Versatile Diffusion: Text, Images and Variations All in One Diffusion Model

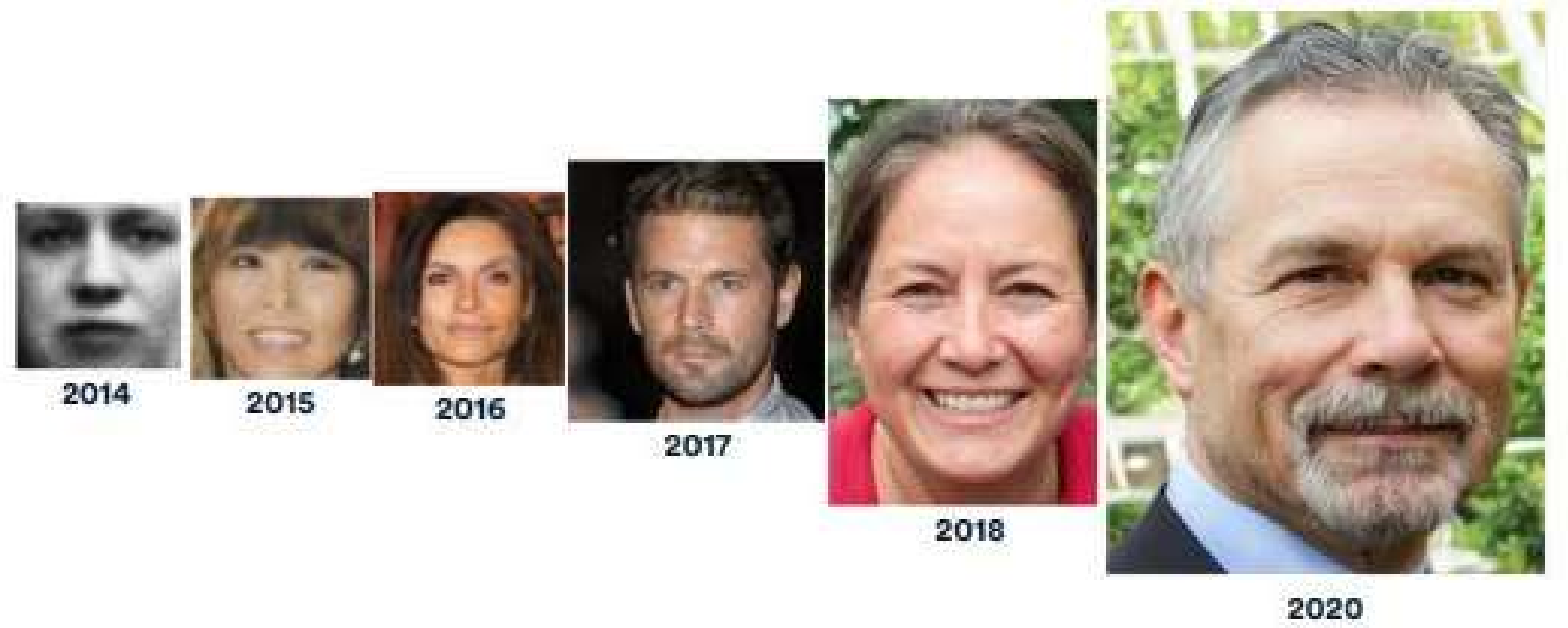
Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, Humphrey Shi

Outline

- History of Generative Models
- Versatile Diffusion
- Network Architecture
- Disentanglement of Style and Semantic
- Dual Context Blender
- Questions

