Hierarchical Text-Conditional Image Generation with CLIP Latents

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, Mark Chen

Group 6:

Jeffrey Chan, Qingyuan Li, Kevin Samms, Zhaoning Wang

Outline

- **Background/Motivation**
- **Method**
  - Overall Method
  - Prior
  - Decoder
- **Image Manipulation**
- **Text-to-Image Generation Analysis**
  - Why the prior matters?
  - GLIDE vs Dalle-2/unCLIP (Human Evaluation)
  - Diversity-Fidelity Trade-off with Guidance
- **Limitations**

Background/Motivation

Text to Image Generation

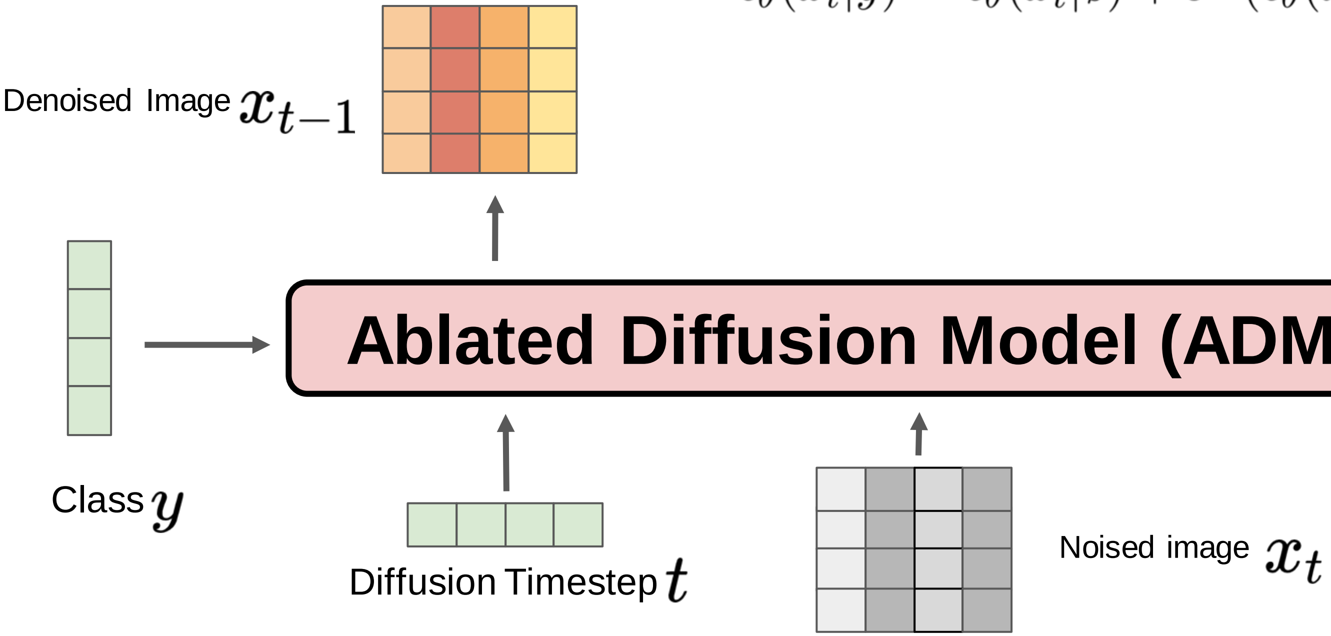
“an espresso machine that makes coffee from human souls, artstation”

“panda mad scientist mixing sparkling chemicals, artstation”

“a corgi’s head depicted as an explosion of a nebula”

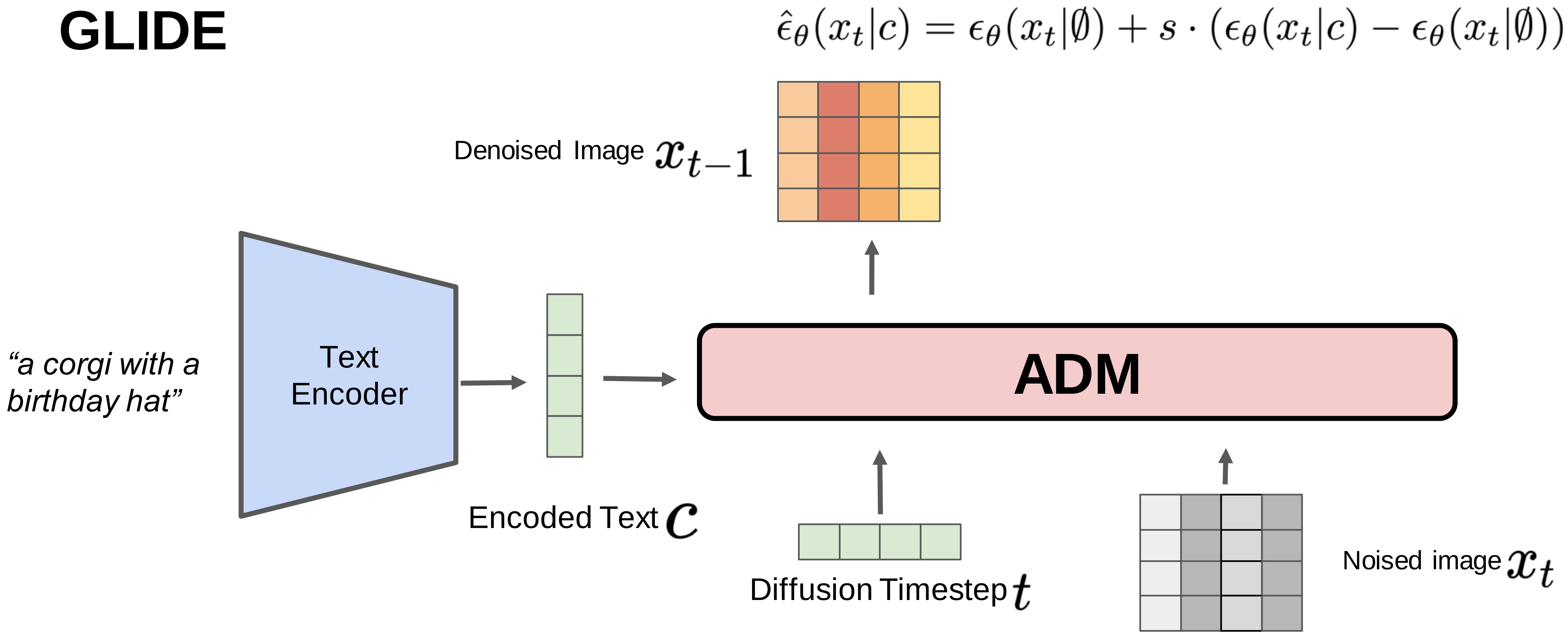
Conditioned Diffusion Model

$$\hat{\epsilon}_\theta(x_t|y) = \epsilon_\theta(x_t|\emptyset) + s \cdot (\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset))$$



