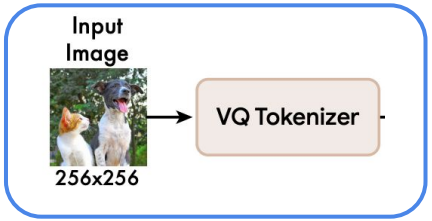
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Raffel et. al., 2019

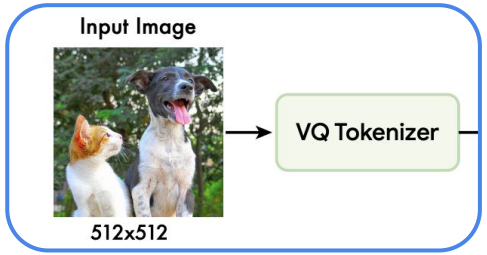
Muse Model

Four components:

1. Pre-trained Text Encoder
2. **Image Tokenizer**
3. Base text-to-image generative transformer
4. SuperRes text-to-image generative transformer







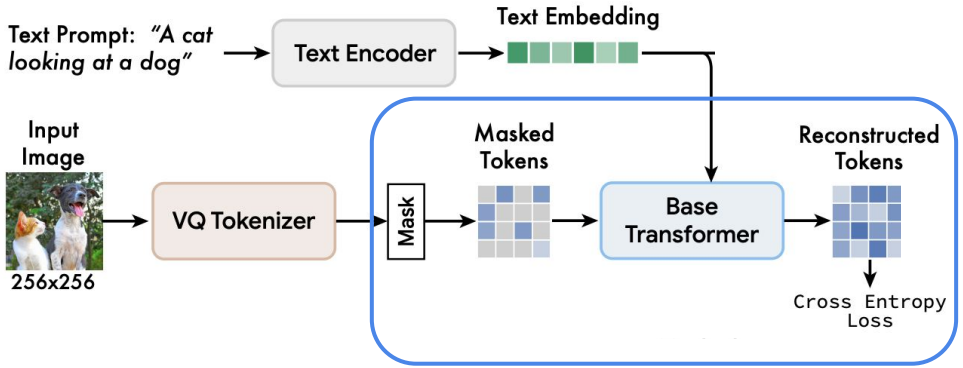
- We pre-train two VQ tokenizers at different resolution with image reconstruction loss, quantization loss, and GAN loss
- Base: Compress 256x256 images into 16x16 perceptual discrete image tokens
- Super-res: Compress 512x512 or 1024x1024 images into 64x64 perceptual discrete image tokens

MaskGIT: Masked Generative Image Transformer, Chang et. al., 2022
Taming Transformers for High-Resolution Image Synthesis, Esser et. al., 2021

Muse Model

Four components:

1. Pre-trained Text Encoder
2. Image Tokenizer
3. **Base text-to-image generative transformer**
4. SuperRes text-to-image generative transformer



