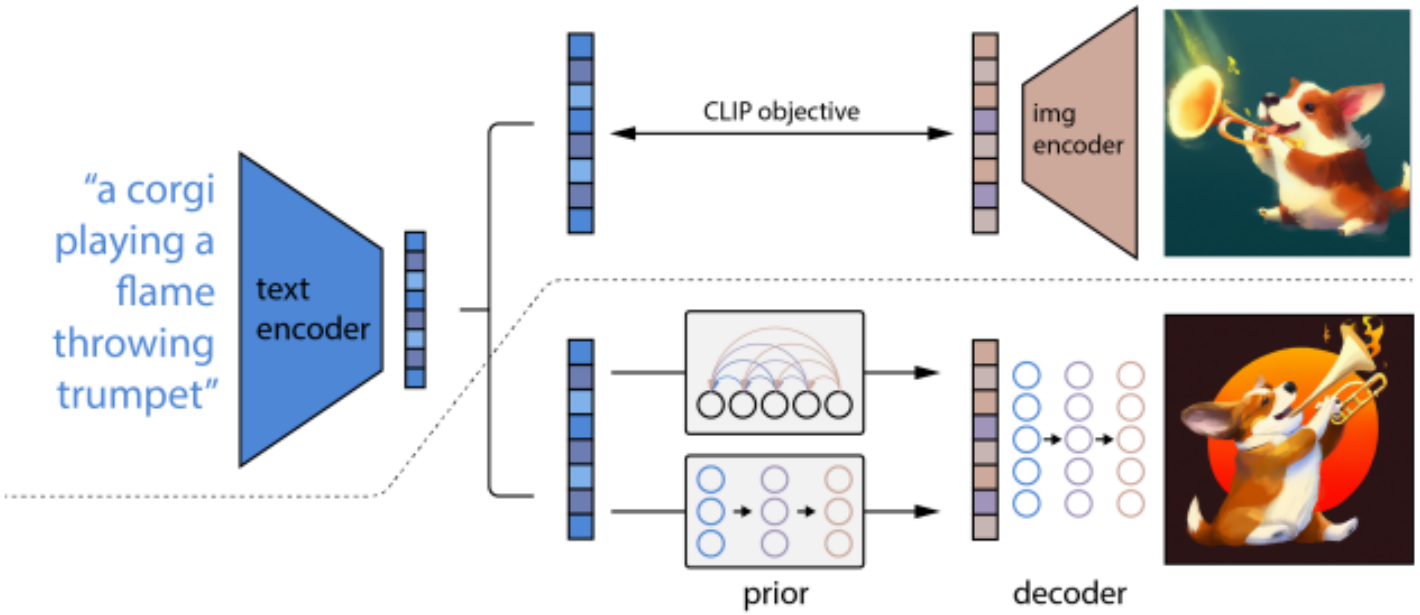
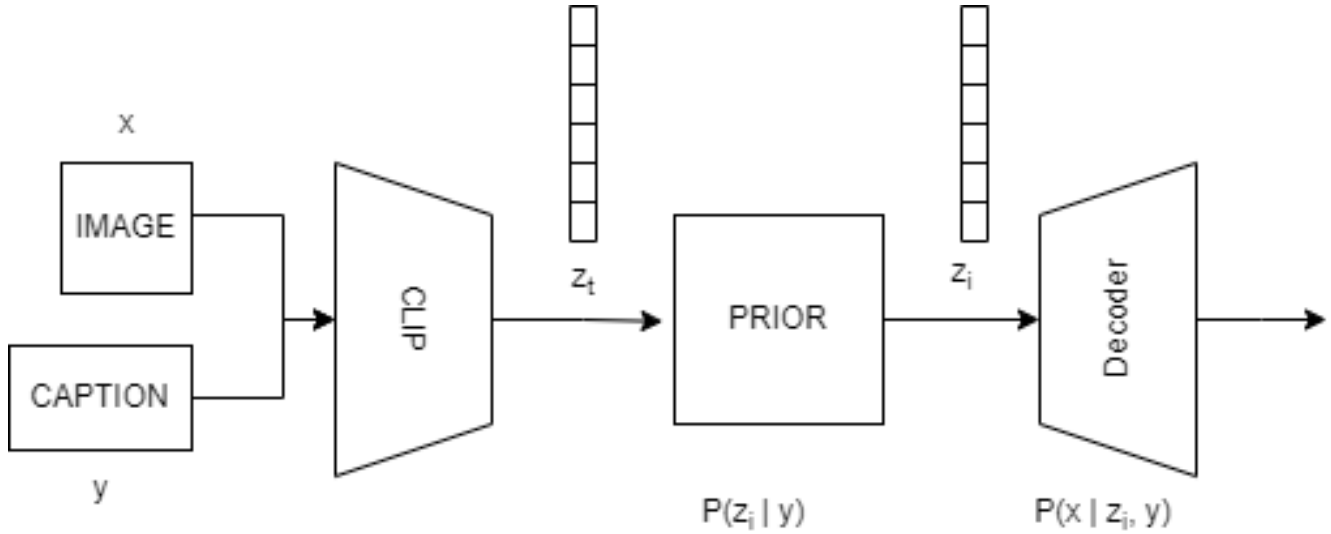
Hierarchical Text-Conditional Image Generation with CLIP Latents