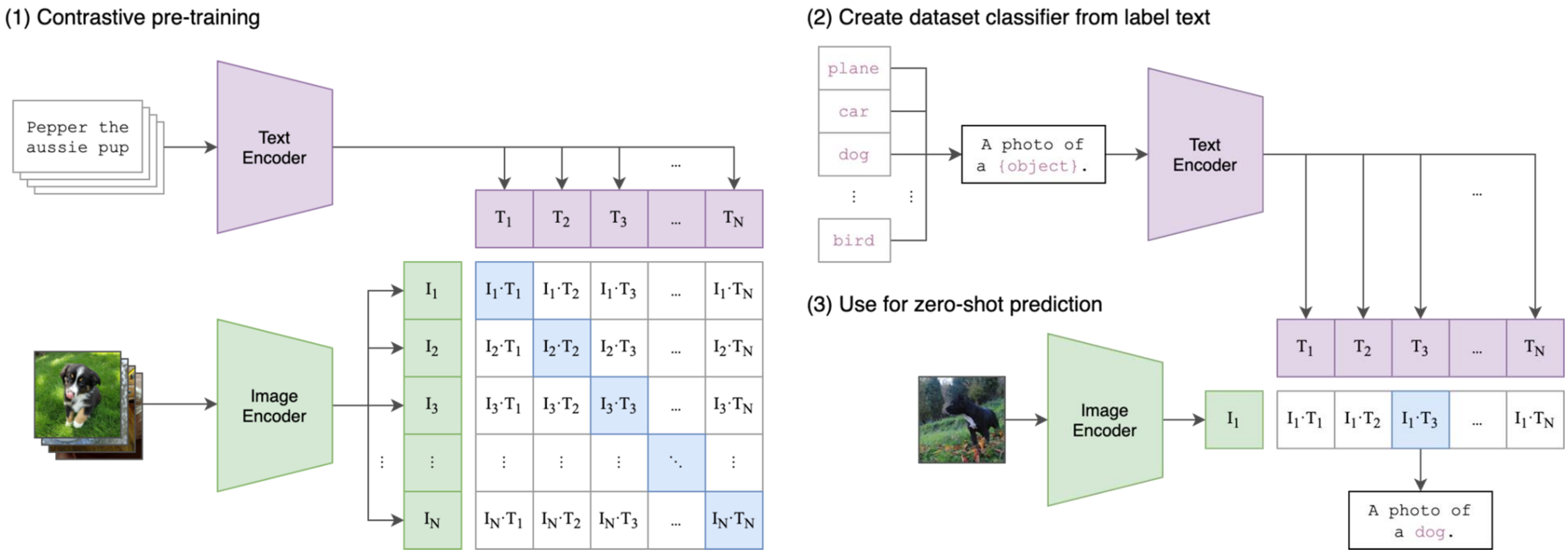
Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." *Advances in Neural Information Processing Systems* 34 (2021): 8780-8794.

GLIDE



Nichol, Alex, et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." *arXiv preprint arXiv:2112.10741* (2021).

