

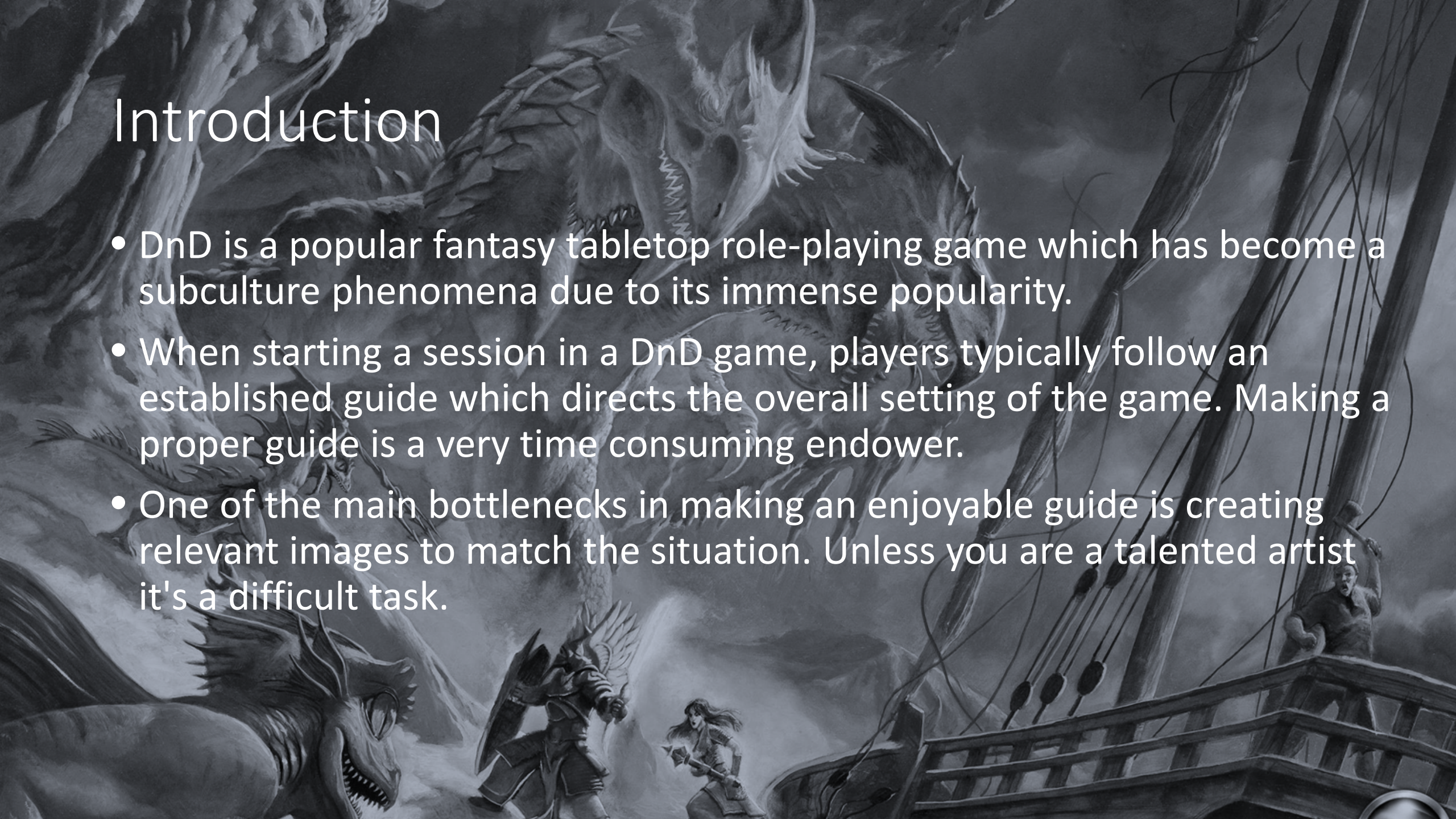
Fantasy Image Generation Relevant to the Context Of a Given D&D Adventure

Table of content

- Introduction
- Literature Review
- Other Notable Research Papers
- Methodology
- Test Results
- Reference

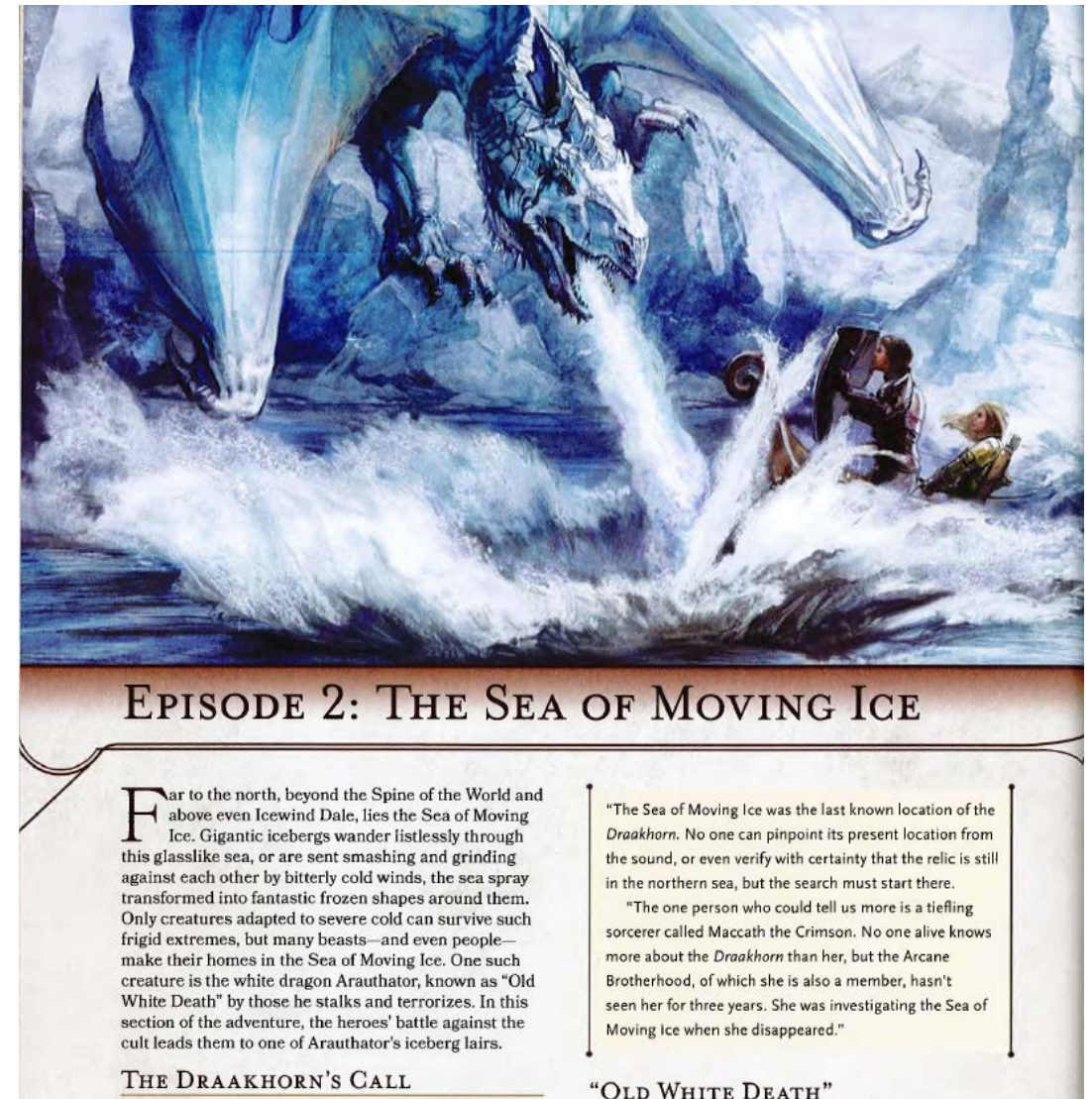
Introduction

- DnD is a popular fantasy tabletop role-playing game which has become a subculture phenomena due to its immense popularity.
- When starting a session in a DnD game, players typically follow an established guide which directs the overall setting of the game. Making a proper guide is a very time consuming endeavor.
- One of the main bottlenecks in making an enjoyable guide is creating relevant images to match the situation. Unless you are a talented artist it's a difficult task.



Introduction

- The guide is in a narrative format and consists mainly of text accompanied by images to influence the mood and help players immerse themselves.
- The purpose of this research is to generate coherent and contextually consistent images by taking the pre-made narrative as a text input.
- The generated images should not conflict with the mood the guide tries to establish.



Introduction

- The field of Image generation has been evolving rapidly with time from GAN models to current diffusion based models.
- Recent introduction of stable diffusion, a 860 million parameter open source text to image diffusion model, has opened up many paths for individual and small team based developers to experiment with diffusion models without massive hardware requirements typically associated with these kinds of models.

Objective

- The goal is to generate cohesive images in the style of D&D modules given the text of an adventure.
- The generated images should adhere to the lore and context of the adventure.

Vector Quantized Diffusion Model for Text-to-Image Synthesis

Shuyang Gu

Dong Chen

Jianmin Bao

Fang Wen

Bo Zhang

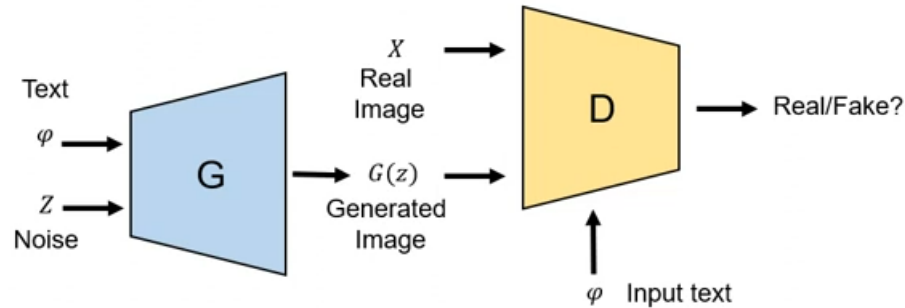
Dongdong Chen

Lu Yuan

Baining Guo

Related Work

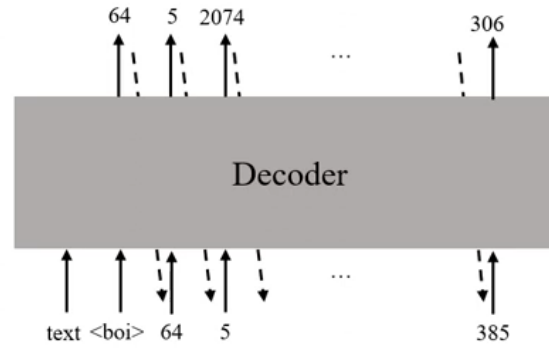
- GAN-based Text-to-image generation.



- Limitations
 - Single domain images
 - Cannot handle complex images

Related Work

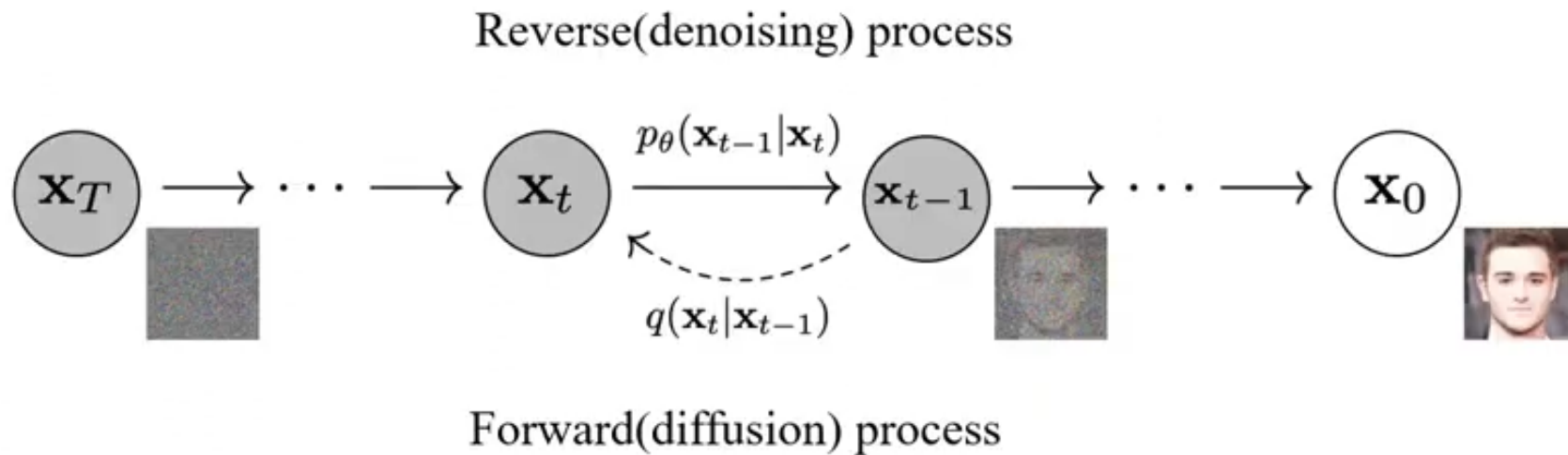
- Auto regressive model.



- Limitations
 - Uni directional bias
 - Accumulated errors
 - Slow in inference

Related Work

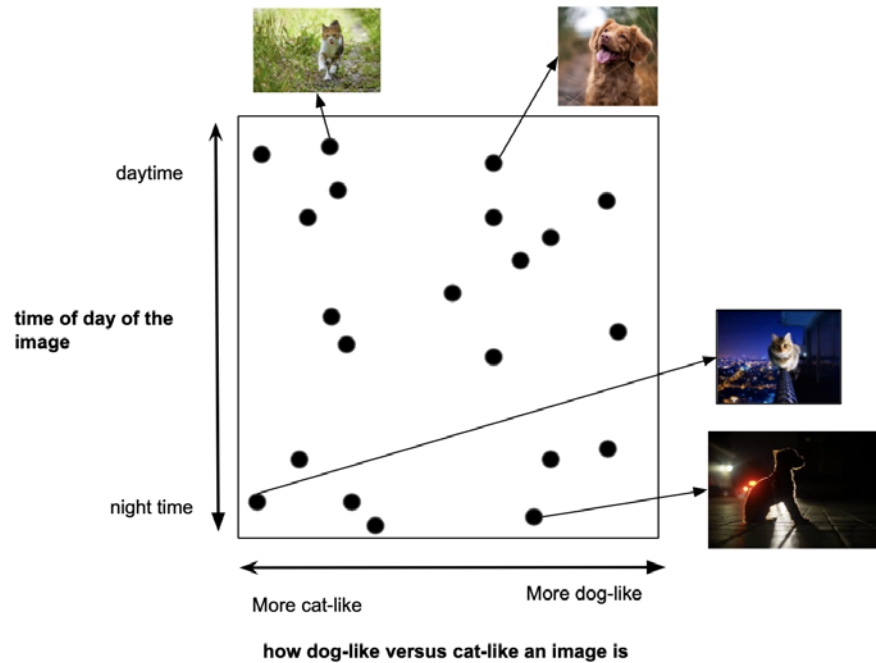
- Denoising Diffusion Probabilistic Models.



