Alignments in Text-to-Image Generation

Zhengyuan Yang

Microsoft
Alignments in Text-to-Image Generation

Q1: how to learn image representations?
Q2: how to extend vision models with more flexible, promptable interfaces?
Q3: how to do image generation?
Q4: how to train multimodal LLM?
Q5: how to chain vision experts with LLM?

Image Encoder

LLM for language understanding and generation

Image Generation

Produce visual data: aligned with human intentions

General-purpose interface

Consume visual data

(a dog is running through the grass)
Text-to-Image Generation

- Text-to-image (T2I)
- Aligning with human intentions

Text prompt: a yellow fire hydrant with a cartoon face drawn on it.

StackGAN  StyleGAN  CogView  VQ-Diff.  GLIDE  Make-A-Scene  Imagen  NUWA-Infinity  eDiff-I Bing image creator  IF


AttnGAN  DALL-E  NUWA  Latent-Diff.  MaskGiT  DALL-E 2  CogView2  Parti  SD  MUSE GigaGAN Midjourney
Alignments in Text-to-Image Generation

Controllable generation

- Image-level: a yellow fire hydrant with a cartoon face drawn on it.
- a truck is parked next to a trash can.
- a red truck is parked in a parking lot.
- a yellow fire hydrant with a face on it and black eyes.

Editing

- “Swap sunflowers with roses”
- “Add fireworks to the sky”

Better following prompts

- “A horse and a dog”
- “A painting of an elephant with glasses”

Concept customization

- Input images
- in the Acropolis
- in a doghouse
- in a bucket

Image credit: ReCo, InstructPix2Pix, Attend-and-Excite, DreamBooth
Agenda

- **Text-to-image (T2I) basics**
- **Aligning human intentions in T2I generation**
  - Controllable generation
  - Editing
  - Better following prompts
  - Concept customization
- **Summary and discussion**
Text-to-Image Basics

- GAN
- Auto-regressive
- Non-AR Transformer
- Diffusion

Text prompt: a yellow fire hydrant with a cartoon face drawn on it.
Text-to-Image Basics

Generative Adversarial Networks (GAN)

Auto-regressive (AR)

Non-AR Transformer

Diffusion

Stable Diffusion (SD) Basics

- SD overview
  - Variational autoencoder (VAE)
  - Condition encoder
  - Conditional denoising U-Net
Stable Diffusion (SD) Basics

- Inference flow
  - Variational autoencoder (VAE)
  - Condition encoder
  - Conditional denoising U-Net

Image credit: https://huggingface.co/blog/stable_diffusion
Stable Diffusion (SD) Basics

- Zooming into conditional U-Net:
  How text condition operates on image?
  - Image-text cross attention

\[
\begin{align*}
\text{Latent: } & b \times 320 \times 64 \times 64 \\
\text{text: } & b \times 77 \times 768 \\
\text{t: } & b \times 1280 \\
\text{Latent': } & b \times 320 \times 32 \times 32
\end{align*}
\]

Q: \( \text{latent} + \text{duplicate(linear(t))} \)

\[
=> b \times 4096 \times 320
\]

K, V: \text{text} \n
\[
=> b \times 77 \times 768
\]

Attention(Q, K, V) = softmax \( \frac{QK^T}{\sqrt{d}} \) \cdot V, with

\[
Q = W_Q^{(i)} \cdot \varphi_i(z_t), \quad K = W_K^{(i)} \cdot \tau_\theta(y), \quad V = W_V^{(i)} \cdot \tau_\theta(y).
\]

Image-text attention map of HW*77
Agenda

• Text-to-image (T2I) basics

• **Aligning human intentions in T2I generation**
  – Controllable generation
  – Editing
  – Better following prompts
  – Concept customization

• Summary and discussion
Controllable Generation

• Text+layout/box: localized description control

• Text+dense control (e.g., mask, edge, scribble, etc.)

