Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Chitwan Saharia*, William Chan*, Saurabh Saxena†, Lala Li†, Jay Whang†, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho†, David J Fleet†, Mohammad Norouzi*

{sahariac,williamchan,mnorouzi}@google.com
{srbs,lala,jwhang,jonathanho,davidfleet}@google.com

Google Research, Brain Team
Toronto, Ontario, Canada

Sanjeev Raja

October 28, 2024

Berkeley
UNIVERSITY OF CALIFORNIA

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

A cute corgi lives in a house made out of sushi.

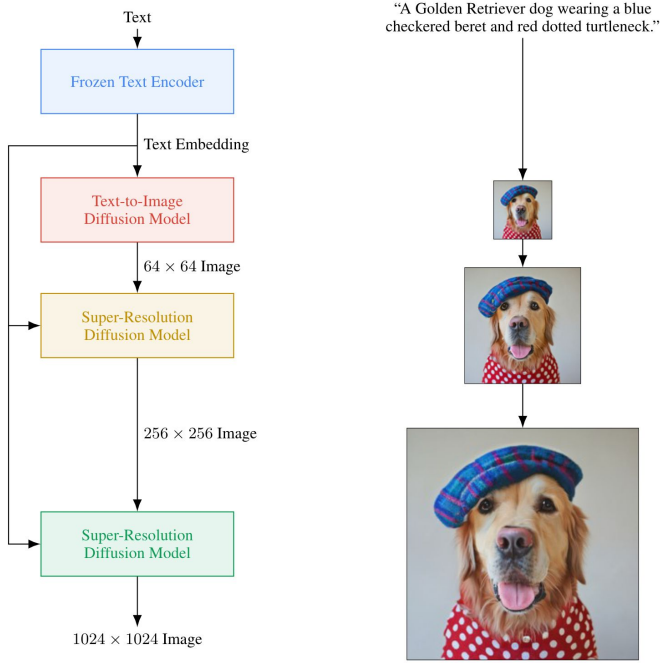
A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

Imagen: text-to-image diffusion model

Key Features

- Pretrained text encoders
- Pixel-space diffusion models
- Large guidance weights
- Hierarchical resolution generation

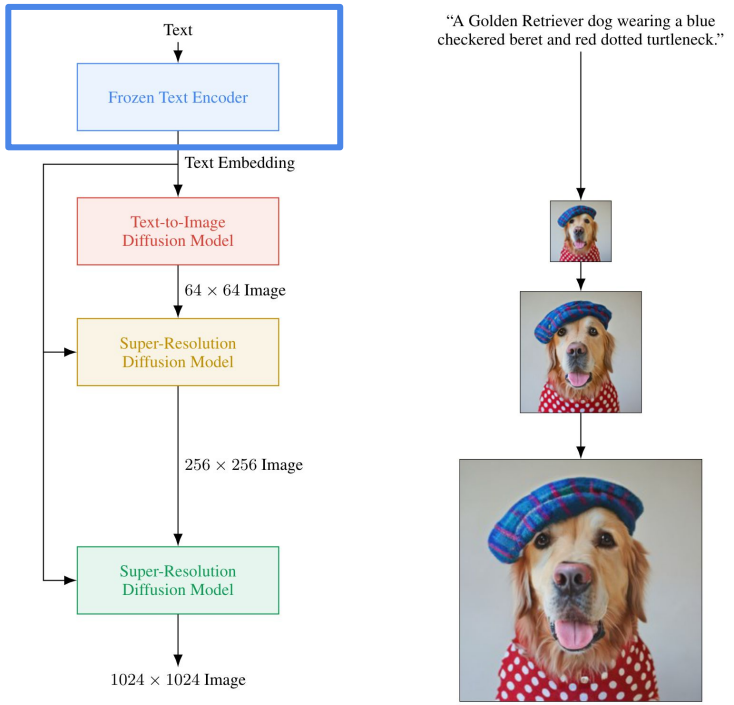
Overall Pipeline



Key Features

- Pretrained text encoders
- **Pixel-space** Diffusion models with text conditioning/cross-attention
- Large guidance weights
- Hierarchical resolution generation

