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.”
Imagen

“panda mad scientist mixing sparkling chemicals, digital art.”
DALL-E 2

“Astronaut in a jungle, cold color palette, muted colors, detailed, 8k”
SDXL

Diffusion models in a jiffy

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

Data

Noise

https://nvlabs.github.io/denoising-diffusion-gan/

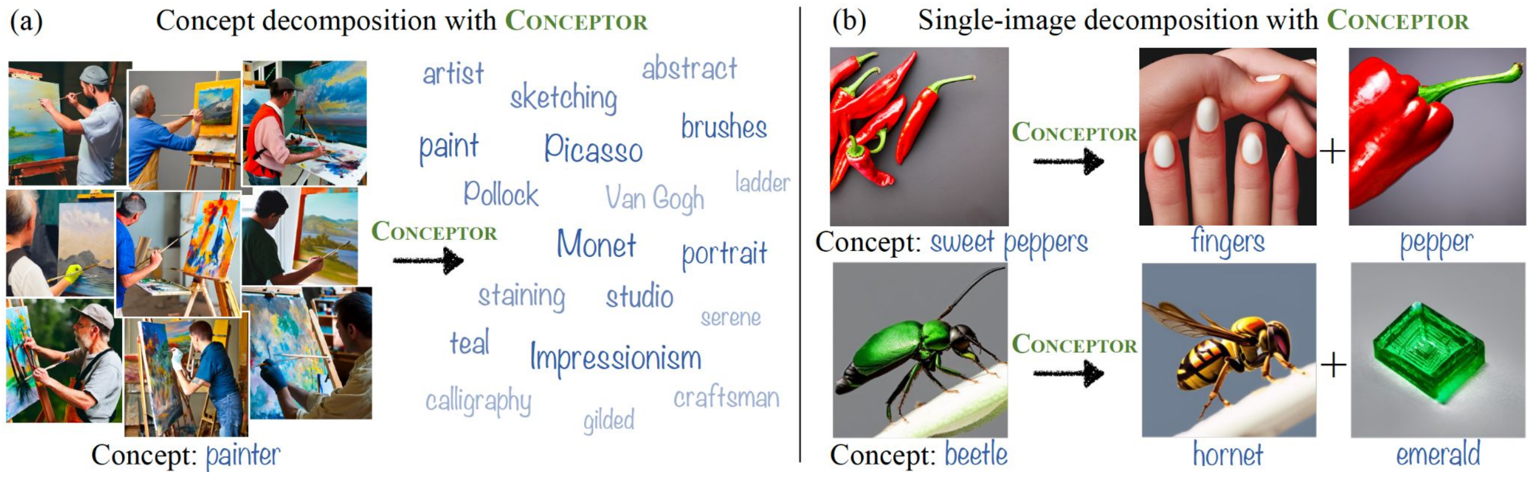
Diffusion models in a jiffy

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:



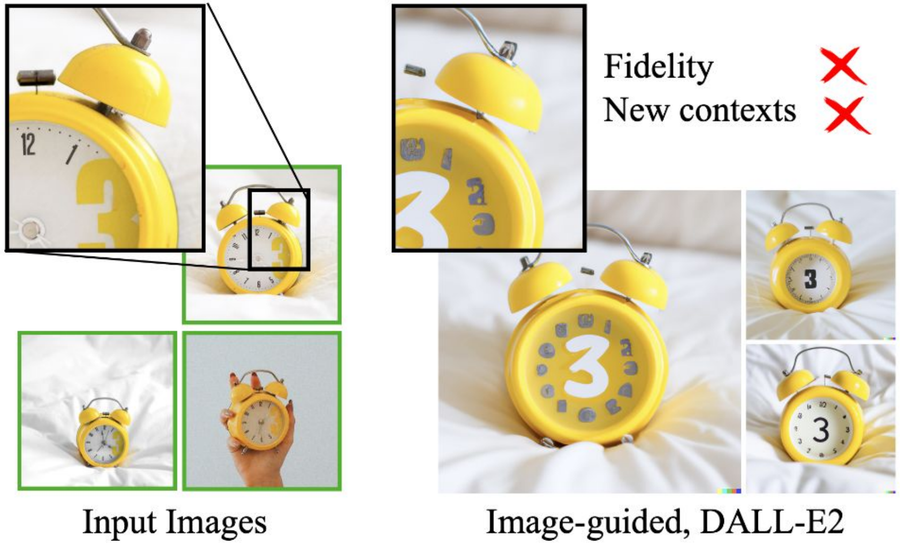
Conceptor; Chefer et al., 2023.

