

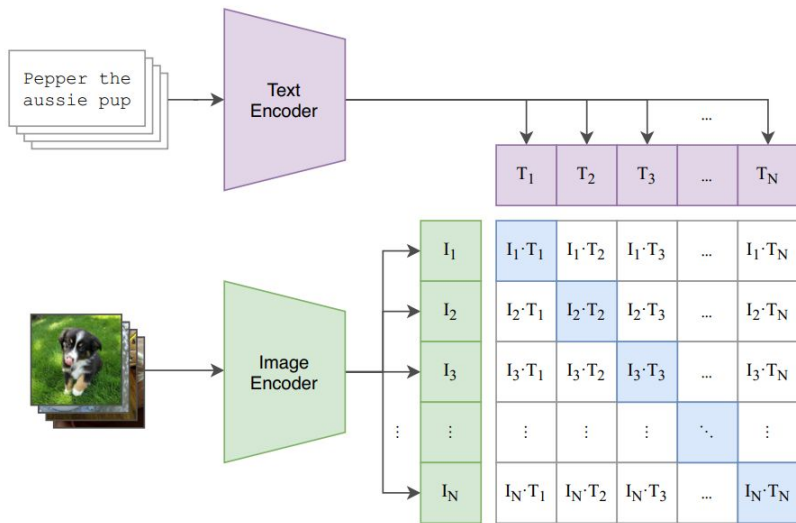
Generative Diffusion Models

Mehmet Saygin Seyfioglu
02/27/24

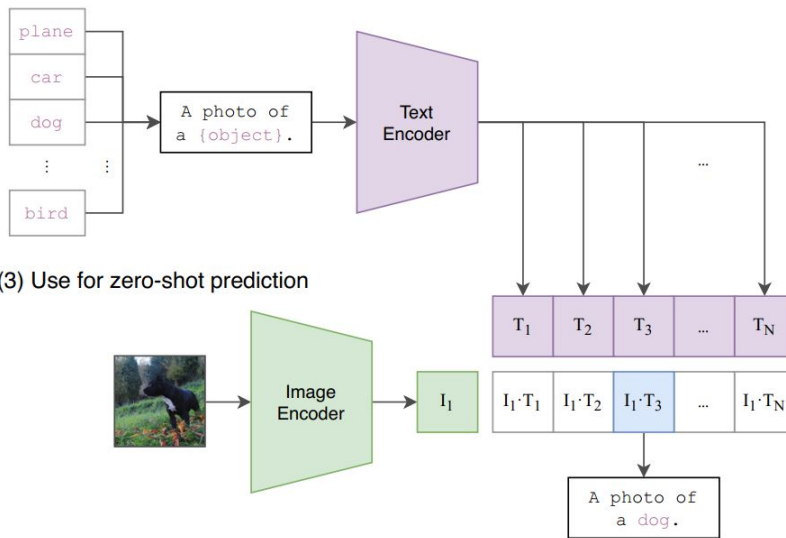
Recap CLIP

- We talked about image-text (multi modal) learning with CLIP.
 - This is very useful in other domains!

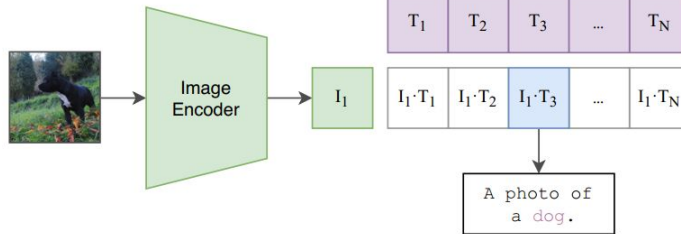
(1) Contrastive pre-training



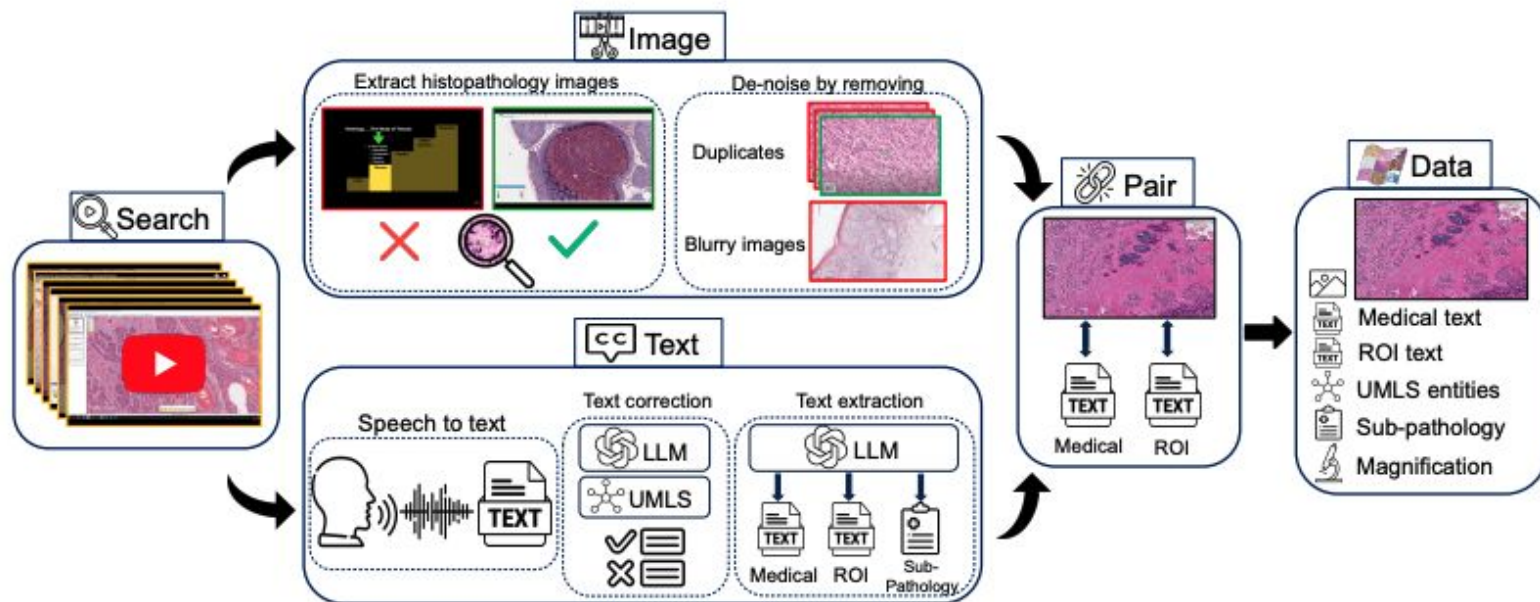
(2) Create dataset classifier from label text



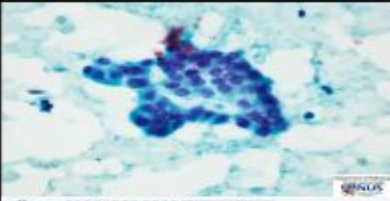
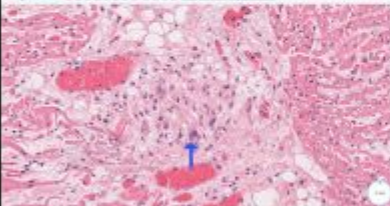
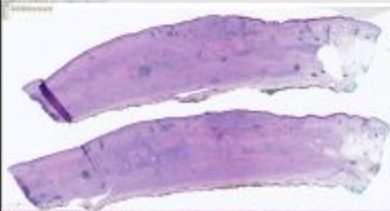
(3) Use for zero-shot prediction