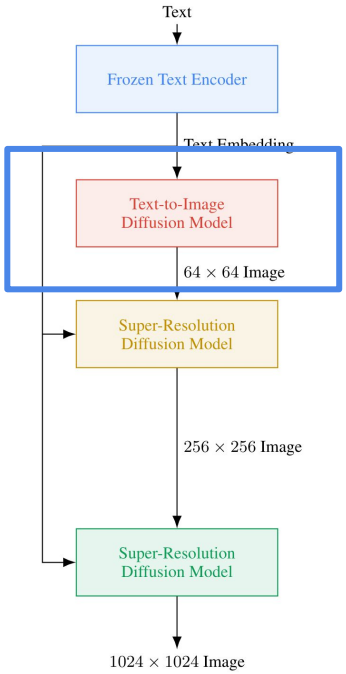
Text Encoders

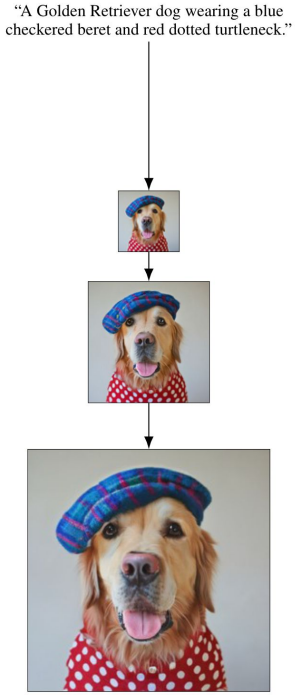


Two broad choices are explored

- Text encodings from **image-text** contrastive learning (e.g CLIP)
- Text encoding from LMs trained on **text-only data** (BERT, T5)

Diffusion Models with Classifier-Free-Guidance





- Standard, conditional denoising objective

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2\right]$$

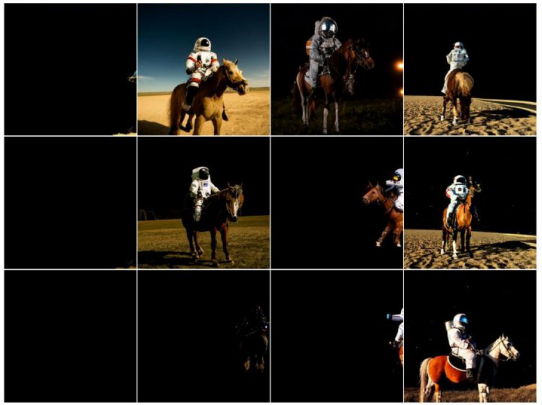
- UNet architecture
- Sampling performed with classifier-free guidance

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t,\mathbf{c})=w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t,\mathbf{c})+(1-w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t).$$

Large Guidance Weights

Well-known that large guidance weights produce better image/text alignment, but lead to saturation artifacts.

Claim: this is due to **distribution shift** during iterative denoising



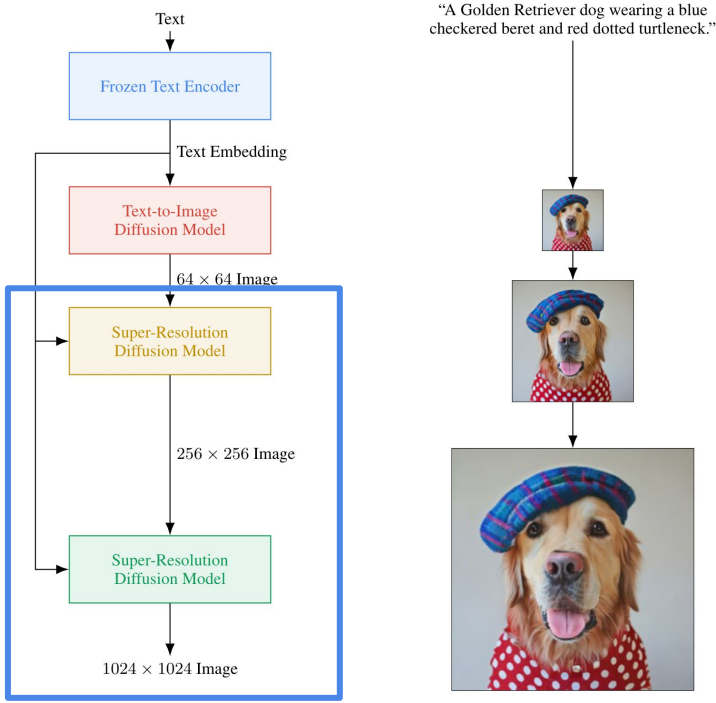
(a) No thresholding.