🧠 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



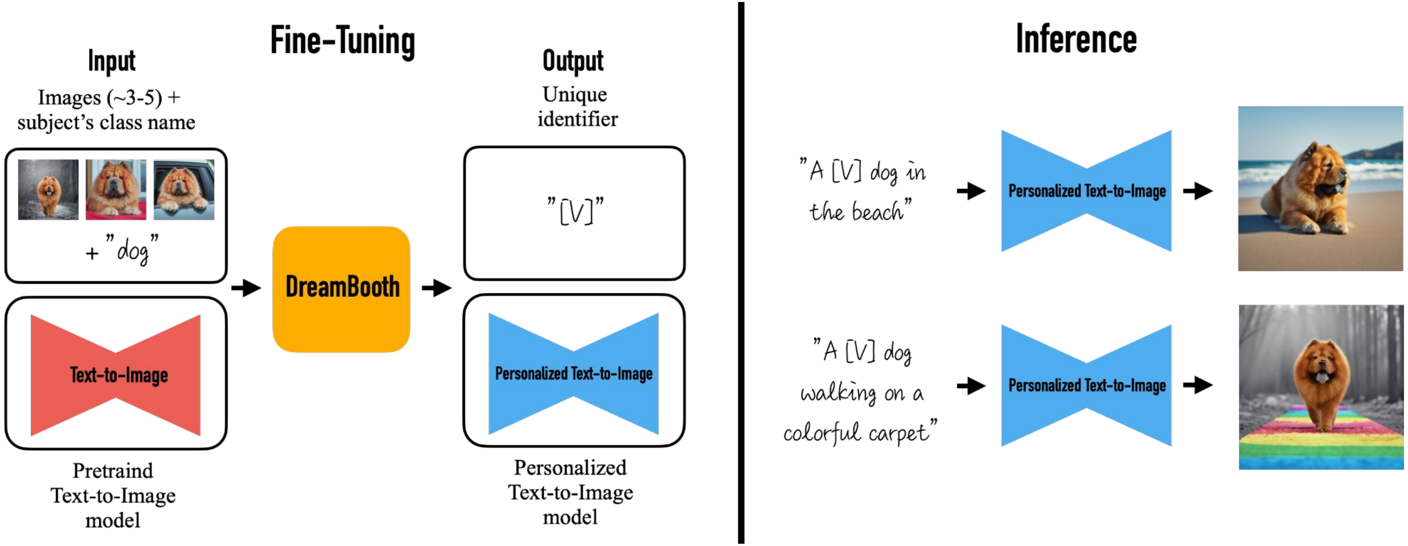
https://dreambooth.github.io/

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

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/

DreamBooth; Ruiz et al., 2022.

Subject-driven generation for personalization

Without the loss of generality, let:

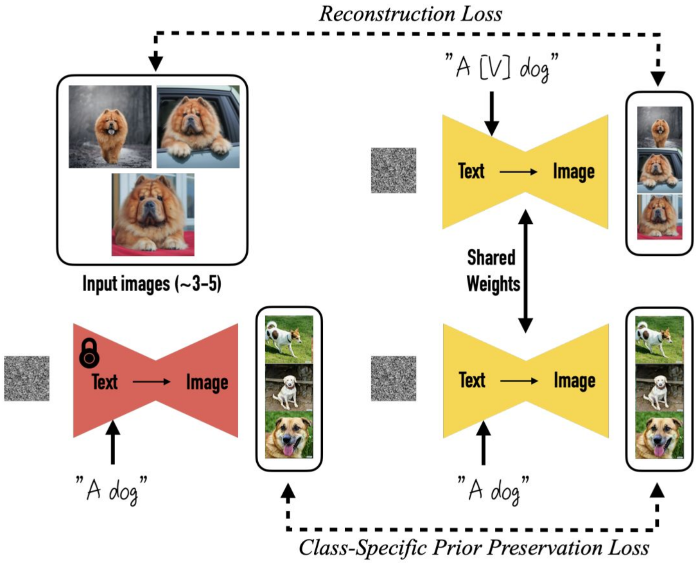
- $\boldsymbol{x}$: original image
- $\boldsymbol{\epsilon}$: noise; $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- t: diffusion process time; $t \sim \mathcal{U}([0, 1])$
- α_t, σ_t, w_t: terms controlling noise schedule and sample quality
- $\mathbf{c}$: conditioning vector (prompt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

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

Inference $$\mathbf{x}_{\text{gen}} = \hat{\mathbf{x}}_\theta(\boldsymbol{\epsilon}, \mathbf{c})$$