Latent Spaces

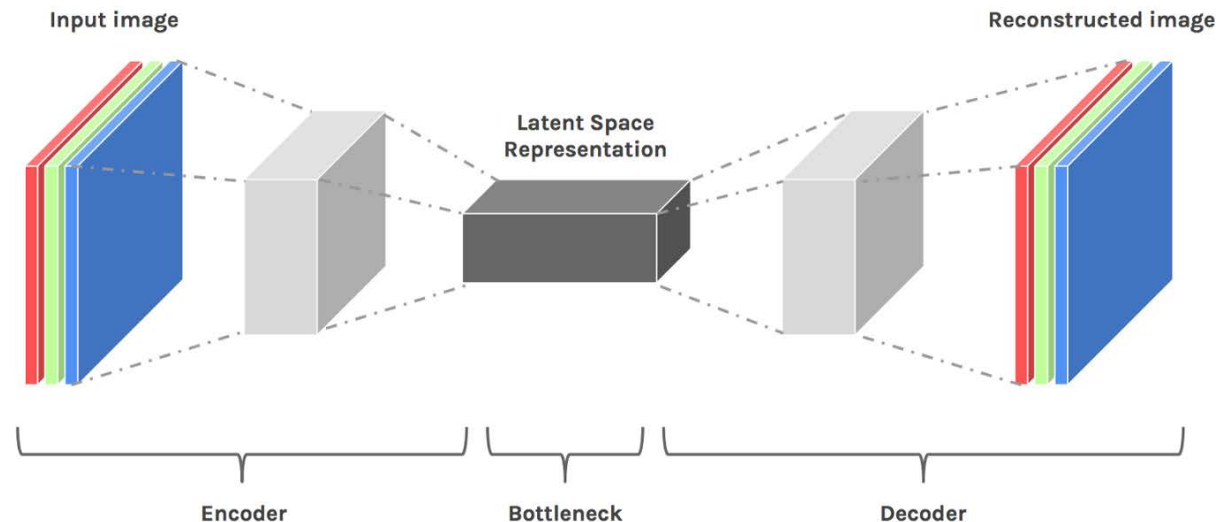
- A latent space is some underlying “hidden” representation for a given distribution of raw data.

An Oversimplified Example of a Cat/Dog Image Latent Space



Autoencoders

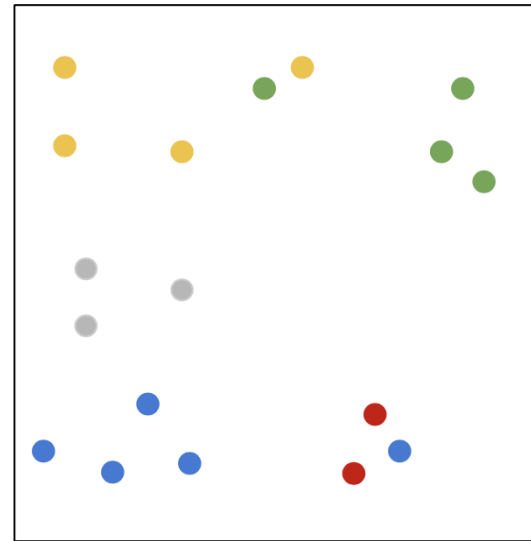
- An autoencoder is an unsupervised learning technique that uses neural networks to find non-linear latent representations for a given data distribution.
- The neural network consists of two parts, an encoder network and a decoder network.



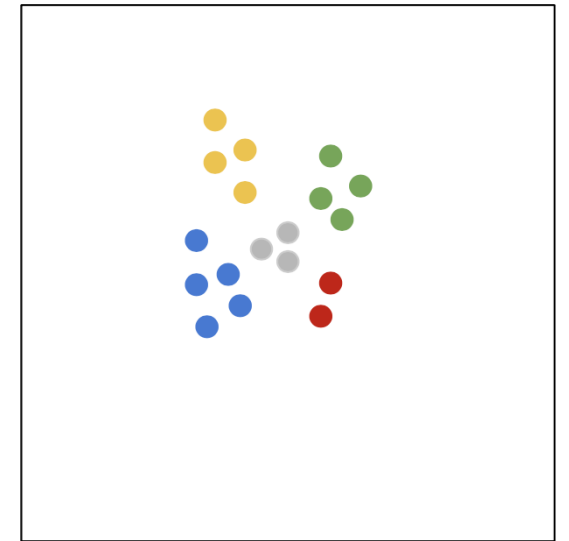
Variational Autoencoders (VAE)

- The main issue with vanilla autoencoders is that the model can learn any latent space it wants, so it often ends up memorizing individual data points by placing them in their own far out pockets of latent space.
- Variational autoencoders overcome this problem by enforcing a probabilistic prior (Typically a gaussian) on the latent space.

Messy Autoencoder Latent Space

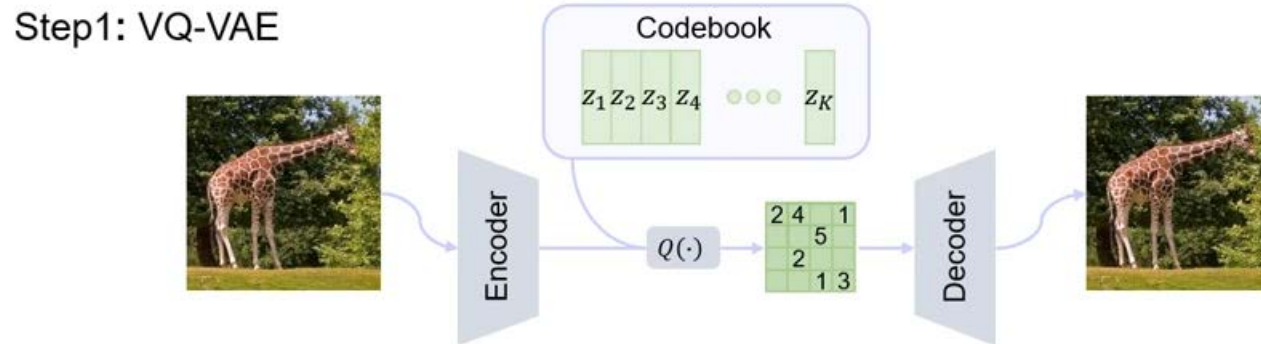


Well Distributed VAE Latent Space



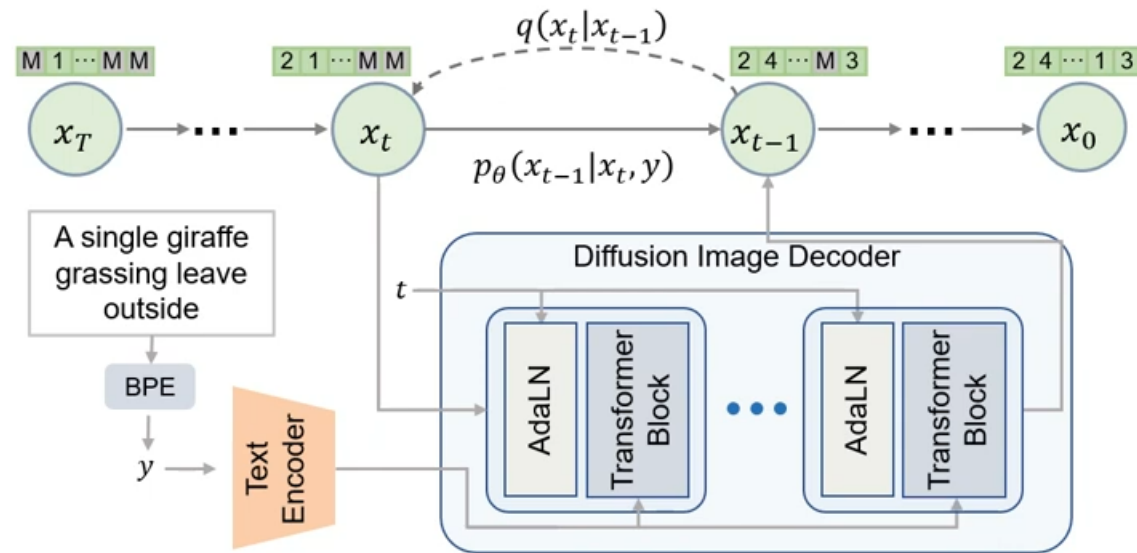
Vector Quantized Diffusion Model

- Two step process
 - VQ-VAE, which reduces context length
 - VQ-Diffusion, which models the discrete latent space



VQ-Diffusion

- Add noise in discrete space (mask & replace)
- Recover the original sample from noise with transformer



Mask & Replace diffusion strategy

- Corrupt the tokens by stochastically masking some of them so that the corrupted locations can be explicitly known by the reverse network.
- Specifically, we introduce an additional special token, [MASK] token, so each token now has $(K + 1)$ discrete states.
- each ordinary token has a probability of γ_t to be replaced by the [MASK] token and has a chance of $K\beta_t$ to be uniformly diffused, leaving the probability of $\alpha_t = 1 - K\beta_t - \gamma_t$ to be unchanged, whereas the [MASK] token always keeps its own state.

Mask & Replace diffusion strategy

- Hence, we can formulate the transition matrix $Q_t \in \mathbb{R}^{(K+1) \times (K+1)}$ as,

$$Q_t = \begin{bmatrix} \alpha_t + \beta_t & \beta_t & \beta_t & \dots & 0 \\ \beta_t & \alpha_t + \beta_t & \beta_t & \dots & 0 \\ \beta_t & \beta_t & \alpha_t + \beta_t & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_t & \gamma_t & \gamma_t & \dots & 1 \end{bmatrix}$$

Benefit of Mask & Replace diffusion strategy

- the corrupted tokens are distinguishable to the network, which eases the reverse process.
- The random token replacement forces the network to understand the context rather than only focusing on the [MASK] tokens.
- The computation cost of $q(x_t|x_0)$ is reduced from $O(tK^2)$ to $O(K)$.

Results

A handsome man with thick eyebrows and moustache



Snow mountain and tree reflection in the lake



The man with sunglasses has a big beard



A green heart with shadow



A red bus is driving on the road



A picture of a very tall stop sign



The sunset on the beach is wonderful



A bare kitchen has wood cabinets and white appliances



A movie poster of mountain and sea



A cartoon boy is smiling



Two smiling beautiful ladies are standing together



Black and white icon of man and woman



A man wears black suit and a tie



Icon of a red heart



A woman is talking in an interview



A man with beard in 1920s



A giraffe walking through a green grass covered field



A group of people gather for a photo



A mountain near the lake



Sunset over the skyline of a city



Two girls in cartoon style



The face of Bill Gates



A woman with curly hairs and brown skin



A picture of some food in the plate



Qualitative comparison

CUB-200 dataset

MSCOCO dataset



FID comparison on different models

	MSCOCO	CUB-200	Oxford-102
StackGAN	74.05	51.89	55.28
StackGAN++	81.59	15.30	48.68
AttnGAN	35.49	23.98	-
DM-GAN	32.64	16.09	-
DF-GAN	21.42	14.81	-
DALLE	27.50	56.10	-
Cogview	27.10	-	-
VQ-Diffusion-S	-	12.97	14.95
VQ-Diffusion-B	19.75	11.94	14.88
VQ-Diffusion-F	13.86	10.32	14.10

Ablation studies on fast inference strategies

	Training steps					
Inference steps		10	25	50	100	200
	10	32.35	27.62	23.47	19.84	20.96
	25	-	18.53	15.25	14.03	16.13
	50	-	-	13.82	12.45	13.67
	100	-	-	-	11.94	12.27
	200	-	-	-	-	11.80

Text to Image Generation with Semantic-Spatial Aware GAN

Wentong Liao

Kai Hu

Michael Ying Yang

Bodo Rosenhahn

Text to Image Generation

**Input
Text**

This is a gray bird with black wings and white wingbars light yellow sides and yellow eyebrows.