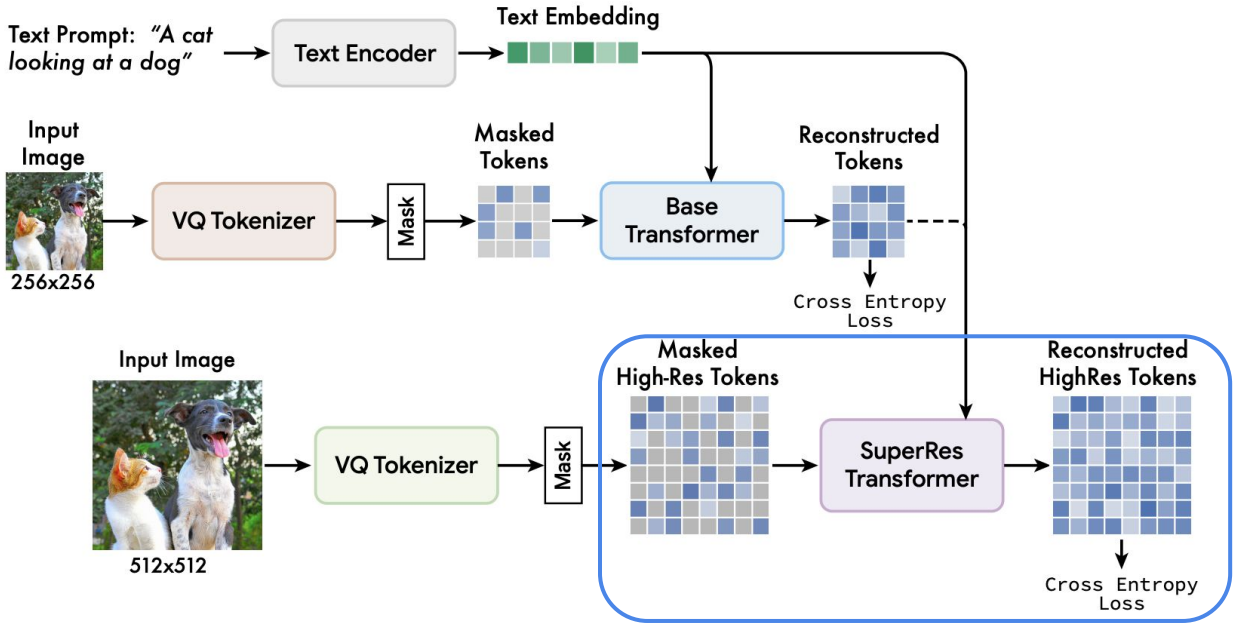
- We adopt the masked image modeling (MIM) from MaskGit
- During training, we first randomly sample a masking ratio in (0,1], to select masked tokens from the sequence
- Text embedding is projected and then attended to image tokens in the cross attention
- We use cross entropy loss on masked tokens

MaskGIT: Masked Generative Image Transformer, Chang et. al., 2022
Taming Transformers for High-Resolution Image Synthesis, Esser et. al., 2021

Muse Model

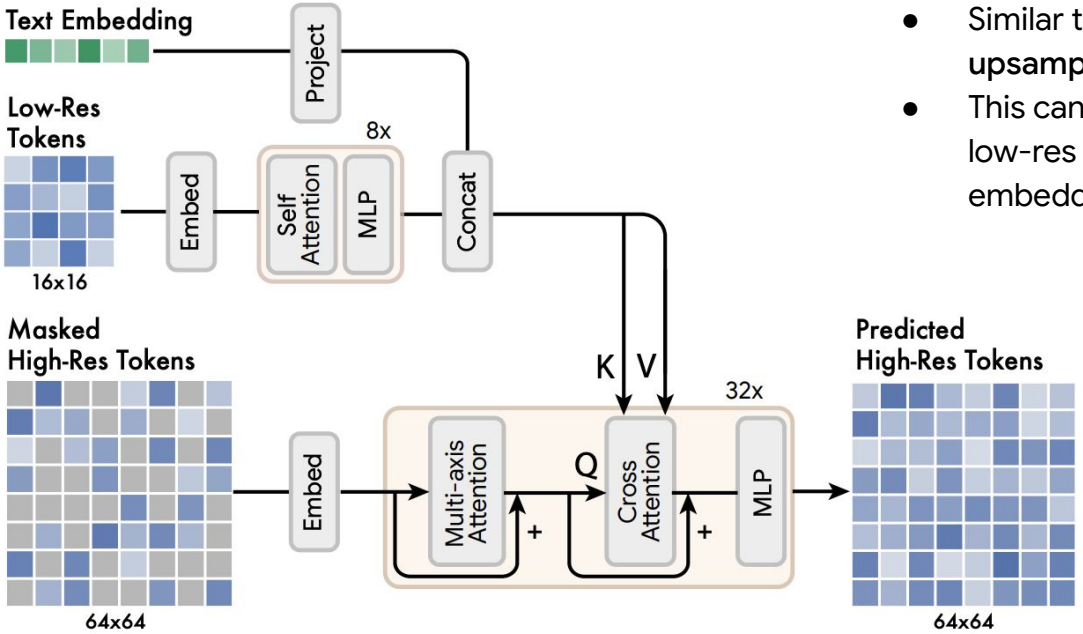
Four components:

1. Pre-trained Text Encoder
2. Image Tokenizer
3. Base text-to-image generative transformer
4. **SuperRes text-to-image generative transformer**



We find that directly predicting 64x64 tokens leads to worse text alignment, and cascaded models seem to mitigate this issue

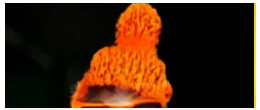
Muse SuperRes (Zoomed-in)

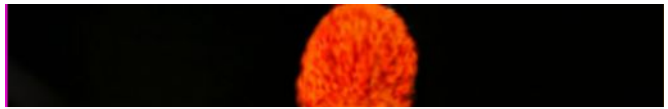


- Similar to our base model, we adopt the MIM tasks for **upsampling tokens** from 16x16 to 64x64
- This can be thought of as a translation task, from low-res tokens to high-res tokens, given text embedding.