• Inference-time guidance

Image credit: “ReCo: Region-Controlled Text-to-Image Generation”

[1] GLIGEN: Open-Set Grounded Text-to-Image Generation
[2] ReCo: Region-Controlled Text-to-Image Generation
[4] Adding Conditional Control to Text-to-Image Diffusion Models
[5] Composer: Creative and Controllable Image Synthesis with Composable Conditions
[9] Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models
[10] UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild
[12] Training-Free Layout Control with Cross-Attention Guidance
ReCo: Region-Controlled T2I Generation

Text: global image text description

Text: grounded global and regional descriptions (Grounded Region-Controlled texts)

Image credit: “ReCo: Region-Controlled Text-to-Image Generation”
ReCo: Region-Controlled T2I Generation

- **Input sequence expansion**: box tokens
- **Grounded**: box tokens operate on the text to follow
- **Finetune T2I to understand box tokens**

Image credit: “ReCo: Region-Controlled Text-to-Image Generation”
GLIGEN: Open-Set Grounded T2I Generation

- Grounding tokens: grounded text entity + spatial location
- Gated self-attention layer with original layers frozen

\[ \mathbf{v} = \mathbf{v} + \beta \cdot \tanh(\gamma) \cdot \text{TS}(\text{SelfAttn}([\mathbf{v}, h^e])) \]
GLIGEN: Open-Set Grounded T2I Generation

- Bounding box grounding
- Keypoint grounding
- Spatially-aligned dense conditions

Image credit: “GLIGEN: Open-Set Grounded Text-to-Image Generation”
Text+Dense Control

- Dense conditions:
  - Canny Edge
  - Hough Line
  - HED Boundary
  - User Sketching
  - Human Pose
  - Semantic Segmentation
  - Depth
  - Normal Maps
  - Cartoon Line Drawing

Image credit: “Adding Conditional Control to Text-to-Image Diffusion Models“ (ControlNet)
Uni-ControlNet, UniControl

- Unified models for different conditions
- Condition composition

Image credit: “Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models”
**Inference-time guidance**

- **Universal Guidance for Diffusion Models:** extending classifier guidance [1] to accept any general guidance function

\[ \hat{e}_\theta(z_t, t) = e_\theta(z_t, t) + s(t) \cdot \nabla_z \ell(c, f(\hat{z}_0)) \]

E.g., detection:

- Anchor classification, bounding box regression, and region label classification loss
- Box and class labels
- Faster-RCNN
- Predicted “noisy” clean image

---


Image credit: “Universal Guidance for Diffusion Models”
Editing

- Latents spatial blend
- Image-text attention edit
- Edit instruction
- External models

[1] Blended Diffusion for Text-driven Editing of Natural Images
[6] Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting
[7] iEdit: Localised Text-guided Image Editing with Weak Supervision
[8] EDICT: Exact Diffusion Inversion via Coupled Transformations
[9] Prompt-to-Prompt Image Editing with Cross Attention Control
[10] Imagic: Text-Based Real Image Editing with Diffusion Models
[12] InstructPix2Pix Learning to Follow Image Editing Instructions
[15] Instruct-X-Decoder
[16] Grounded-SAM Inpainting
[17] Inpaint Anything

Image credit: “InstructPix2Pix: Learning to Follow Image Editing Instructions”
Latents Spatial Blend

• Spatial editing with mask
• Image, text prompt, user input or segmented mask

\[
z_t \leftarrow z_{fg} \odot m_{\text{latent}} + z_{bg} \odot (1 - m_{\text{latent}})
\]

from text original bg image
Image-text Attention Edit

- Edit generated images
- Manipulate image-text cross-attention map
- Word swap, adding new phrase, attention re-weighting

Image credit: “Prompt-to-Prompt Image Editing with Cross Attention Control”
Image-text Attention Edit

- Maintaining two sets of cross-attention maps for edit:
  Original prompt: $M_t$  Edited prompt: $M^*_t$

Goal

Image credit: “Prompt-to-Prompt Image Editing with Cross Attention Control”
Image-text Attention Edit