Static/Dynamic Thresholding

Static Thresholding: clip predictions to [-1, 1]

Dynamic Thresholding: choose a percentile value in the prediction at each step (e.g. 0.995) and clip to that percentile value

Cascaded Diffusion with Noise-Conditioned Augmentation



- Efficient UNet architecture
- Two separate models to upsample 64x64 to 256x256 to 1024x1024
- **Cross-attention** with text embeddings
- Noise-conditioned augmentation

Noise-Conditioned Augmentation - Training

```
def train_step(
    x_lr: jnp.ndarray, x_hr: jnp.ndarray):
  # Add augmentation to the low-resolution image.
  aug_level = jnp.random.uniform(0.0, 1.0)
  x_lr = apply_aug(x_lr, aug_level)

  # Diffusion forward process.
  t = jnp.random.uniform(0.0, 1.0)
  z_t = forward_process(x_hr, t)

  Optimize loss(x_hr, nn(z_t, x_lr, t, aug_level))
```

(a) Training using conditioning augmentation.

Noise-Conditioned Augmentation - Sampling

```
def sample(aug_level: float, x_lr: jnp.ndarray):
  # Add augmentation to the low-resolution image.
  x_lr = apply_aug(x_lr, aug_level)

  for t in reversed(range(T)):
    x_hr_t = nn(z_t, x_lr, t, aug_level)

    # Sampler step.
    z_tm1 = sampler_step(x_hr_t, z_t, t)
    z_t = z_tm1
  return x_hr_t
```

Efficient UNet Architecture

- More parameters allocated to lower resolutions
- Scaling skip connections by 1/ sqrt(2)
- Reverse the order of downsampling/upsampling and convolutions

Training Details

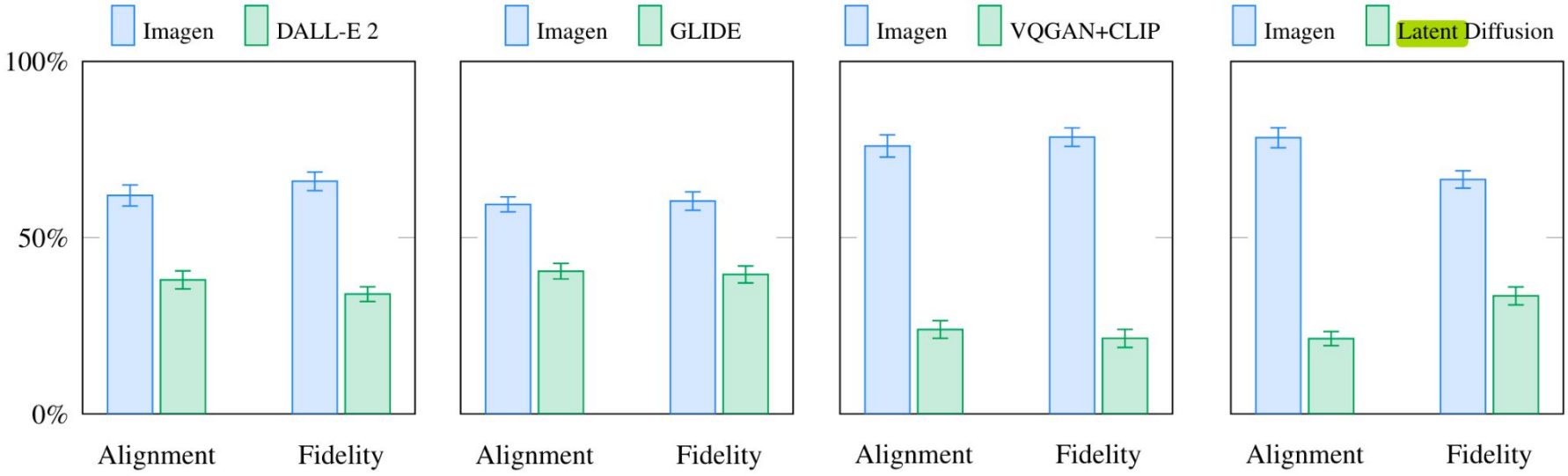
- Base diffusion model: 2B params
- Super-resolution models: 600M and 400M params
- **Training data:** Internal (460M image-text pairs) + Laion (400M image-text pairs)