Quilt-1M

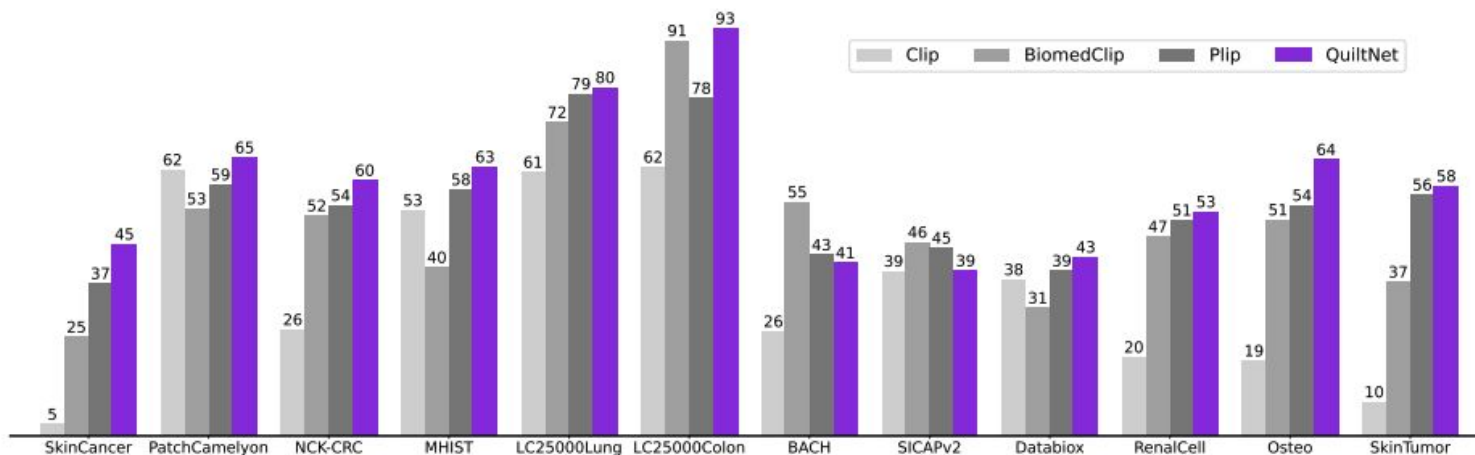


Quilt-1M: Dataset

Image	Medical TEXT	ROI Text	Sub-pathology Classification
	['There are clusters of cells with micro-follicular formations.', 'Nuclear pseudo-inclusions, oval nuclei, nuclear grooves, and small nucleoli are present in some cells.']	['clusters of cells', 'micro-follicular formations', 'nuclear pseudo-inclusions', 'oval nuclei', 'nuclear grooves', 'small nucleoli']	['Endocrine', 'Cytopathology', 'Head and Neck']
	['Cluster of macrophages and T cells is characteristic of acute rheumatic fever.', 'Aschoff body is a characteristic feature of acute rheumatic fever.', 'Macrophages with elongated chromatin are called Anitchkow cells and are commonly seen in Aschoff bodies.', 'Pancarditis with Aschoff bodies is present.']	['Cluster of macrophages and T cells', 'Aschoff body', 'Macrophages with elongated chromatin', 'Anitchkow cells', 'Pancarditis']	['Cardiac', 'Hematopathology', 'Endocrine']
	['An 80-year-old man has a scar-like plaque on the scalp that has been called malignant on a biopsy.', 'The tissue affected by the plaque extends from the epidermis to the galea aponeurotica, near the periosteum of the skull.', 'The skin, dermis, and subcutis are all affected by the process.']	['scar-like plaque on the scalp', 'malignant on a biopsy', 'skin, dermis, and subcutis affected by the process']	['Dermatopathology', 'Soft tissue', 'Hematopathology']

QuiltNet

$$\mathcal{L} = -\frac{1}{2N} \left(\sum_{i=1}^N \log \frac{e^{\cos(I_i, T_i)}}{\sum_{j=1}^N e^{\cos(I_i, T_j)}} + \sum_{i=1}^N \log \frac{e^{\cos(I_i, T_i)}}{\sum_{j=1}^N e^{\cos(I_j, T_i)}} \right)$$



Diffusion Models

An image of a husky surfing in the space



A horse formula 1 driver



Inpainting



Prompt: a white cat, blue eyes, wearing a sweater, lying in park



Outpainting



Taken from <https://blog.segmind.com/exploring-the-magic-of-outpainting-with-stable-diffusion-uncropping-the-creative-possibilities/>

Outpainting



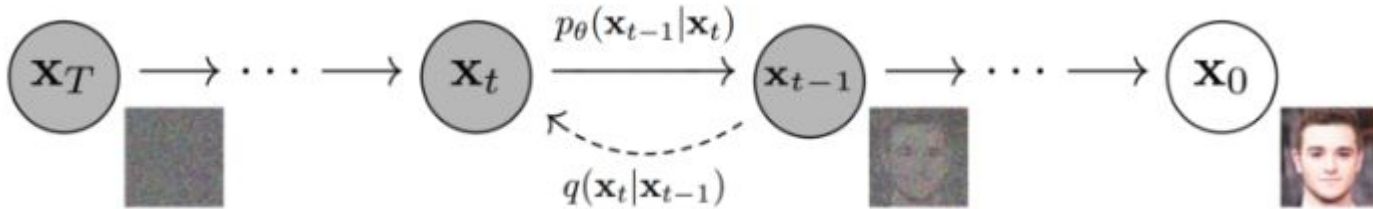
Taken from <https://blog.segmind.com/exploring-the-magic-of-outpainting-with-stable-diffusion-uncropping-the-creative-possibilities/>

Diffusion Models

- Is actually a self-supervised framework.
- This time, instead of aligning image and text embeddings, like in the case of CLIP. We do something more advanced.
- We learn the distribution of images, then use text (or whatever other modality) to generate them from noise.
- How?

Diffusion Models

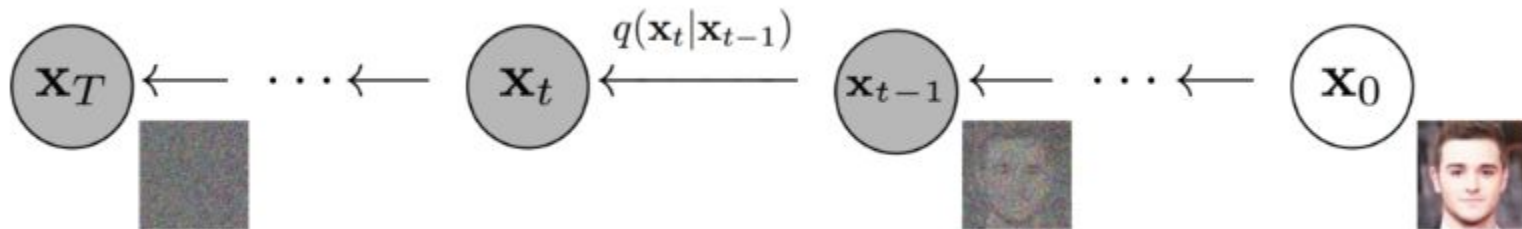
- What we aim is to, create a self-supervised paradigm, where we gradually add noise to an image until it becomes an isotropic gaussian noise.
- Then use a neural network to predict the added noise and gradually decrease it.



$q(\mathbf{x}_t|\mathbf{x}_{t-1})$ Pdf of an image at timestep t given image \mathbf{x}_{t-1}

$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ Pdf of \mathbf{x}_{t-1} given \mathbf{x}_t parameterized by the model (θ)

The Forward Process



The probability density function

The distribution q in the forward diffusion process is defined as *Markov Chain* given by:

$$q(x_1, \dots, x_T | x_0) := \prod_{t=1}^T q(x_t | x_{t-1}) \quad \dots (1)$$

$$q(x_t | x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I) \quad \dots (2)$$

Adding noise

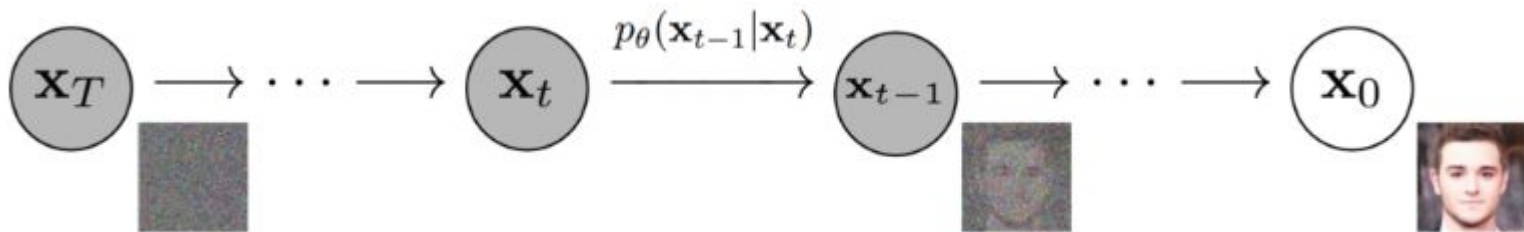
x_t is generated from x_{t-1} adding noise. In this way, starting from x_0 , the original image is iteratively corrupted from $t=1 \dots T$

$$x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} \epsilon \quad \dots (3)$$

