Controlling Text-to-Image Diffusion Models

BigMAC Workshop @ ICCV’23
Sayak Paul
Hugging Face 😊
💡 **Disclaimer**: This talk is NOT an exhaustive overview of all possible methods.
Era of text-to-image diffusion models!

“A transparent sculpture of a duck made out of glass.”

“Panda mad scientist mixing sparkling chemicals, digital art.”

“Astronaut in a jungle, cold color palette, muted colors, detailed, 8k”
Diffusion models in a jiffy

What happens when you refine a noise vector to become a realistic image?

https://nvlabs.github.io/denoising-diffusion-gan/
Diffusion models in a jiffy

When you “condition” the denoising process with text:

DALL-E 2 prompt: “A photo of a white fur monster standing in a purple room”
Extracting visual connections from textual concepts

Concept discovery in text-to-image diffusion models:

(a) Concept decomposition with Conceptor

(b) Single-image decomposition with Conceptor

Conceptor; Chefer et al., 2023.
We’ll focus on “latent-space” diffusion models throughout this talk. More specifically, the Stable Diffusion family.
Limitations and solutions
Part I
Subject-driven generation for personalization

- Render concepts/subjects *new* to the model in interesting contexts.
- Introduce *personalization*.

https://dreambooth.github.io/

DreamBooth; Ruiz et al., 2022.
Subject-driven generation for personalization

Embedding a new subject in the output domain of the (pre-trained) model: 
DreamBooth!

https://dreambooth.github.io/

[Image: Diagram illustrating the process of embedding a new subject in the output domain of the (pre-trained) model: DreamBooth.]

DreamBooth; Ruiz et al., 2022.
Subject-driven generation for personalization

Without the loss of generality, let:

- \( \mathbf{x} \): original image
- \( \mathbf{\epsilon} \): noise; \( \mathbf{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \)
- \( t \): diffusion process time; \( t \sim \mathcal{U}([0, 1]) \)
- \( \alpha_t, \sigma_t, \mathbf{w}_t \): terms controlling noise schedule and sample quality
- \( \mathbf{c} \): conditioning vector (prompt embeddings, for example)
- \( \hat{\mathbf{x}}_\theta \): diffusion model to be learned

**Training**

\[
\mathbb{E}_{\mathbf{x}, \mathbf{c}, \mathbf{\epsilon}, t} \left[ w_t \left\| \hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \mathbf{\epsilon}, \mathbf{c}) - \mathbf{x} \right\|_2^2 \right]
\]

**Inference**

\[
\mathbf{x}_{\text{gen}} = \hat{\mathbf{x}}_\theta(\mathbf{\epsilon}, \mathbf{c})
\]
Subject-driven generation for personalization

Prior-preservation loss to preserve the class-specific semantic prior:

\[
\mathbb{E}_{x,c,\epsilon,\epsilon',t} \left[ w_t \| \hat{x}_\theta (\alpha_t x + \sigma_t \epsilon, c) - x \|^2_2 + \lambda w_t' \| \hat{x}_\theta (\alpha_t' x_{pr} + \sigma_t' \epsilon', c_{pr}) - x_{pr} \|^2_2 \right]
\]

DreamBooth; Ruiz et al., 2022.
One framework, multiple use cases

General subject-driven generation

Input images

A [V] dog in the Versailles hall of mirrors
A [V] dog in the gardens of Versailles
A [V] dog in Coachella
A [V] dog in mountain Fuji
A [V] dog with Eiffel Tower in the background

DreamBooth; Ruiz et al., 2022.
One framework, multiple use cases

Art rendition

DreamBooth; Ruiz et al., 2022.
One framework, multiple use cases

Property modification

Hybrids (“A cross of a [V] dog and a [target species]”)

Input  Bear  Panda  Koala  Lion  Hippo

DreamBooth; Ruiz et al., 2022.
Pushing the extremes with DreamBooth
Further reads

- **BLIP-Diffusion**; Li et al., 2023 (zero-shot subject-driven generation).
- **Custom Diffusion**; Kumari et al., 2022.
- **Pivotal Tuning**; Roich et al., 2021 (in SD context it’s Textual Inversion + DreamBooth).
Limitations and solutions
Part II
Going beyond text conditioning

What if we wanted to condition the generation process on a pose image along with language supervision?