DrawBench Evaluation Suite

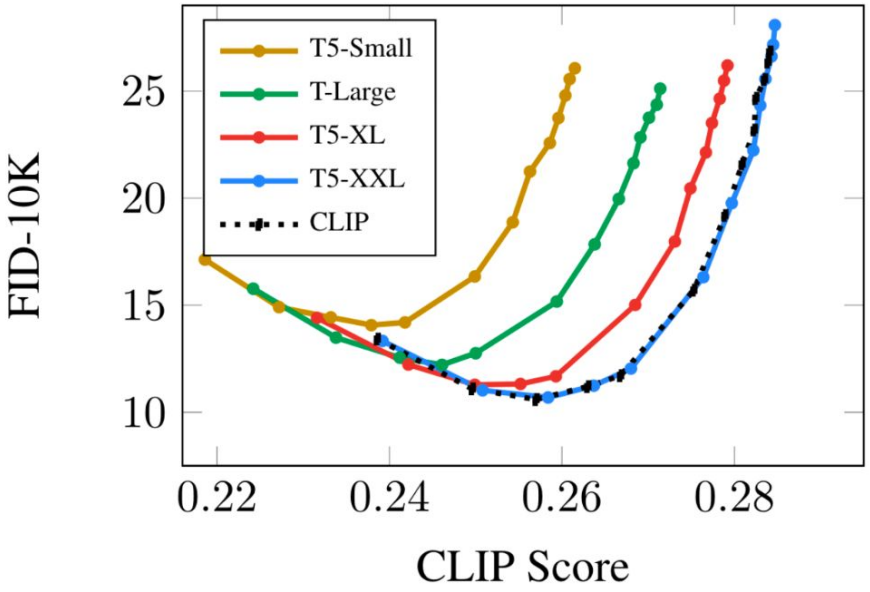
Category	Description	Examples
Colors	Ability to generate objects with specified colors.	"A blue colored dog." "A black apple and a green backpack."
Counting	Ability to generate specified number of objects.	"Three cats and one dog sitting on the grass." "Five cars on the street."
Conflicting	Ability to generate conflicting interactions b/w objects.	"A horse riding an astronaut." "A panda making latte art."
DALL-E [53]	Subset of challenging prompts from [53].	"A triangular purple flower pot." "A cross-section view of a brain."
Description	Ability to understand complex and long text prompts describing objects.	"A small vessel propelled on water by oars, sails, or an engine." "A mechanical or electrical device for measuring time."
Marcus et al. [38]	Set of challenging prompts from [38].	"A pear cut into seven pieces arranged in a ring." "Paying for a quarter-sized pizza with a pizza-sized quarter."
Misspellings	Ability to understand misspelled prompts.	"Rbefraigerator." "Tcennis rpacket."
Positional	Ability to generate objects with specified spatial positioning.	"A car on the left of a bus." "A stop sign on the right of a refrigerator."
Rare Words	Ability to understand rare words[3].	"Artophagous." "Octothorpe."
Reddit	Set of challenging prompts from DALLE-2 Reddit[4].	"A yellow and black bus cruising through the rainforest." "A medieval painting of the wifi not working."
Text	Ability to generate quoted text.	"A storefront with 'Deep Learning' written on it." "A sign that says 'Text to Image'."