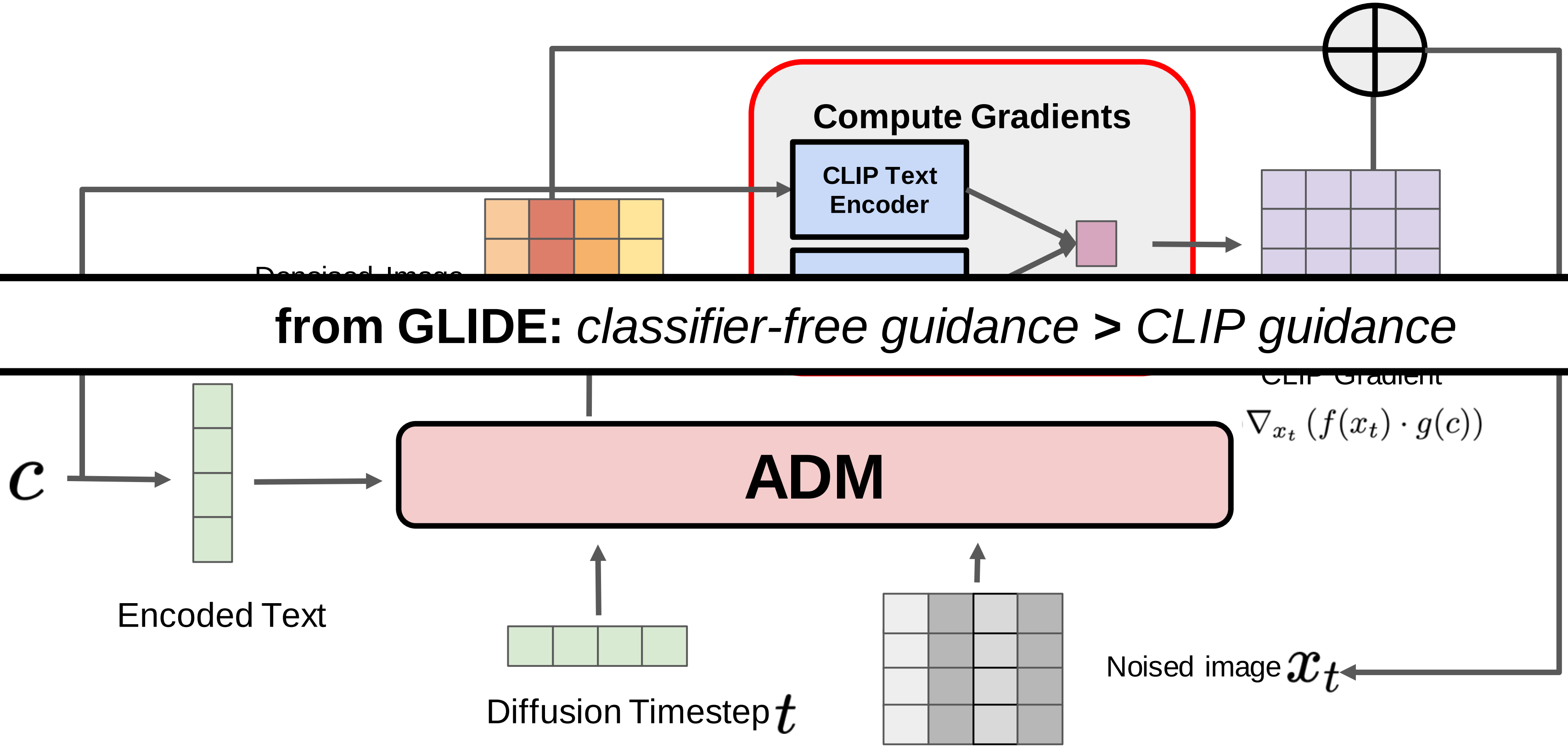
CLIP



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

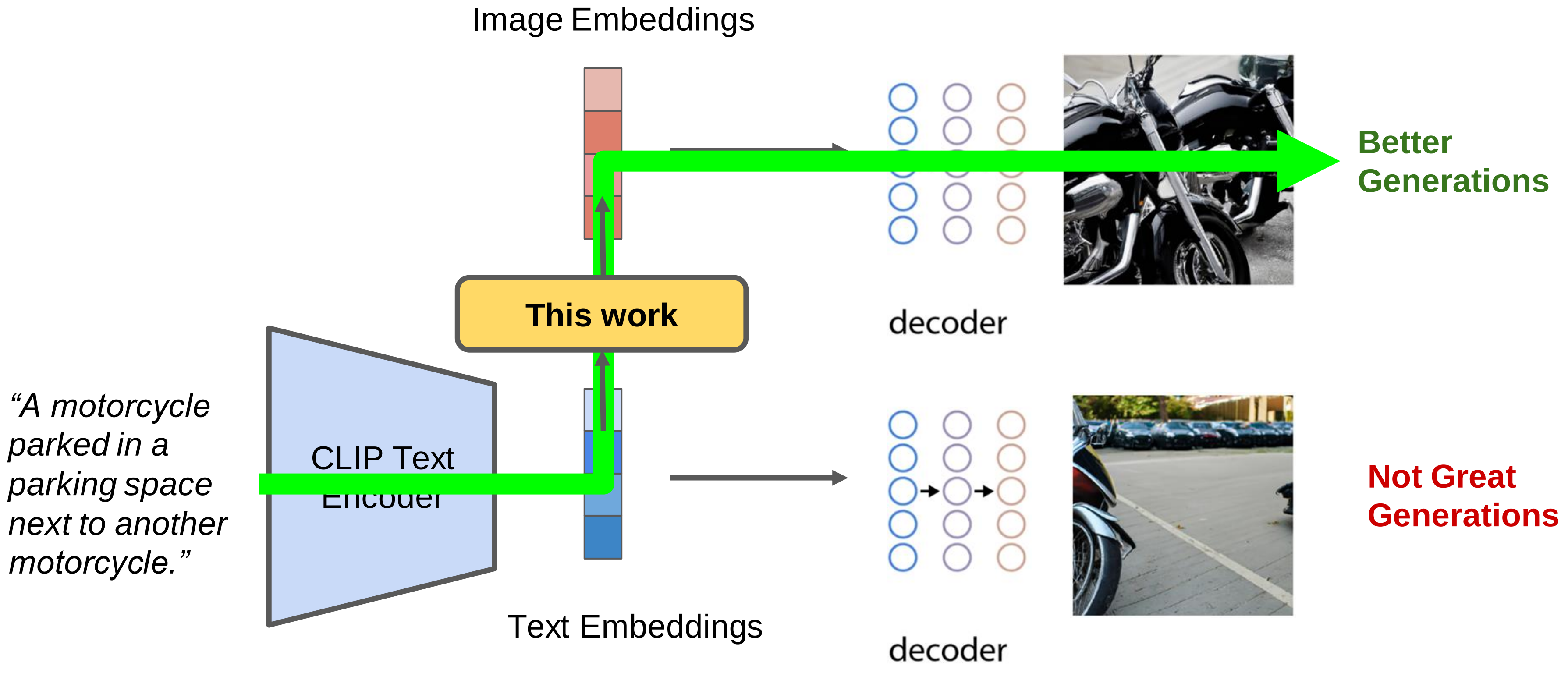
CLIP Guided Diffusion Model

$$\hat{\mu}_\theta(x_t|c) = \mu_\theta(x_t|c) + s \cdot \Sigma_\theta(x_t|c) \underline{\nabla_{x_t}(f(x_t) \cdot g(c))}$$



from GLIDE: *classifier-free guidance* **>** *CLIP guidance*

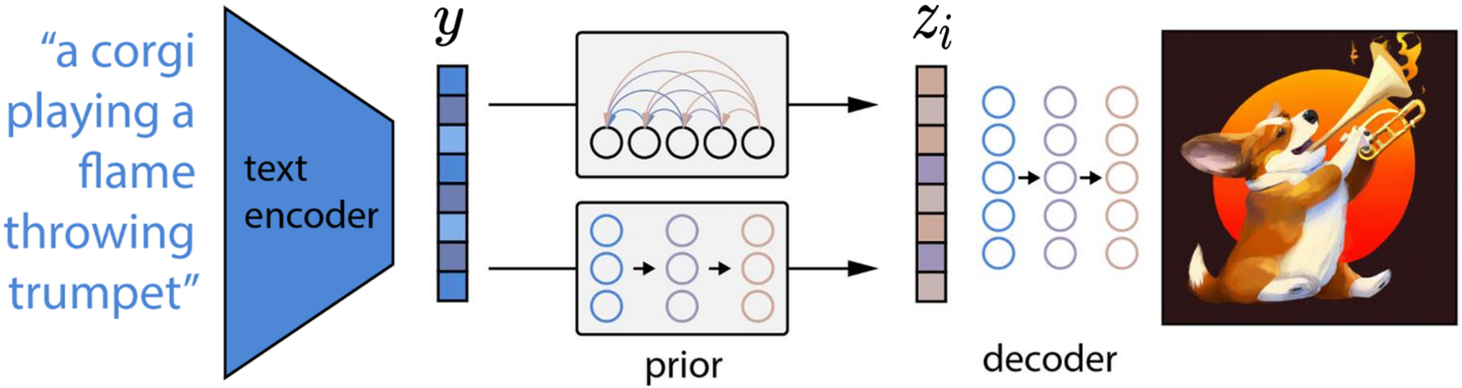
How use CLIP more effectively to improve generations?



Method

unCLIP/DALL-E-2 architecture

- Prior
  - Given CLIP Text encoder output (text embedding) y, generate corresponding Image Embedding z_i
- Decoder
  - Produces the image from Image embedding z_i

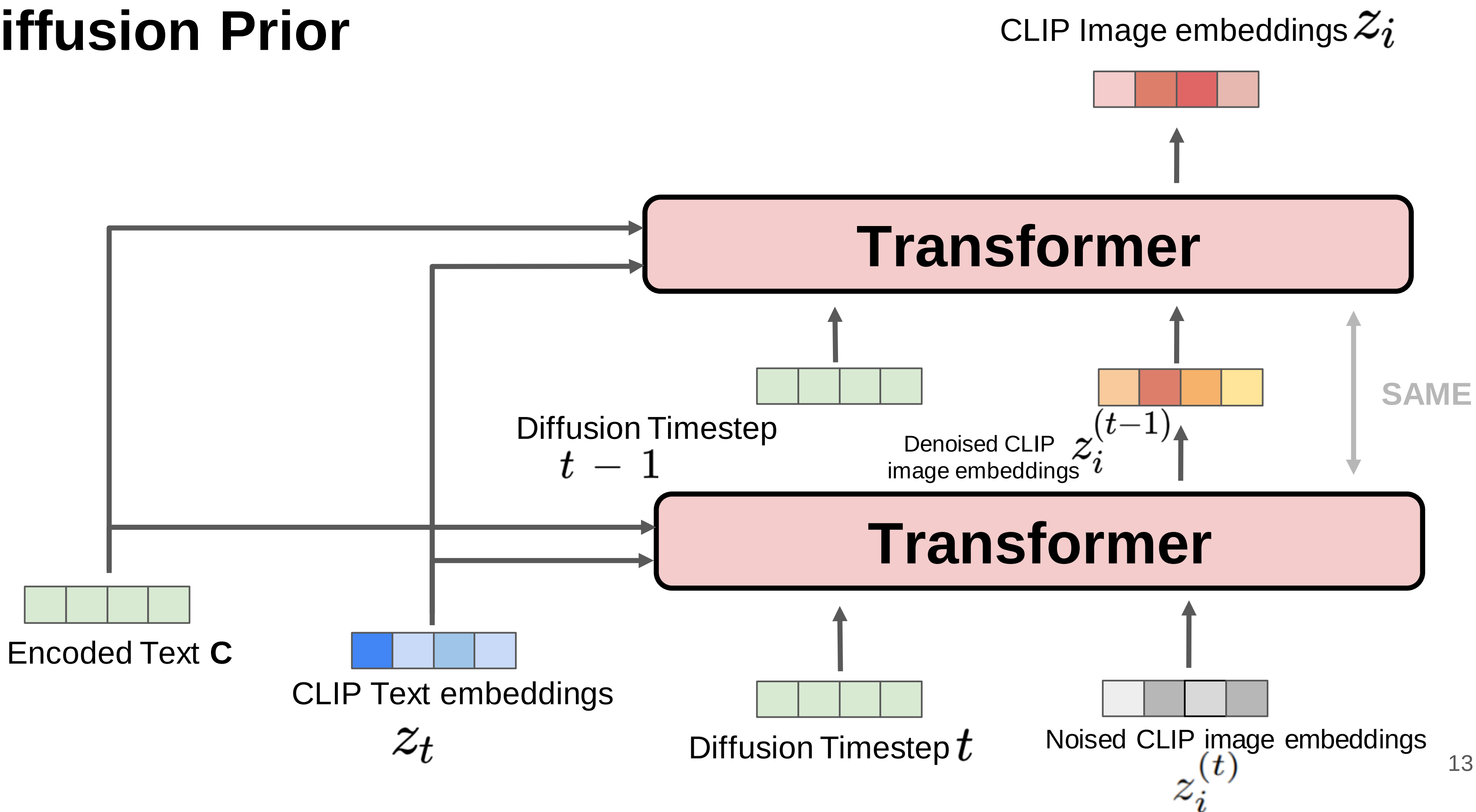


Prior

- Autoregressive (AR) prior:
  - AR models predict a sequence of data on a previous data sequence
  - Use a transformer to predict Image embedding sequence from the Text embedding sequence.