; where $\epsilon \sim \mathcal{N}(0, I)$

Reverse Diffusion Process

Reverse Markov Chain -> We want this because if we follow the forward trajectory in reverse, we may return to the original data distribution

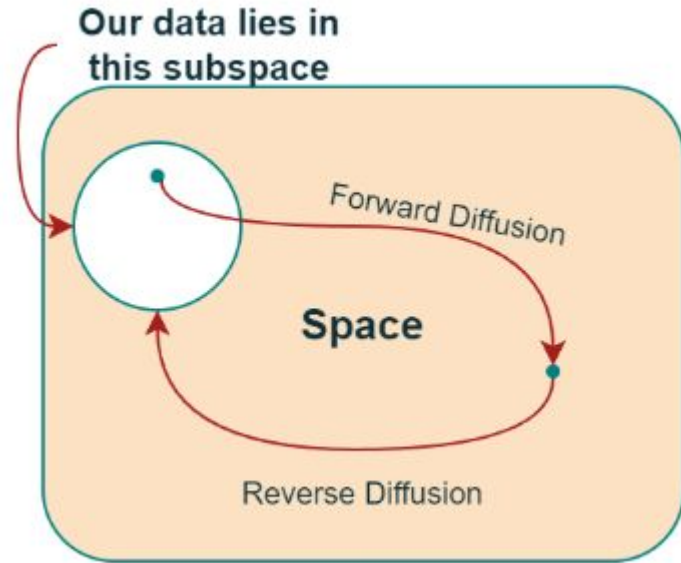


At step \mathbf{x}_{t-1} , the network predicts the mean of the noise that is added at \mathbf{x}_t

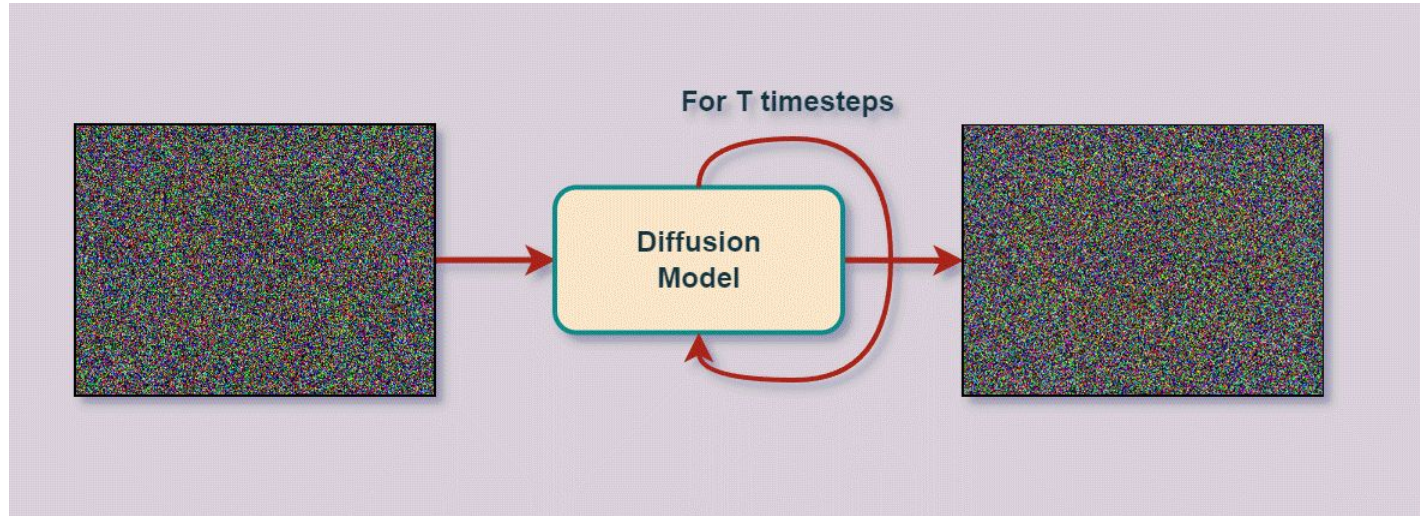
In doing so, we would also learn how to generate new samples that closely match the underlying data distribution, starting from a pure gaussian noise

$$L_{\text{simple}} = E_{t, x_0, \epsilon} [||\epsilon - \epsilon_\theta(x_t, t)||^2]$$

A high-level conceptual overview of the entire image space



After Learning The Model



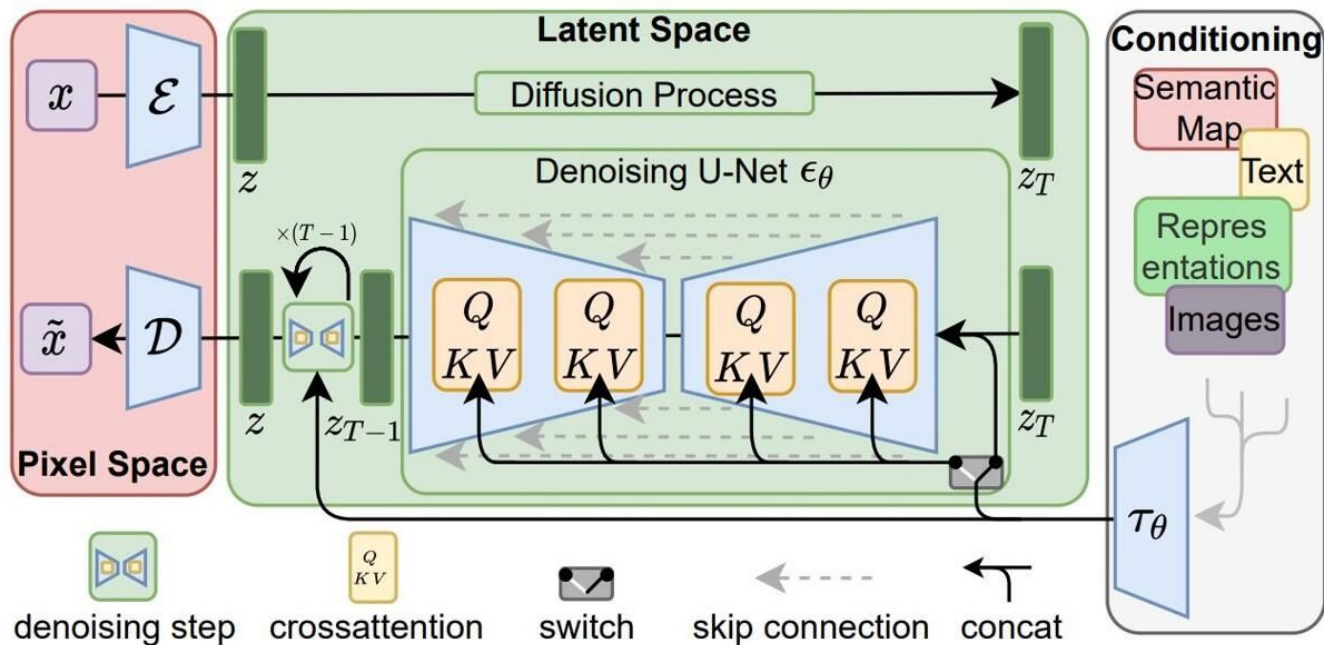
Conditional Diffusion

Better yet, instead of just learning the underlying data distribution to sample new stuff of that certain category, we can guide the diffusion process. This is great because we can now then mix concepts together, which the model has not even seen before!