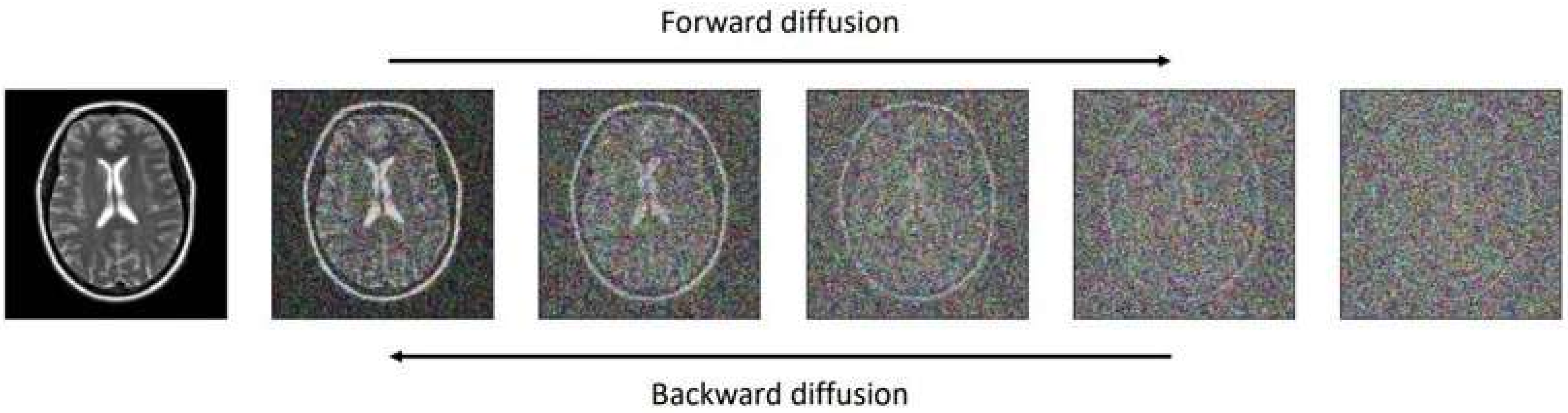
Generative Adversarial Networks

GAN focus on specific domains and specific tasks i.e. faces



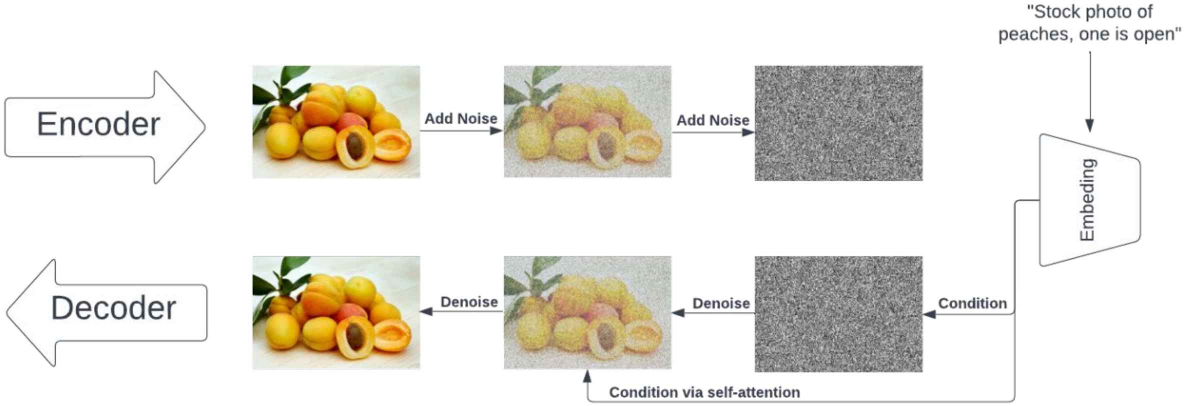
New Horizons: Diffusion Model

Likelihood based models that gradually restore image contents from gaussian corruptions.



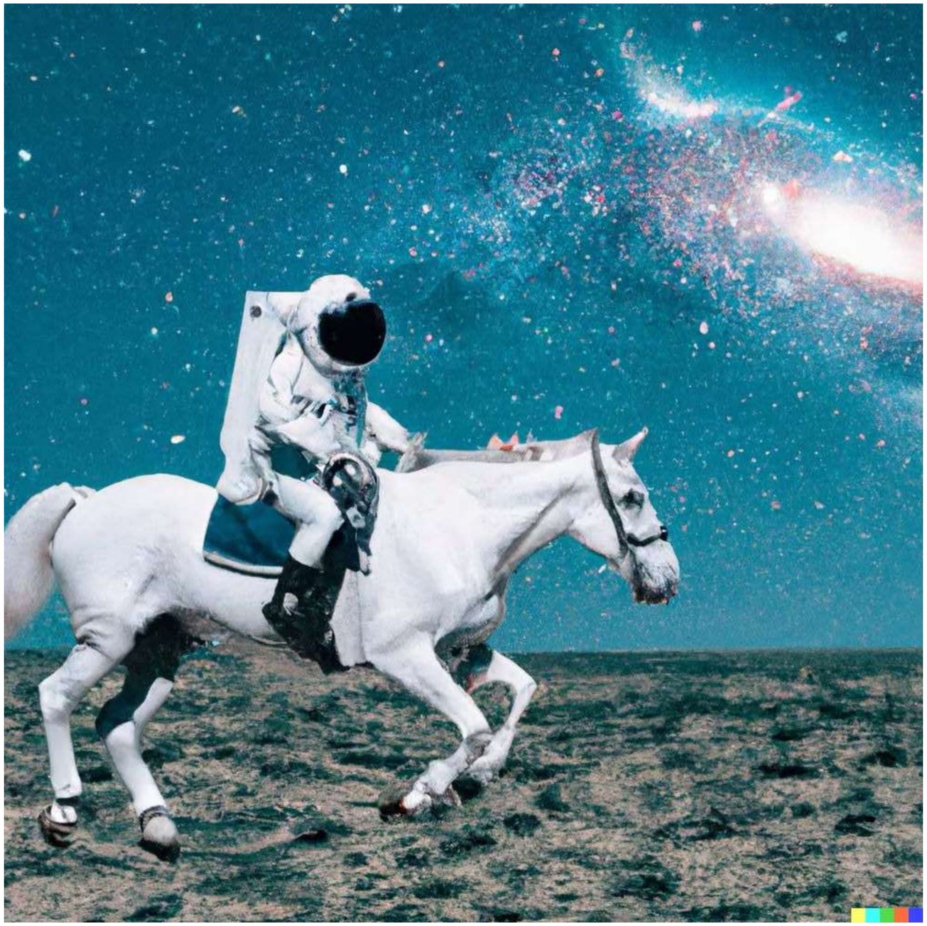
New Horizons: Diffusion Model

It has proven to be effective in bridge modalities and tasks.