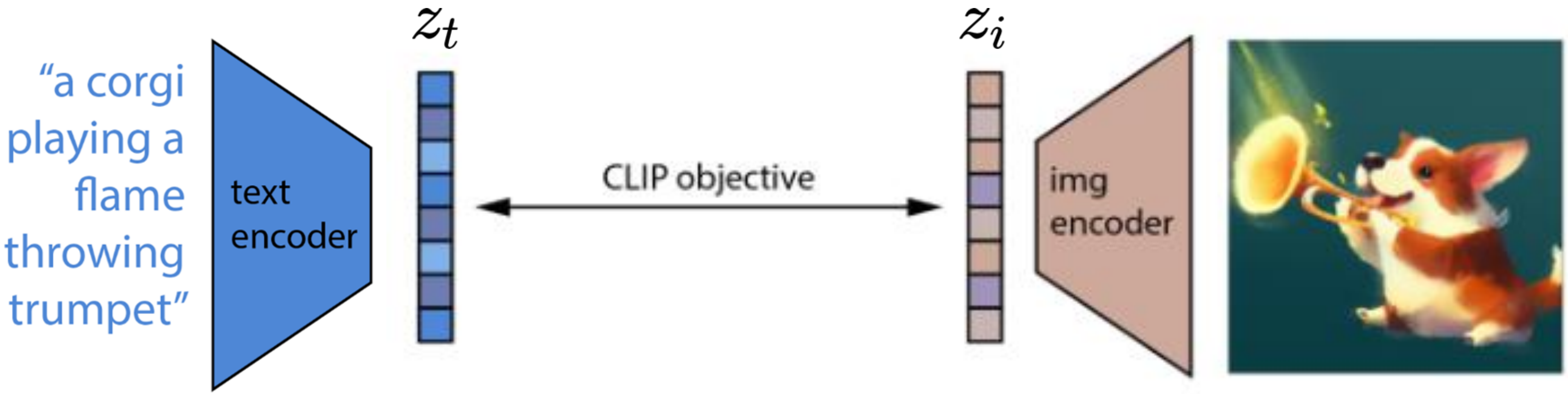
- Diffusion prior:
  - Diffusion model on CLIP Image Embedding
  - Input:
    - Encoded text
    - CLIP text embedding
    - Timestep
    - Noised CLIP Image Embedding

Diffusion Prior



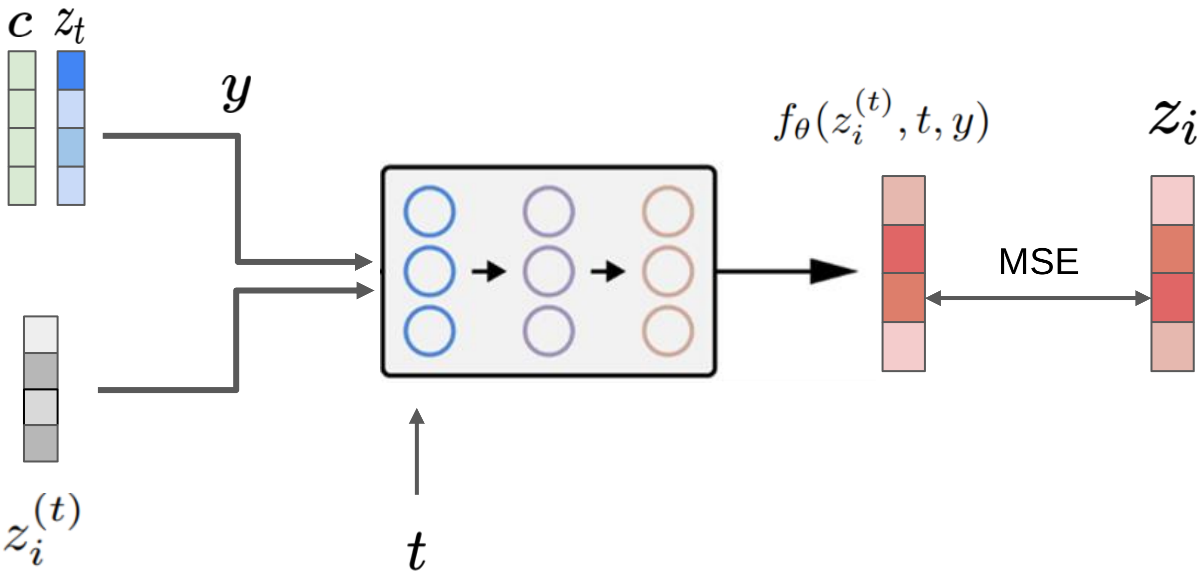
Training

- Using CLIP to get input and ground-truth while training the prior.



Training Loss

$$L_{\text{prior}} = \mathbb{E}_{t\sim[1,T], z_i^{(t)}\sim q_t}\left[\|f_\theta(z_i^{(t)}, t, y) - z_i\|^2\right]$$