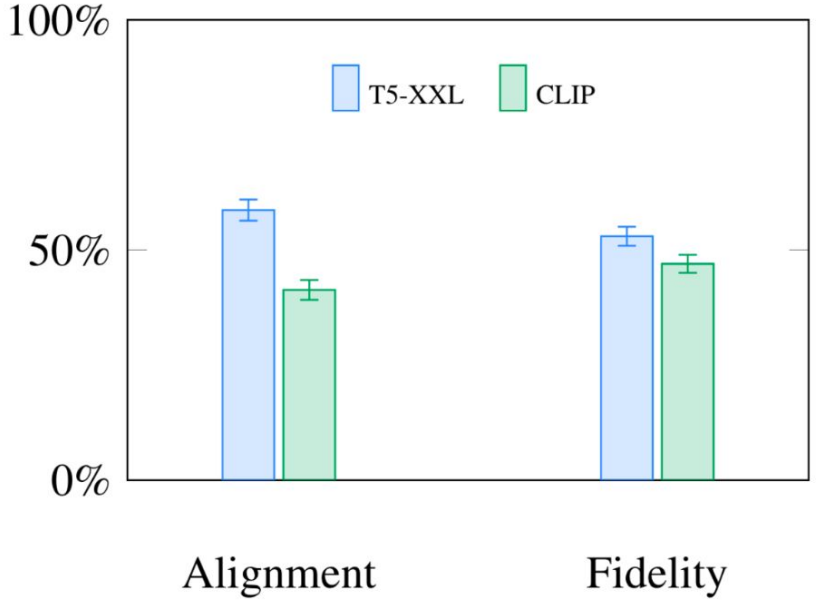
Key Takeaways

Comparison to other T2I Models



Text-Only Encodings are Surprisingly Good

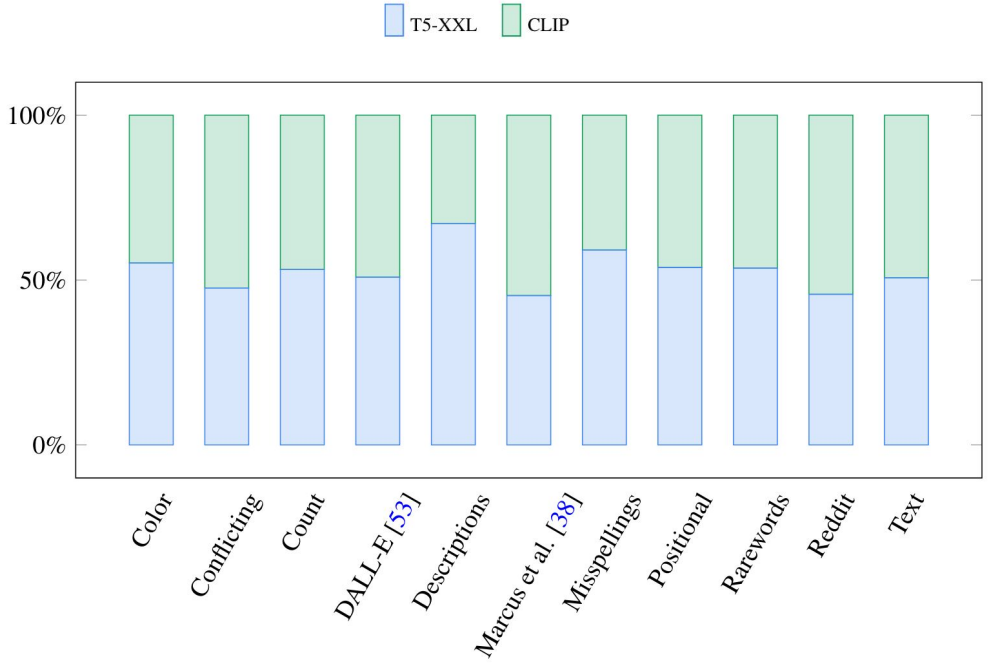




Text-Only Encodings are Surprisingly Good

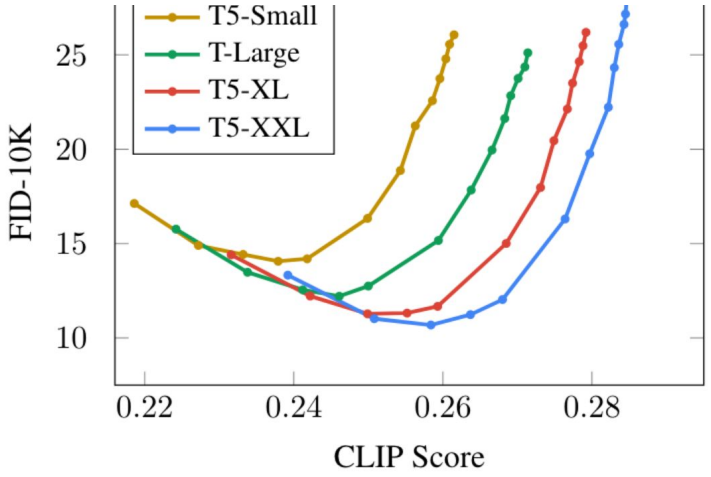
(a) Alignment

Text-Only Encodings are Surprisingly Good