Why can’t we borrow existing superRes model?

Prompt: *A high contrast portrait photo of a fluffy hamster wearing an orange beanie and sunglasses holding a sign that says "Lets paint"*





Note: This doesn’t imply that Muse SuperRes model is better than Imagen SuperRes model. The comparison is between Muse SR and Imagen SR results on the MUSE low res outputs. Unlike traditional SR model upsampling images in pixel space, Muse SR model upsamples coarse tokens to finer ones. Hence, it is specifically designed for Muse, as Muse operates in token space.

Muse SuperRes (In Token Space)

Imagen SuperRes (In Pixel Space)


Step = 2


Step = 4


Step = 6


Step = 8


Step = 12


Step = 14


Step = 16


Step = 18


Step = 2


Step = 4


Step = 8

Classifier-Free Guidance

- $\ell_g = (1+t)\ell_c - t\ell_u$
- Higher guidance scale → Better prompt alignment, less sample diversity

guidance scale = 1.0

guidance scale = 7.0

guidance scale = 25.0

Prompt: A pikachu drinking coffee, photorealistic.

Negative Prompting

- We can exploit guidance to “move away” from certain concepts/ideas in the generation
- Negative prompting is more effective than including negative word prompting such as “without, no, less”
- Implemented as negative guidance value

Prompt: A cat without ear; photorealistic.

\+ *Negative Prompt: cat ear.*

Experiments and Evaluation

- Metrics
  - Automated
    - FID ↓: sample fidelity and diversity
    - CLIP Score ↑: text-image alignment and fidelity
  - Human Preferences

- Datasets:
  - CC3M train and val set
  - MS-COCO validation set for zero-shot evaluation

Quantitative Eval on CC3M

Approach	Model Type	Params	FID	CLIP
VQGAN (Esser et al., 2021b)	Autoregressive	600M	28.86	0.20
ImageBART (Esser et al., 2021a)	Diffusion+Autogressive	2.8B	22.61	0.23
LDM-4 (Rombach et al., 2022)	Diffusion	645M	17.01	0.24
RQ-Transformer (Lee et al., 2022a)	Autoregressive	654M	12.33	0.26
Draft-and-revise (Lee et al., 2022b)	Non-autoregressive	654M	9.65	0.26
Muse(base model)	Non-autoregressive	632M	6.8	0.25
Muse(base + super-res)	Non-autoregressive	632M + 268M	6.06	0.26

Table 1. Quantitative evaluation on CC3M (Sharma et al., 2018); all models are trained and evaluated on CC3M.