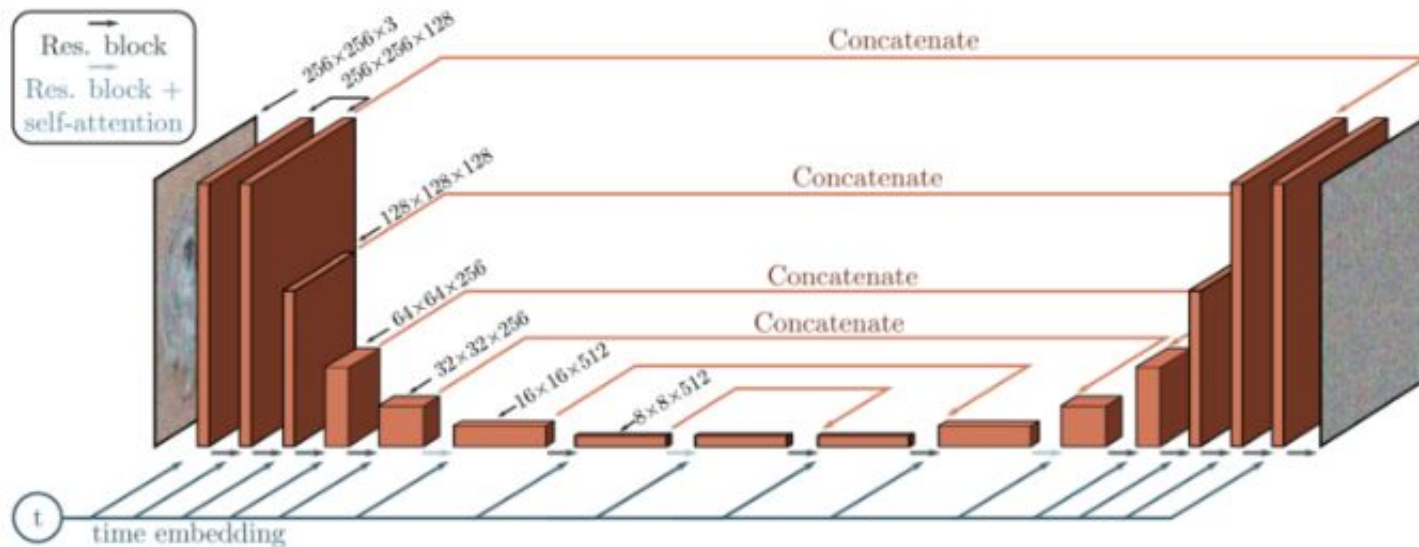
Remember LAION 5b? That's really handy to train this model (image caption pairs)

Also the CLIP model? That's a tool we could leverage

The Only Open-Source Diffusion Model: Stable Diffusion



Architecture



Stable Diffusion Examples

A Dog In A Hat Looking
Like A Vintage Portrait



A Giant Panda In
Between A Celestial War



Some State of the Art Diffusion Applications

- In the last year, some cool methods have been proposed using Stable Diffusion.

DreamBooth



Input images



in the Acropolis



swimming



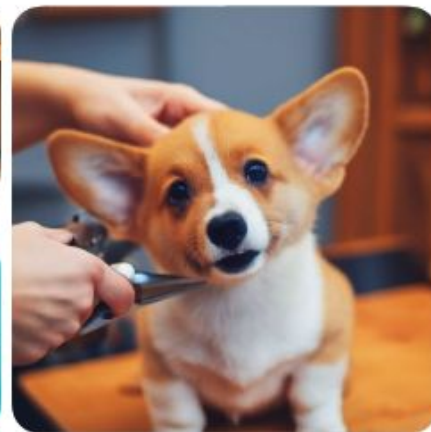
sleeping



in a doghouse

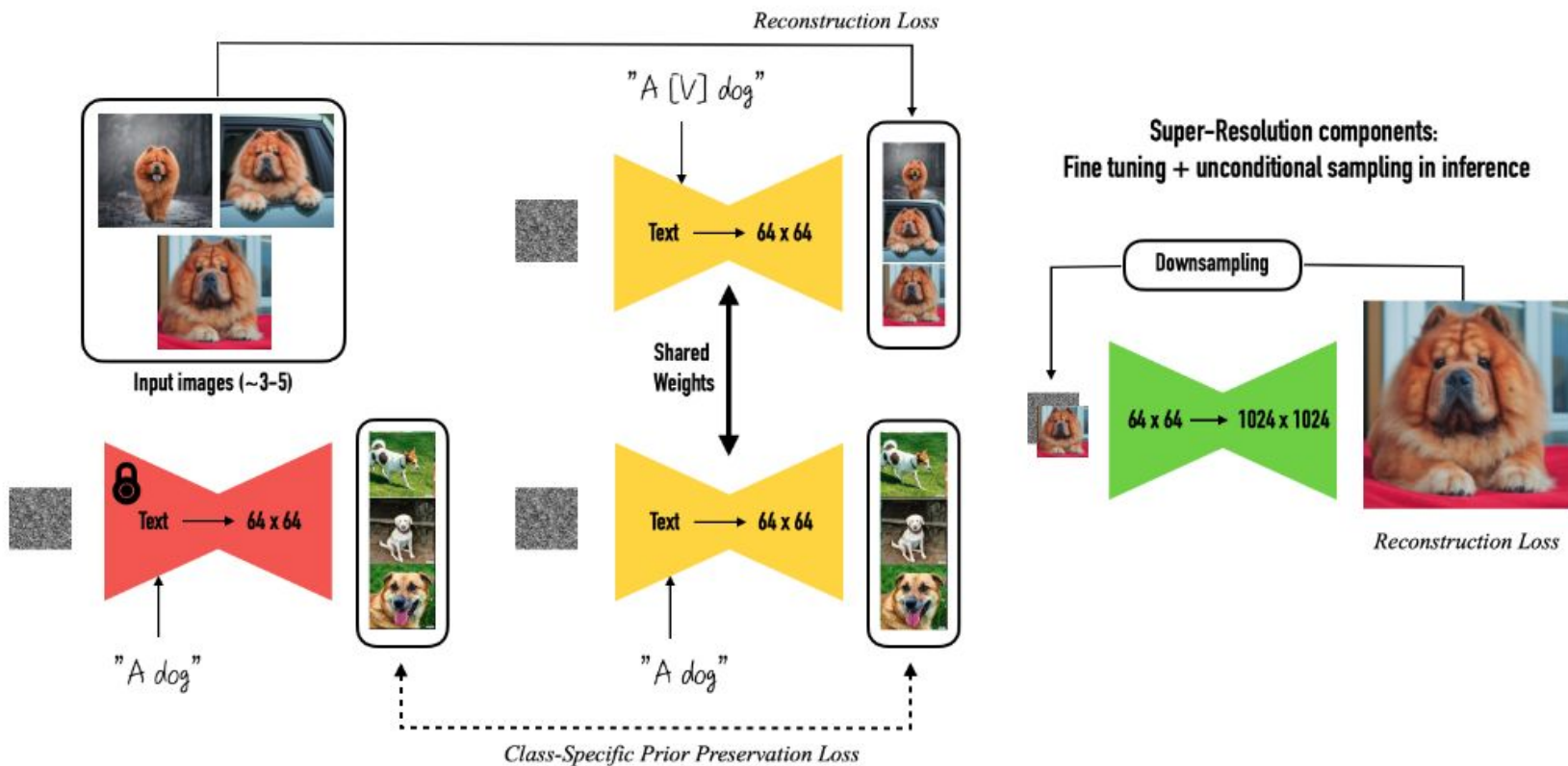


in a bucket



getting a haircut

DreamBooth



InstructPix2Pix

"Swap sunflowers with roses"



"Add fireworks to the sky"



"Replace the fruits with cake"



"What would it look like if it were snowing?"



"Turn it into a still from a western"



"Make his jacket out of leather"




InstructPix2Pix

Training Data Generation

(a) Generate text edits:

Input Caption: "photograph of a girl riding a horse" → **GPT-3** → Instruction: "have her ride a dragon"
Edited Caption: "photograph of a girl riding a dragon"

(b) Generate paired images:

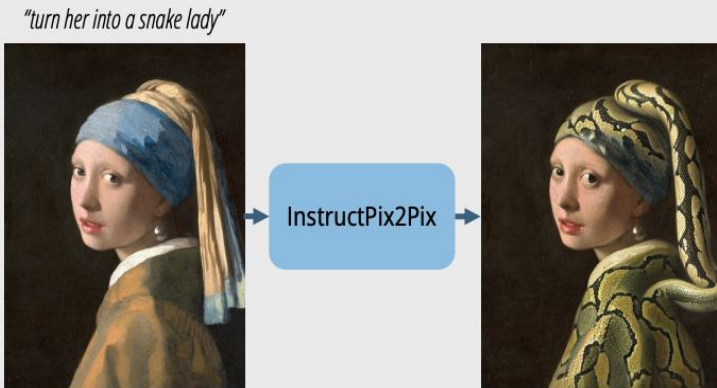
Input Caption: "photograph of a girl riding a horse"
Edited Caption: "photograph of a girl riding a dragon" → **Stable Diffusion + Prompt2Prompt** → 

(c) Generated training examples:

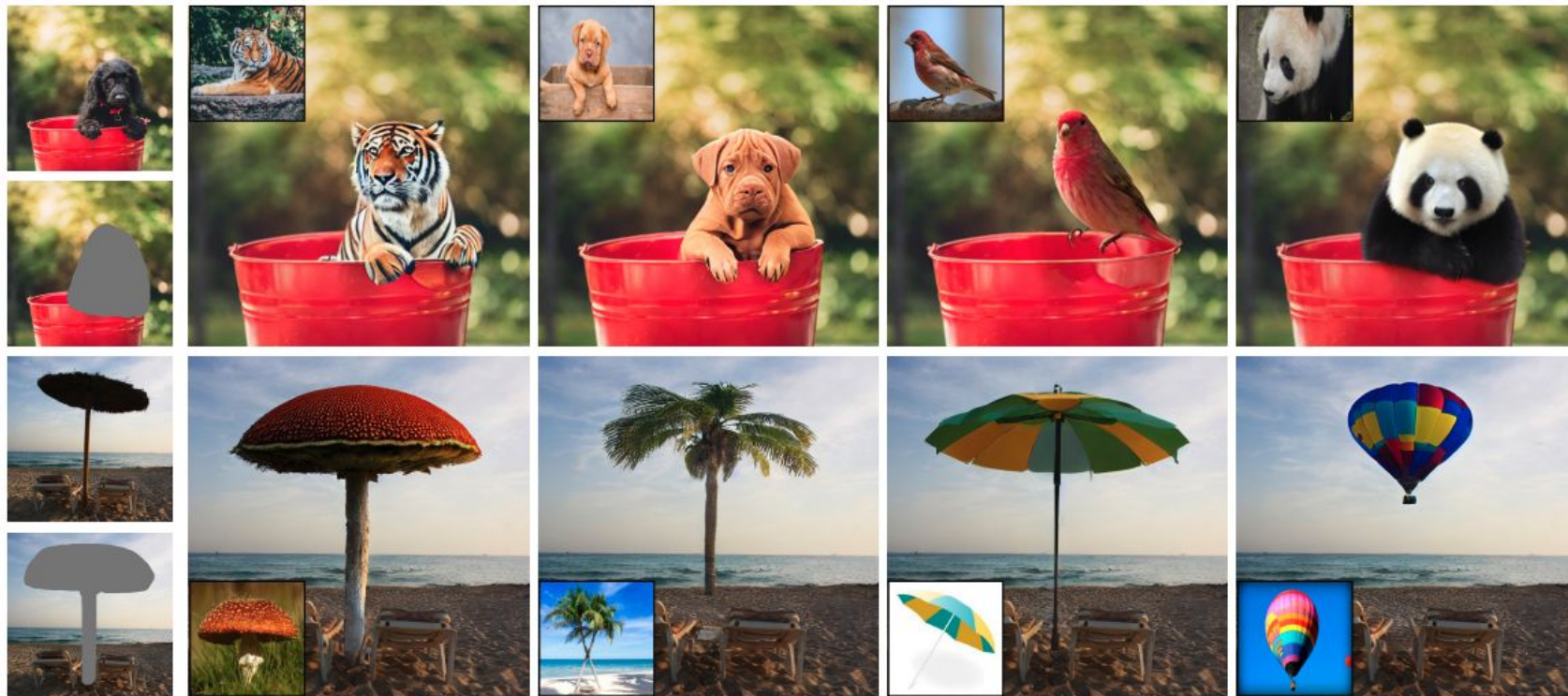


Instruction-following Diffusion Model

(d) Inference on real images:

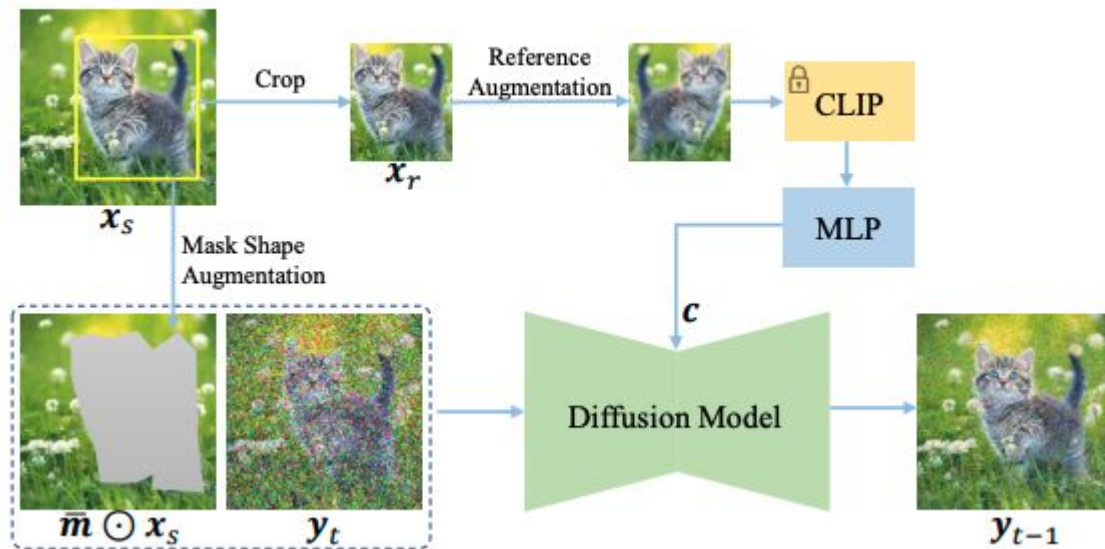


Paint by Example



Yang, Binxin, et al. "Paint by example: Exemplar-based image editing with diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Paint by Example



My Diffusion Research

- I try to achieve Virtual Try-All

Allowing shoppers to virtually 'try' any product from any category within their personal environments (in the wild examples).

Virtual Try-All cont'd

amazon prime

Deliver to Mehmet
Seattle 98103

All couch

Q

EN - Hello, Mehmet
Account & List

All

Early Black Friday Deals

Medical Care

Prime Video

Household, Health & Baby Care

Amazon Home

Coupons

Pet Supplies

Beauty & Personal Care

Subscribe & Save

Amazon Basics

Buy Again

Handmade

Amazon Home

Shop by Room

Discover

Shop by Style

Home Décor


Furniture

Kitchen & Dining

Bed & Bath

Garden & Outdoor

Signature
DESIGN BY
ASHLEY




Signature Design by Ashley Alessio Modern Transitional...

★★★★☆ 70

\$369.99 prime

Back to results



Share

Koorlian Sofa Couch, 2 Seater Fabric Loveseat, Mid Century Modern Couches for Living Room, Button Tufted Seat Cushion, Square Armrest, 2 Throw Pillows, Fit for Small Spaces, Dorm, Apart, Beige

Visit the Koorlian Store

3.9 ★★★★★ 314 ratings

400+ bought in past month

Deal


-20% \$175⁹⁹


Typical price: \$219.99


Thank you for being a Prime member. Get \$200 off: Pay \$0.00 \$175.99 upon approval for Prime Visa.