Subject-driven generation for personalization

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

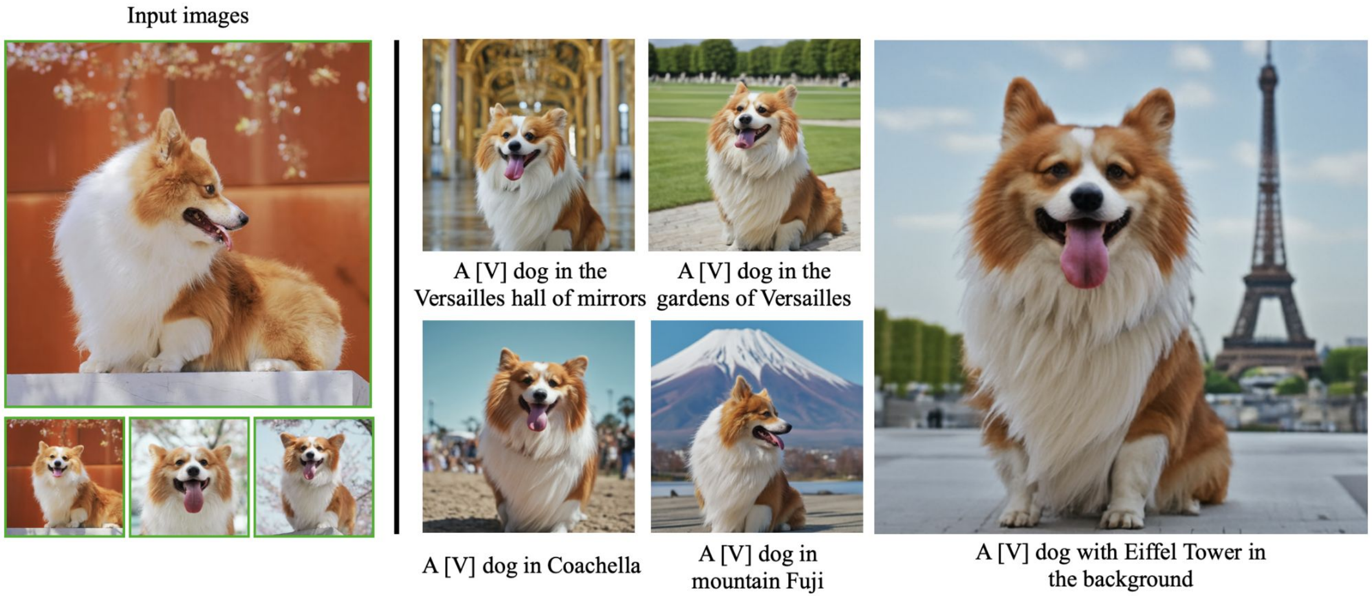


$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}\big[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2+ \lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2\big]$$

DreamBooth; Ruiz et al., 2022.

One framework, multiple use cases

General subject-driven generation



DreamBooth; Ruiz et al., 2022.

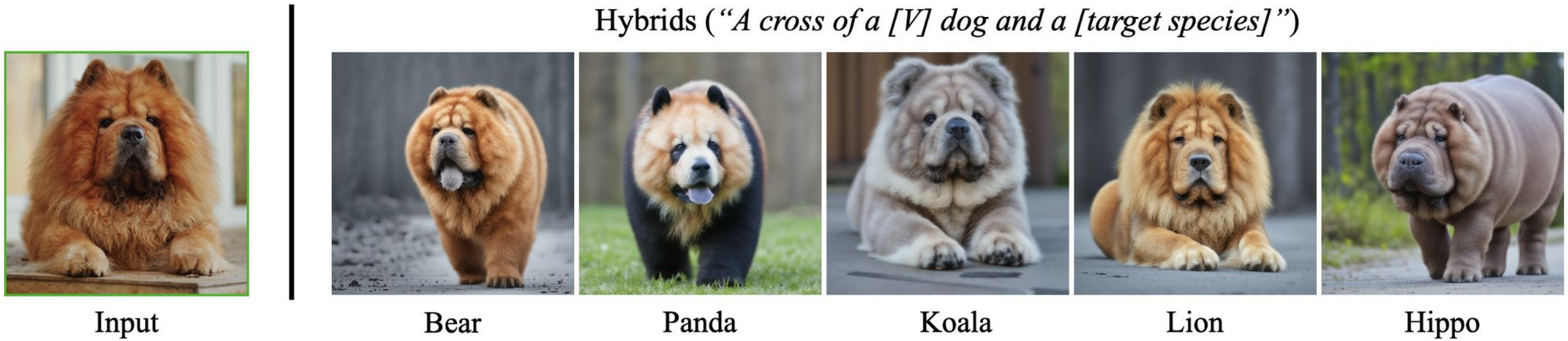
One framework, multiple use cases

Art rendition

DreamBooth; Ruiz et al., 2022.