Quantitative Eval on COCO

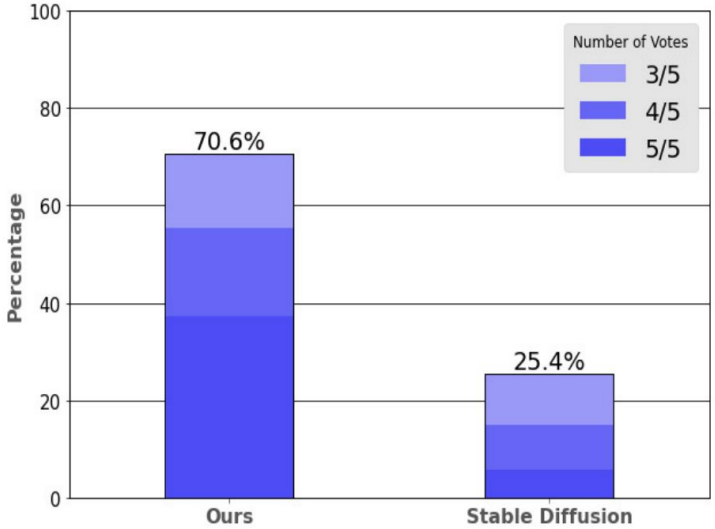
Approach	Model Type	FID-30K	Zero-shot FID-30K
AttnGAN (Xu et al., 2017)	GAN	35.49	-
DF-GAN (Tao et al., 2020)	GAN	21.42	-
XMC-GAN (Zhang et al., 2021)	GAN	9.33	-
LAFITE (Zhou et al., 2021)	GAN	8.12	-
Make-A-Scene (Gafni et al., 2022)	Autoregressive	7.55	-
DALL-E (Ramesh et al., 2021)	Autoregressive	-	17.89
CogView (Ding et al., 2021)	Autoregressive	-	27.1
LAFITE (Zhou et al., 2021)	GAN	-	26.94
VQ-Diffusion (Gu et al., 2022)	Diffusion	13.86^{F}	19.75
LDM (Rombach et al., 2022)	Diffusion	-	12.63
GLIDE (Nichol et al., 2021)	Diffusion	-	12.24
DALL-E 2 (Ramesh et al., 2022)	Diffusion	-	10.39
Imagen-3.4B (Saharia et al., 2022)	Diffusion	-	7.27
Parti-3B (Yu et al., 2022b)	Autoregressive	-	8.10
Parti-20B (Yu et al., 2022b)	Autoregressive	3.22^{F}	7.23
Muse-3B-512	Non-Autoregressive	-	7.88
Muse-3B-1024	Non-Autoregressive	-	7.39

Runtime on TPU-v4

Approach	Resolution	Time
Imagen	256×256	9.1s
Parti-3B	256×256	6.4s
Muse-3B	256×256	**0.5s**
LDM (250 steps)	512×512	8.2s
LDM (50 steps)	512×512	1.7s
Muse-3B	512×512	**1.3s**
Imagen	1024×1024	13.3s
Muse-3B	1024×1024	**1.4s**

Qualitative Evals

- Conduct a user study on 1650 prompts (P2)
- Muse matches prompt better than Stable Diffusion-1.0 **70.6%** of the time
- New versions of SD are much better due to new data/larger models



Composition

Three small yellow boxes on a large blue box.

A large present with a red ribbon to the left of a Christmas tree.

Two baseballs to the left of three tennis balls.

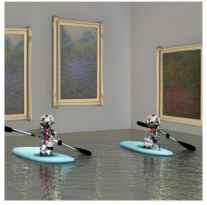
Text Rendering

A t-shirt with Carpe Diem written on it.

High-contrast image of the word "WOMBAT" written with thick colored graffiti letters on a white wall with dramatic splashes of paint.

The saying "BE EXCELLENT TO EACH OTHER" written in a stained glass window.