**Output
Image**



Single object

A horse in a grassy field set against a foggy mountain range

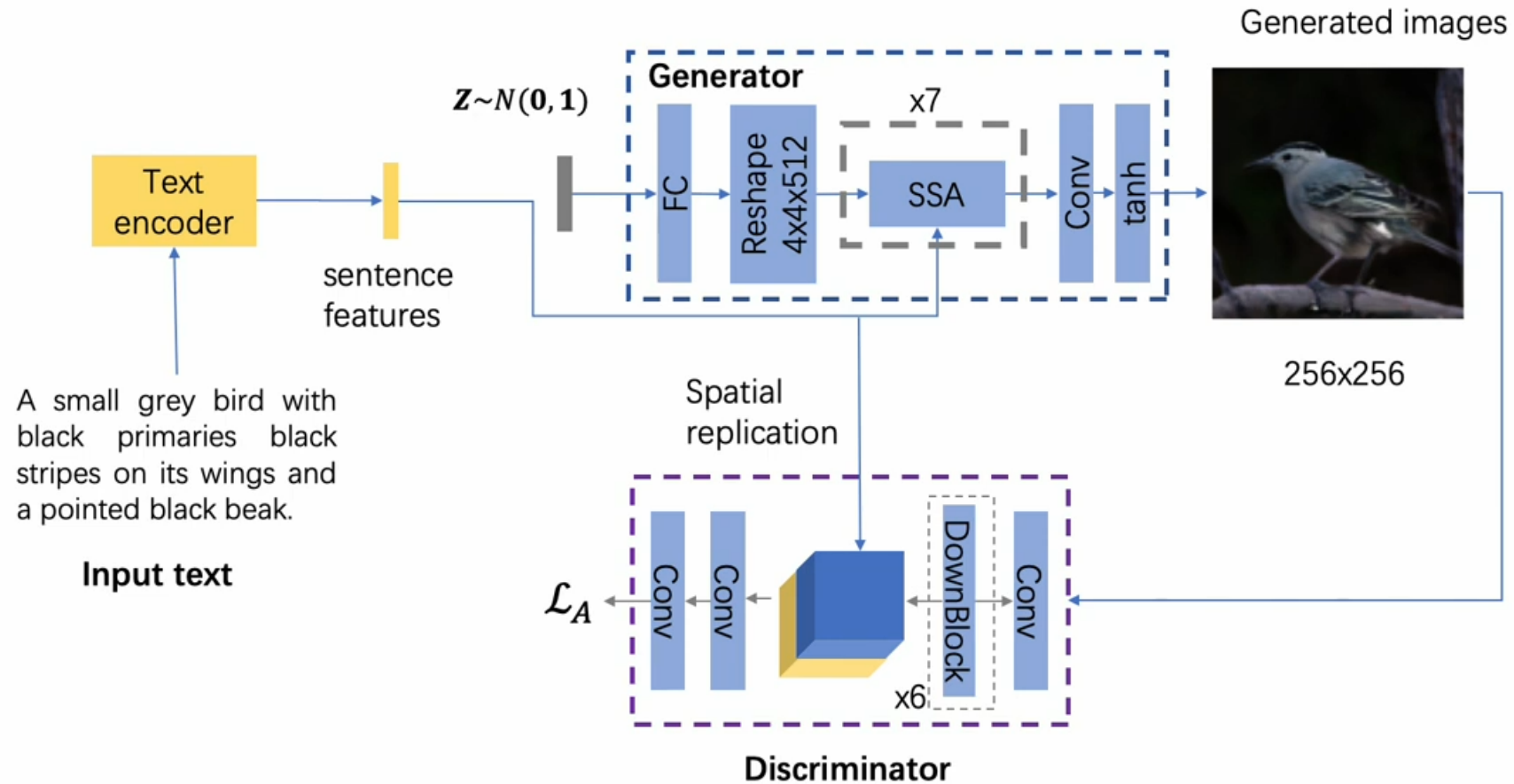


Complex scene with multiple objects

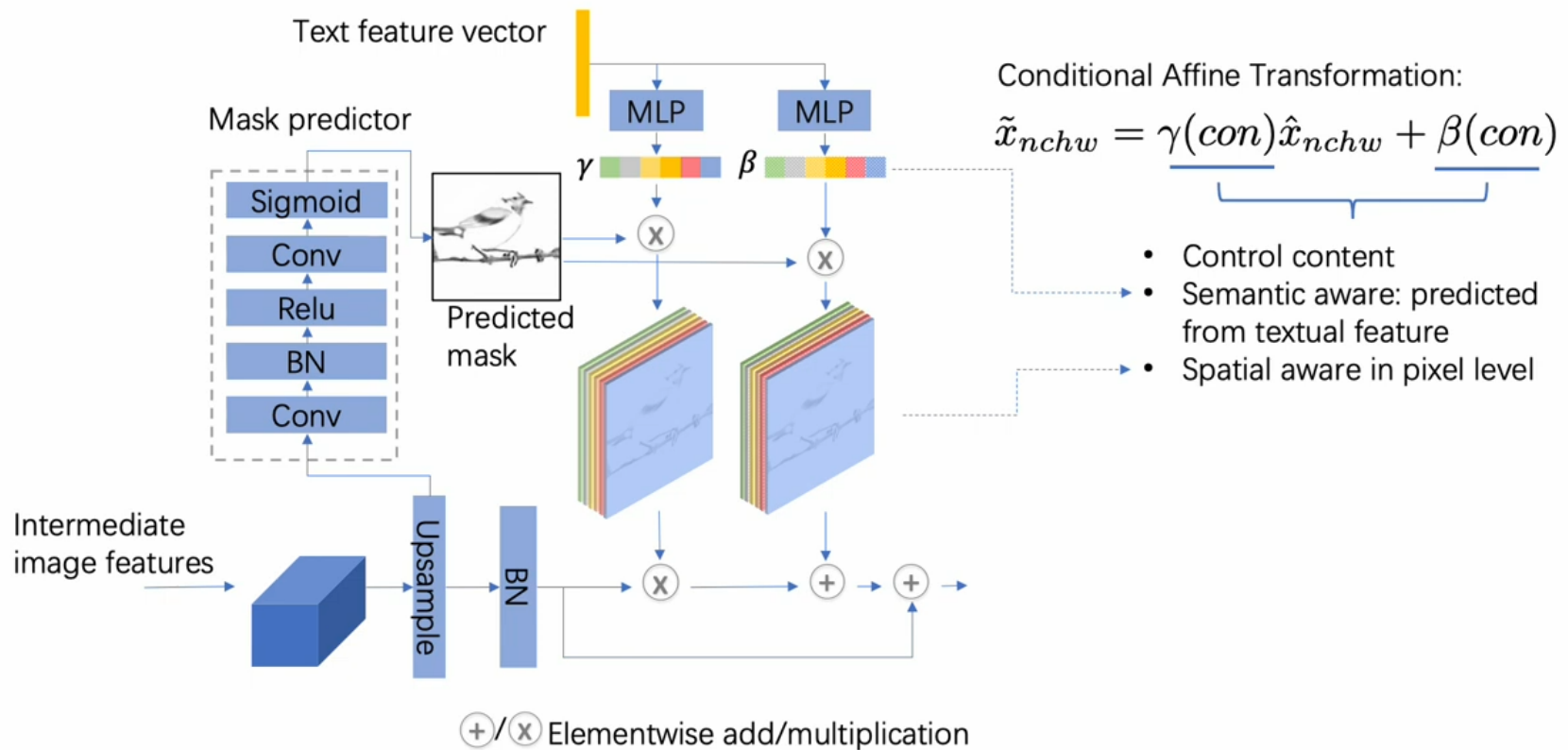
Challenges

- **Cross domain problem** – effective fusion of natural language information (text) and visual information (image)
- **Semantic consistency** – generated image should be holistically/locally consistent with the text
- **Abstractive textual information** – text information is abstract and lack explicit spatial information

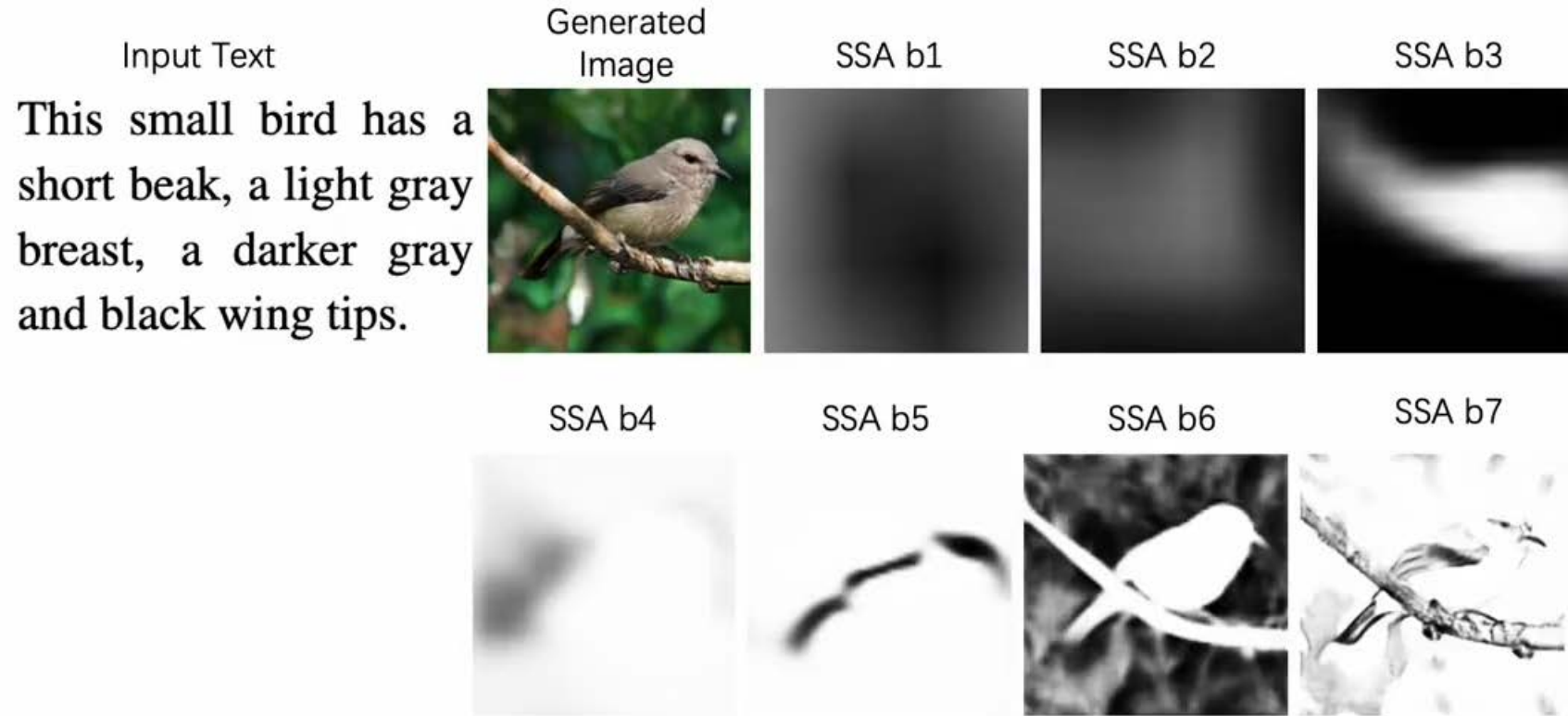
Semantic special aware GAN



Semantic special aware block



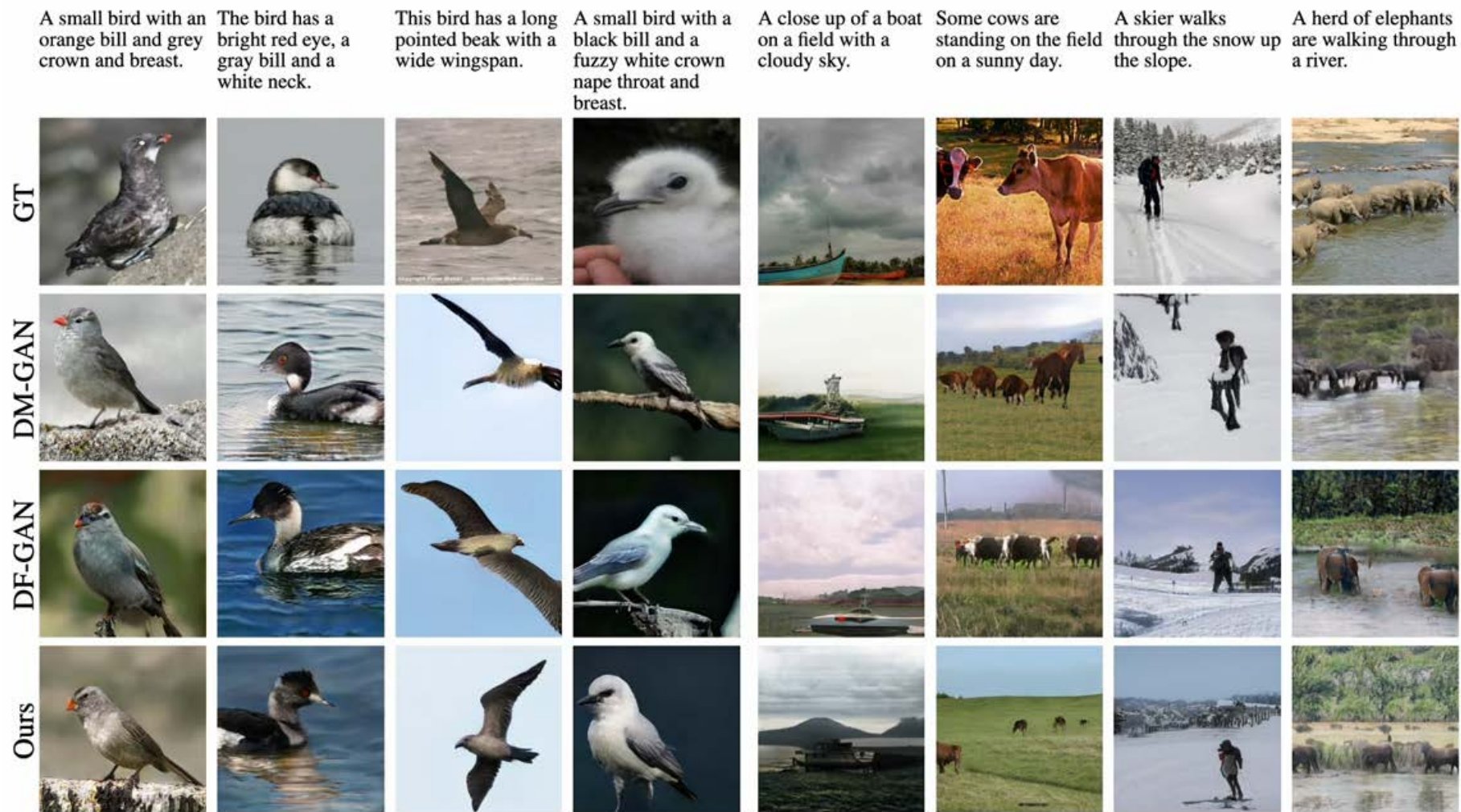
Visualization of predicted masks



Quantitative results

Methods	IS \uparrow	FID \downarrow		R-precision \uparrow	
	CUB	CUB	COCO	CUB	COCO
StackGAN++ [35]	4.04 \pm 0.06	15.30	81.59	-	-
AttnGAN [30]	4.36 \pm 0.03	23.98	35.49	67.82 \pm 4.43	85.47 \pm 3.69
ControlGAN [16]	4.58 \pm 0.09	-	-	69.33 \pm 3.23	82.43 \pm 2.43
SD-GAN [31]	4.67 \pm 0.09	-	-	-	-
DM-GAN [39]	4.75 \pm 0.07	16.09	32.64	72.31 \pm 0.91	88.56 \pm 0.28
DF-GAN [28]	4.86 \pm 0.04	19.24	28.92	-	-
DF-GAN \ddagger [28]	5.10	14.81	21.42	-	-
DAE-GAN [24]	4.42 \pm 0.04	15.19	28.12	85.4\pm0.57	92.6\pm0.50
Ours	5.17\pm0.08	15.61	19.37	75.9 \pm 0.92	90.6 \pm 0.71

Qualitative result



Scaling Autoregressive Models for Content-Rich Text-to-Image Generation

- Jiahui Yu
- Yuanzhong Xuy
- Jing Yu Kohy
- Thang Luongy
- Gunjan Baidy
- Zirui Wang
- Vijay Vasudevany
- Alexander Kuy
- Yinfei Yang
- Burcu Karagol Ayan
- Ben Hutchinson
- Wei Han
- Zarana Parekh
- Xin Li Han Zhang
- Jason Baldridgey
Yonghui Wu

Examples



Figure 1: Example images generated by Parti. **Top row:** “Oil-on-canvas painting of a blue night sky with roiling energy. A fuzzy and bright yellow crescent moon shining at the top. Below the exploding yellow stars and radiating swirls of blue, a distant village sits quietly on the right. Connecting earth and sky is a flame-like cypress tree with curling and swaying branches on the left. A church spire rises as a beacon over rolling blue hills.” (a 67-word description of the *Starry Night* by Vincent van Gogh). **Middle row:** “A close-up high-contrast photo of Sydney Opera House sitting next to Eiffel tower, under a blue night sky of roiling energy, exploding yellow stars, and radiating swirls of blue”. **Last row:** Similar to the middle row, but with “anime illustration” and different landmarks (the Great Pyramid and the Parthenon).



A. A photo of a frog reading the newspaper named “Toaday” written on it. There is a frog printed on the newspaper too.



B. A portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket. The city of Los Angeles is in the background. Hi-res DSLR photograph.



C. A high-contrast photo of a panda riding a horse. The panda is wearing a wizard hat and is reading a book. The horse is standing on a street against a gray concrete wall. Colorful flowers and the word “PEACE” are painted on the wall. Green grass grows from cracks in the street. DSLR photograph, daytime lighting.



D. A giant cobra snake made from X . $X \in \{\text{“salad”, “pancakes”, “sushi”, “corn”}\}$



E. A wombat sits in a yellow beach chair, while sipping a martini that is on his laptop keyboard. The wombat is wearing a white panama hat and a floral Hawaiian shirt. Out-of-focus palm trees in the background. DSLR photograph. Wide-angle view.



F. The saying “BE EXCELLENT TO EACH OTHER” ..., (a) brick wall and alien (b) driftwood. (c) old wooden boat with reflection. (d) stained glass. (See text for full prompts.)

Parti Model

- Parti is a two-stage model, composed of an image tokenizer and an autoregressive model
- The first stage involves training a tokenizer that turns an image into a sequence of discrete visual tokens for training and reconstructs an image at inference time.
- The second stage trains an autoregressive sequence-to-sequence model that generates image tokens from text tokens.

Model Overview

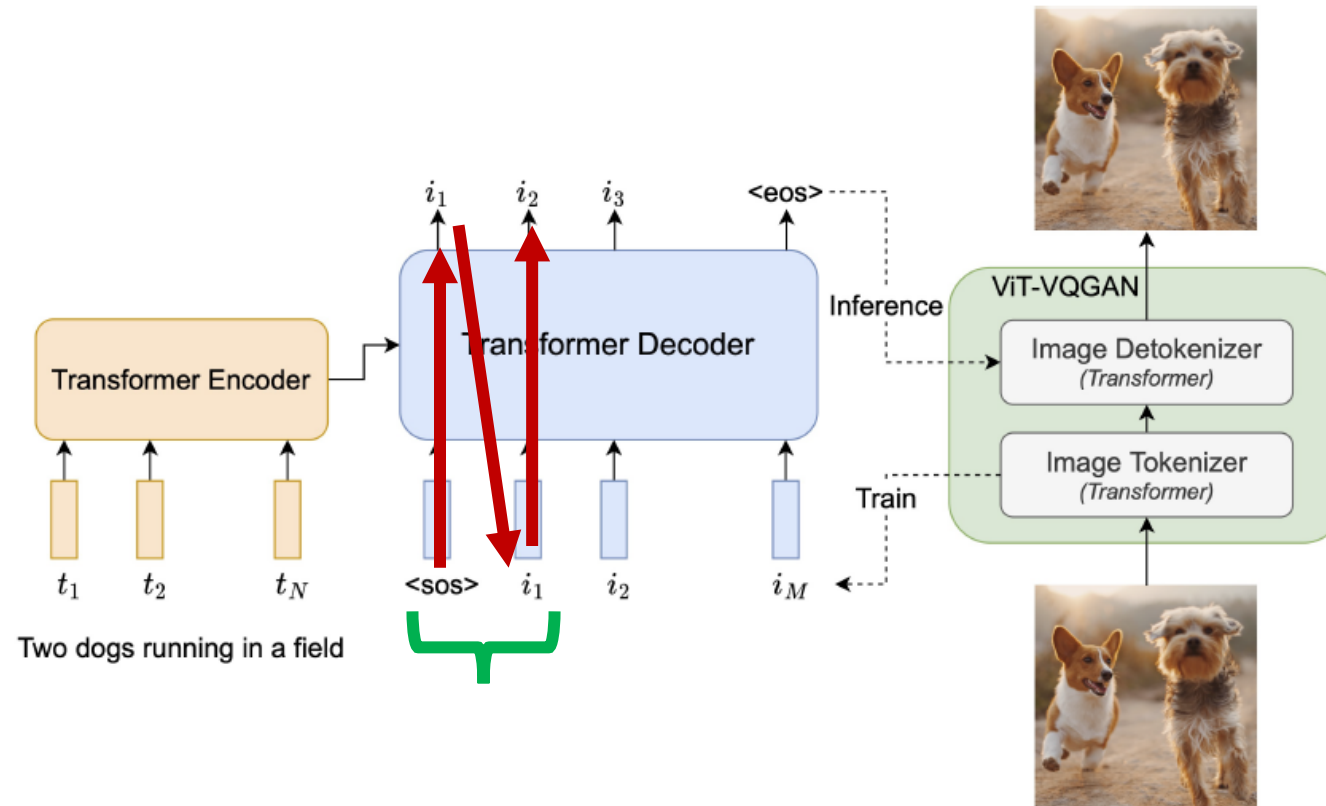


Figure 3: Overview of Parti sequence-to-sequence autoregressive model (left) for text-to-image generation with ViT-VQGAN as the image tokenizer [21] (right).