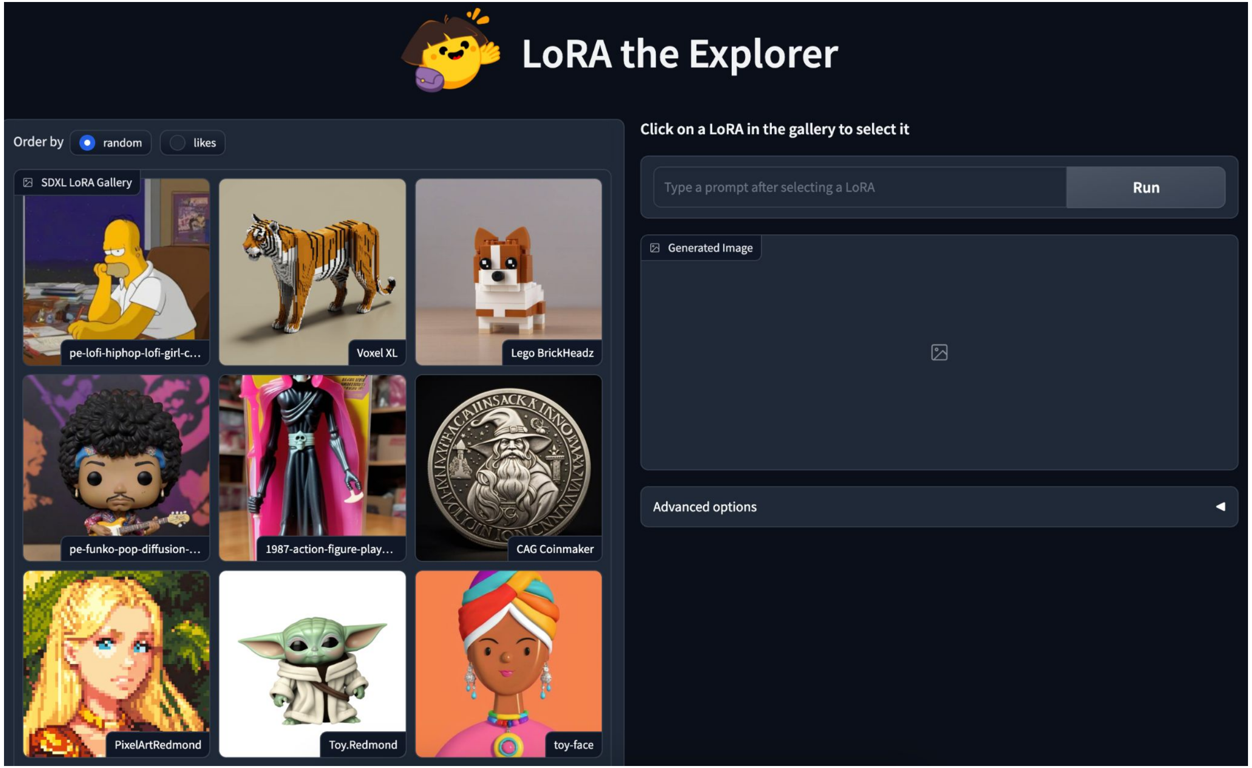
One framework, multiple use cases

Property modification



DreamBooth; Ruiz et al., 2022.

Pushing the extremes with DreamBooth



Further reads

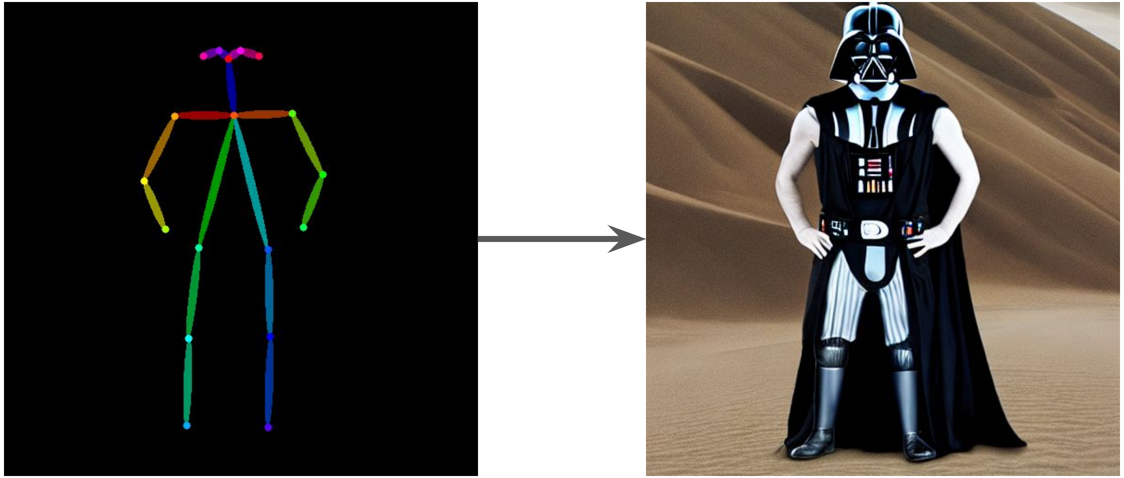
- **BLIP-Diffusion**; Li et al., 2023 (zero-shot subject-driven generation).
- **Custom Diffusion**; Kumari et al., 2022.
- **Pivotal Tuning**; Roich et al., 2021 (in SD context it's Textual Inversion + DreamBooth).

Limitations and solutions

Part II

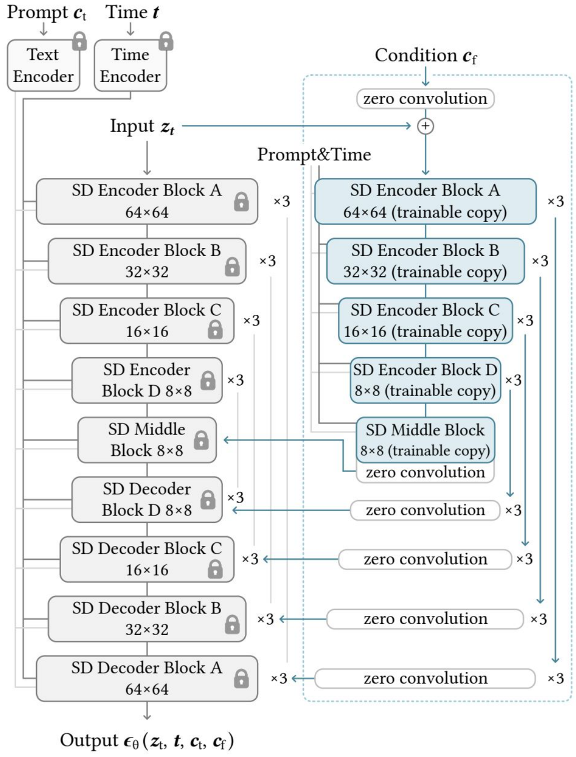
Going beyond text conditioning

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



"Darth Vader dancing in a desert"