Usage of Entire Prompt

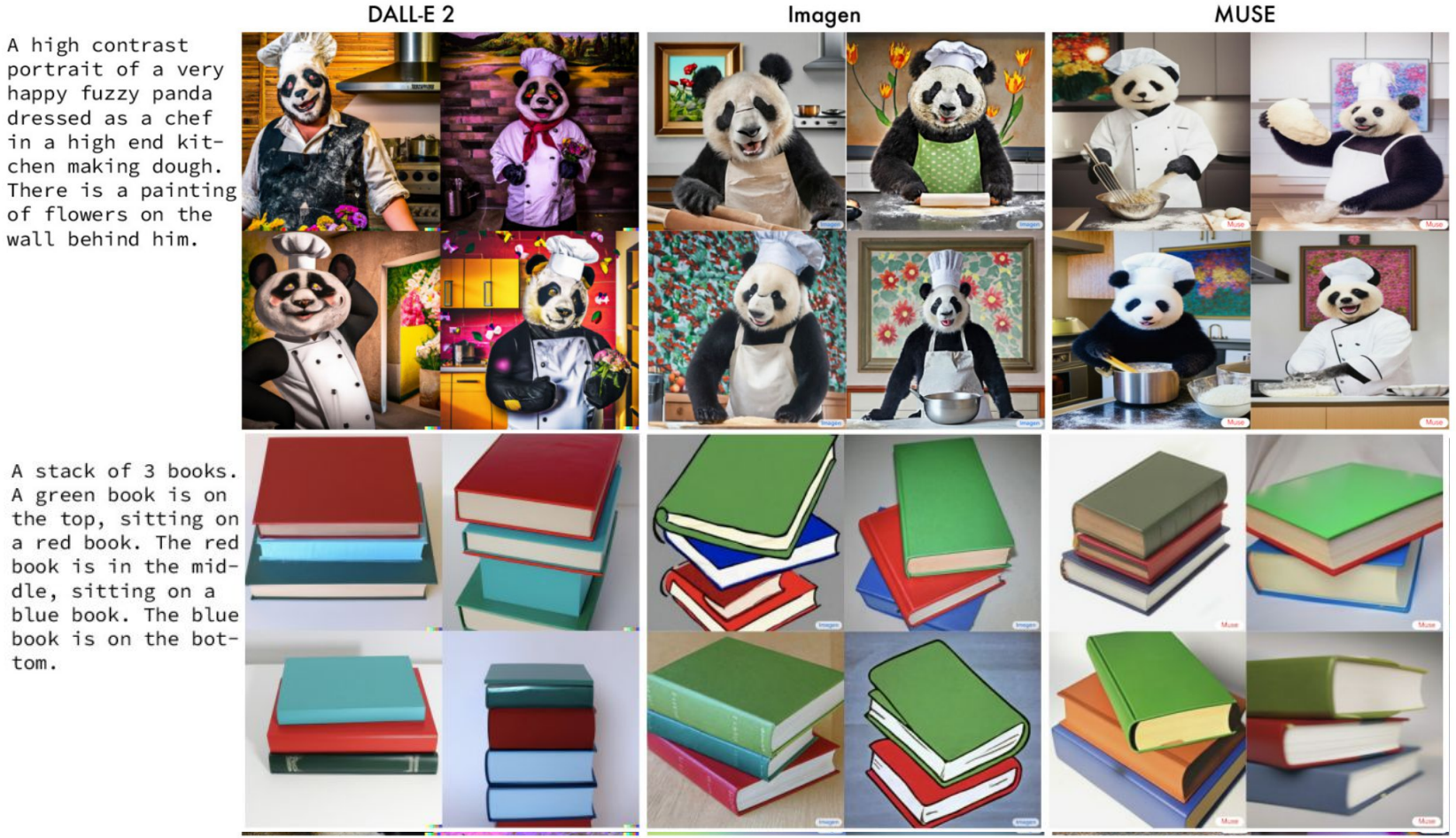


An art gallery displaying Monet paintings. The art gallery is flooded. Robots are going around the art gallery using paddle boards.

A photograph of the inside of a subway train. There are raccoons sitting on the seats. One of them is reading a newspaper. The window shows the city in the background.

Two cups of coffee, one with latte art of yin yang symbol. The other has latter art of a heart.

Subjective comparisons to other SOTA models



Study on Super Resolution Comparison

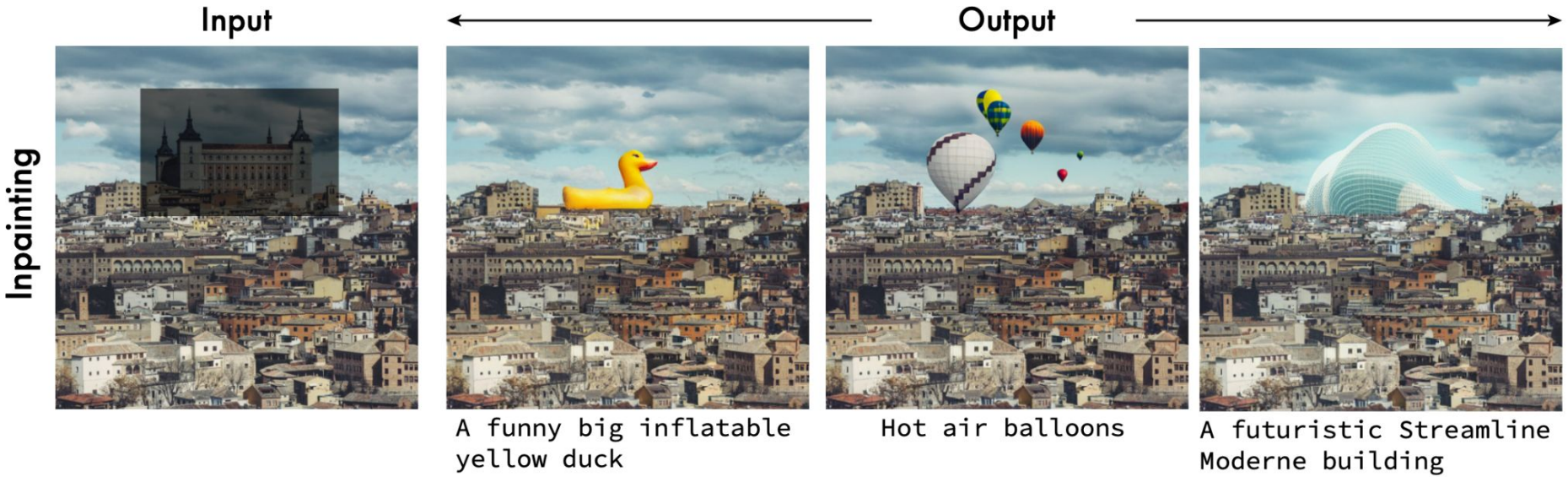
Prompt: *A Welsh corgi holding a sign in its mouth that says 'Muse' on a sunny day. Award winning.*