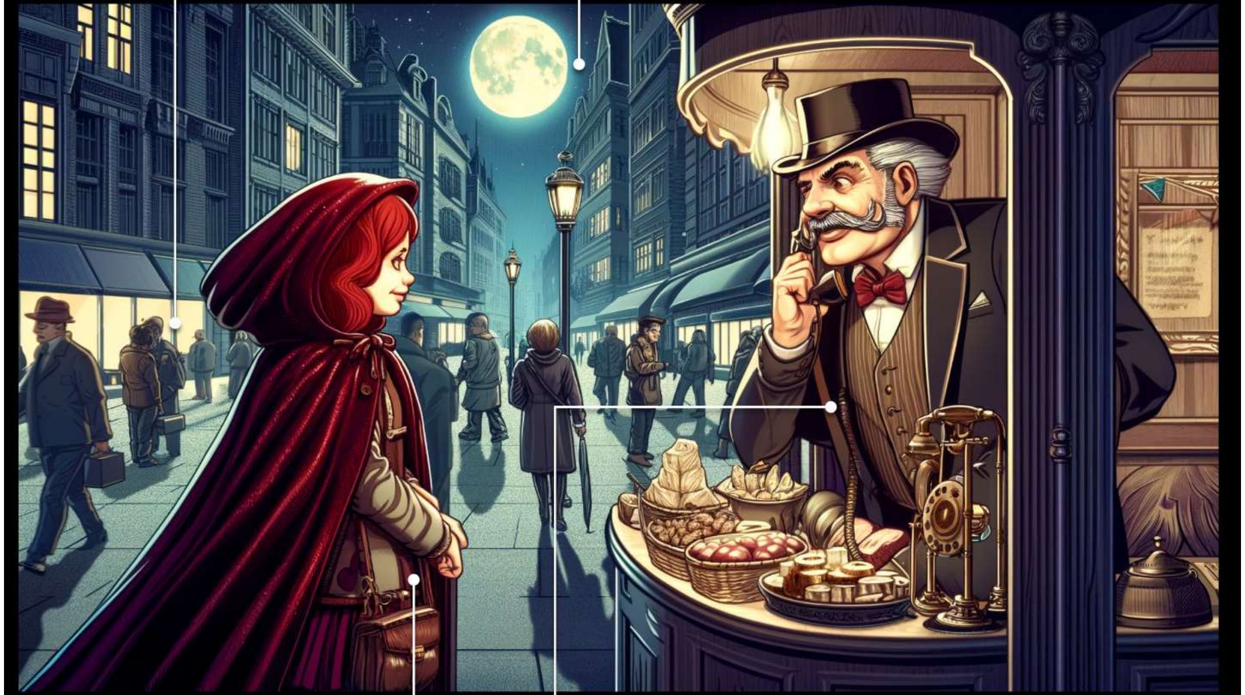

"Stock photo of peaches, one is open"
Encoder
Add Noise
Add Noise
Embeding
Decoder
Denoise
Denoise
Condition
Condition via self-attention

DALL-E2 (2021)

Input: An astronaut riding a horse in photorealistic style.



DALL-E3 (2023)



The sidewalks bustling with **pedestrians enjoying the nightlife.**

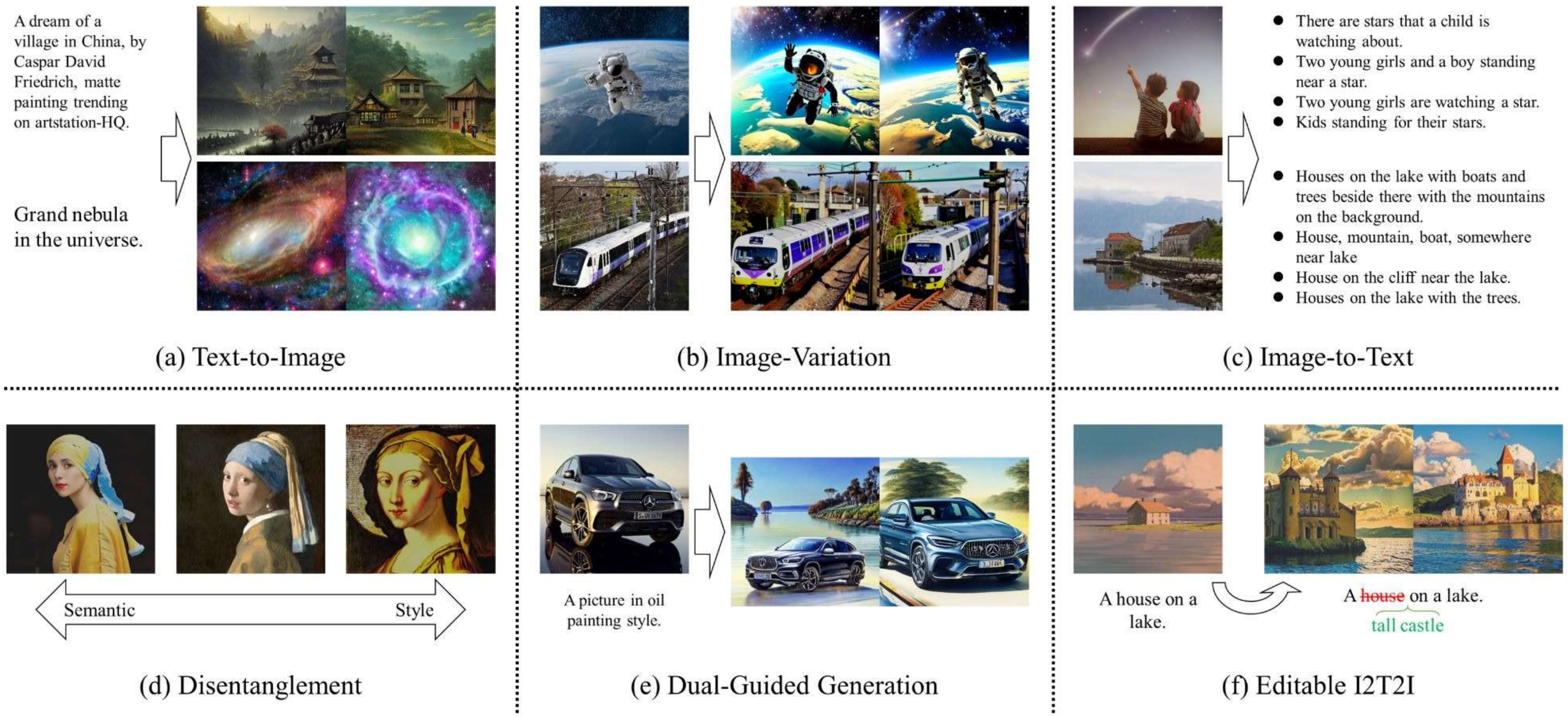
A bustling city street under the shine of a **full moon.**

At the corner stall, a **young woman** with fiery red hair, dressed in a signature velvet cloak, is **haggling with the grumpy old vendor.**

The grumpy vendor, a **tall, sophisticated man,** is wearing a sharp suit, sports a **noteworthy moustache** and is animatedly conversing on his **steampunk telephone.**

What’s Next?