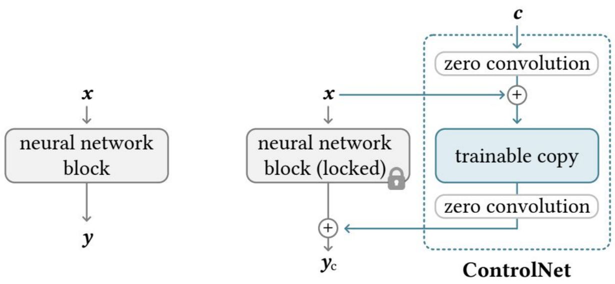
Going beyond text conditioning - ControlNets



(a) Stable Diffusion (b) ControlNet

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

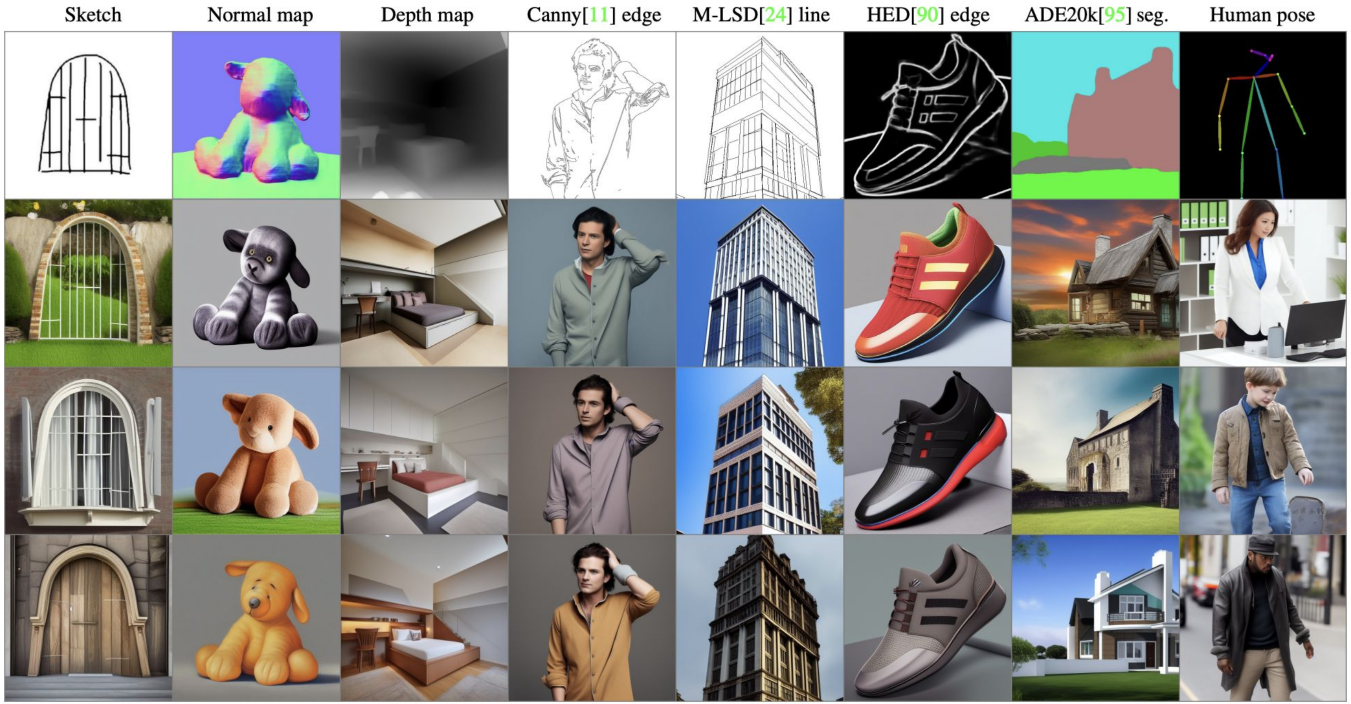
Image-space conditioning vector for ControlNet



(a) Before (b) After

ControlNet; Zhang et al., 2023.

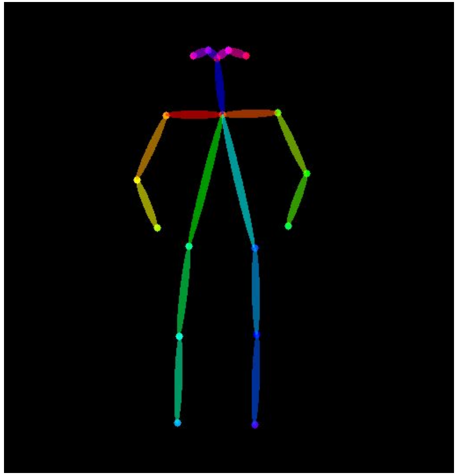
ControlNets - a powerful framework to inject additional control



ControlNet; Zhang et al., 2023.

Or shall I say ***controls***?

Canny map



Pose

Final Image

"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.