Method

- CLIP
  - Image to text
- unCLIP
  - Text-conditional image generation
  - Prior (AR and diffusion)
  - Decoder (diffusion)





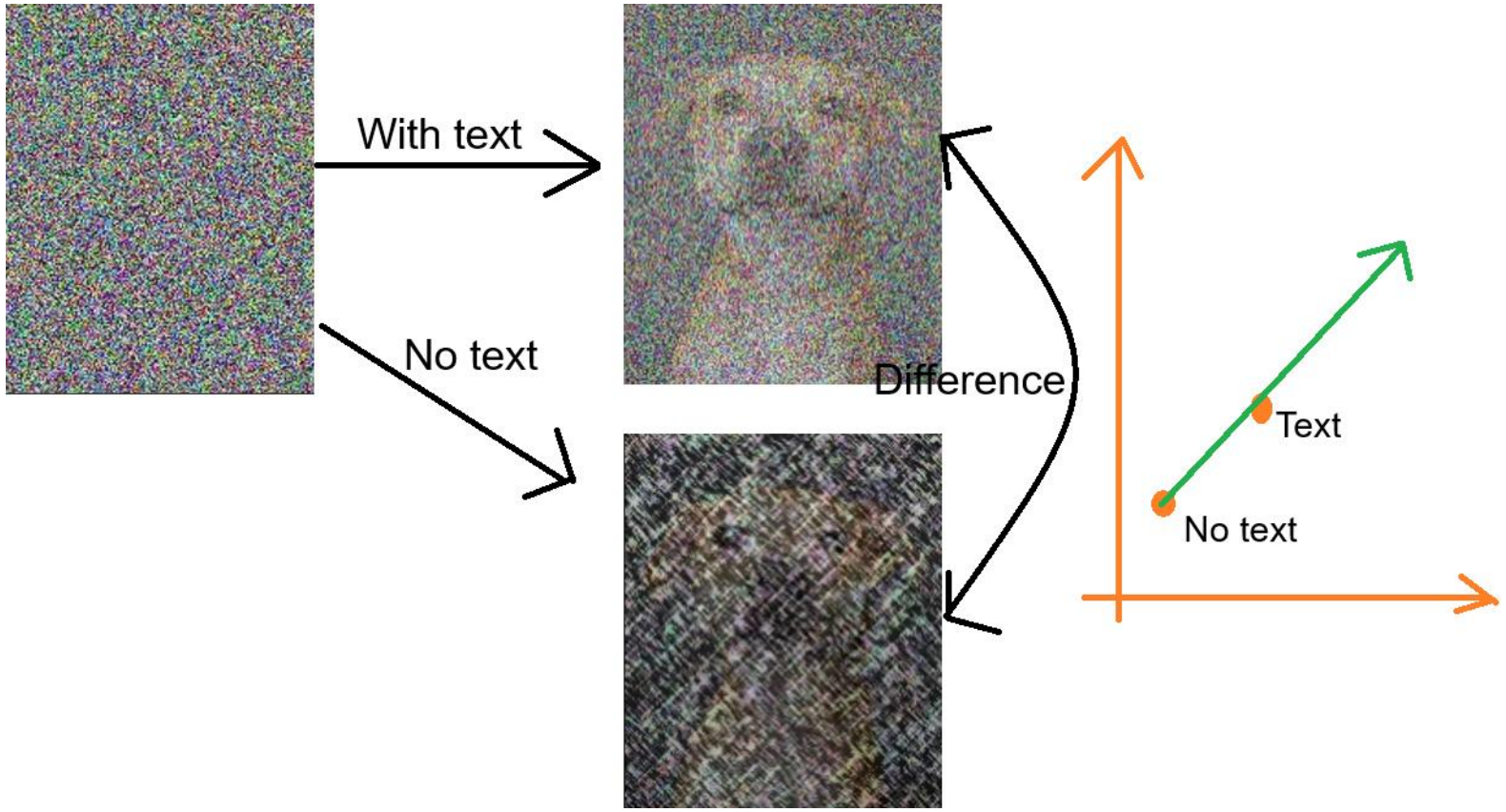
Diffusion

- Inspired by non-equilibrium thermodynamics
  - Like paint in water
  - Reversing is unachievable

- Diffusion models try to reverse this process
  - Adding noise to images
  - Markov chain (up to 1000s times)
  - Gaussian noise
  - Uses several NNs

Decoder

- GLIDE
  - Text guided
  - Uses text prompt
  - Classifier-free guidance
- Modified GLIDE
  - Uses CLIP embeddings
  - 2 upsamplers



Prior

- Autoregressive prior (AR)
- Diffusion prior
  - Decoder only transformer
  - Input
    - Encoded text
    - CLIP embedding
    - Timestamp
    - Noised CLIP image embedding
  - Output:
    - Denoised CLIP image embedding
  - Computationally more efficient and higher quality samples