Delivery & Support


Select to learn more


 Ships from Do I Want


 Returnable until Jan 31, 2024


 Customer Support

Color: Beige


 \$175.99


 \$185.99


 \$185.99


 \$185.99


Roll over image to zoom in













5 VIDEOS





Qty: 1

Add to Cart

Buy Now

Ships from Do I Want

Sold by Do I Want

Returns Returnable until Jan 31, 2024

Payment Secure transaction

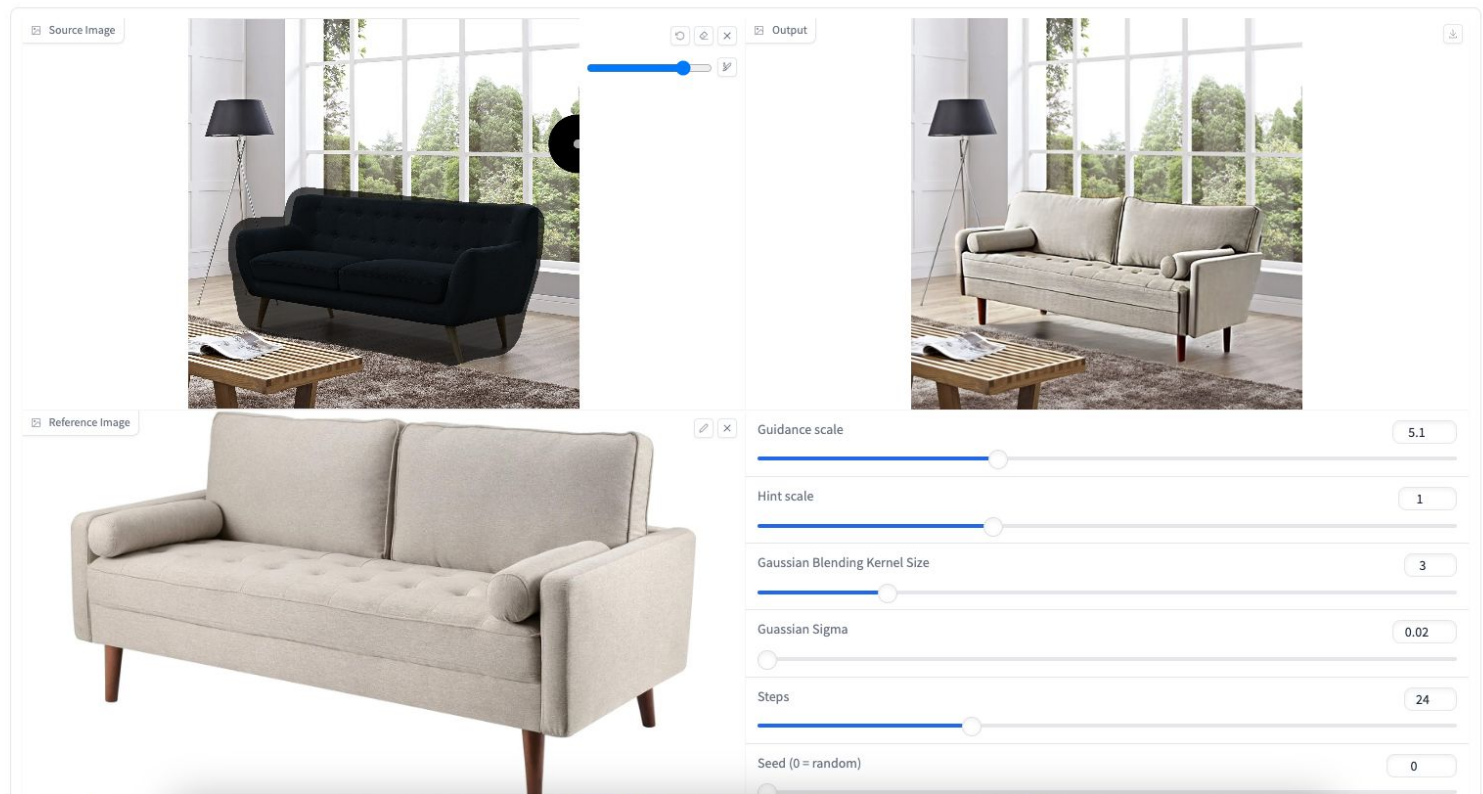
Add a Protection Plan:

☐ 2 Year Furniture Protection Plan for \$31.99

☐ 3 Year Furniture Protection Plan for \$39.99

Add to List

Virtual Try-All cont'd



How?

For Virtual Try-All model to be effective, it must fulfill three primary conditions:

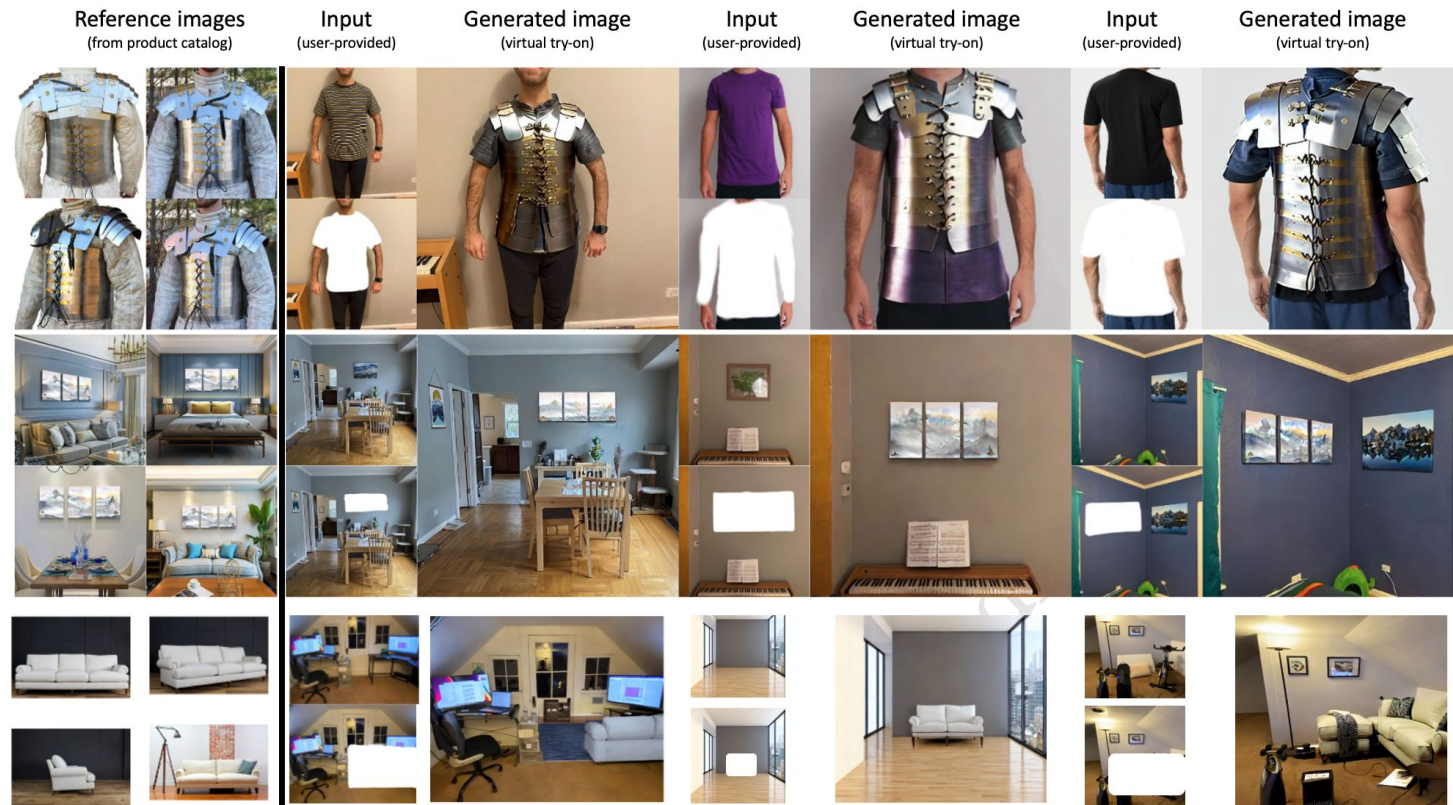
1. Operate in any 'in-the-wild' user image, and reference image,
2. Integrate the reference product harmoniously with the surrounding context while maintaining the product's identity
3. Perform fast inference to facilitate real-time usage across billions of products and millions of users.

DreamPaint

Previously we implemented DreamPaint [1] (Dreambooth-Inpaint), which is a framework to intelligently inpaint any e-commerce product on any user-provided context image without requiring any expensive 3D AR/VR inputs.

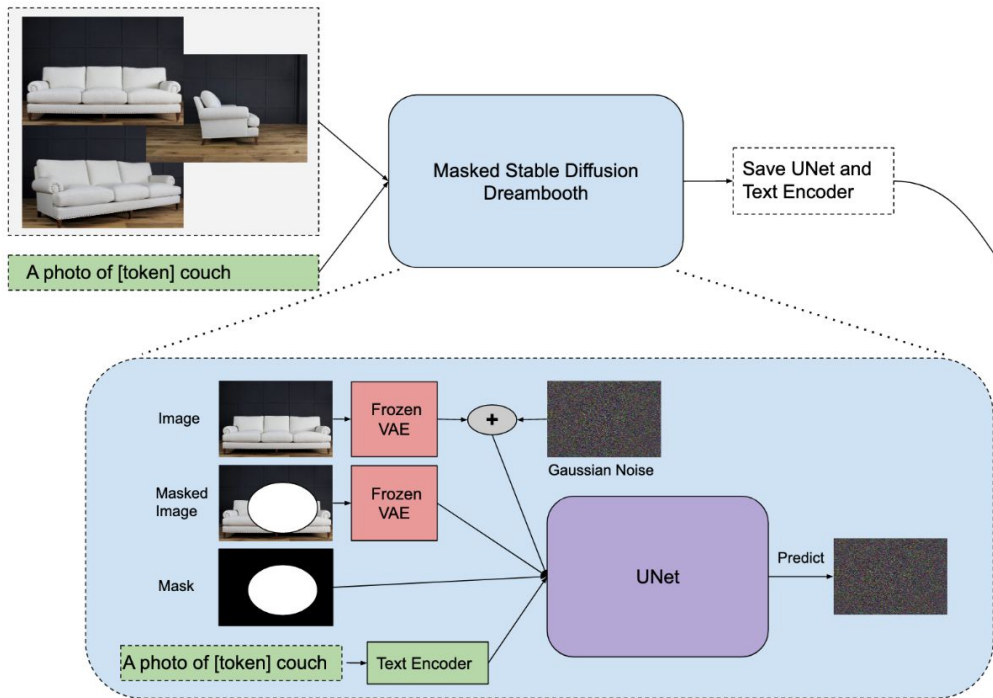
[1] Seyfioglu, Mehmet Saygin, et al. "DreamPaint: Few-Shot Inpainting of E-Commerce Items for Virtual Try-On without 3D Modeling." *arXiv preprint arXiv:2305.01257* (2023).