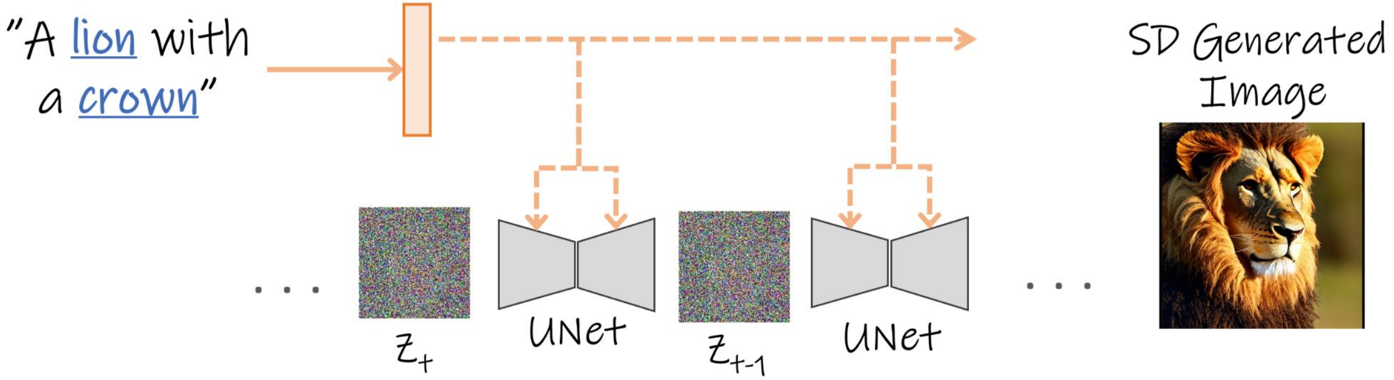

y tho?

🥷 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.



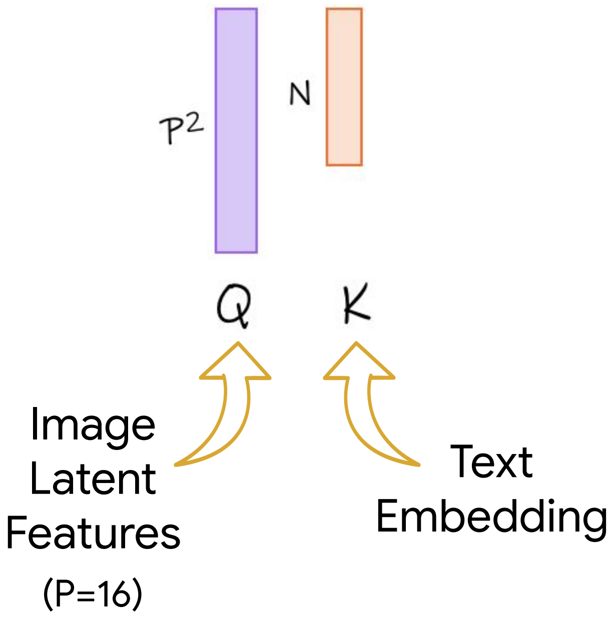
Attend and Excite; Chefer et al., 2023.

Cross Attention

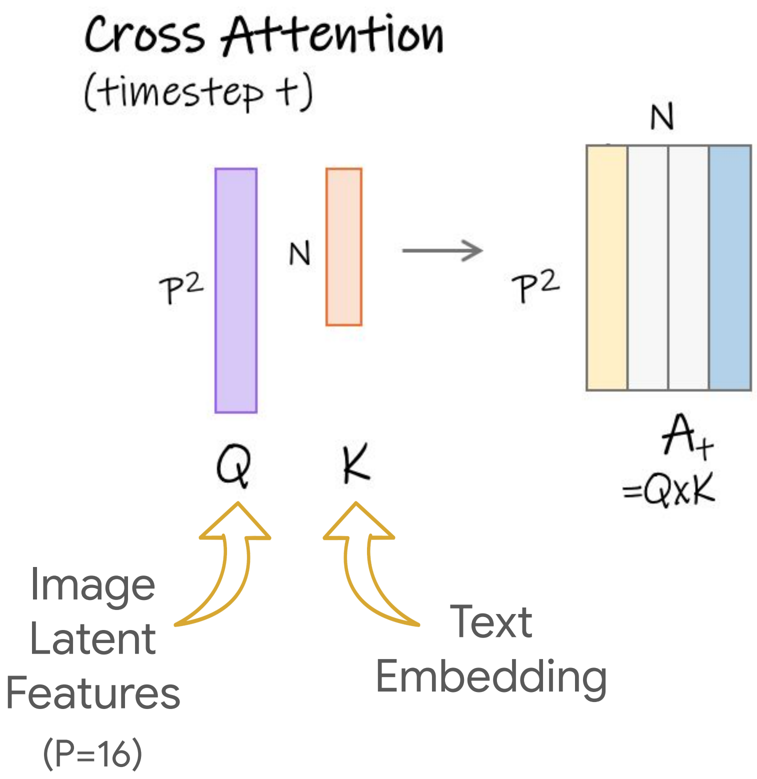
(timestep t)

Slide courtesy: Hila Chefer

Cross Attention
(timestep t)