Versatile Diffusion (2023)


A dream of a village in China, by Caspar David Friedrich, matte painting trending on artstation-HQ.
Grand nebula in the universe.
● There are stars that a child is watching about.
● Two young girls and a boy standing near a star.
● Two young girls are watching a star.
● Kids standing for their stars.
● Houses on the lake with boats and trees beside there with the mountains on the background.
● House, mountain, boat, somewhere near lake
● House on the cliff near the lake.
● Houses on the lake with the trees.
Semantic
Style
A picture in oil painting style.
A house on a lake.
A ~~house~~ on a lake.
tall castle

(a) Text-to-Image

(b) Image-Variation

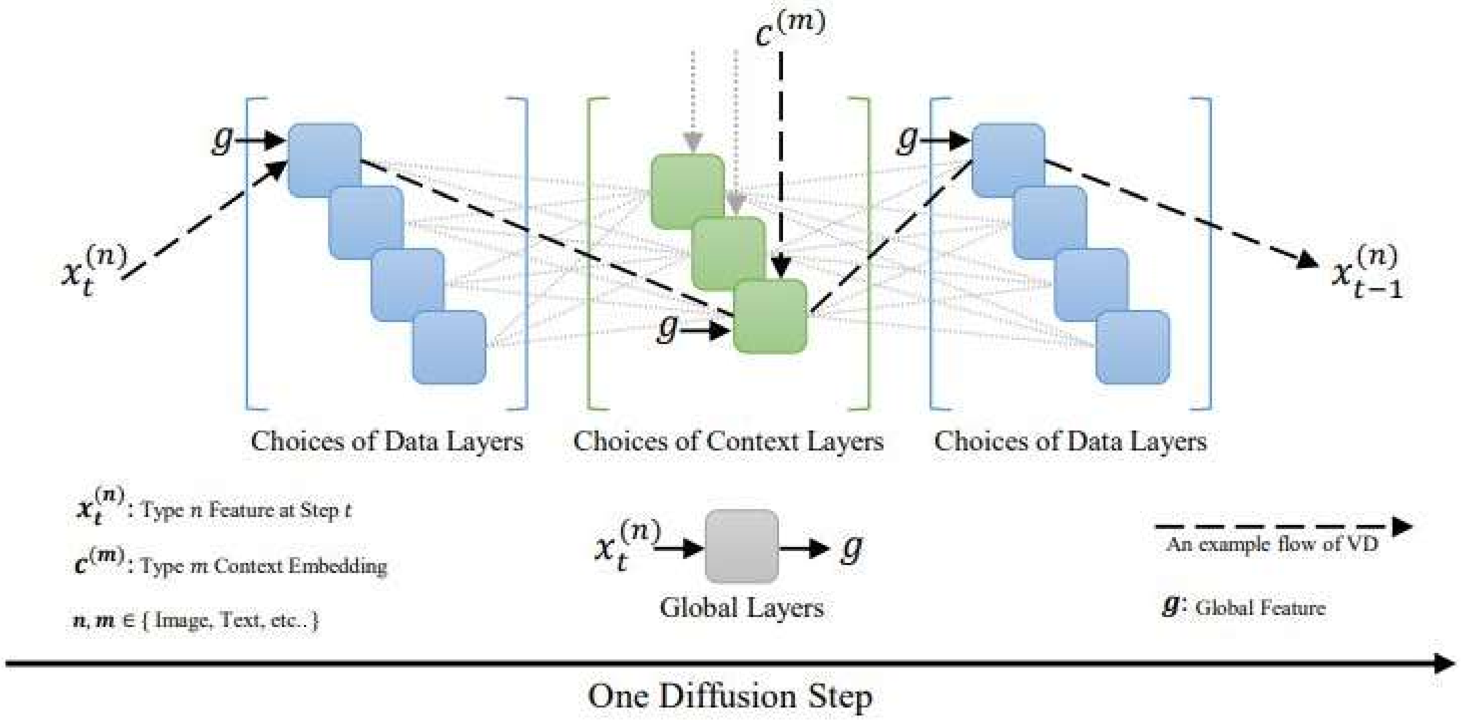
(c) Image-to-Text

(d) Disentanglement

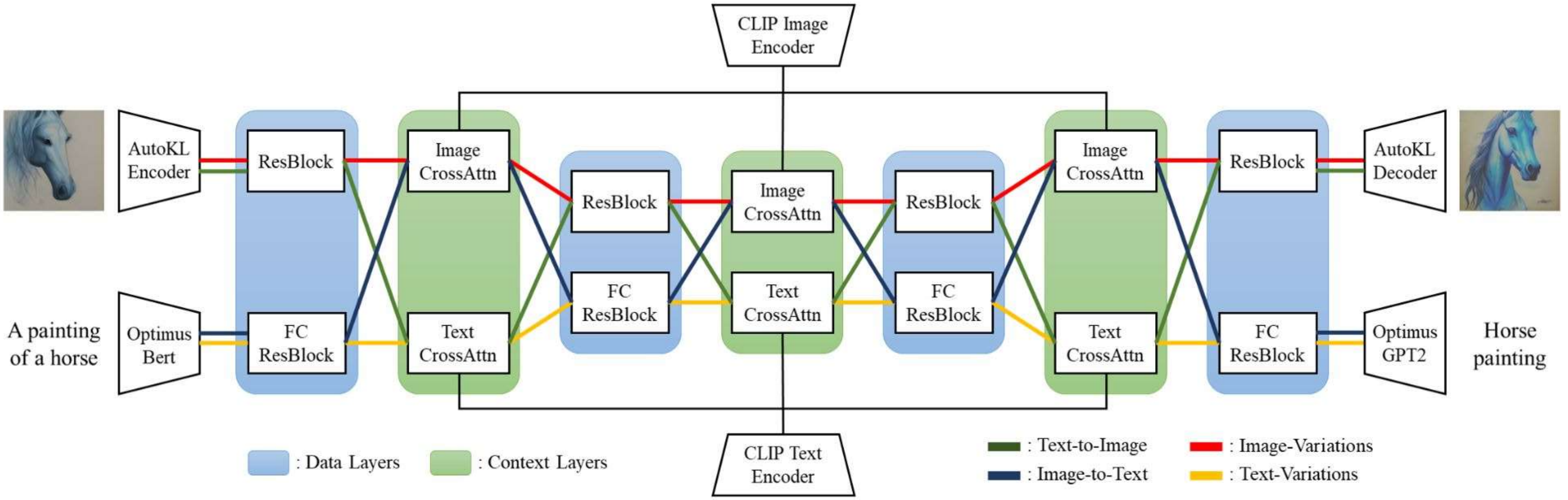
(e) Dual-Guided Generation

(f) Editable I2T2I

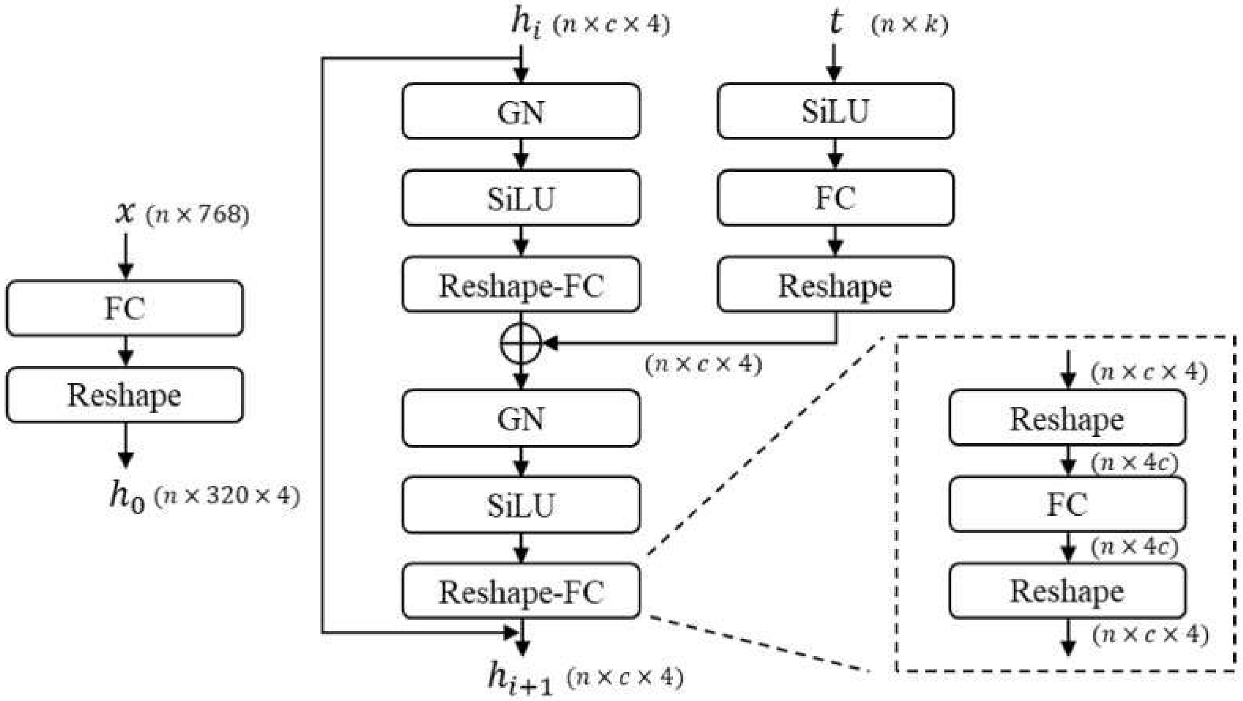
One Diffusion Step



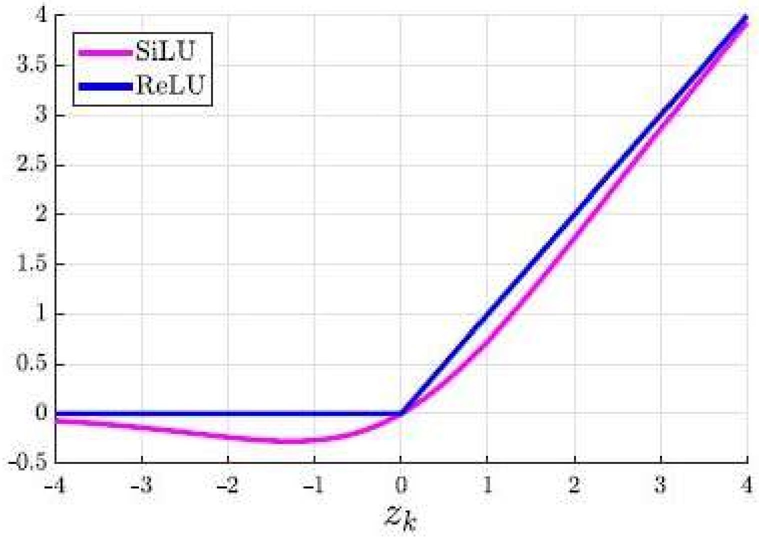