DreamPaint Examples

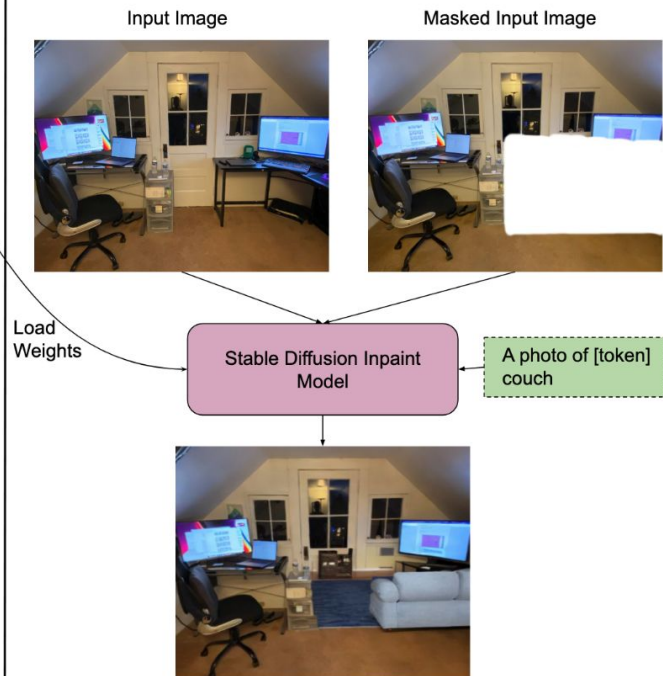


DreamPaint Model

Masked Dreambooth Fine-Tuning



Inference Inpainting

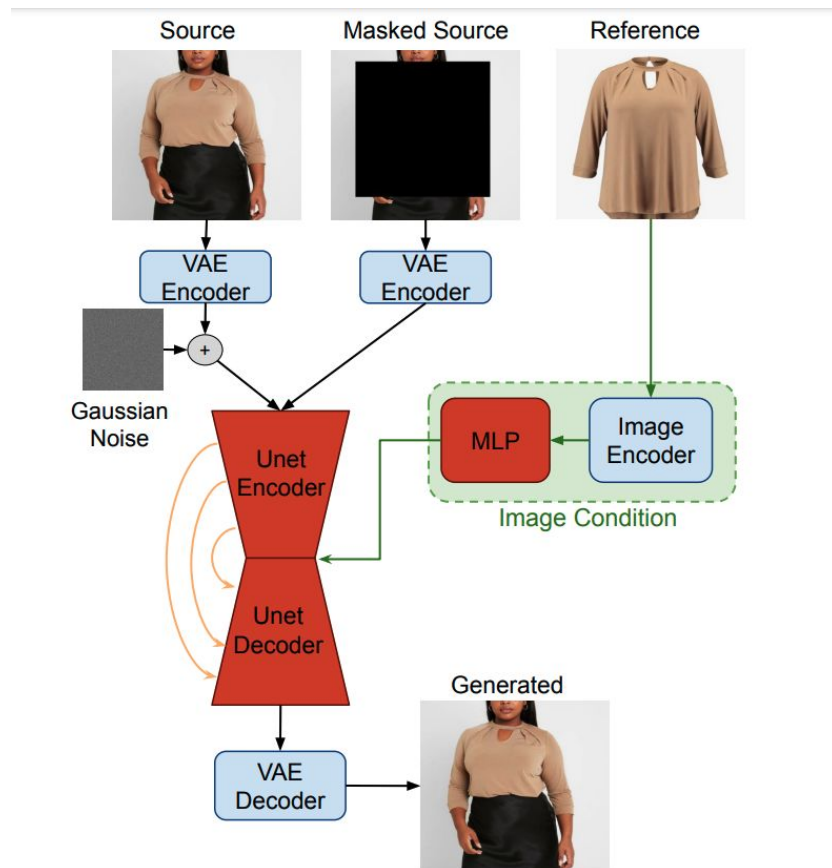


DreamPaint

1. DreamPaint is pretty good at operating with in-the-wild images. ✓
2. DreamPaint can preserve most of the product details, and can semantically blend the product image with its context. ✓
3. DreamPaint requires 40 minutes of fine-tuning with few-shot examples for each product. ✗
 - a. We can do LORA + save only cross attention weights to save space (which reduces model size from 10GB to 30MB) But we still have to train individual models for each asin.

Paint by Example (PBE)

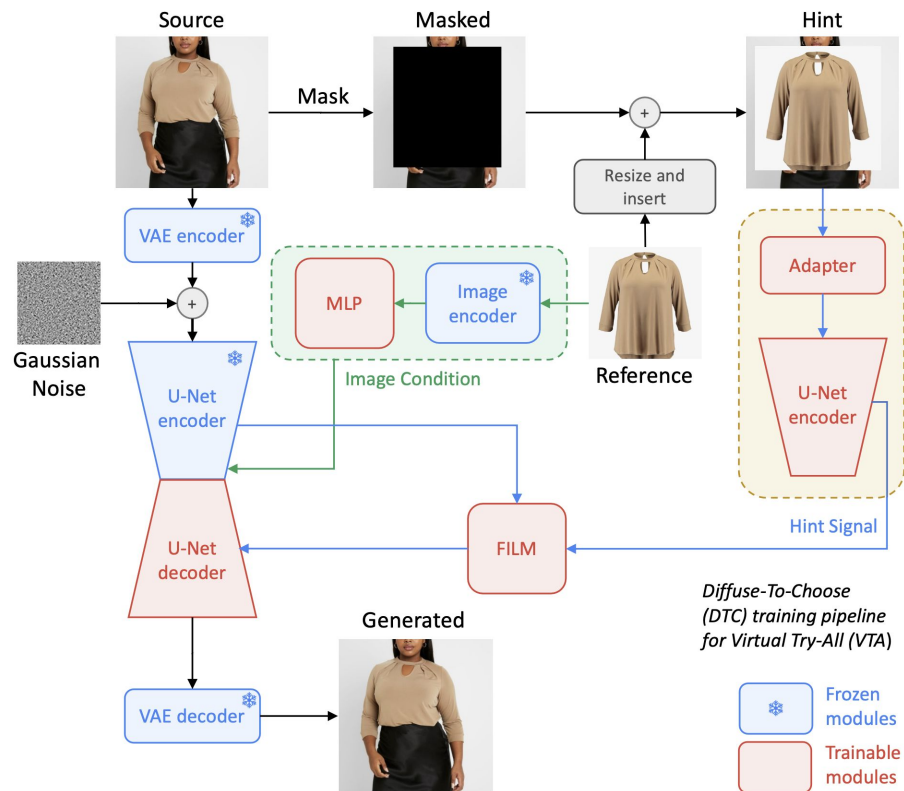
- For catalog items, we don't have to constraint ourselves with self-referencing.
 - Thus no need for the information bottleneck and aggressive augmentations.
- How far can we go with this approach?