Model Overview

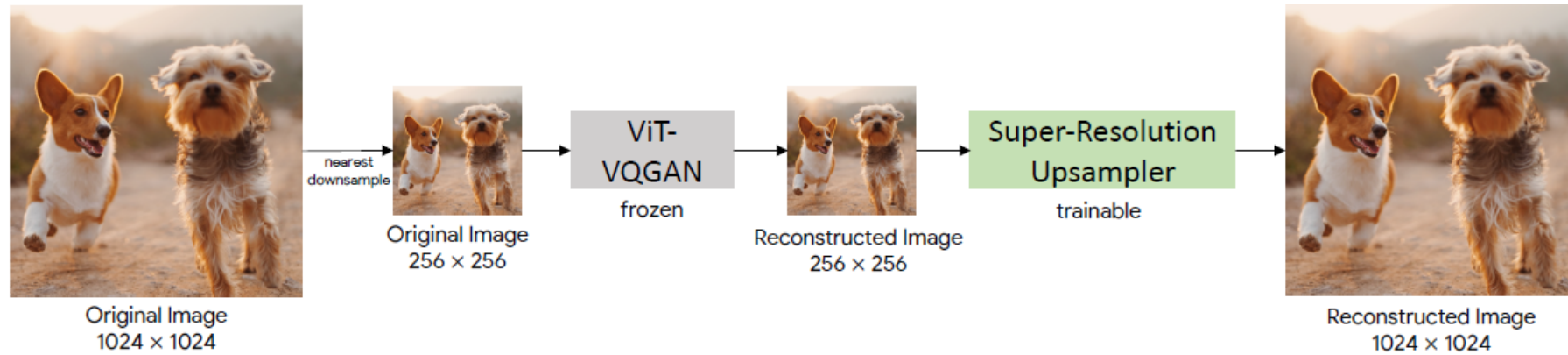


Figure 4: A learned super-resolution module to upsample 256×256 images to higher-resolution 1024×1024 ones based on a frozen ViT-VQGAN image tokenizer. The super-resolution module takes 256×256 images as inputs without conditioning on text inputs.

Size variants of Parti

Model	Encoder Layers	Decoder Layers	Model Dims	MLP Dims	Heads	Total Params
Parti-350M	12	12	1024	4096	16	350M
Parti-750M	12	36	1024	4096	16	750M
Parti-3B	12	36	2048	8192	32	3B
Parti	16	64	4096	16384	64	20B

Scaling

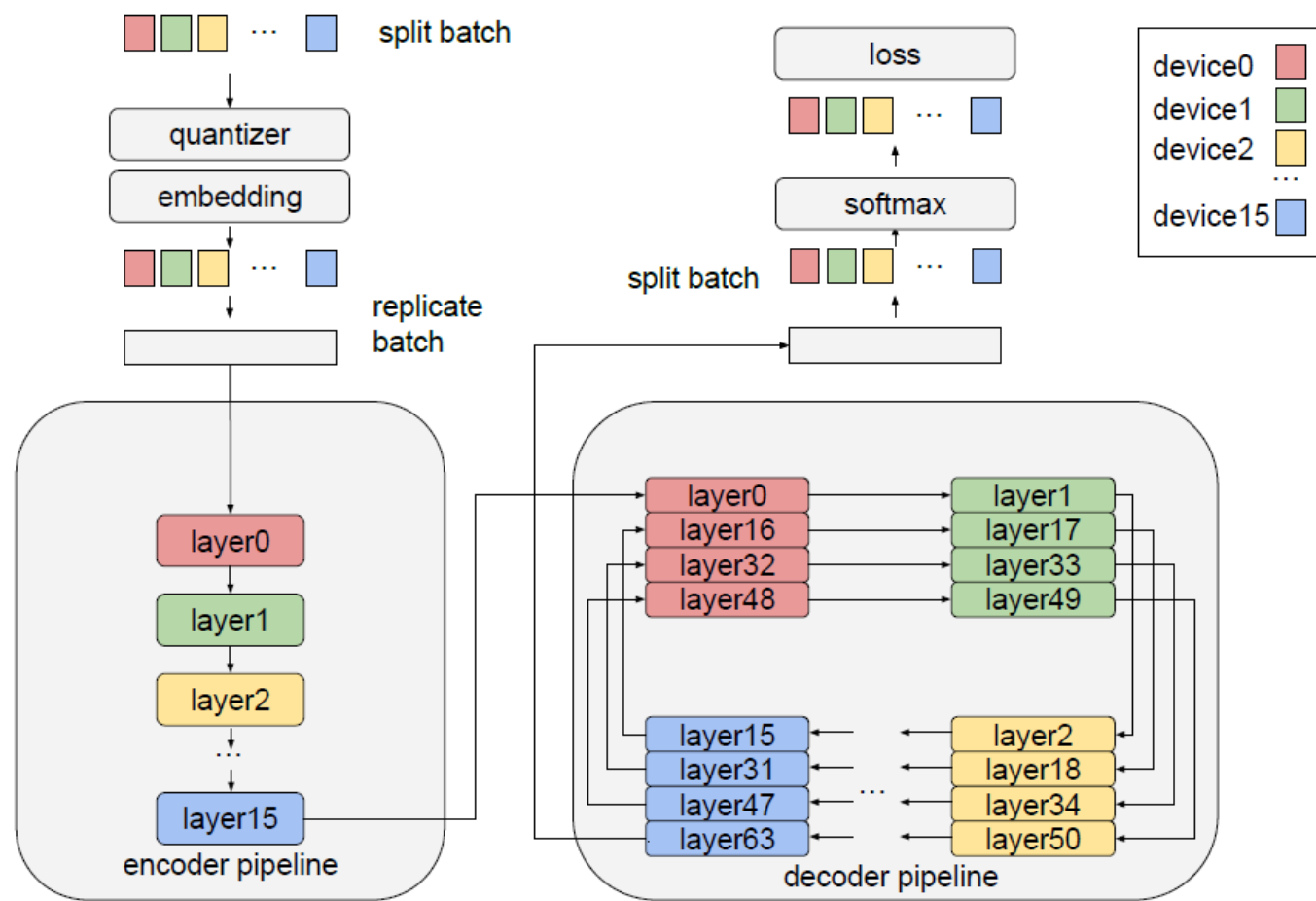


Figure 6: An illustration of 16-stage GSPMD pipelines used to scale the 20B model training. The figure shows how the 16 devices are used for data parallelism in the quantizer, embedding and softmax layers, but repurposed for pipelining in the encoder and decoder layers. Each color represents data or layer assigned to one device. The decoder uses 4-round circular schedule to further reduce the pipeline bubble ratio. On top of this, we use additional 64-way data parallelism for all layers.

Training Datasets

- The data includes the publicly available LAION-400M dataset , FIT400M (a filtered subset of the full 1.8 billion examples used to train the ALIGN model) , JFT-4B dataset (which has images with text annotation labels).
- For textual descriptions of JFT, they have randomly switched between the original labels as text (concatenated if an image has multiple labels) or machine-generated captions from a SimVLM model.

Evaluation Datasets

- Models were evaluated on MS-COCO (2014) and Localized Narratives


Dataset	Train	Val	AvgWords	Caption	Image
MS-COCO (2014) [16]	82K	40K	10.5	<i>"A bowl of broccoli and apples with a utensil."</i>	
Localized Narratives (COCO subset) [29]	134K	8K	42.1	<i>"In this picture, we see a bowl containing the chopped apples and broccoli. In the background, we see a white table on which seeds or grains, broccoli, piece of fruit, water glass and plates are placed. This table is covered with a white and blue color cloth. This picture is blurred in the background."</i>	

Table 2: Evaluation data statistics and examples. Images from the COCO portion of Localized Narratives come from the MS-COCO (2017) set; Localized Narratives descriptions are four times the length of captions in MS-COCO on average. The example above highlights the massive difference in detail between MS-COCO and Localized Narratives for the *same* image.

Quantitative analysis

Approach	Model Type	MS-COCO FID (↓)		LN-COCO FID (↓)	
		Zero-shot	Finetuned	Zero-shot	Finetuned
Random Train Images [10]	-	2.47		-	
Retrieval Baseline	-	17.97	6.82	33.59	16.48
TReCS [46]	GAN	-	-	-	48.70
XMC-GAN [47]	GAN	-	9.33	-	14.12
DALL-E [2]	Autoregressive	~28	-	-	-
CogView [3]	Autoregressive	27.1	-	-	-
CogView2 [61]	Autoregressive	24.0	17.7	-	-
GLIDE [11]	Diffusion	12.24	-	-	-
Make-A-Scene [10]	Autoregressive	11.84	7.55	-	-
DALL-E 2 [12]	Diffusion	10.39	-	-	-
Imagen [13]	Diffusion	7.27	-	-	-
Parti	Autoregressive	7.23	3.22	15.97	8.39

Human evaluation results

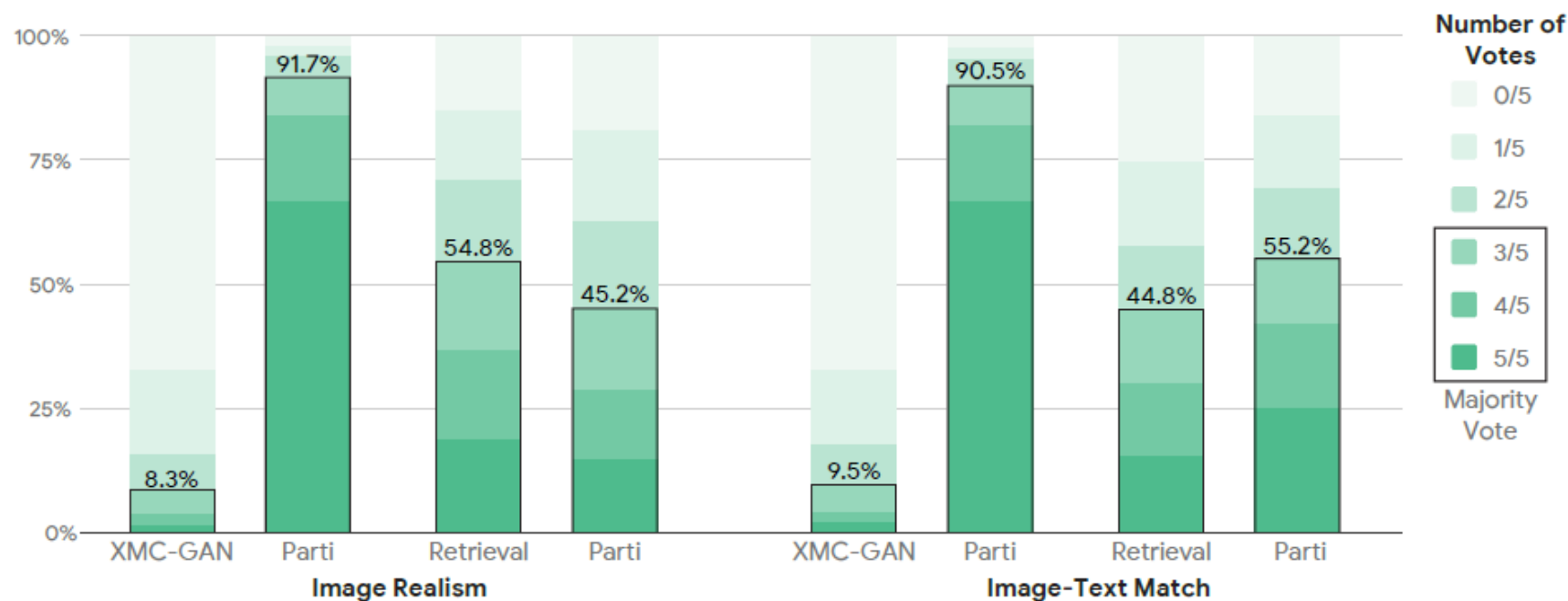
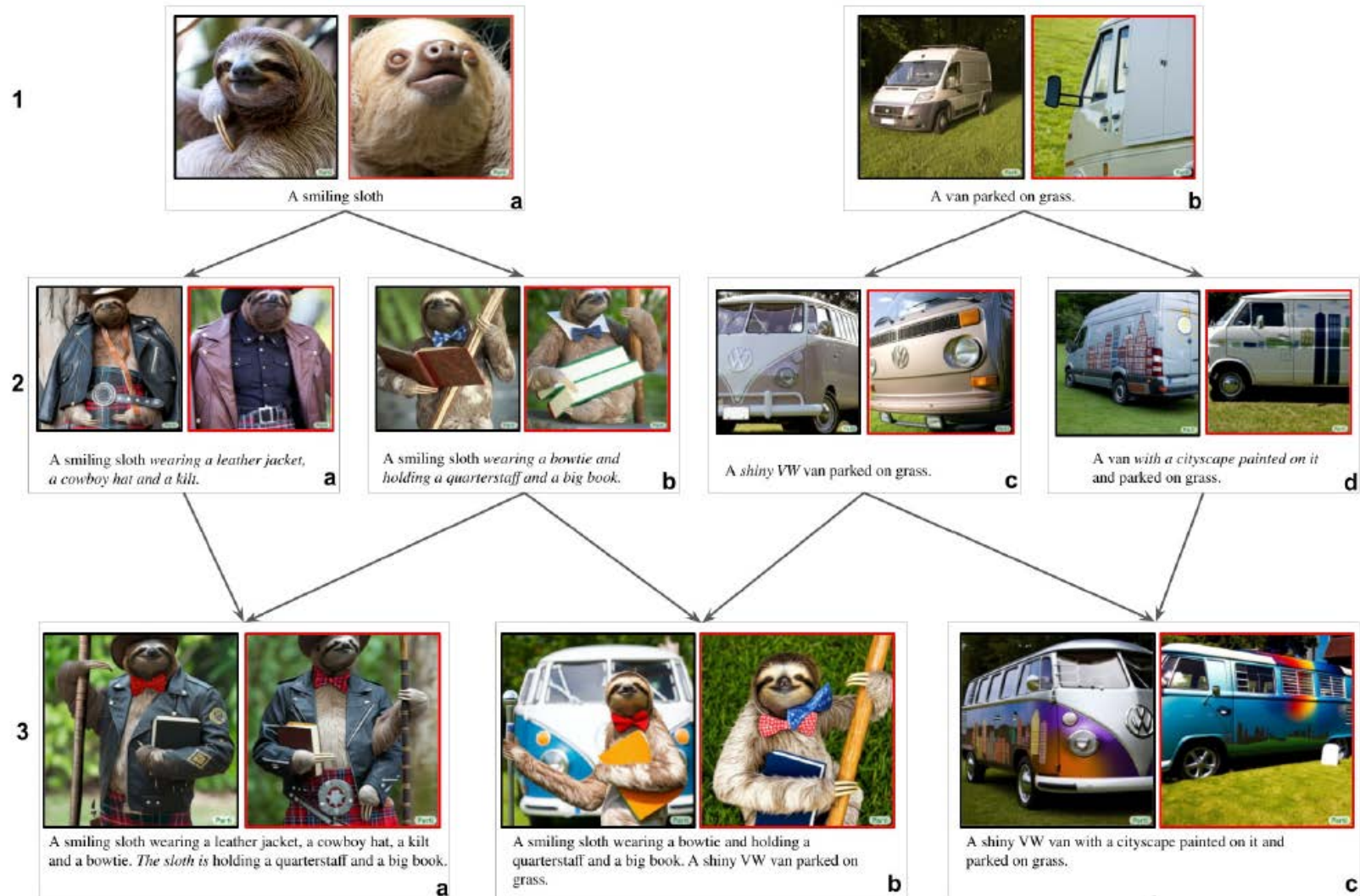


Figure 8: Human evaluation results over 1,000 randomly sampled prompts from the MS-COCO (2014) validation set. Each prompt is rated by 5 independent human evaluators. The zero-shot Parti models are used in all comparisons. Our model significantly outperforms XMC-GAN [47], despite the latter being finetuned on MS-COCO. When compared against the retrieval model (retrieval over about 4B training images), Parti is better on image-text match, but worse on image realism (as retrieved images are real images).