"Darth Vader dancing in a desert"
Going beyond text conditioning - ControlNets

\[ \mathbb{E}_{\mathbf{x}, \mathbf{c}_t, \mathbf{c}_f, \mathbf{t}, \mathbf{c}_t} \left[ w_t \left\| \hat{x}_\theta (\alpha_t \mathbf{x} + \sigma_t \mathbf{c}, \mathbf{c}_t, \mathbf{c}_f) - \mathbf{x} \right\|_2^2 \right] \]

Image-space conditioning vector for ControlNet

(a) Before
(b) After

ControlNet; Zhang et al., 2023.
ControlNets - a powerful framework to inject additional control

ControlNet; Zhang et al., 2023.
Or shall I say **controls**?

"a giant standing in a fantasy landscape, best quality"
Further reads

- **T2I-Adapters**: Mou et al., 2023.
- **IP-Adapters**: Ye et al., 2023.
- **InstructPix2Pix**: Brooks et al., 2022.
Limitations and solutions
Part III
Catastrophic neglect & incorrect attribute binding

“A yellow bowl and a blue cat”
Neglects one or more objects in the generation.

“A yellow bow and a brown bench”
Fails to properly bind attributes to objects.

Attend and Excite; Chefer et al., 2023.
y tho?
Going to steal a couple of slides from Hila Chefer here.
Why Does the Model Fail?

DDPM process:

Given an input text prompt, the DDPM gradually denoises a pure noise latent to obtain the output image.
Cross Attention
(timestep $t$)
Cross Attention
(timestep t)

Image Latent Features
(P=16)

Text Embedding

Attend and Excite; Chefer et al., 2023.
Cross Attention (timestep $t$)

$A_+[i,n] = \text{presence of the token } n \text{ in patch } i$

Attend and Excite; Chefer et al., 2023.
Cross Attention (timestep t)

Image Latent Features (P=16)

Text Embedding

$p^2 \rightarrow p^2$

$Q$ \rightarrow $K$

$A_+ = Q \times K$

Reshape

$A_+ \rightarrow A_+^1, A_+^2, A_+^3, A_+^4, A_+^5$

Attend and Excite; Chefer et al., 2023.

Attend and Excite; Chefer et al., 2023.
**Problem:** The crown gets low attention values for all patches.

*Attend and Excite; Chefer et al., 2023.*
Generative semantic nursing

We want to:

● Encourage the model to better consider the semantic information passed from the input text prompt.
● Ensure all tokens are attended to by some image patch meaningfully.
How can we fix this?

💡 **Intuition**: a generated subject should have an image patch that significantly attends to the subject’s token.

💡 **Idea**: strengthen the activation of the most neglected token.

*Loss:* $L = \max(L_2, L_5)$

*Update:* $z_+ = z_+ - \alpha \nabla_{z_+} L$

How close are we to having a strong patch?

💡 **Idea**: strengthen the activation of the *most neglected* token.
Putting It All Together

Attend to and Excite all subject tokens!

Attend and Excite; Chefer et al., 2023.
Results

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

Stable Diffusion

Attend-and-Excite

Attend and Excite; Chefer et al., 2023.
Results

“A grizzly bear catching a salmon in a crystal clear river surrounded by a forest”

Stable Diffusion

Attend-and-Excite

Attend and Excite; Chefer et al., 2023.
Notable mentions

Controlling semantic attributes (training-free):

- Semantic Guidance; Brack et al., 2023.
- LEDITS; Tsaban et al., 2023.

Controlling using “rich-text” (training-free):

- Expressive Text-to-Image Generation with Rich Text; Ge et al., 2023.

Improving discriminative performance:

- Synthetic Data from Diffusion Models Improves ImageNet Classification; Azizi et al., 2023.
**IF prompt:** A cute panda standing amidst a mountain and holding a placard saying “Thank you!”