GLIDE SIGNIFICANCE

- GLIDE performs better than DALL-E for photorealism and caption similarity.
- DALL-E 2 uses a modified GLIDE model that incorporates projected CLIP text embeddings in two ways.

"a hedgehog using a calculator"

"a corgi wearing a red bowtie and a purple party hat"

"robots meditating in a vipassana retreat"

"a fall landscape with a small cottage next to a lake"

MAPPING FROM TEXT TO IMAGE

- Mapping from Textual Semantics to Corresponding Visual Semantics
- In addition to our image encoder, CLIP also learns a text encoder. DALL-E 2 uses another model, which the authors call the prior, in order to map from the text encodings of image captions to the image encodings of their corresponding images