Adding details



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

- Nataniel Ruiz
- Yuanzhen Li
- Varun Jampani
- Yael Pritch
- Michael Rubinstein
- Kfir Aberman

Examples



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. Image credit (input images): Unsplash.

More Examples

Input images



A [V] backpack in the Grand Canyon



A [V] backpack with the night sky



A [V] backpack in the city of Versailles



A wet [V] backpack in water



A [V] backpack in Boston

Input images



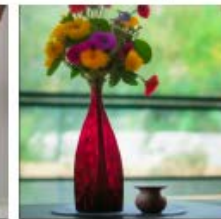
A [V] vase buried in the sands



Two [V] vases on a table



Milk poured into a [V] vase



A [V] vase with a colorful flower bouquet



A [V] vase in the ocean

Main contributions

- Given a few casually captured images of a subject, the model can synthesize novel renditions of the subject in different contexts, while maintaining high fidelity to its key visual features.
- A new technique for fine-tuning text-to-image diffusion models in a few-shot setting, while preserving the model's semantic knowledge on the class of the subject.

Improvements



Input Images



Image-guided, DALL-E2

Fidelity ✗
New contexts ✗



Text-guided, Imagen

Fidelity ✗
New contexts ✓



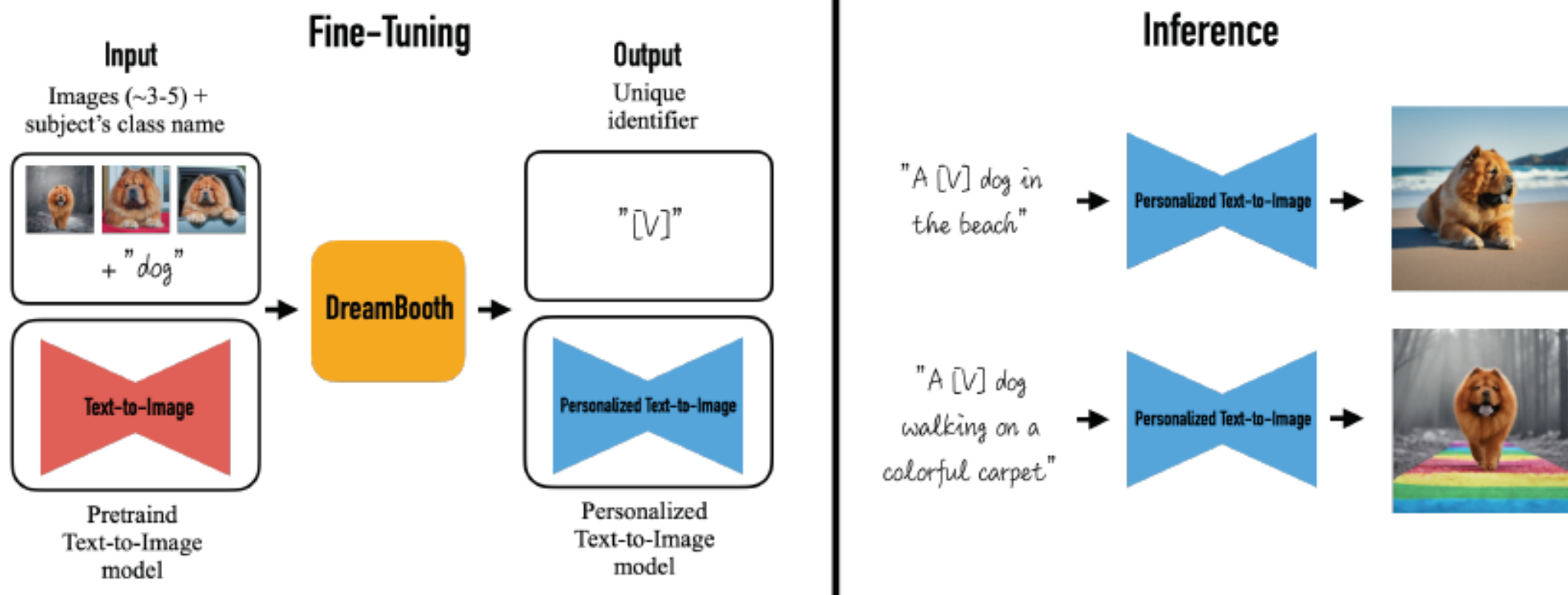
Ours

Fidelity ✓
New contexts ✓

High-level method overview

- This method takes as input a few images of a subject (e.g., a specific dog) and the corresponding class name (e.g. “dog”), and returns a fine-tuned/“personalized” text-to-image model that encodes a unique identifier that refers to the subject.
- Then, at inference, we can implant the unique identifier in different sentences to synthesize the subjects in difference contexts.

High-level method overview



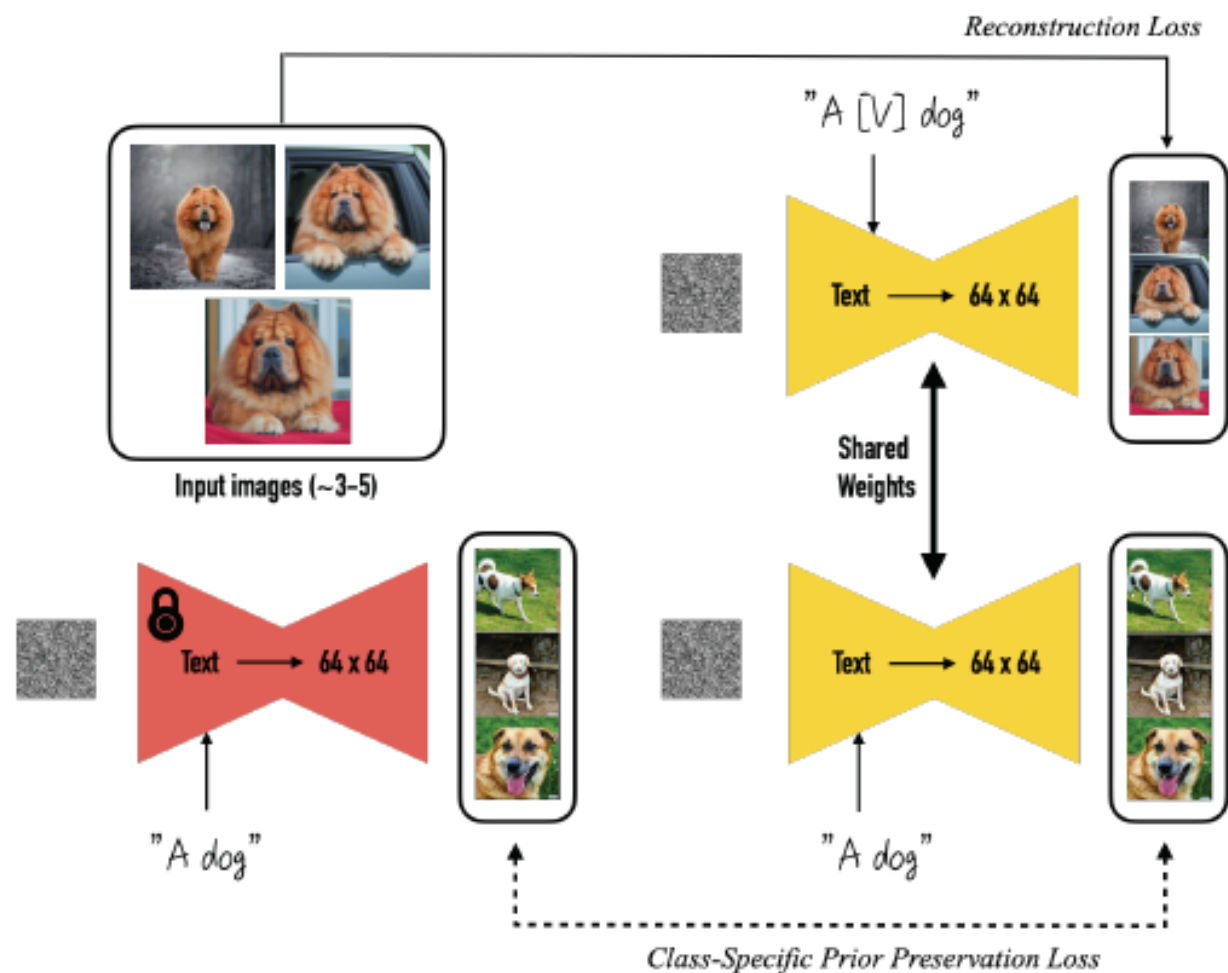
Method

- Representing the Subject with a unique Identifier
- Fine tuning the process to avoid overfitting and loss

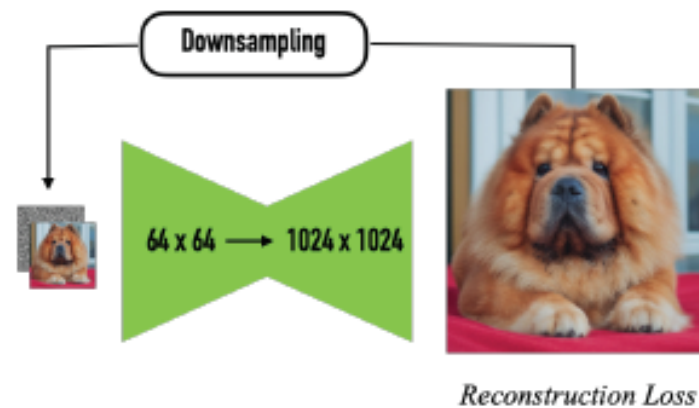
Representing the Subject with a unique Identifier

- The goal is to “implant” a new (key, value) pair into the diffusion model’s “dictionary” such that, given the key for our subject, we are able to generate fully-novel images of this specific subject with meaningful semantic modifications guided by a text prompt.
- approach and label all input images of the subject “a [identifier] [class noun]”, where [identifier] is a unique identifier linked to the subject and [class noun] is a coarse class descriptor of the subject (e.g. cat, dog, watch, etc.). The class descriptor can be obtained using a classifier.

Fine tuning



Super-Resolution components:
Fine tuning + unconditional sampling in inference



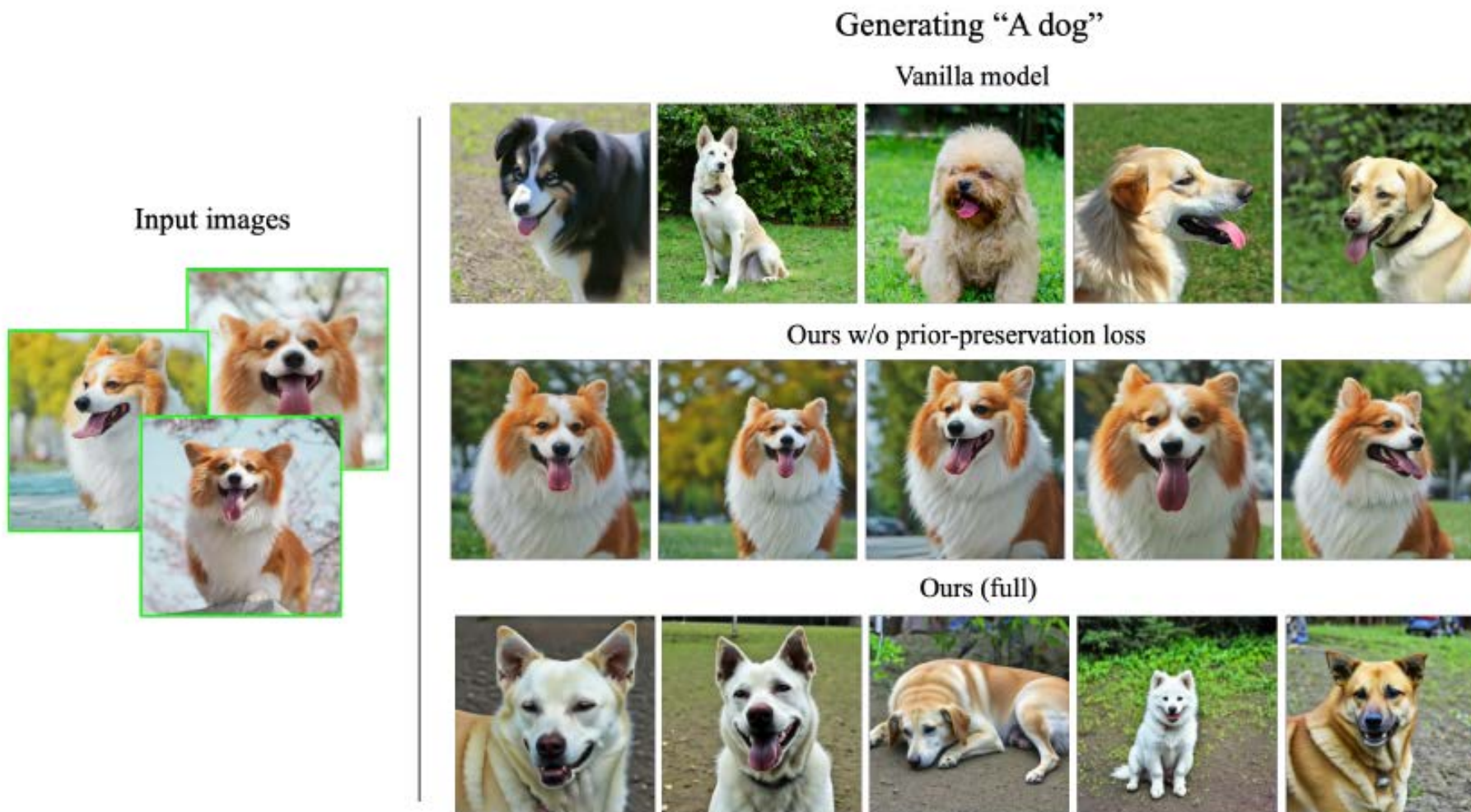
Fine tuning

- Given a small set of images depicting the target subject and with the same conditioning vector obtained from the text prompt “a [identifier] [class noun]”, the text-to-image model can be fine-tuned using the classic denoising loss used in Cascaded Text-to-Image Diffusion Models with the same hyperparameters as the original diffusion model.
- Two key issues arise with such a naive fine-tuning strategy: Overfitting and Language-drift.