Network Structure



Text Data Layers



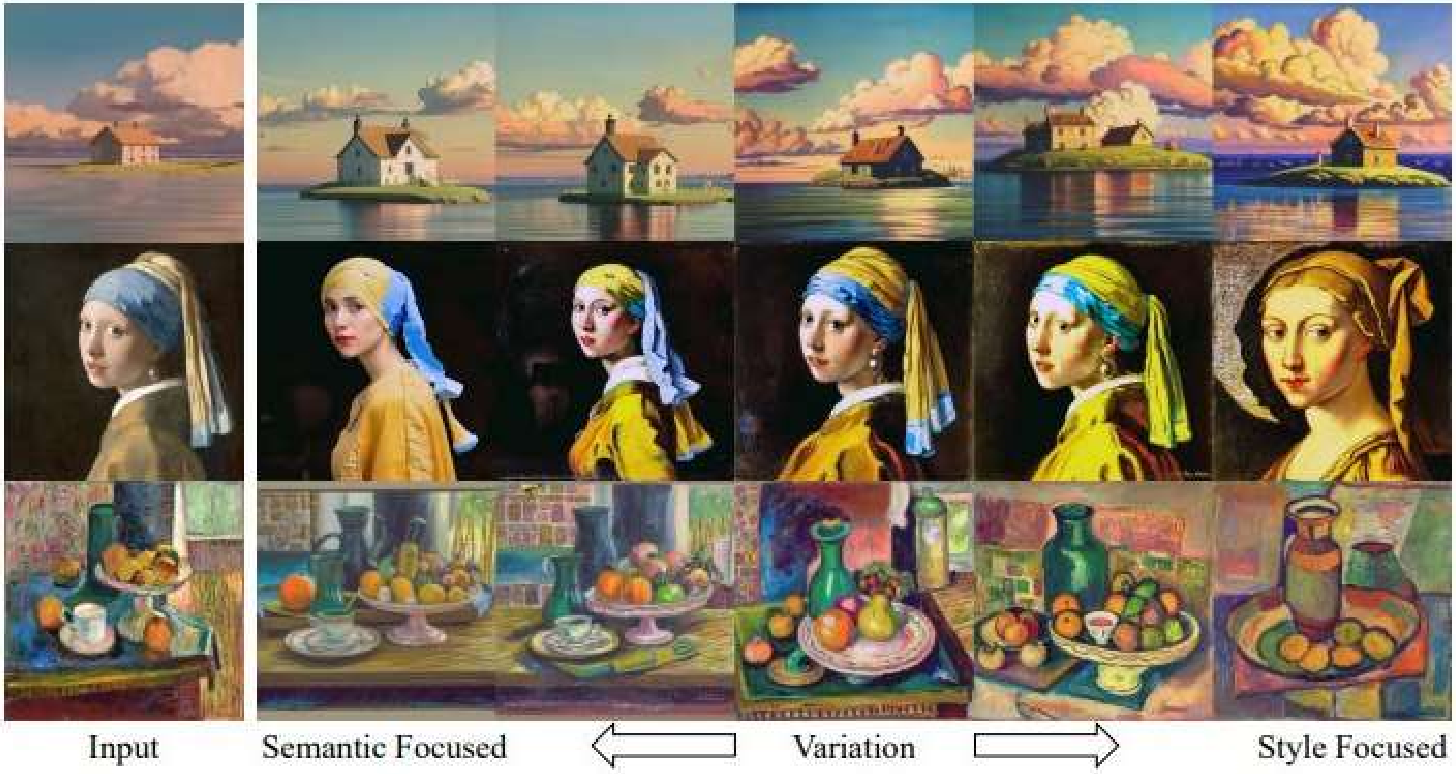
SiLu



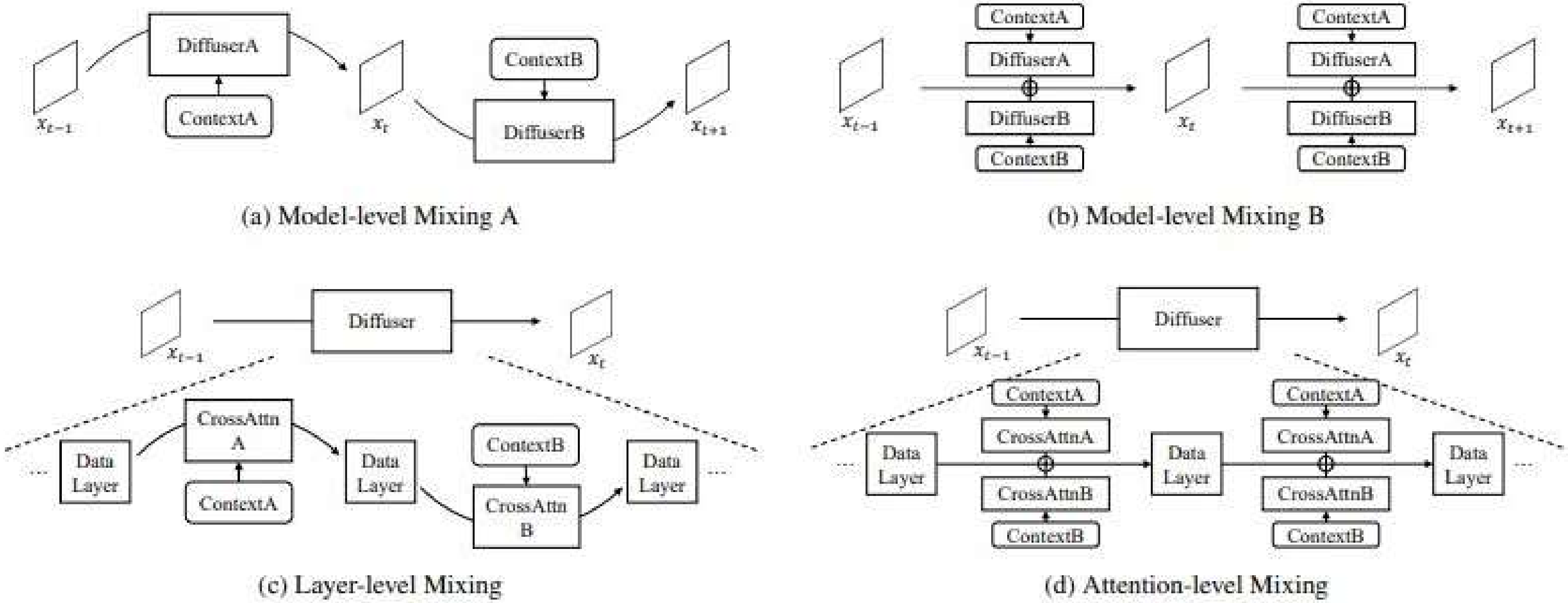
Training

- For each of the flows compute the variational weighted losses and do regular backpropagation.
- Update model weights when the gradient in all flows are accumulated.
- VD is trained progressively in three settings:
  - Single Flow: Image Variation
  - Dual Flow: Text-to-Image and Image-variation