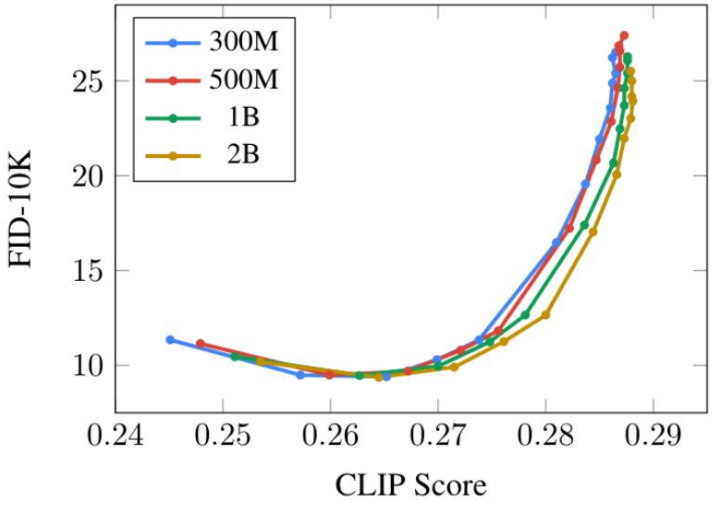


(b) Fidelity

Scaling Text Encoders is More Effective than Scaling the Diffusion Model



(a) Impact of encoder size.



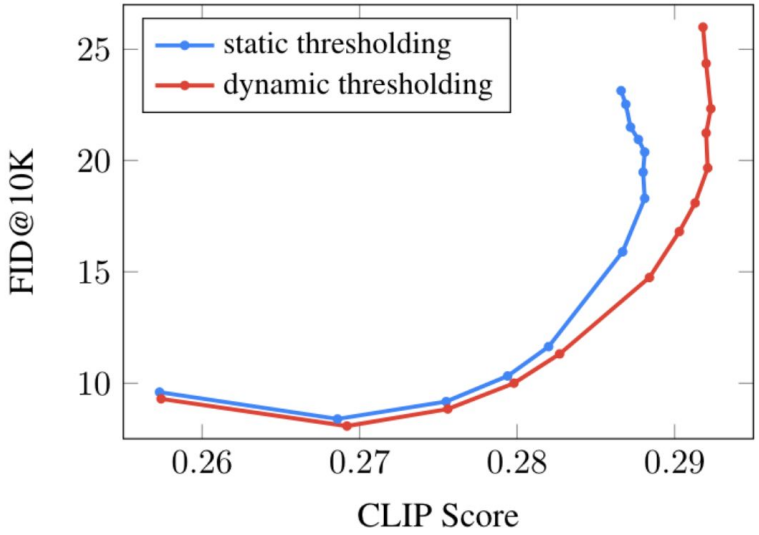
(b) Impact of U-Net size.

Dynamic Thresholding is Critical

(a) Samples using static thresholding.

