VLP for High-Resolution Visual Synthesis

Chenfei Wu
MSRA NCL Group
Overview

GAN

- from global latents (LocoGAN, SinGAN, Infinite-GAN, MS-PIE, TileGAN)
  - Repetitive

- from coordinate latents (ALIS, InfiniteGAN)
  - Non-repetitive
  - But rough transition

AR

- token by token (Taming Transformers)
  - Smooth transition between tokens
  - Rely on sketches
  - Fixed Size

NAR

- mask by mask (MaskGIT)
  - Smooth transition between masks

Diffusion

- Vague to Clear (Latent Diffusion)
  - Fixed Size

SR

- from low-resolution images (CogView, DALLE-2, Imagen)
  - Fixed Size

AR over AR

- NUWA-Infinity
  - Infinite Size
GAN-based Models

[1] from global latents (LocoGAN, SinGAN, Infinite-GAN, MS-PIE, TileGAN)

$z \sim N(0, I) \rightarrow w$

LocoGAN generates different patches with the same global latents, which limits the diversity and leads to repetitive results.
To address the repetition issue, ALIS generates different patches with different coordinate latents. However, this sometimes brings rough transition between patches when the coordinate latents are too different.
Auto-Regressive Models are naturally beneficial for generating smooth transition results. Taming Transformer views images as discrete tokens and generate a high-resolution image token by token, which considers the smooth transition between nearby tokens. As a result, it fails to model long-range consistency and has a slow inference speed.
MaskGIT generates masked tokens based on visible tokens inside a sliding window. The inference speed is improved by generating masked tokens in parallel in a fixed steps. Although it considers the smooth transition between the generated masked tokens, it still fails to model long-range consistency.
**Algorithm 1 Training**

1: repeat  
2: \( x_0 \sim q(x_0) \)  
3: \( t \sim \text{Uniform}([1, \ldots, T]) \)  
4: \( \epsilon \sim \mathcal{N}(0, I) \)  
5: Take gradient descent step on  
   \[ \nabla_{\theta} \| \epsilon - \epsilon(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, t) \|^2 \]  
6: until converged

**Algorithm 2 Sampling**

1: \( x_T \sim \mathcal{N}(0, I) \)  
2: for \( t = T, \ldots, 1 \) do  
3: \( z \sim \mathcal{N}(0, I) \) if \( t > 1 \), else \( z = 0 \)  
4: \( x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z \)  
5: end for  
6: return \( x_0 \)
DALL-E

Stage One

image

Image Encoder

Image Decoder

dVAE

Stage Two

BPE Encoder

Concatenate

Transformer

Sample Generation

Rerank using CLIP

caption

a small red block sitting on a large green block
Generate images from text

What do you want to see?

generate multiple three-dimensional geometric squares with four main colors of red, yellow, blue and green

Design multiple three-dimensional geometric squares with four main colors of red, yellow, blue and green

These results have been obtained using model 207003c from an ongoing training run.

Again!

Generate images from text

What do you want to see?

colors of red, yellow, blue and green, to have a sense of composition and a sense of depth in the space.

Design multiple three dimensional geometric squares with four main colors of red, yellow, blue and green, to have a sense of composition and a sense of depth in the space.

These results have been obtained using model 207003c from an ongoing training run.

Again!
[1] A unified multimodal pretrained model that can generate both images and videos
Task 1: Text-To-Image (T2I)

NÜWA achieves a significant high FID score of \textbf{12.9}, outperforms \textbf{27.1} of DALL-E from Open AI.
Task 2: Sketch-To-Image (S2I)

**NÜWA** generates a diverse realistic results, even the reflection of the bus window is clearly visible.
Task 3: Image Completion (I2I)

**NÜWA** shows strong imagination and can even complete an image with only 5% tokens.
Task 4:
Image Manipulation (I2I)

NÜWA can add, remove or change any parts of an image to whatever you want.
Task 5: Text to Video (T2V)

NÜWA can generate unseen videos like “Running on the sea.”
Task 6: Sketch to Video (S2V)

**NÜWA** generates temporally consistent open-domain videos.
Task 7:
Video Prediction (V2V)

**NÜWA** predicts the future of an image.
Task 8: Video Manipulation (TV2V)

Raw Video: Manipulation1: The diver is swimming to the surface. Manipulation2: The diver is swimming to the bottom. Manipulation3: The diver is swimming to the sky.

