Generative Diffusion Models

Mehmet Saygin Seyfioglu

11/20/23

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

Taken from https://aws.amazon.com/tr/blogs/machine-learning/inpaint-images-with-stable-diffusion-using-amazon-sagemaker-jumpstart/

Outpainting

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

Outpainting

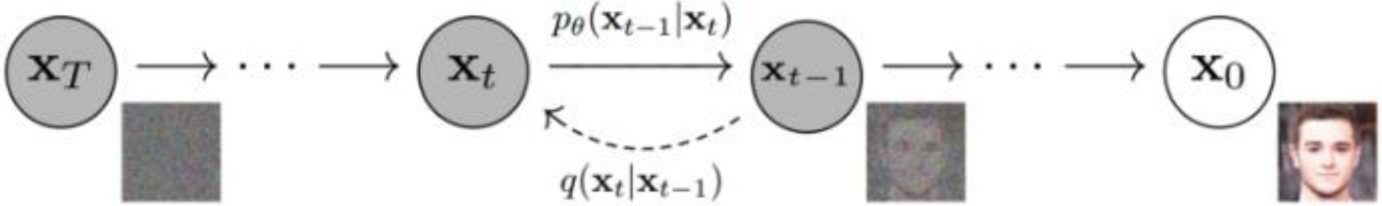
Taken from https://blog.segmind.com/exploring-the-magic-of-outpainting-with-stable-diffusion-uncropping-the-creative-possibilities/

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 noise and gradually decrease it.

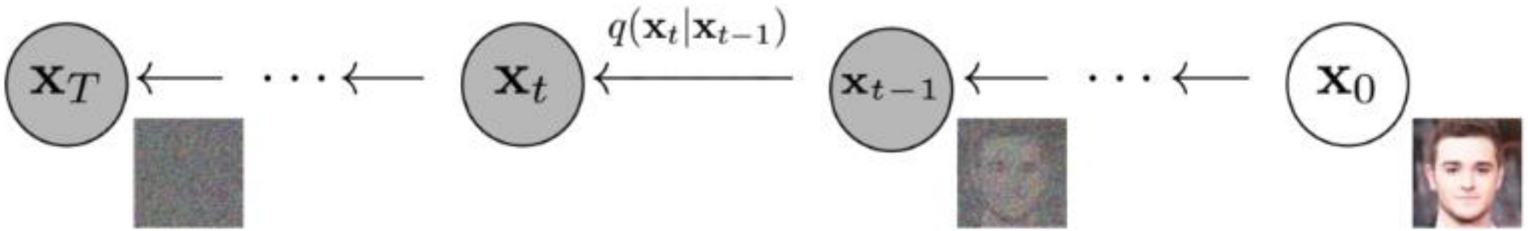


$q(x_t|x_{t-1})$ Pdf of an image at timestep t given image x_{t-1}

$p_\theta(x_{t-1}|x_t)$ Pdf of x_{t-1} given x_t parameterized by the model (theta)

Taken from https://learnopencv.com/denoising-diffusion-probabilistic-models/

The Forward Process



The probability density function

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

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

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

Adding noise

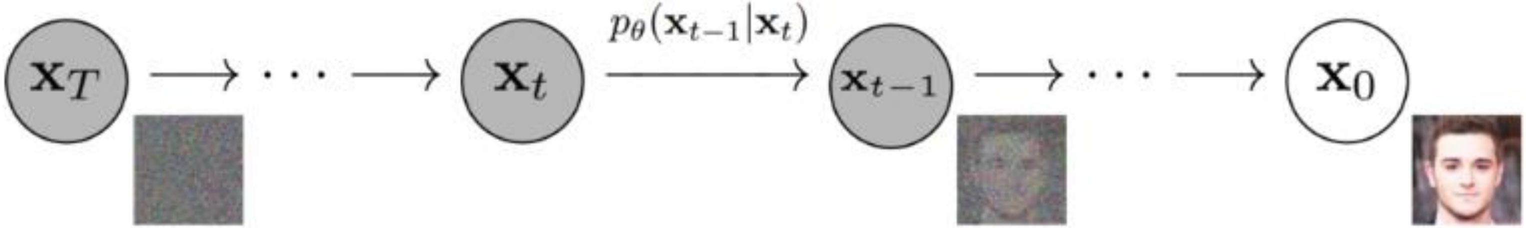
x_t is generated from x_{t-1} adding noise. In this way, starting from x0, the original image is iteratively corrupted from t=1…T