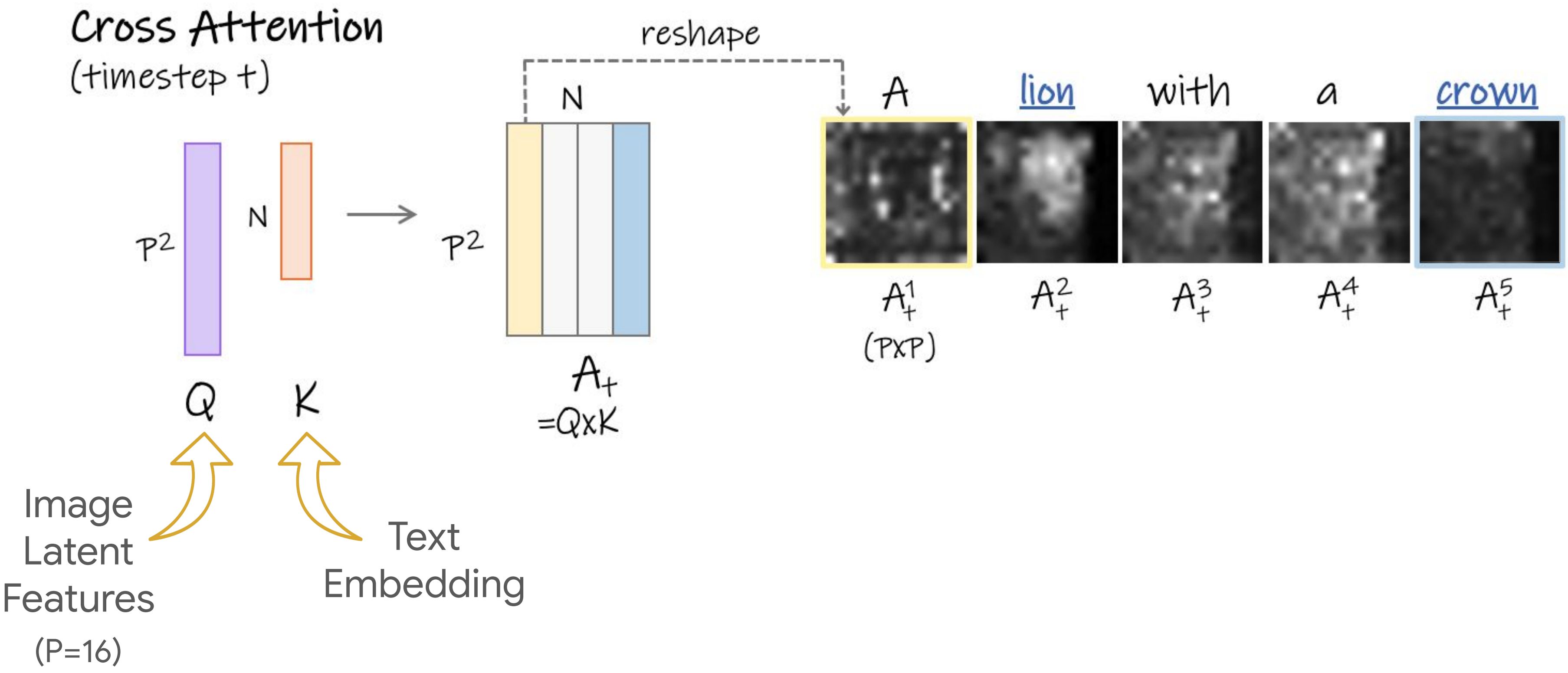
Attend and Excite; Chefer et al., 2023.



$A_t[i,n]$ = presence of the token n in patch i

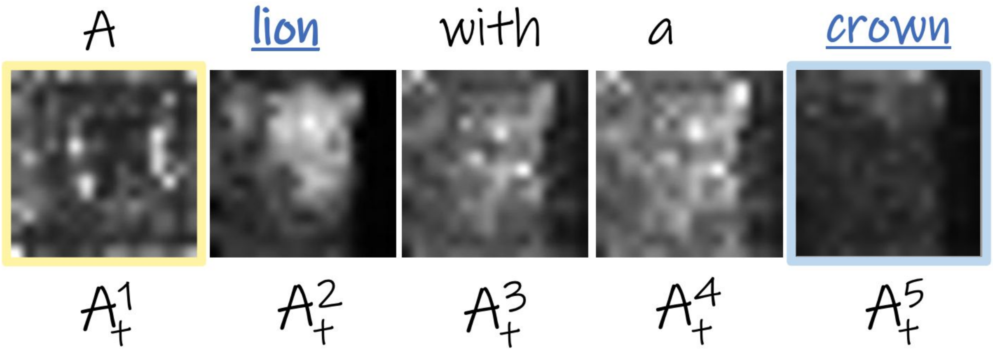
Attend and Excite; Chefer et al., 2023.

Slide courtesy: Hila Chefer



Attend and Excite; Chefer et al., 2023.

Slide courtesy: Hila Chefer



Problem: **crown** gets low attention values for all patches

Attend and Excite; Chefer et al., 2023.