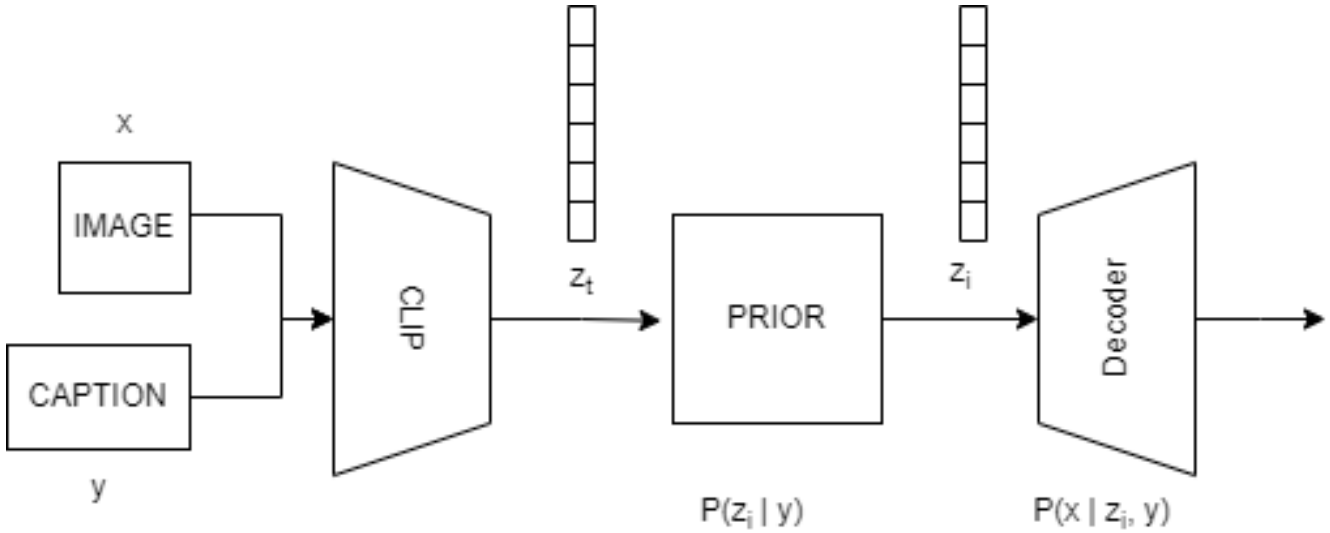
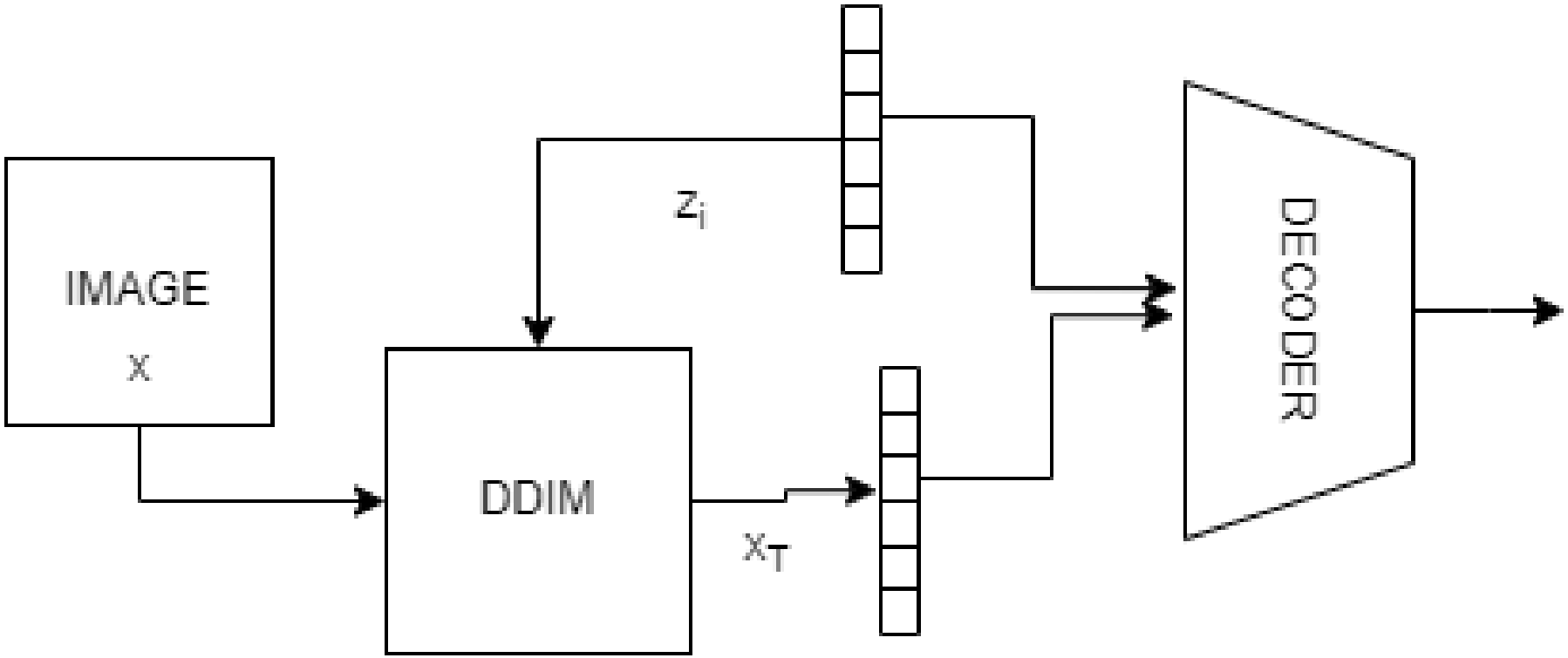


Image manipulation



Variation manipulations

Interpolation manipulations

Language-guided manipulations

a photo of a cat → an anime drawing of a super saiyan cat, artstation

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

a photo of an adult lion → a photo of lion cub

a photo of a landscape in winter → a photo of a landscape in fall

Why need a prior?