  - Four Flow: All the tasks together

Disentanglement of Style and Semantic



Dual-context Blender



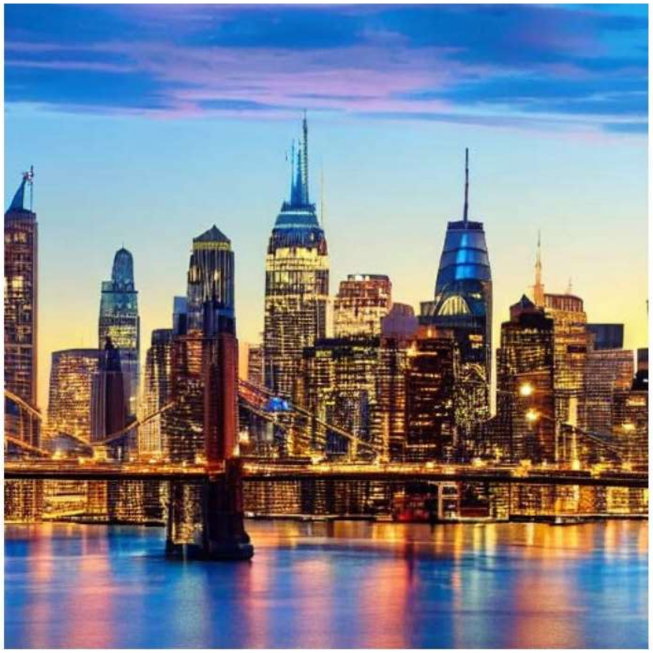
(a) Model-level Mixing A

(b) Model-level Mixing B

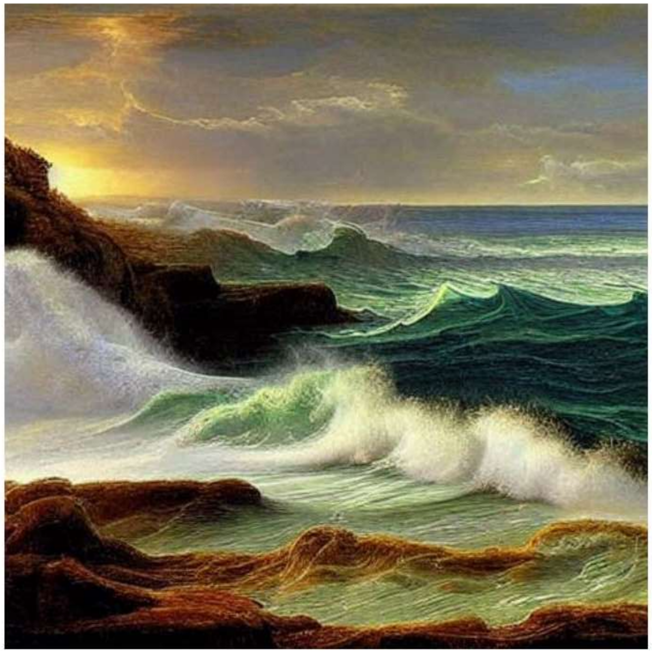
(c) Layer-level Mixing

(d) Attention-level Mixing

Results: Text to Image

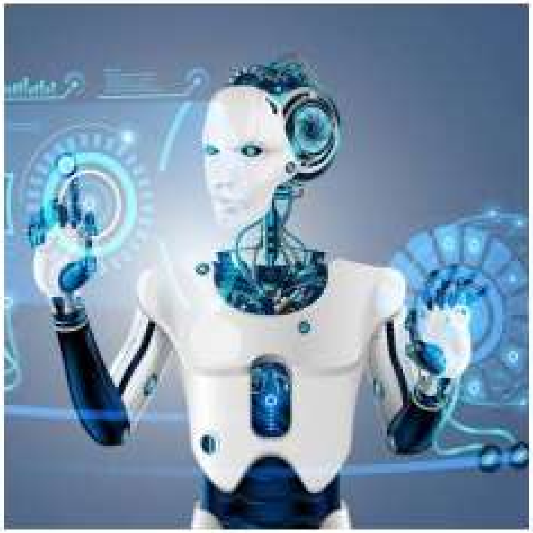


“A wonderful evening in New York City with a great view of Brooklyn Bridge and a magnificent city view of Manhattan, HD 8K”