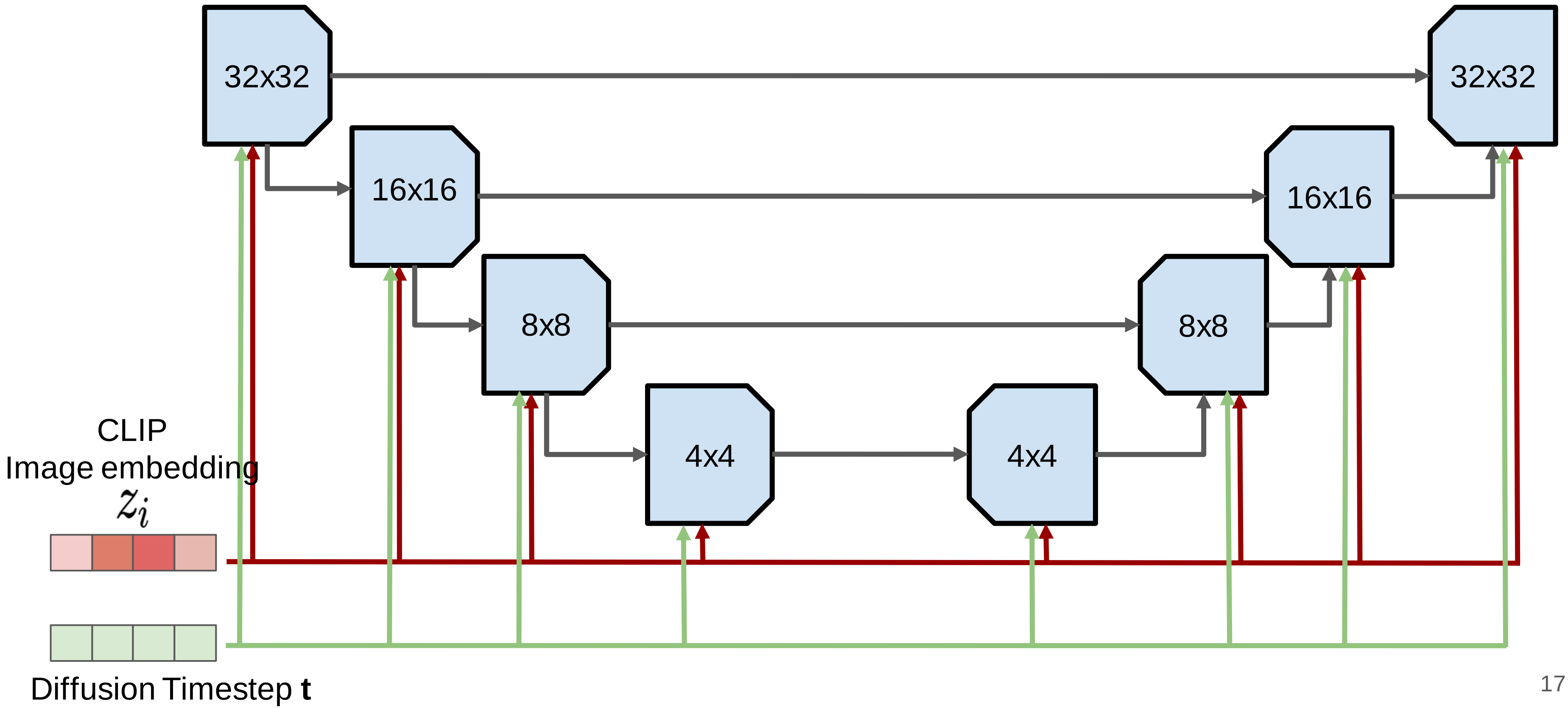


* y *is the combination of encoded text* ***C*** *and CLIP Text Embedding* z_t

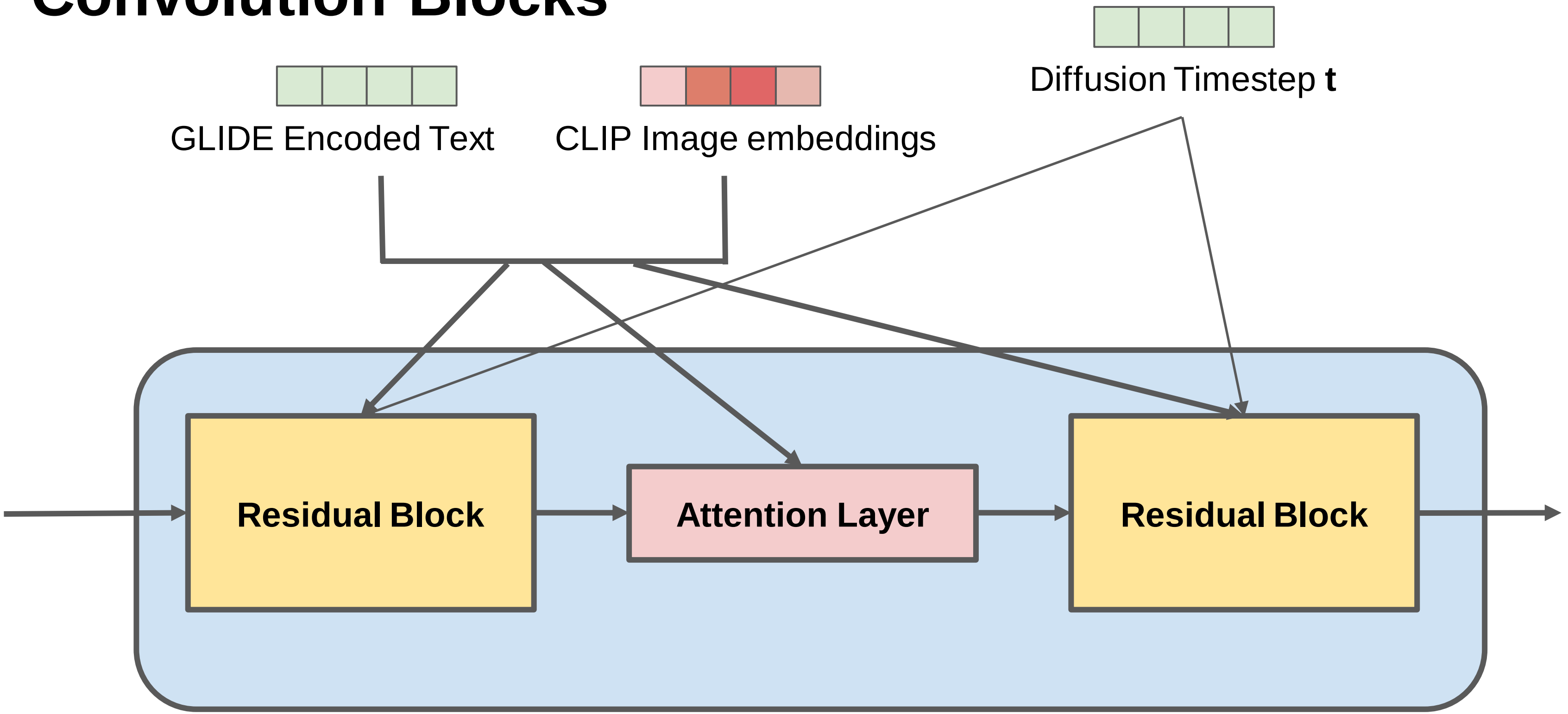
Decoder

- Diffusion model based on GLIDE
  - GLIDE uses a transformer to embedding the input text
  - Dall-E-2 put CLIP embedding into the process
- Upsampler
  - Used to generate higher-resolution Images
  - No conditioning, and no guidance