256x256

256x256 -> 512x512

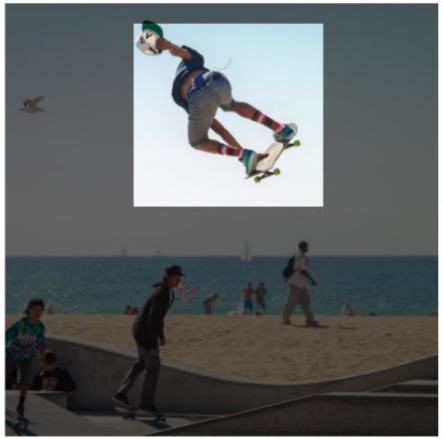
256x256 -> 1024x1024

Text Guided Inpainting (no fine-tuning)



Text Guided Outpainting (Uncropping)

Outpainting

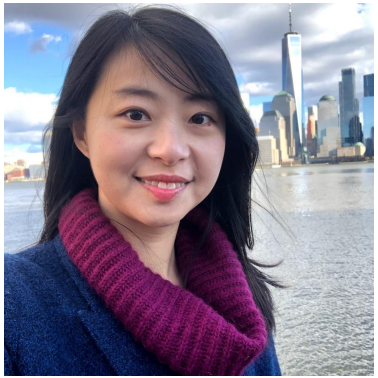


London skyline

A wildflower bloom at Mountain Rainier

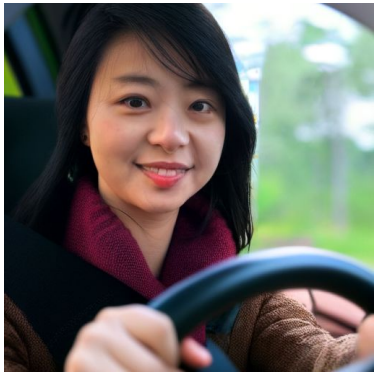
On the ring of Saturn

Text Guided Outpainting (Uncropping)



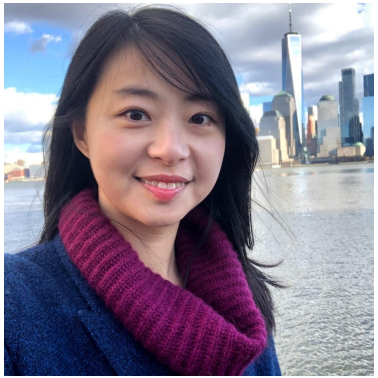
[Huiwen] in a bar

[Huiwen] as a sushi chef



[Huiwen] driving

Text Guided Outpainting (Uncropping)



[Huiwen] in a bar

[Huiwen] and Einstein drinking beer in a bar

Mask-free Editing on real image inputs

Mask-free Editing

A man wearing a blue t-shirt with “hello world” written on it

A man wearing a christ- mas sweater.

A woman wearing a dress

Mask-free Editing

Conclusions and Future Work

- Fast, high-performing discrete flow parallel decoding model
- Out-of-the-box editing capabilities
- Easy to integrate into multimodal LLM models
- Tends to be lower performing than diffusion based models, especially for higher frequencies and details
  - Can potentially be overcome with combining diffusion and quantization approaches

https://muse.github.io/