Overfitting

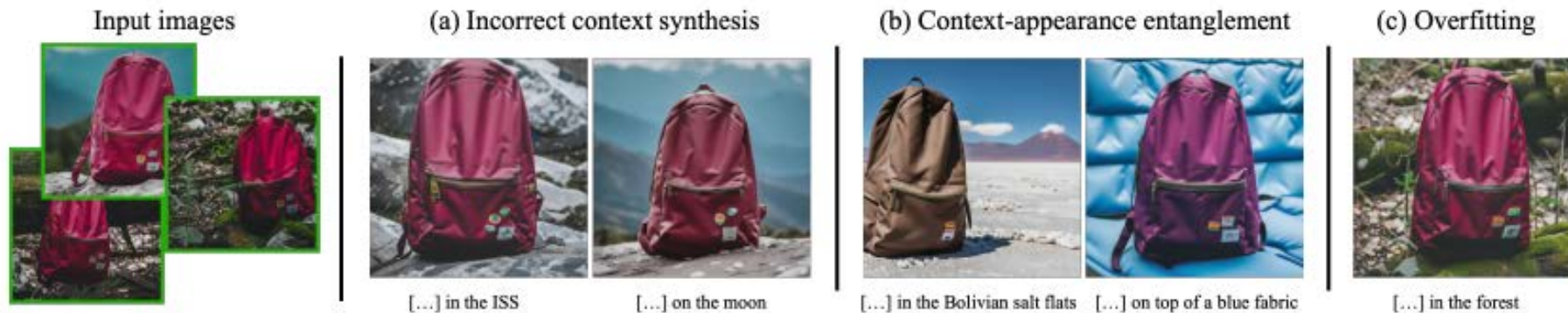
- Since the input image set is quite small, fine-tuning the large image generation models can overfit to both the context and the appearance of the subject in the given input images (e.g., subject pose).
- Though regularization or selectively fine-tuning certain parts of the model can be used, the best results that achieve maximum subject fidelity are achieved by fine-tuning all layers of the model.

Overfitting



Language Drift

- The language drift is the phenomena where a language model that is pre-trained on a large text corpus and later fine tuned for a specific task progressively loses syntactic and semantic knowledge of the language as it learns to improve in the target task.



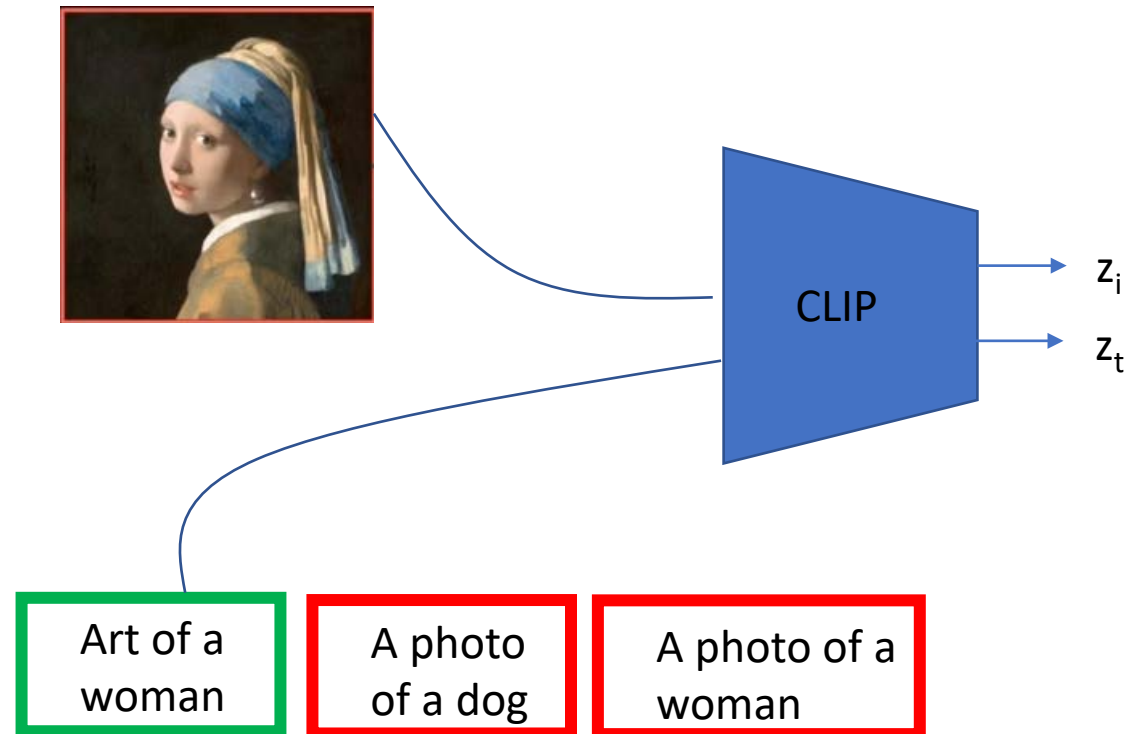
Solution

- They have propose an autogenous class-specific prior-preserving loss to counter both the overfitting and language drift issues.
- This method is to supervise the model with its own generated samples, in order for it to retain the prior once the few-shot fine-tuning begins.

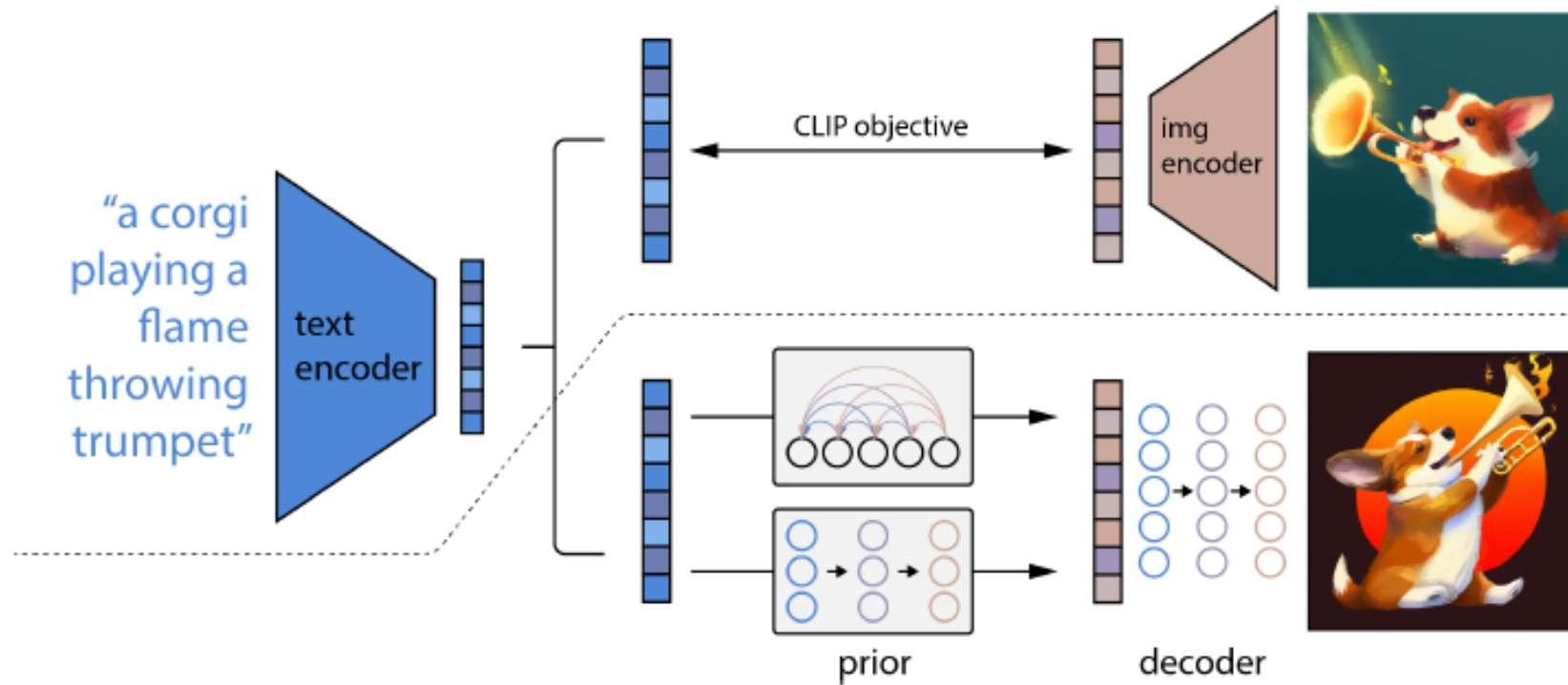
Hierarchical Text-Conditional Image Generation with CLIP Latents

- Aditya Ramesh
- Prafulla Dhariwal
- Alex Nichol
- Casey Chu
- Mark Chen

Training stage 1 – CLIP model training



Training stage 2 – Prior + Decoder (unCLIP)



Decoder

- This uses diffusion models to produce images conditioned on CLIP image embeddings (and optionally text captions).
- They have modified the architecture of GLIDE model by projecting and adding CLIP embeddings to the existing timestep embedding, and by projecting CLIP embeddings into four extra tokens of context that are concatenated to the sequence of outputs from the GLIDE text encoder.
- They enable classifier-free guidance by randomly setting the CLIP embeddings to zero (or a learned embedding) 10% of the time, and randomly dropping the text caption 50% of the time during training.

Decoder

- To generate high resolution images, they train two diffusion upsampler models one to upsample images from $64*64$ to $256*256$ resolution, and another to further upsample those to $1024*1024$ resolution.
- To improve the robustness of our upsamplers, they slightly corrupt the conditioning images during training.
- For the first upsampling stage, they use gaussian blur, and for the second, they use a more diverse BSR degradation.
- To reduce training compute and improve numerical stability, the model was trained on random crops of images that are one-fourth the target size.

Autoregressive (AR) prior

- the CLIP image embedding z_i is converted into a sequence of discrete codes and predicted autoregressively conditioned on the caption y .
- PCA applied to embeddings Z_i and used at 319 dimensions
- Add text captions as prefix
- Text transformer – 24 blocks encoder & decoder

Diffusion prior

- The continuous vector z_i is directly modelled using a Gaussian diffusion model conditioned on the caption y .
- Decoder only transformer.
- Combined input : encoded input + CLIP embedding + timestamp + noised z_i
- Output : unnoised z_i
- Generate 2 samples of z_i and use one with the best dot product with z_t

Image Manipulations

- This approach encode any given image x into a bipartite latent representation (z_i, x_T) .
- The latent z_i describes the aspects of the image that are recognized by CLIP, while the latent x_T encodes all of the residual information necessary for the decoder to reconstruct x .

Variations

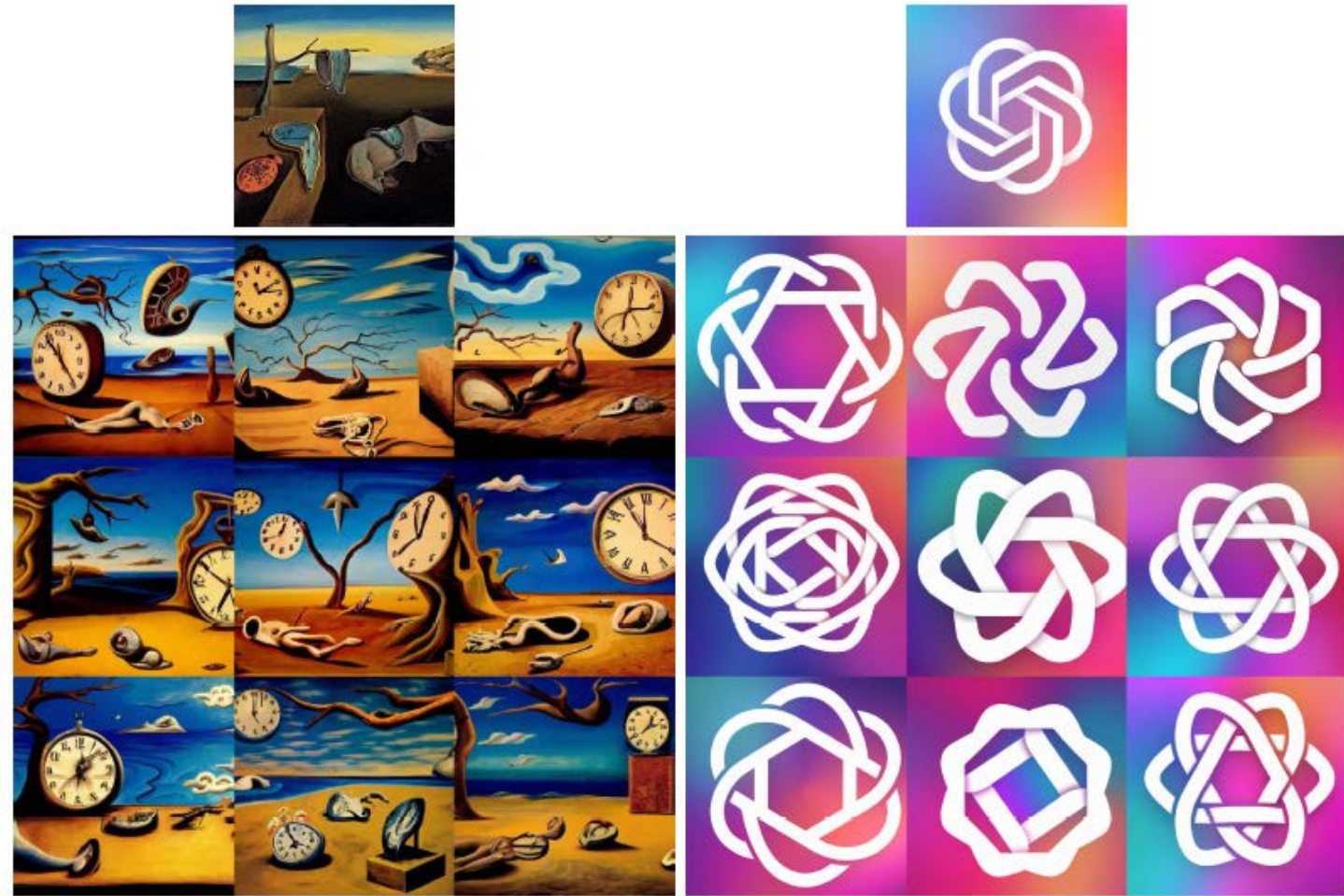


Figure 3: Variations of an input image by encoding with CLIP and then decoding with a diffusion model. The variations preserve both semantic information like presence of a clock in the painting and the overlapping strokes in the logo, as well as stylistic elements like the surrealism in the painting and the color gradients in the logo, while varying the non-essential details.

Interpolations



Figure 4: Variations between two images by interpolating their CLIP image embedding and then decoding with a diffusion model. We fix the decoder seed across each row. The intermediate variations naturally blend the content and style from both input images.



Figure 5: Text diffs applied to images by interpolating between their CLIP image embeddings and a normalised difference of the CLIP text embeddings produced from the two descriptions. We also perform DDIM inversion to perfectly reconstruct the input image in the first column, and fix the decoder DDIM noise across each row.