Decoder U-Net detail

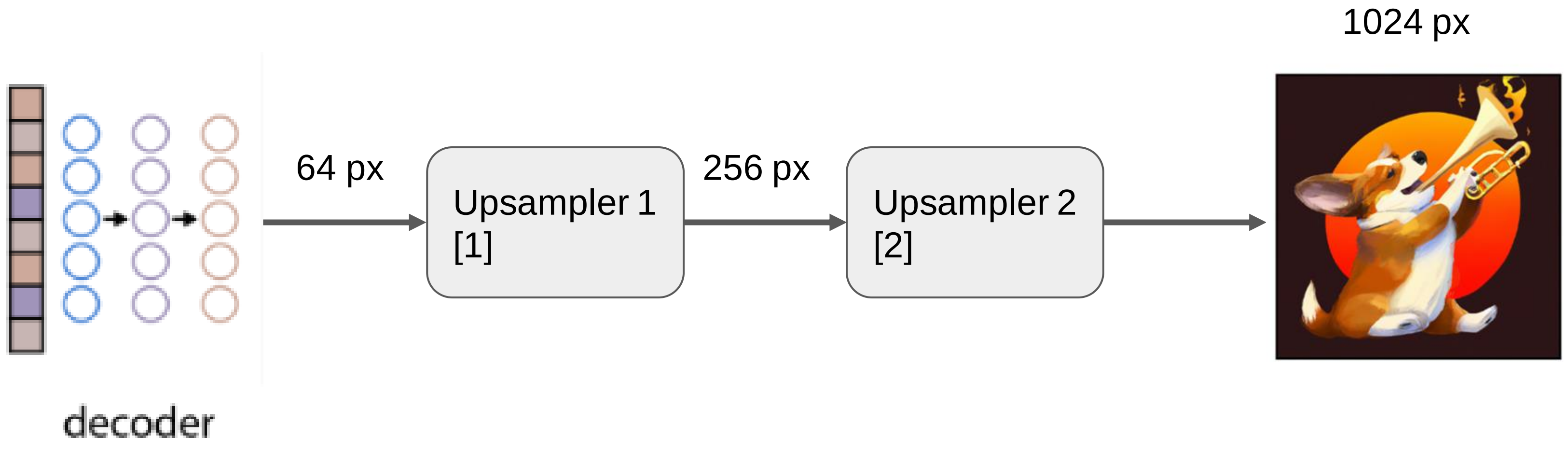


Convolution Blocks



Upsampler

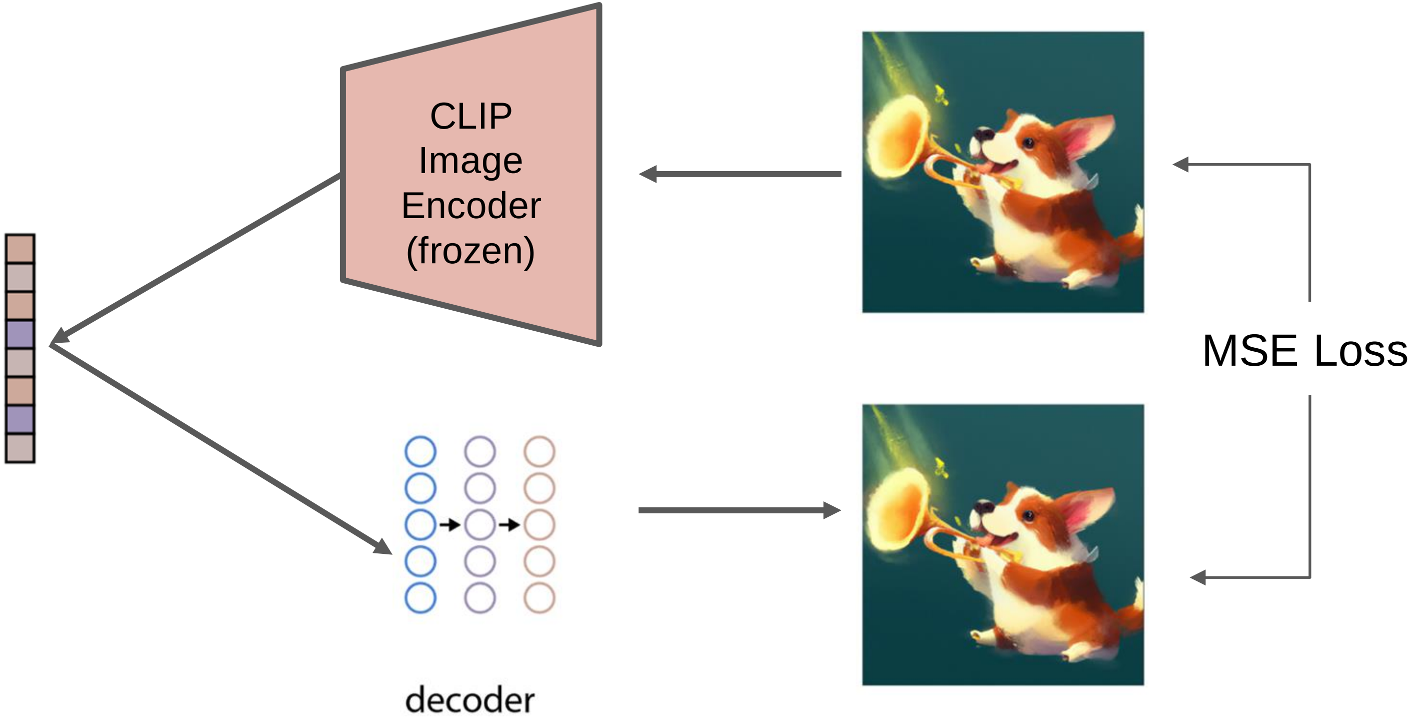
2 unconditional off-the-shelf upsamplers to create images in higher resolution



[1] Alex Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. arXiv:2102.09672, 2021.

[2] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. arXiv:arXiv:2104.07636, 2021.

Training the decoder with CLIP encoder



Inference

- Prior
  - Convert the CLIP Text Embedding to CLIP Image Embedding z
- Decoder
  - Produces the image from Image embedding z and optionally with text embedding y.

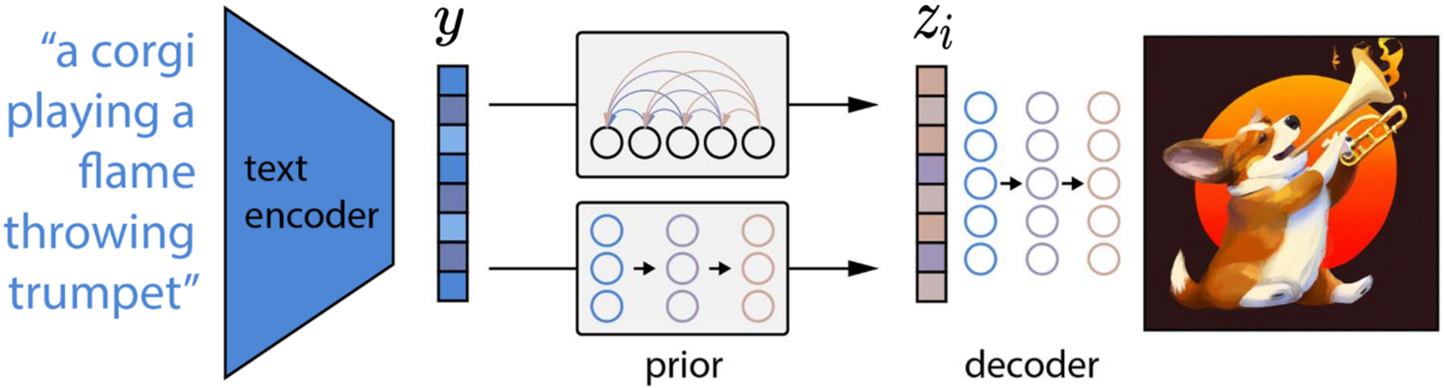


Image Manipulations

What is Latent space

Bipartite latent representation (z_i , X_t)

Encode with CLIP image encoder

DDIM inversion [1]

[1]Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233, 2021

Variation

Fix z_i

Vary X_t

Input Image:

Generation:

Interpolation

z_{i1} z_{i2}

Modify image embedding:

$$z_{i_\theta} = \text{slerp}(z_{i_1}, z_{i_2}, \theta)$$

Text Diff

"a photo of a victorian house" → "a photo of a modern house"

z_i $\quad z_d = \text{norm}(\ z_t \quad - \quad z_{t_0}\ )$

Compute Difference CLIP Embeddings

$$z_\theta = \text{slerp}(z_i, z_d, \theta)$$

Apply text diff direction to the image embedding

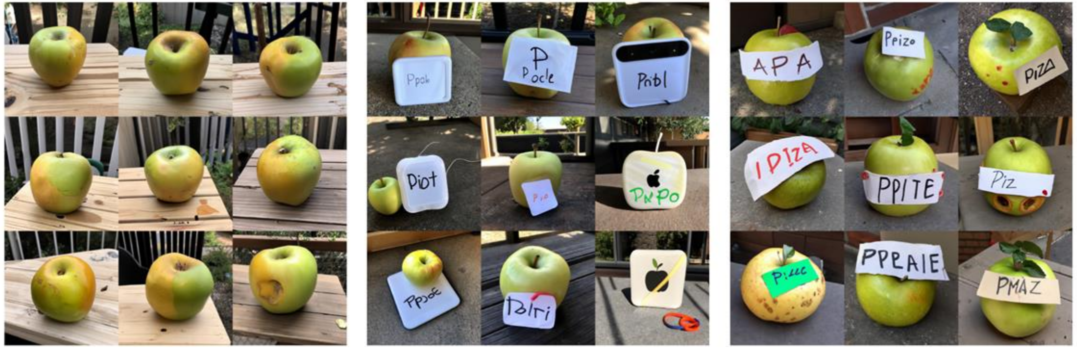
Typographic Attacks

Attack:

Clip Image Prediction:

Granny Smith: 100% iPod: 0% Pizza: 0%	Granny Smith: 0.02% iPod: 99.98% Pizza: 0%	Granny Smith: 94.33% iPod: 0% Pizza: 5.66%

Generation Image Embedding :



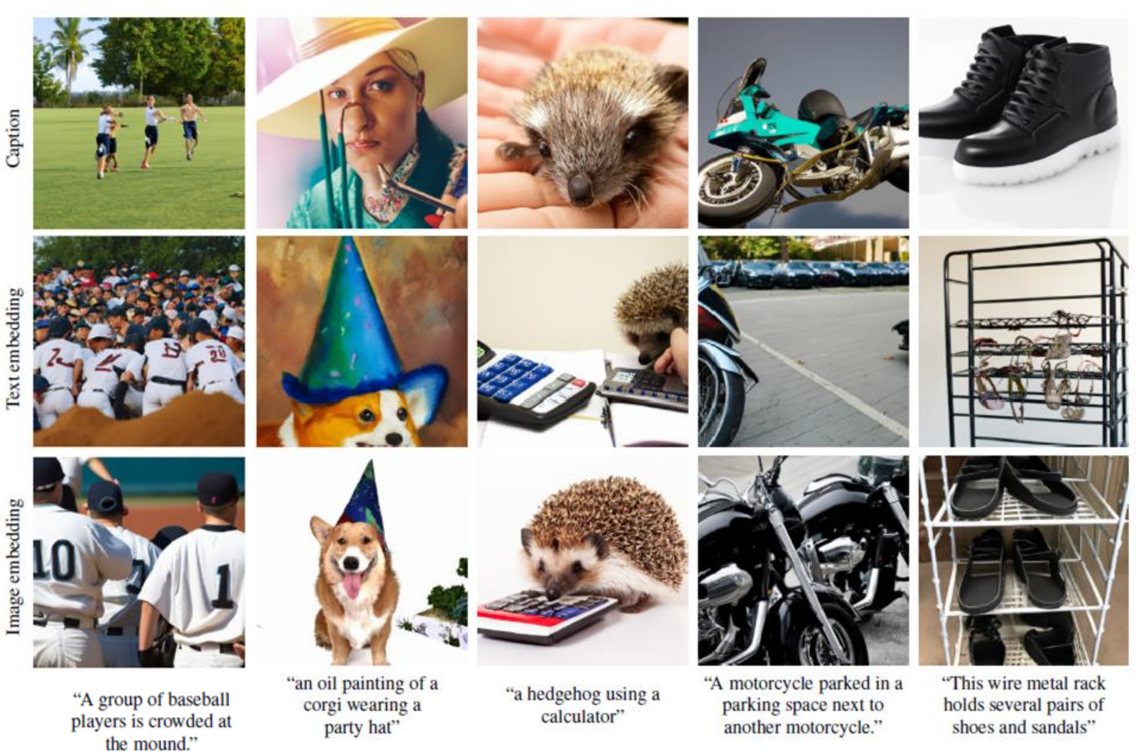
Text-to-Image Generation Analysis

Why the prior matters?

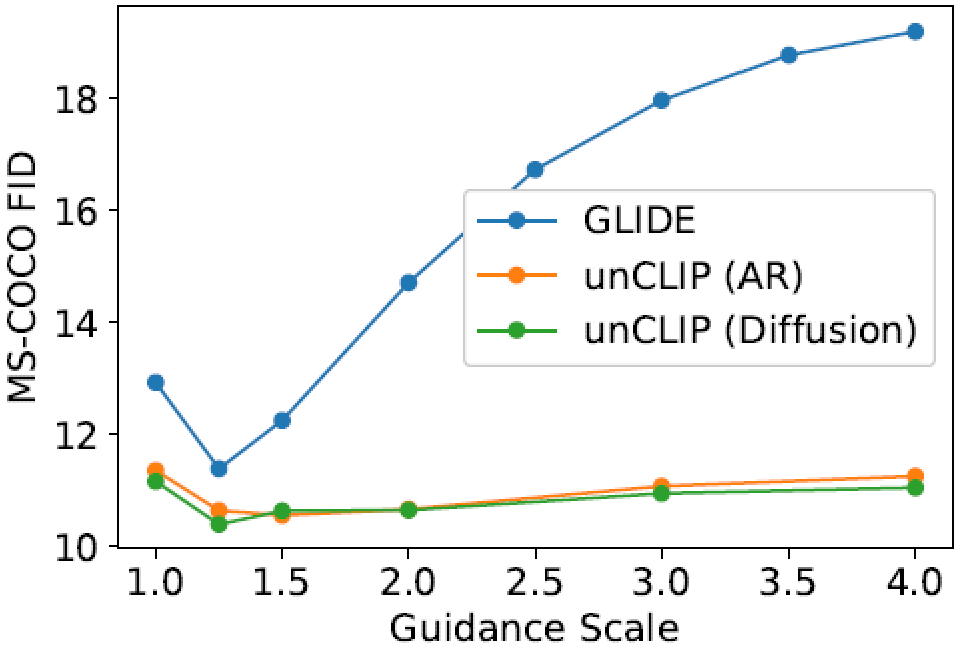
Condition decoder on captions alone ✗

Condition decoder on Caption + text embedding impersonating image embeddings ✗

Prior + CLIP image embedding ✓



MS COCO FID SCORE



GLIDE vs unCLIP
(MS-COCO)

MS-COCO - standard evaluation:

- Zero-shot FID score 10.39 - beats GLIDE & DALL-E in 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**

GLIDE vs unCLIP

(Human Evaluations)

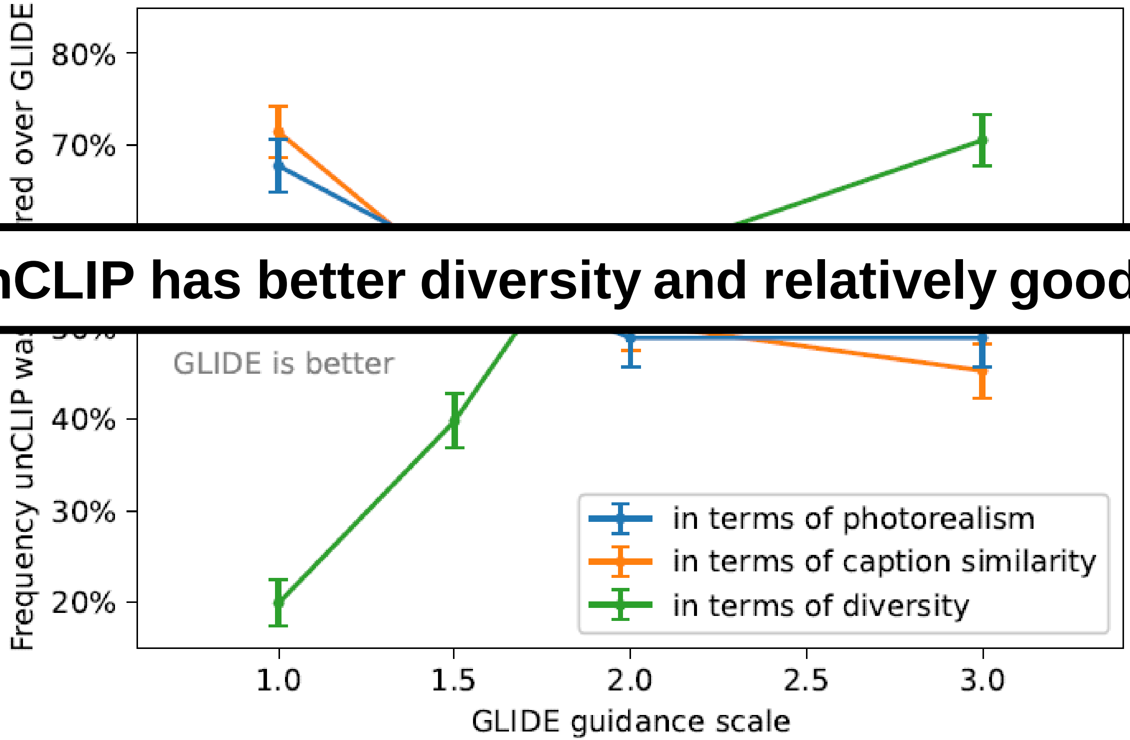
FID not always in agreement with human evaluation