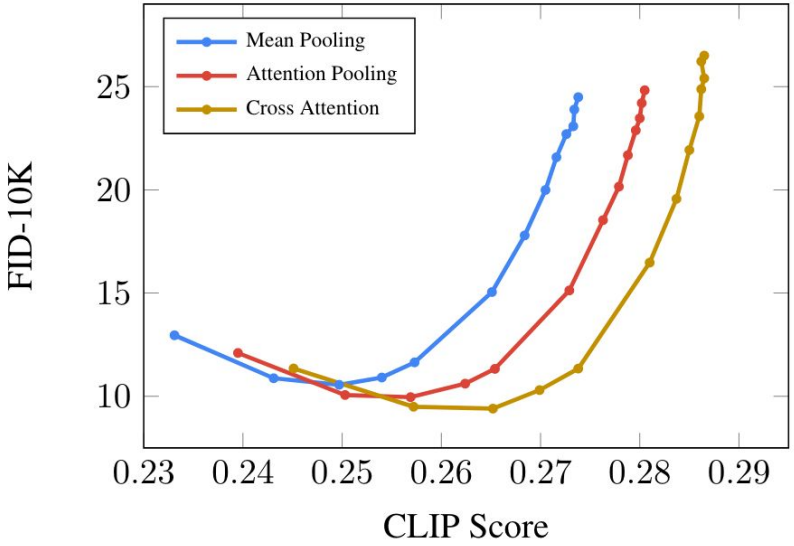
(b) Samples using dynamic thresholding ($p = 99.5$)

Dynamic Thresholding is Critical



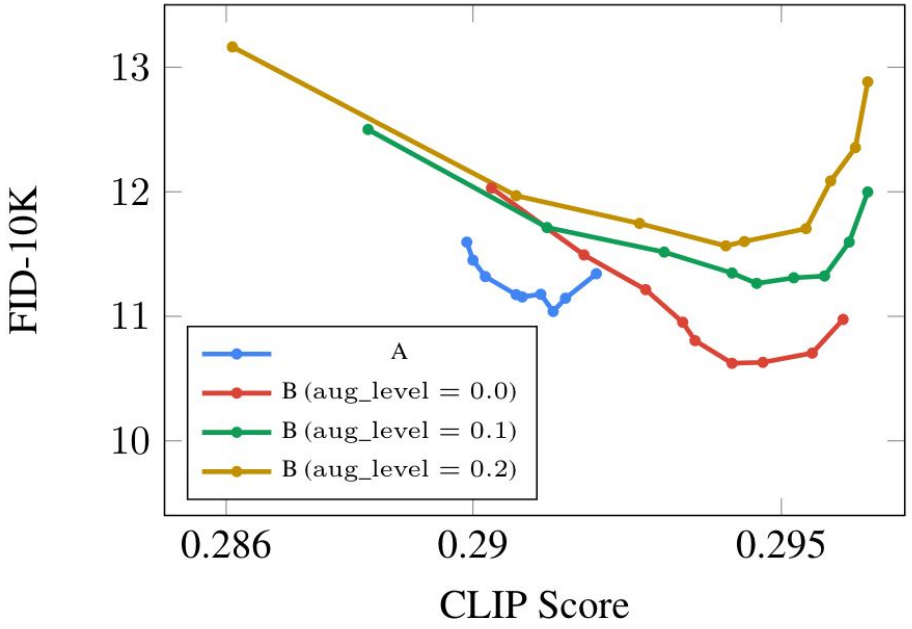
(c) Impact of thresholding.

Cross-attention is critical



(a) Comparison between different text encoders.

Noise conditioning augmentation is critical



Cascaded Diffusion + Noise Augmentation Enables Diversity

Key Limitation: Face Generation

"Our human evaluations found Imagen obtains significantly **higher preference rates** when evaluated on images that **do not portray people**, indicating a degradation in image fidelity."

Discussion - Frozen Language Models

1. The paper demonstrates that large **frozen language models** trained only on text perform better than multimodal models like CLIP for text encoding. Does this suggest that **pure language understanding** is more fundamental than visual-semantic alignment for text-to-image generation?
2. How might the design choice of **freezing the text encoder** affect the model's generalization capabilities?

Discussion - Video Generation + Scaling

1. Does this paradigm readily extend to **video generation**? Can we really capture the evolution of temporal semantics just through text prompts?

2. For scaling to video/larger images, is l**atent or cascaded diffusion** a more promising paradigm?

Thanks!