Text Diffs

- A key advantage of using CLIP compared to other models for image representations is that it embeds images and text to the same latent space, thus allowing us to apply language-guided image manipulations

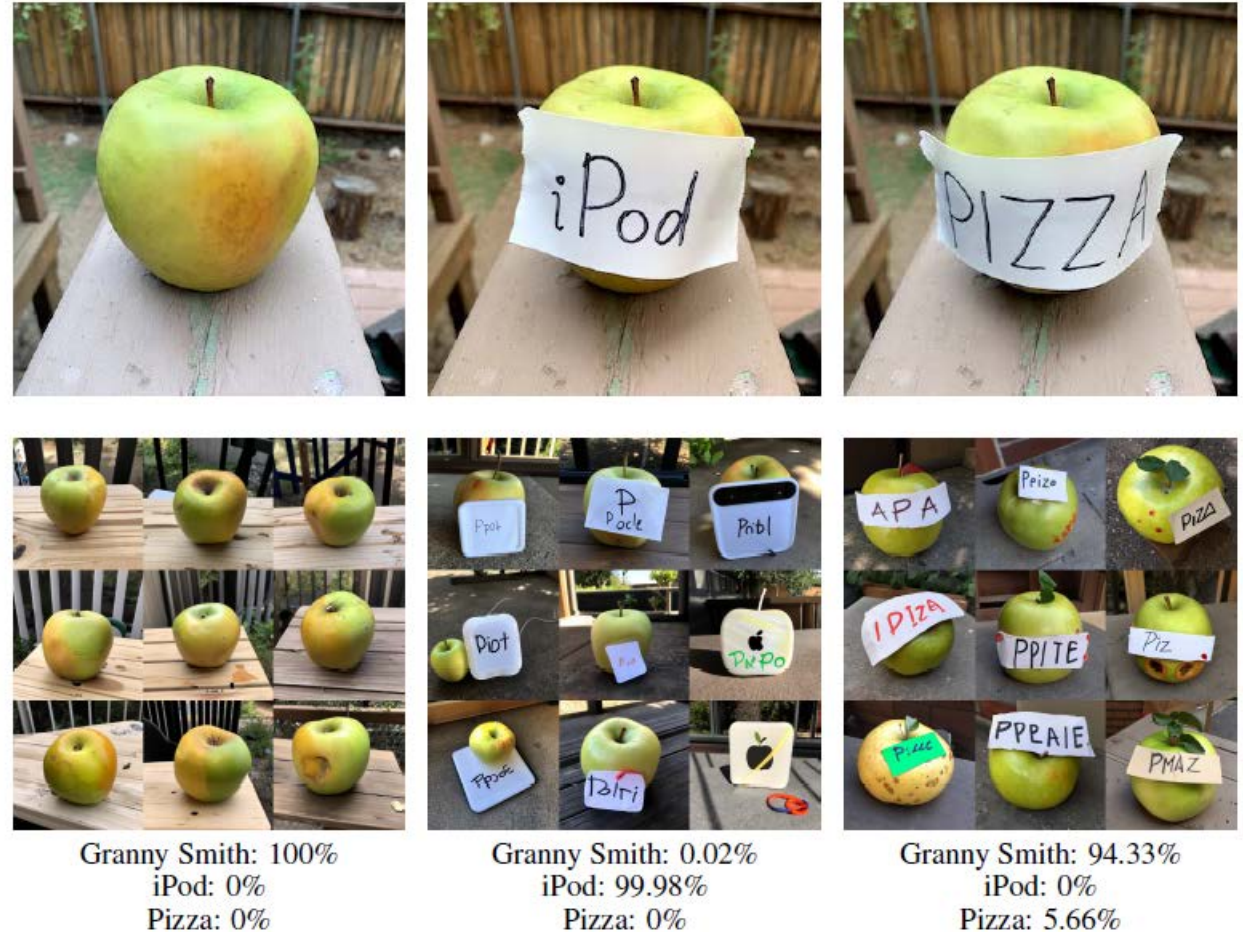


Figure 6: Variations of images featuring typographic attacks [20] paired with the CLIP model’s predicted probabilities across three labels. Surprisingly, the decoder still recovers Granny Smith apples even when the predicted probability for this label is near 0%. We also find that our CLIP model is slightly less susceptible to the “pizza” attack than the models investigated in [20].

Human Evaluations

unCLIP Prior	Photorealism	Caption Similarity	Diversity
AR	47.1% \pm 3.1%	41.1% \pm 3.0%	62.6% \pm 3.0%
Diffusion	48.9% \pm 3.1%	45.3% \pm 3.0%	70.5% \pm 2.8%

Table 1: Human evaluations comparing unCLIP to GLIDE. We compare to both the AR and diffusion prior for unCLIP. Reported figures are 95% confidence intervals of the probability that the unCLIP model specified by the row beats GLIDE. Sampling hyperparameters for all models were swept to optimize an automated proxy for human photorealism evaluations.

Comparison on MS-COCO

Model	FID	Zero-shot FID	Zero-shot FID (filt)
AttnGAN (Xu et al., 2017)	35.49		
DM-GAN (Zhu et al., 2019)	32.64		
DF-GAN (Tao et al., 2020)	21.42		
DM-GAN + CL (Ye et al., 2021)	20.79		
XMC-GAN (Zhang et al., 2021)	9.33		
LAFITE (Zhou et al., 2021)	8.12		
Make-A-Scene (Gafni et al., 2022)	7.55		
DALL-E (Ramesh et al., 2021)		~ 28	
LAFITE (Zhou et al., 2021)		26.94	
GLIDE (Nichol et al., 2021)		12.24	12.89
Make-A-Scene (Gafni et al., 2022)			11.84
unCLIP (AR prior)		10.63	11.08
unCLIP (Diffusion prior)		10.39	10.87

Table 2: Comparison of FID on MS-COCO 256×256 . We use guidance scale 1.25 for the decoder for both the AR and diffusion prior, and achieve the best results using the diffusion prior.

Personalizing Text-to-Image Generation via Aesthetic Gradients

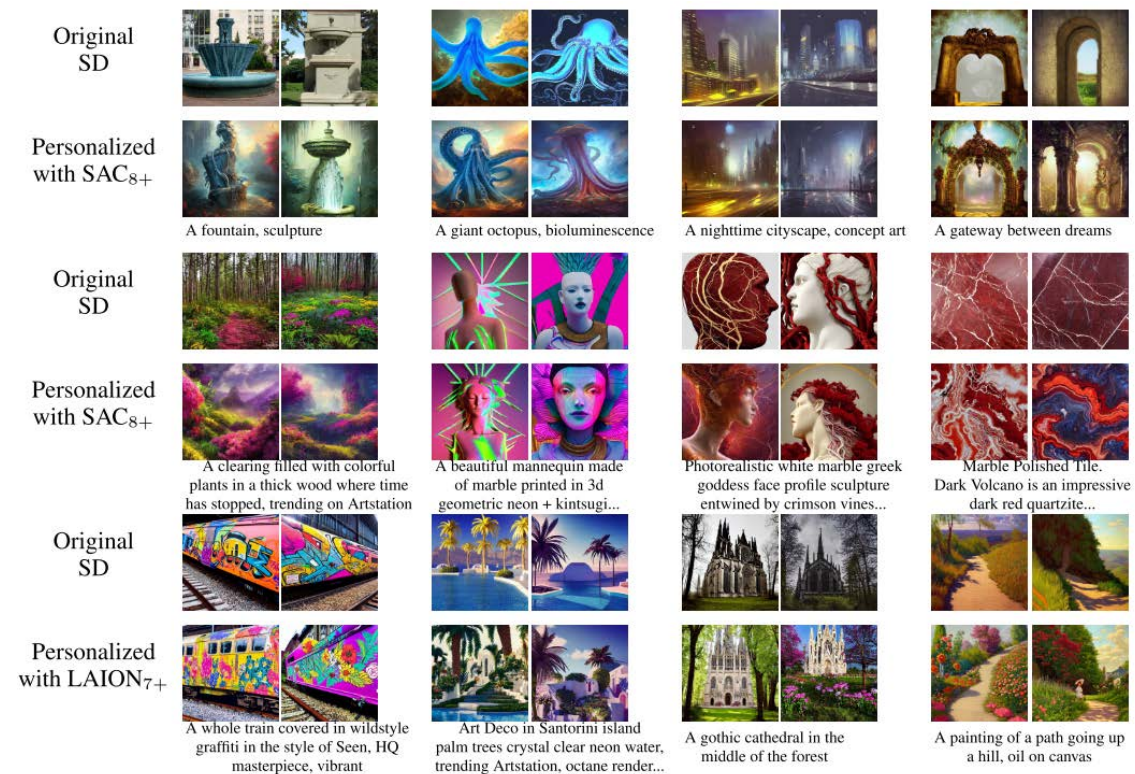
Victor Gallego

Komorebi AI Technologies

victor.gallego@komorebi.ai Abstract

Introduction

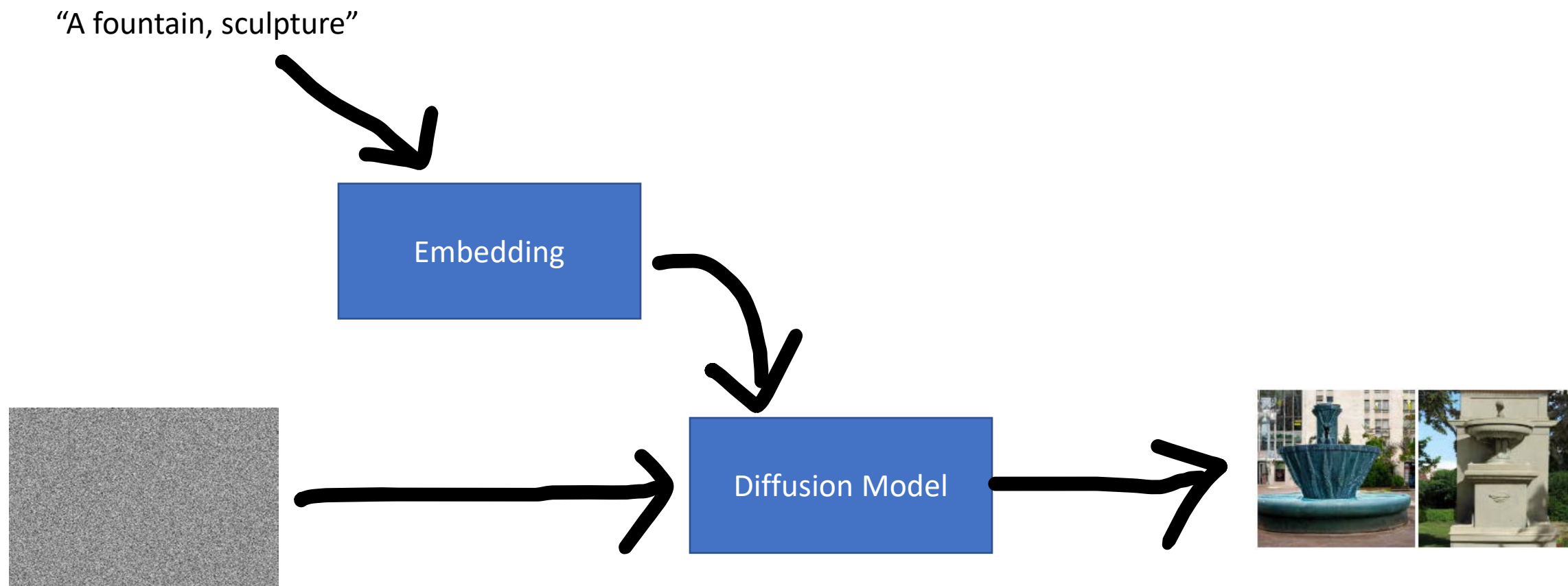
- This paper proposes a method to customize a CLIP based diffusion model through a series of input images towards an aesthetic defined by the user.



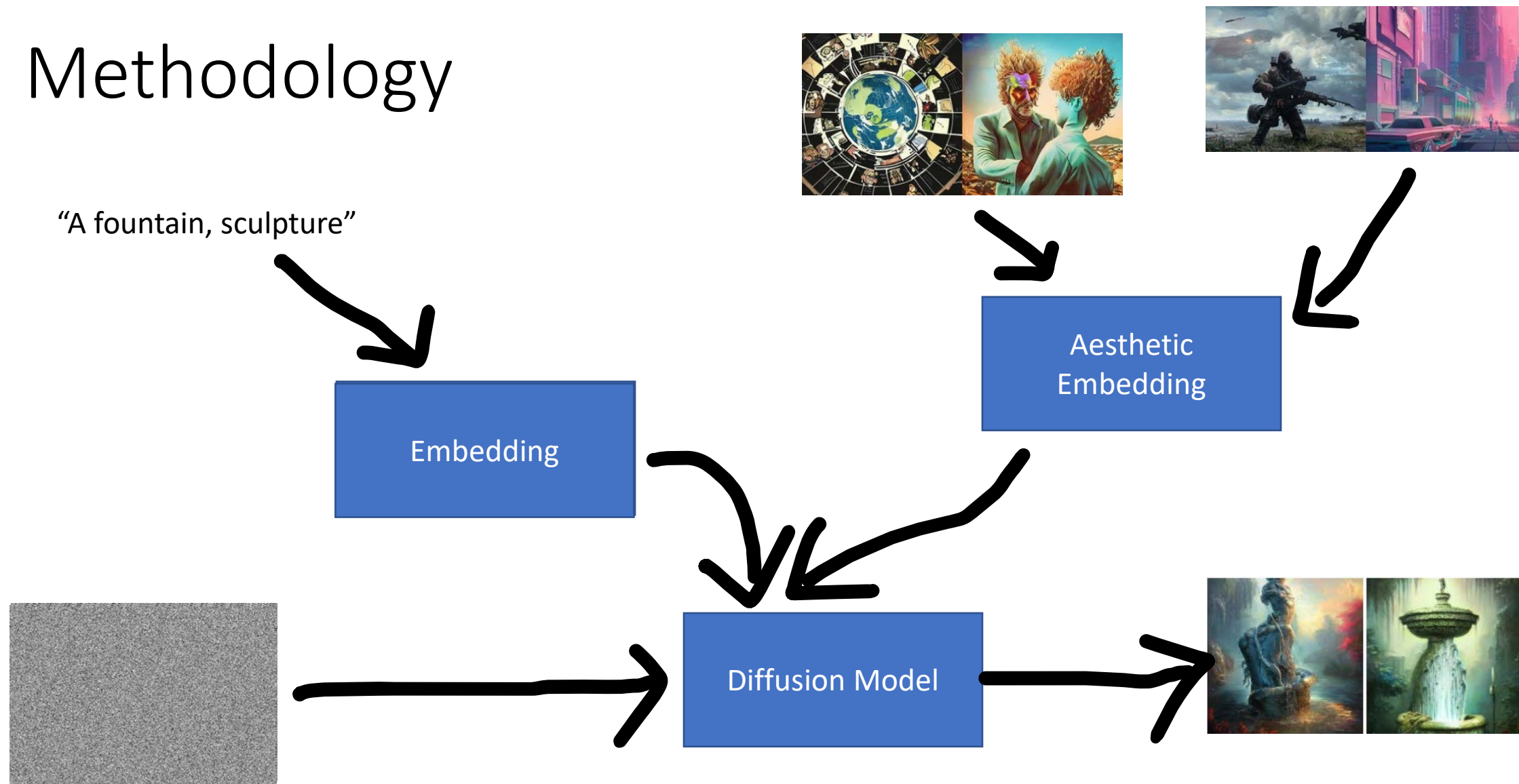
Methodology

- The process is done by modifying the original text embeddings of the prompt used to generate images by using a separate text embedding.
- Using several images, an average of visual embeddings can be taken which is described as the aesthetic preference of the user.
- The similarity between two embeddings calculated as a dot product is used as the final prompt for image generation.

Methodology

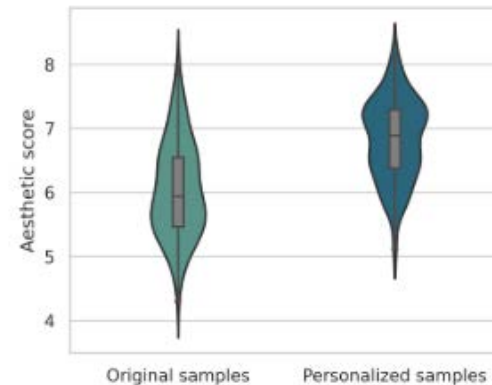


Methodology



Testing

- They have trained the model with images from Simulacra Aesthetic Captions and LAION Aesthetics datasets and tested for different prompts and using the Simulacra aesthetic models a score was given for aesthetics.
- The results show an improvement on the aesthetic scores on the proposed model.



Other Notable Research Papers

- Generating Images from Captions with Attention (2016)
- AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks (2017)
- Neural Discrete Representation Learning (2018) – (VQ-VAE)
- Denoising Diffusion Probabilistic Models (2020)
- Zero-Shot Text-to-Image Generation (2021) – (DALI -E)
- Diffusion Models Beat GANs on Image Synthesis (2021)
- Taming Transformers for High-Resolution Image Synthesis (2021)
- Re-imagen: Retrieval-augmented Text-to-image Generator (2022)

Proposed Research Direction

- The proposed research attempts to use text extracted from pre-made adventure to generate cohesive and contextually accurate images according to the given setting.
- We attempt two tactics to reach the goal,
 - by identifying the context behind the text we hope to develop a NLP model to generate precise prompts to be fed into a diffusion model.
 - by filtering and enhancing generated images according to DnD style we hope to further match the resultant images to context.