Photorealism → winner: GLIDE - by **small** margin; 48.9%CI

Caption Similarity → winner: GLIDE - by **small** margin; 45.3%CI

Sample Diversity (4 x 4 grid) → winner: unCLIP stack by **wide** margin; 70.5%CI

Diversity-Fidelity Trade-off with Guidance



unCLIP has better diversity and relatively good fidelity

Image aesthetics improved for both unCLIP and GLIDE

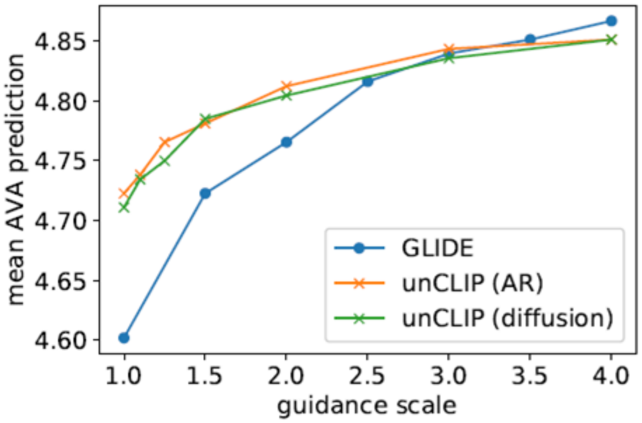
GLIDE vs unCLIP

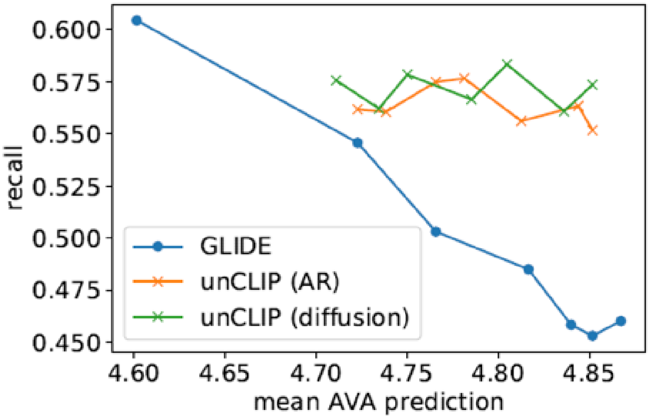
Aesthetic Quality

CLIP Linear Probe + AVA Dataset → Human aesthetic judgement

Result:

- Guidance improves GLIDE, and CLIP decoder (negative effect on CLIP prior)
- GLIDE sacrifices Recall for aesthetic quality improvement, unCLIP does not



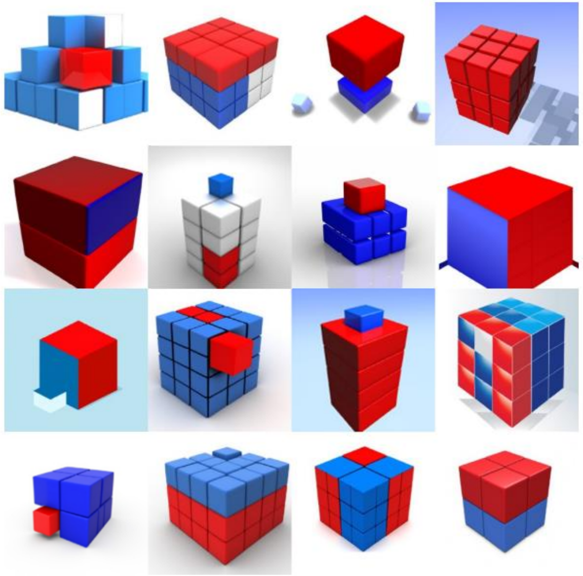


Limitation of the model

Attribute Binding

- Suffer prompt where it must bind two separate objects (cubes) to two separate attributes (colors).
- Reconstructions mix up objects and attributes

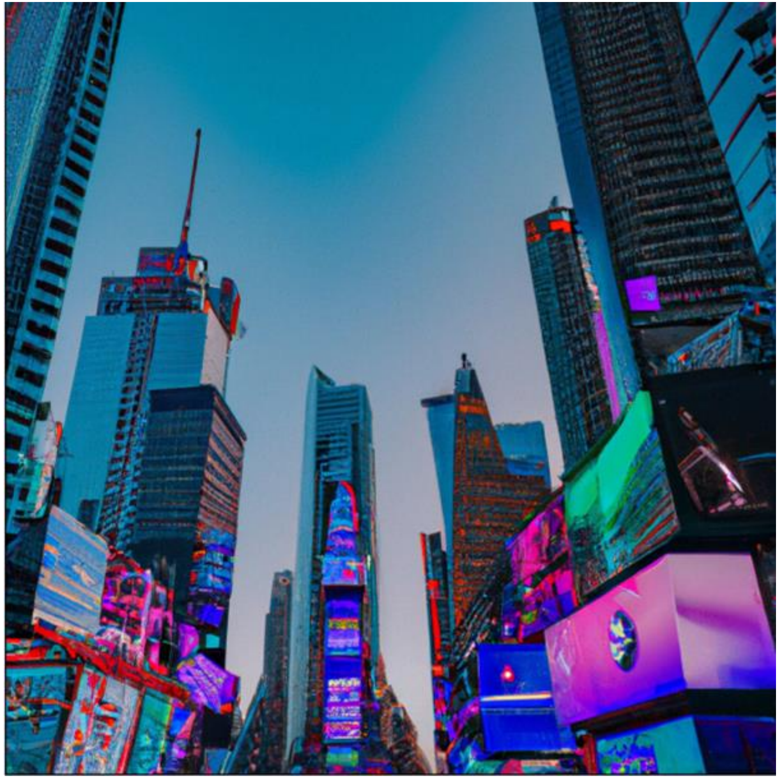
“a red cube on top of a blue cube”.



Coherent Text

A sign that says deep learning

Complex Scene



Conclusion

- Image embedding creates better generation than text embeddings.
- CLIP embedding Z_i holds image content information; meanwhile X_t holds the style of image generation.
- Diffusion prior (Text-to-Image embeddings) increases the fidelity of image generation.
- unCLIP has limitations with attribute binding, text generation, and complex scenes.