PBE

1. PBE is pretty good at operating with in-the-wild images. ✓
2. PBE in its proposed form cannot preserve most product details. (?)
3. After trained, PBE can operate in zero-shot setting, only takes about 5 seconds to generate an image on a low-end GPU with 12GB of RAM. ✓

Diffuse to Choose



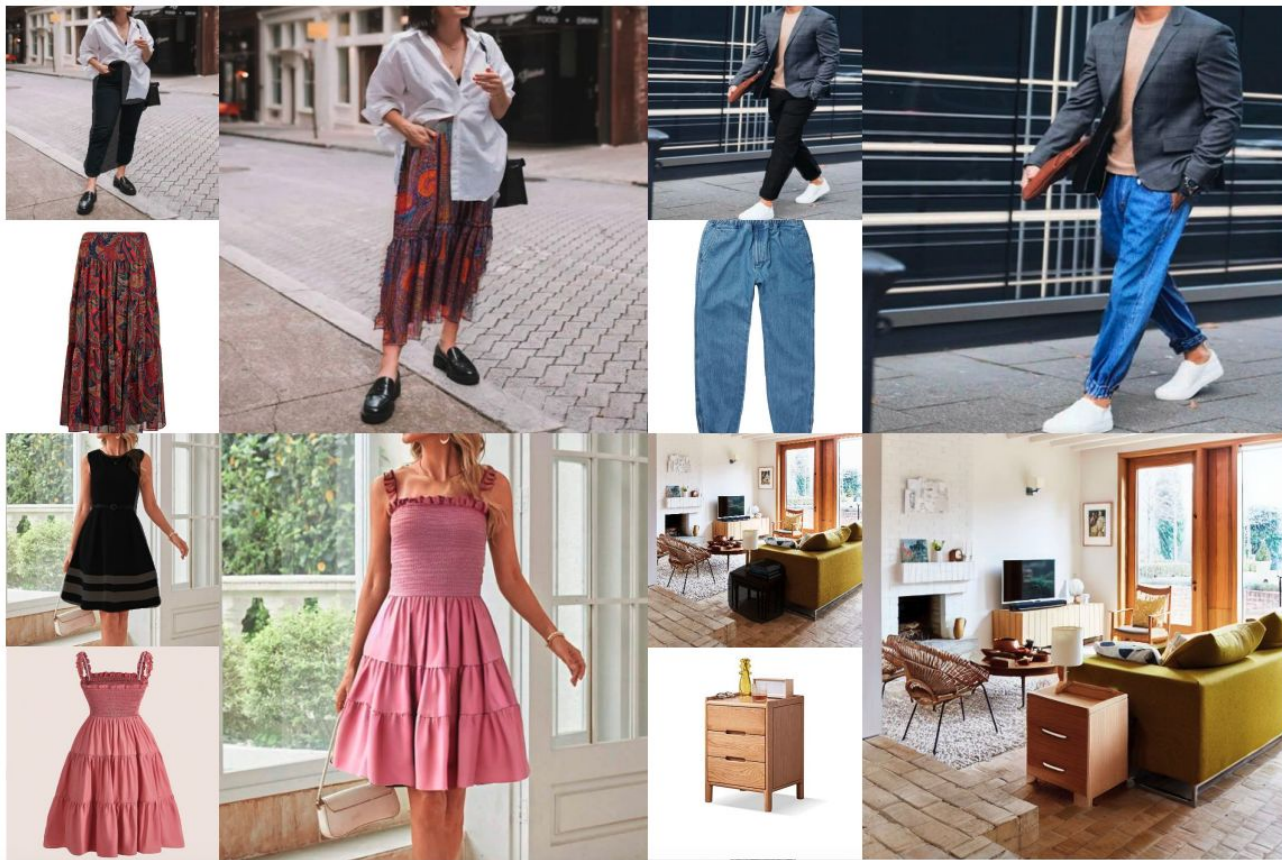
Diffuse to Choose

<https://7cb9c8c71416f5be57.gradio.live/>

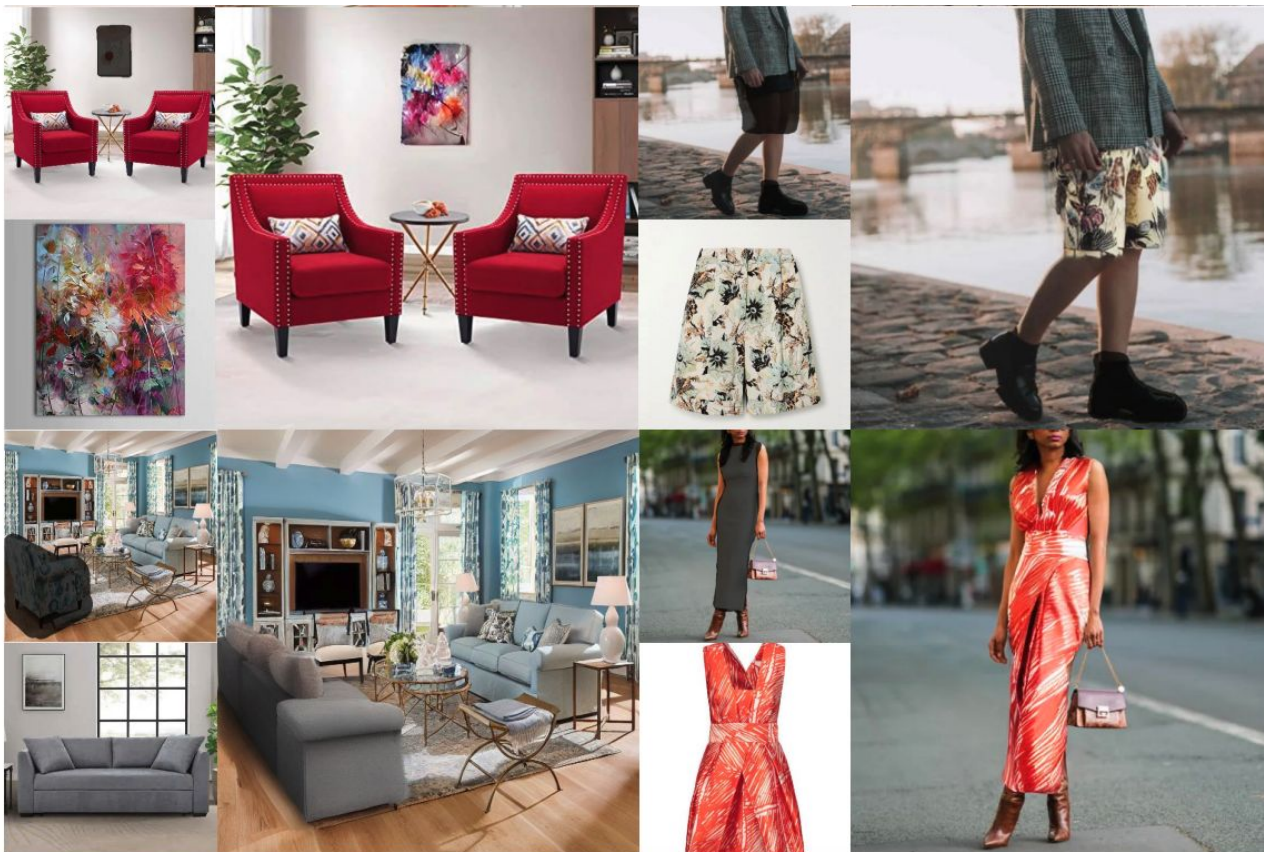
Visual Results



Visual Results



Visual Results



Iterative Decoration



Cool Masking Effect



The Best DTC variant

- Directly stitch the hint image and use FILM on decoding (we computed FID and CLIP scores on our dataset) performs the best. (Cross Attention is really close to FILM)

Table 1. Quantitative comparison between DTC variants and PBE_{best} , which denotes a PBE variant using DINOv2 and perceptual loss. CA denotes Cross-Attention.

Method	CLIP Score (\uparrow)	FID (\downarrow)
PBE_{best}	85.43	6.65
$\text{Ours}_{\text{addition}}$	86.94	6.19
Ours_{CA}	88.01	5.68
$\text{Ours}_{\text{FILM}}$	88.14	5.72

Compare Against PBE variants



Method	CLIP Score (\uparrow)	FID (\downarrow)
PBE CLIP _{cls} [36]	82.43	9.54
+ PBE CLIP _{all}	84.01	8.93
+ PBE DINOv2	87.48	6.18
+ PBE perceptual	87.79	5.93
Ours	90.14	5.39

We further compared against DreamPaint

Source & Reference



PBE Best



DreamPaint



Diffuse to Choose (Ours)



Human Study

Table 3. The average results of the human study. Similarity evaluates the resemblance of the inpainted region to the reference image, while Semantic Blending assesses the accuracy of the reference image’s integration within its context.

Method	Similarity (\downarrow)	Semantic Blending (\downarrow)
PBE _{best}	3.7	3.13
DreamPaint [25]	2.83	2.53
Ours	2.9	2.5