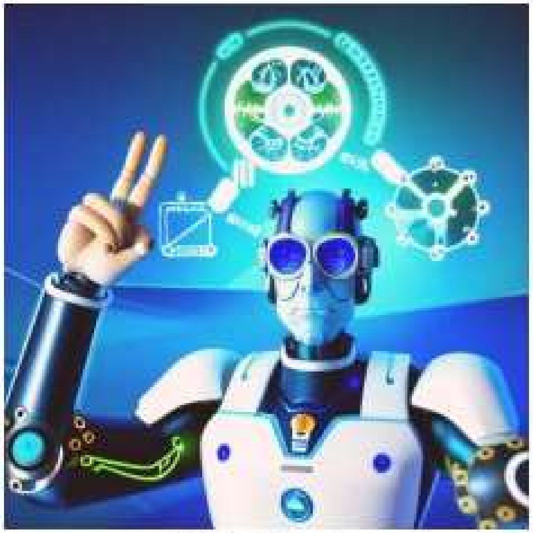


“A beautiful painting of waves crashing on a cliff by Thomas Cole”

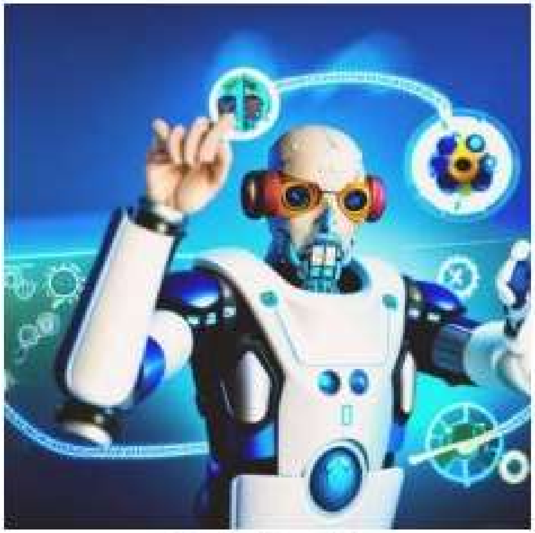
Results: Image Variation



Input

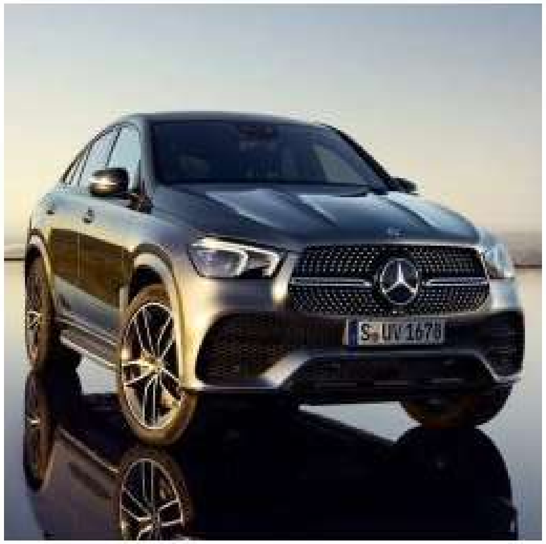


Variation #1

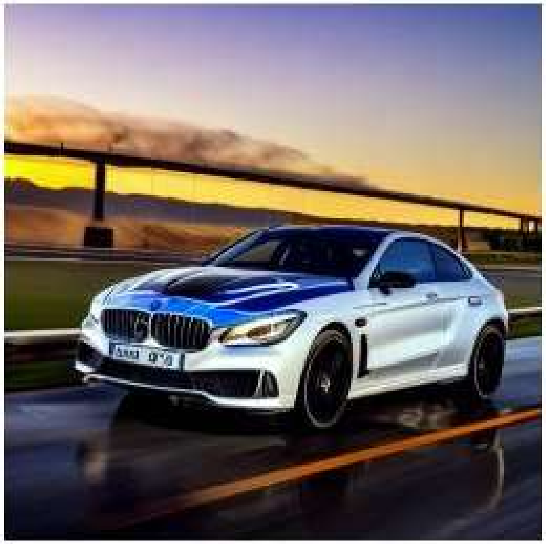


Variation #2

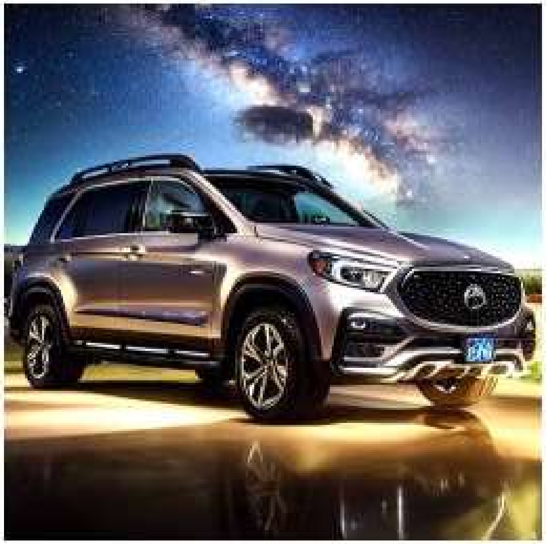
Results: Dual-Context Blender



Input



“100 mph”



“Traveling among the stars”

Limitations

- **Limited Latent Space**: Optimus VAE’s latent vector are 768 single dimension generated using Bert which might be inadequate for long text sentences. It is weak in understanding word locations and orders

- **Imperfect Text Data**: There is domain shift in Optimus VAE’s training data compare to VD’s training data making it difficult to reconstruct certain images.

Question 1

How does the authors disentangle and semantic using the VD?

Question 2

Give three example of basic tasks VD can achieve?