Word Swap

Adding a New Phrase

Attention Re-weighting

Image credit: “Prompt-to-Prompt Image Editing with Cross Attention Control”
Imagic

• Generated => natural image edits
• E.g., different dogs
Imagic

- Obtain original text

\[ L(x, e, \theta) = \mathbb{E}_{t,e} \left[ \| e - f(x_t, t, e) \|_2^2 \right] \]

\[ \bar{e} = \eta \cdot e_{tgt} + (1 - \eta) \cdot e_{opt} \]

Image credit: “Imagic: Text-Based Real Image Editing with Diffusion Models”
InstructPix2Pix

- Obtain original text => Instruction-style text
  “a bird standing”, “a bird spreading wings” => “have wings spread”
Editing Systems with External Models

- Segmentation
- LLM

Better Following Prompts

• Test-time latents
• Test-time attention
• Alignment tuning


Image credit: “Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models”
StructureDiffusion

Stable Diffusion

Ours

A red car and a white sheep.
Attribute leakage

A brown bench sits in front of an old white building
Interchanged attributes

A blue backpack and a brown elephant
Missing objects

StructureDiffusion

- Manipulating values in cross-attention based on linguistic parsing tree to enforce language structure
- Look at all noun phrases

Q: latent + duplicate(linear(t))
=> b*4096*320

K, V: text => b*77*768

Attention(Q, K, V) = softmax \( \frac{QK^T}{\sqrt{d}} \) \cdot V, with

\[ Q = W_Q^{(i)} \cdot \varphi_i(z_i), \quad K = W_K^{(i)} \cdot \tau_\theta(y), \quad V = W_V^{(i)} \cdot \tau_\theta(y) . \]
Attend-and-Excite

- Enhance the maximal attention for the most neglected subject token
- Updates the latent with attention loss

Image credit: “Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models”
Attend-and-Excite

“A horse and a dog”

“A painting of an elephant with glasses”

“A playful kitten chasing a butterfly in a wildflower meadow”

Stable Diffusion

+Attend-and-Excite

Image credit: “Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models”
DDPO

- RL for optimizing diffusion models on different downstream objectives

Image credit: “Training Diffusion Models with Reinforcement Learning”
DDPO

- VLM similarity reward to improve image-prompt alignment
Concept Customization

- Single-concept customization
- Multi-concept customization
- Without test-time finetuning

[8] BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing
[9] Face0: Instantaneously Conditioning a Text-to-Image Model on a Face
[10] FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention
[12] Re-Imagen: Retrieval-Augmented Text-to-Image Generator
[14] Subject-driven text-to-image generation via apprenticeship learning
Single-Concept Customization

Image credit: “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation”
Single-Concept Customization

- Tuning unique identifier [V] for customized subject
- Originally generated samples to alleviate forgetting

Image credit: “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation”
Multi-Concept Customization

- Multi-concept customization $[V1]$, $[V2]$, ... from single image or multiple images

Image credit: “Multi-Concept Customization of Text-to-Image Diffusion”
“Break-A-Scene: Extracting Multiple Concepts from a Single Image”
Without Test-Time Finetuning

- Retrieve-augmented/ In-context generation
- Similar customization, but w/o test-time finetuning

Image credit: “Subject-driven text-to-image generation via apprenticeship learning”
“InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning”
Agenda

• Text-to-image (T2I) basics

• Aligning human intentions in T2I generation
  – Controllable generation
  – Editing
  – Better following prompts
  – Concept customization

• Summary and discussion
Discussion

Open-source v.s. Closed-source

Consuming and producing visual data: Understanding (I2T) and generation (T2I) loop

- Stable Diffusion
- DeepFloyd AI
- Parti
- Imagen
- DALL-E 2
- Adobe Firefly

a close up of a dog near a bowl.
a silver bowl with water in it.
a dog with his tongue out next to a bowl of water.
a close up of a dog near a bowl.
a silver bowl with water in it.
a dog with his tongue out next to a bowl of water.
Thank you!