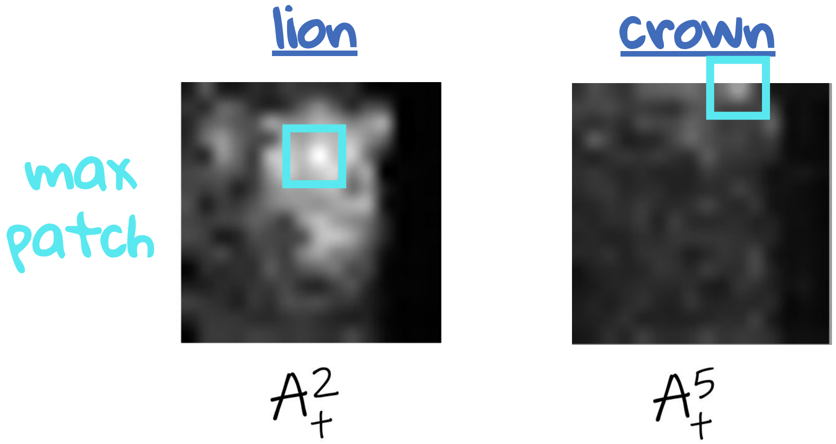
Generative semantic nursing

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.



$L_2 = 1 - \max A_t^2$ $\quad$ $L_5 = 1 - \max A_t^5$

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

Update: $z_t = z_t - \alpha \nabla_{z_t} L$

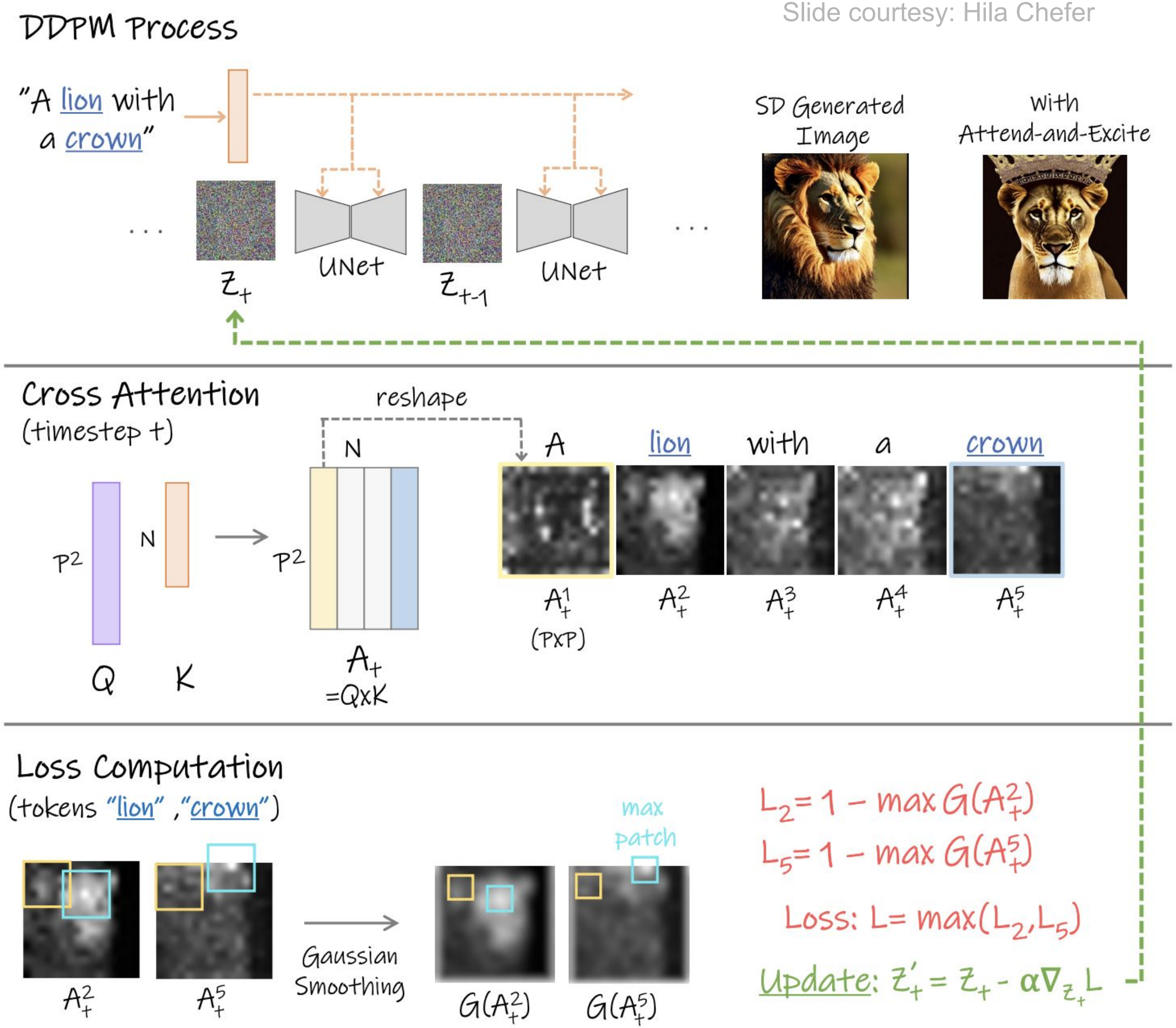
How close are we to having a strong patch?

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

Slide courtesy: Hila Chefer

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!”

Slides