**NÜWA** can even manipulate video futures with the control of language.
1. A prior that generates a CLIP image embedding given a text.
2. A decoder that generates an image conditioned on the image embedding.
Sprouts in the shape of text ‘Imagen’ coming out of a fairytale book.
Figure 3: Comparison between Imagen and DALL-E 2 [54], GLIDE [41], VQ-GAN+CLIP [12] and Latent Diffusion [57] on DrawBench: User preference rates (with 95% confidence intervals) for image-text alignment and image fidelity.

Figure 4: Summary of some of the critical findings of Imagen with pareto curves sweeping over different guidance values. See Appendix D for more details.
Super-Resolution based Models

- DALL-E
  - (256x256 images)
  - Nov 5, 2021

- DALL-E 2
  - (1024x1024 images)
  - April 7, 2022

- Imagen
  - (1024x1024 images)
  - May 17, 2022

- NUWA
  - (256x256 images and videos)
  - June 2022

- NUWA-Infinity
  - (infinite-resolution images and videos)
Why Infinite Resolution important?

Higher resolution implies not only more details, but also **wider views**.
Traditional **Two-Stage Solution**
*(CogView, DALLE-2, Imagen)*

**Stage 1:**
Coconut trees beside the sea.

**Stage 2:**
Fix-sized, limited Resolution.
Supports images only.

---

Our **End-to-End Solution**
*(NUWA-Infinity)*

Coconut trees beside the sea.

**Context-aware Rendering**

**Patch Context Pool**

**Patch Feedback**

Arbitrary-sized, Infinite Resolution.
Supports both images and videos.
Our Solution: NUWA-Infinity

Patch Context Pool

Context: The green mountain reflects on the green lake.

Prompt: The green mountain reflects on the green lake.
White clouds floated in the sky.
PART 1: UIG Task

None → Unconditional Image Generation (UIG)
[1] Unconditional Image Generation Task on Landscape dataset
[2] Unconditional Image Generation Task on Design dataset
Unconditional Image Generation Task on “Riverside Scene at Qingming Festival” *(One-shot Leaning from Scratch)*
[3] Unconditional Image Generation task on “Riverside Scene at Qingming Festival” (2048x6114)
PART 2: IE Task

Image-Extension (IE)
[1] Image-Extension Task on Landscape dataset
[1] Image-Extension Task on Windows Wallpaper (Zero-shot)
[1] Image-Extension Task on “The Starry Night” (Zero-Shot)
[1] Image-Extension Task on “The Gleaners” (Zero-Shot)
PART 3: T2I Task

White clouds floated in the sky.

Text-to-Image (T2I)
[1] Text-to-Image Task on Landscape dataset

Input Text: a road that is going down a hill.

DALL-E (256x256)

NUWA-Infinity (1024x2048)
[1] Text-to-Image Task on Landscape dataset

a snowy forest with trees covered in snow.

a path in a forest with tall trees.

a snowy forest with trees covered in snow.

a path through a forest with fog and trees.
a tree with no leaves on it under a cloudy sky.
[2] Text-to-Image Task on Design dataset

Input Color: Black, Green, Yellow.
PART 4: I2V Task
[1] Image-to-Video Task on Pexels dataset
[1] Image-to-Video Task on Pexels dataset
[2] Image-to-Video Task on Design dataset
PART 5: T2V Task

White clouds floated in the sky.

Text-to-Video (T2V)
Input Story:

1) Fred wants to find treasure.
2) He drives a car. He passes some trees.
3) He stops at a cave. He goes inside.
4) Fred finds the treasure.
5) He is very happy.

Output Cartoon:
From Idea to Design

IDEA

Input

AI

Output

Ideal design
(Object + Style + Color)
Case 1

"I want to design a concept map in Office PowerPoint. There is a blue 3D squares in the background. The overall style should have a sense of three-dimensional composition, and the space should reflect the sense of depth."

Current design: Search or Sketch or Software
Case 1

blue
3D squares
sense of depth

Text → Image Generation → Expand

Stage 1: Text-to-Image
Stage 2: Image Extension
Case 1
Case 2

“I designed a cover for the magazine of MSRA Matrix, I want to animate it.”
Case 3

"I want to put the car ads on Bing, but need to make it adaptive on web pages."
Case 3
Case 4

I am in a Microsoft Teams Meeting. I want to animate my Teams Background wallpaper.
Case 4
Other Scenarios in future

Restoration  Gaming  Story board  Calligraphy  Disability