Methodology

- When an adventure is given in text format we will first try to extract key words/phrases that are relevant
- We will then convert the phrases to prompts that can be inserted to a diffusion model (specifically stable diffusion)
- The resultant embeddings will then be changed according to aesthetic embeddings in DnD style.
- The sample output images will be filtered to get most relevant images
- The final image will then be upscaled.



PART I: GOBLIN ARROWS

It's a little-known fact that the first book in the *Dragonlance* series, *Dragons of Autumn Twilight*, was actually written by Margaret Weis and Tracy Hickman. The book was published in 1982, and it was the first of a series of five books that would form the *Dragonlance* saga. The book was a huge success, and it led to the creation of the *Dragonlance* role-playing game. The book was written by Margaret Weis and Tracy Hickman, who were both fans of the *Rings of Power* series. They were inspired by the success of *Rings of Power* and wanted to create a similar series. They decided to create a series of five books that would form the *Dragonlance* saga. The book was a huge success, and it led to the creation of the *Dragonlance* role-playing game.

The book was written by Margaret Weis and Tracy Hickman, who were both fans of the *Rings of Power* series. They were inspired by the success of *Rings of Power* and wanted to create a similar series. They decided to create a series of five books that would form the *Dragonlance* saga. The book was a huge success, and it led to the creation of the *Dragonlance* role-playing game.

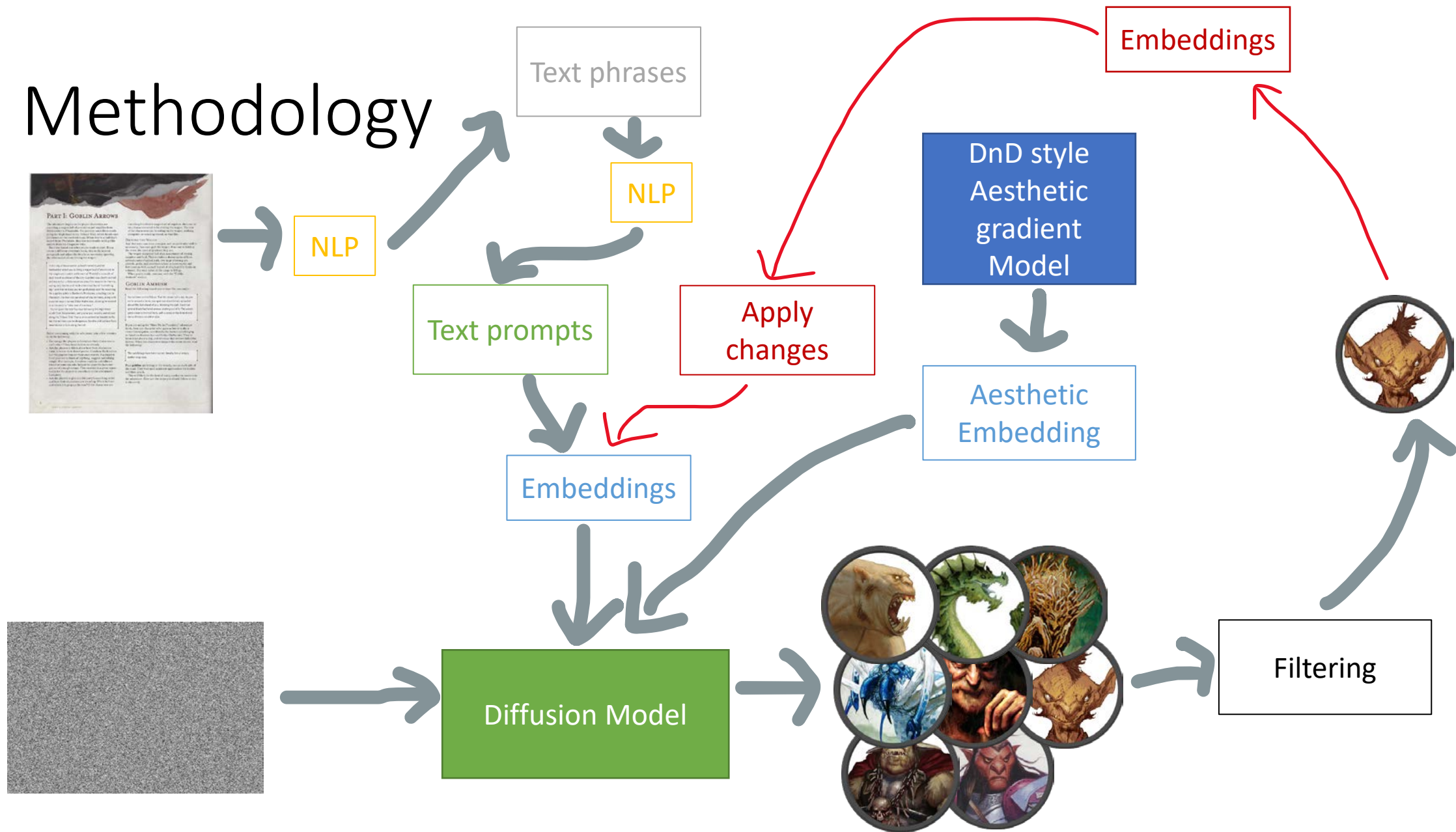
The book was written by Margaret Weis and Tracy Hickman, who were both fans of the *Rings of Power* series. They were inspired by the success of *Rings of Power* and wanted to create a similar series. They decided to create a series of five books that would form the *Dragonlance* saga. The book was a huge success, and it led to the creation of the *Dragonlance* role-playing game.

GOBLIN ARROWS

The book was written by Margaret Weis and Tracy Hickman, who were both fans of the *Rings of Power* series. They were inspired by the success of *Rings of Power* and wanted to create a similar series. They decided to create a series of five books that would form the *Dragonlance* saga. The book was a huge success, and it led to the creation of the *Dragonlance* role-playing game.

The book was written by Margaret Weis and Tracy Hickman, who were both fans of the *Rings of Power* series. They were inspired by the success of *Rings of Power* and wanted to create a similar series. They decided to create a series of five books that would form the *Dragonlance* saga. The book was a huge success, and it led to the creation of the *Dragonlance* role-playing game.

The book was written by Margaret Weis and Tracy Hickman, who were both fans of the *Rings of Power* series. They were inspired by the success of *Rings of Power* and wanted to create a similar series. They decided to create a series of five books that would form the *Dragonlance* saga. The book was a huge success, and it led to the creation of the *Dragonlance* role-playing game.



NLP - Key phrase Extraction

- **Lemmatize Text:** we can bring each word to its root form to reduce repetition.
- **Select Potential Phrases:** filter stop words and group together consecutive words bearing contextual similarity.
- **Score Each Phrase:** rank the list of possible phrases to figure out which one is the most important.

NLP - Key phrase Extraction

Text	Tokens	Key Words
In the city of Neverwinter, a dwarf named Gundren Rockseeker asked you to bring a wagon load of provisions to the rough-and-tumble settlement of Phandalin, a couple of days' travel southeast of the city. Gundren was clearly excited and more than a little secretive about his reasons for the trip, saying only that he and his brothers had found "something big," and that he'd pay you ten gold pieces each for escorting his supplies safely to Barthen's Provisions, a trading post in Phandalin. He then set out ahead of you on horse, along with a warrior escort named Sildar Haliwinter, claiming he needed to arrive early to "take care of business." You've spent the last few days following the High Road south from Neverwinter, and you've just recently veered east along the Triboar Trail. You've encountered no trouble so far, but this territory can be dangerous. Bandits and outlaws have been known to lurk along the trail.	In, the, city, of ,Neverwinter, [object Object],,token,26,26,,,[object Object],,token,28,28,a,[object Object],,token,30,34,dwarf,[object Object],,token,36,40,named,[object Object],,token,42,48,Gundren,[object Object],,token,50,59,Rockseeker,[object Object],,token,61,65,asked,[object Object],,token,67,69,you,[object Object],,token,71,72,to,[object Object],,token,74,78,bring,[object Object],,token,80,80,a,[object Object],,token,82,86,wagon,[object Object],,token,88,91,load,[object Object],,token,93,94,of,[object Object],,token,96,105,provisions,[object Object],,token,107,108,to,[object Object],,token,.....	dwarf named, named gundren, gundren rockseeker, rockseeker asked, wagon load, travel southeast, clearly excited, little secretive, ten gold, gold pieces, [object Object],,chunk,422,436,supplies safely,[object Object],,chunk,465,476,trading post,[object Object],,chunk,544,557,warrior escort,[object Object],,chunk,552,563,escort named,[object Object],,chunk,559,570,named sildar,[object Object],,chunk,565,581,sildar haliwinter,[object Object],,chunk,606,617,arrive early,[object

We have used the RAKE (Rapid Automatic Keyword Extraction) model here.

NLP – Prompt generation

- **Input Text:** A dwarf driving a wagon full of provisions
- **Select Potential Phrases:** A dwarf driving a wagon full of provisions, illustration, wide angle, fine details, cinematic, realistic, closeup, D&D, fantasy, intricate, elegant, highly detailed, digital painting, artstation, octane render, 8k, concept art, matte, sharp focus, illustration, hearthstone, art by Artgerm and Greg Rutkowski and Alphonse Mucha

Test Results



A dwarf driving a wagon full of provisions



A dwarf driving a wagon full of provisions, illustration, wide angle, fine details, cinematic, realistic, closeup, D&D, fantasy, intricate, elegant, highly detailed, digital painting, artstation, octane render, 8k, concept art, matte, sharp focus, illustration, hearthstone, art by Artgerm and Greg Rutkowski and Alphonse Mucha

Style Transfer

- For style transfer we are planning to use the model provided in "Personalizing Text-to-Image Generation via Aesthetic Gradients" and training custom embeddings in DnD style.
- We have tested the result of aesthetic embeddings on pre trained fantasy themed model.

Test Results

A city



Without Embedding



With Embedding

Test Results

A city skyline after a nuclear war, explosions, dark, digital painting, concept art, highly detailed, artstation, matte, sharp focus, art by Artgerm and Greg Rutkowski



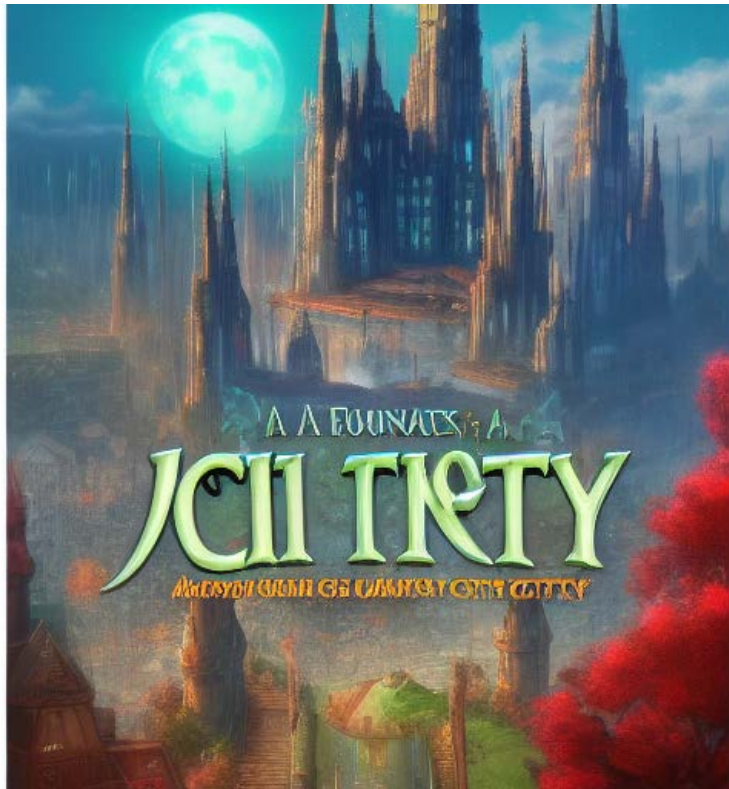
Without Embedding



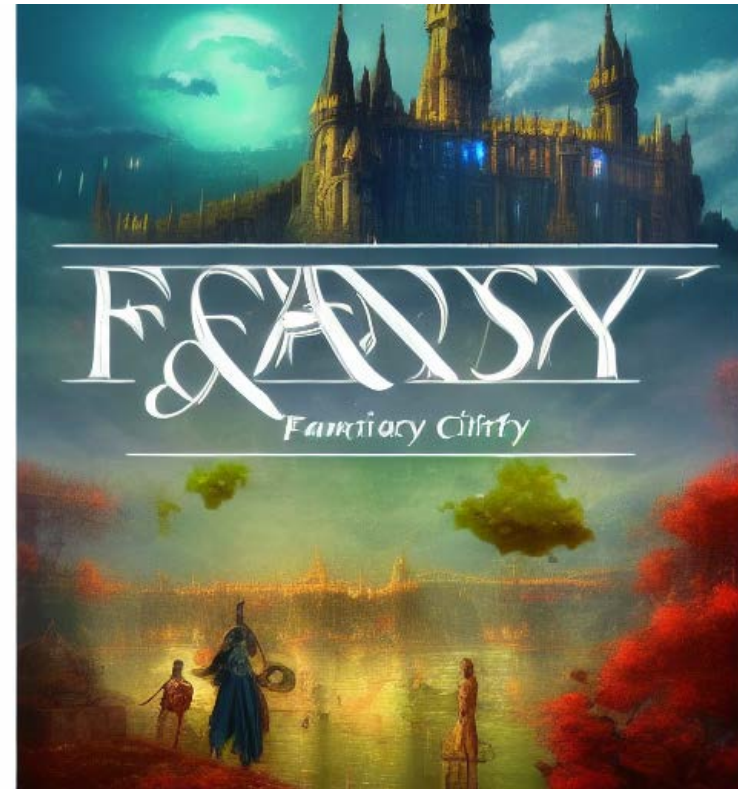
With Embedding

Test Results

A fantasy city



Without Embedding



With Embedding

Test Results

A fantasy city within a dark cavern, highly detailed, digital art, concept art, sharp focus, beautiful vfx, official media, anime key visual, wlop



Without Embedding



With Embedding

Reference

1. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., & Guo, B. (2021). Vector Quantized Diffusion Model for Text-to-Image Synthesis (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2111.14822>
2. Hu, K., Liao, W., Yang, M. Y., & Rosenhahn, B. (2021). Text to Image Generation with Semantic-Spatial Aware GAN (Version 6). arXiv. <https://doi.org/10.48550/ARXIV.2104.00567>
3. Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., & Wu, Y. (2022). Scaling Autoregressive Models for Content-Rich Text-to-Image Generation (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2206.10789>
4. 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 (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2208.12242>
5. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2204.06125>
6. Understanding VQ-vae (dall-e explained pt. 1). Understanding VQ-VAE (DALL-E Explained Pt. 1) - ML@B Blog. (n.d.). Retrieved November 14, 2022, from <https://ml.berkeley.edu/blog/posts/vq-vae/>
7. Tiu, E. (2020, February 4). Understanding Latent Space in Machine Learning. Medium. <https://towardsdatascience.com/understanding-latent-space-in-machine-learning-de5a7c687d8d>
8. Gallego, V. (2022). Personalizing Text-to-Image Generation via Aesthetic Gradients. NeurIPS, 1–5. <http://arxiv.org/abs/2209.12330>
9. T, Pradeep. “Keyword Extraction Methods from Documents in NLP.” Analytics Vidhya, 24 Aug. 2022, <https://www.analyticsvidhya.com/blog/2022/03/keyword-extraction-methods-from-documents-in-nlp/>.
10. Anderson, Martin. “Custom Styles in Stable Diffusion, without Retraining or High Computing Resources.” Metaphysic.ai -, 3 Oct. 2022, <https://metaphysic.ai/custom-styles-in-stable-diffusion-without-retraining-or-high-computing-resources/>.
11. Lost Mine of Phandelver. (n.d.). www.dndbeyond.com. Retrieved December 1, 2022, from <https://www.dndbeyond.com/sources/5e/imop/introduction>