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

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

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



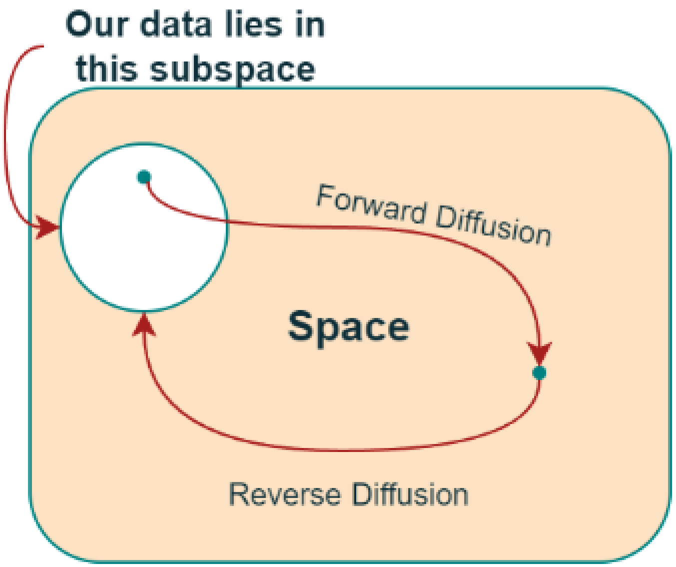
At step x_{t-1}, the network predicts the mean of the noise that is added at x_t

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

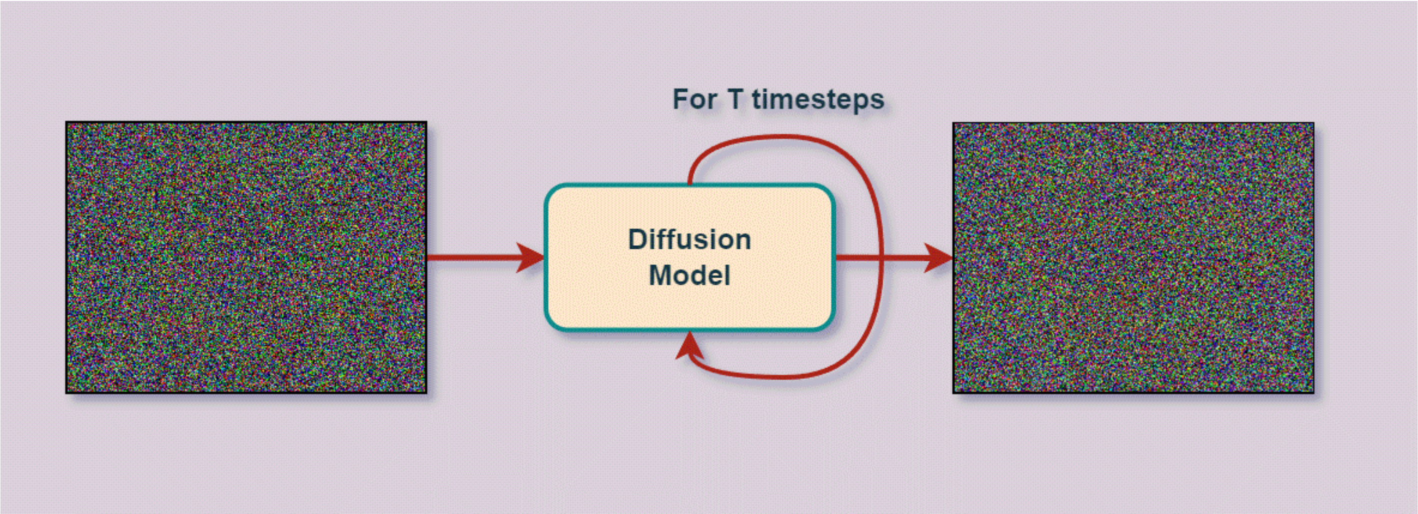
$$L_{\text{simple}} = E_{t,x_0,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right]$$

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020):

A high-level conceptual overview of the entire image space



After Learning The Model



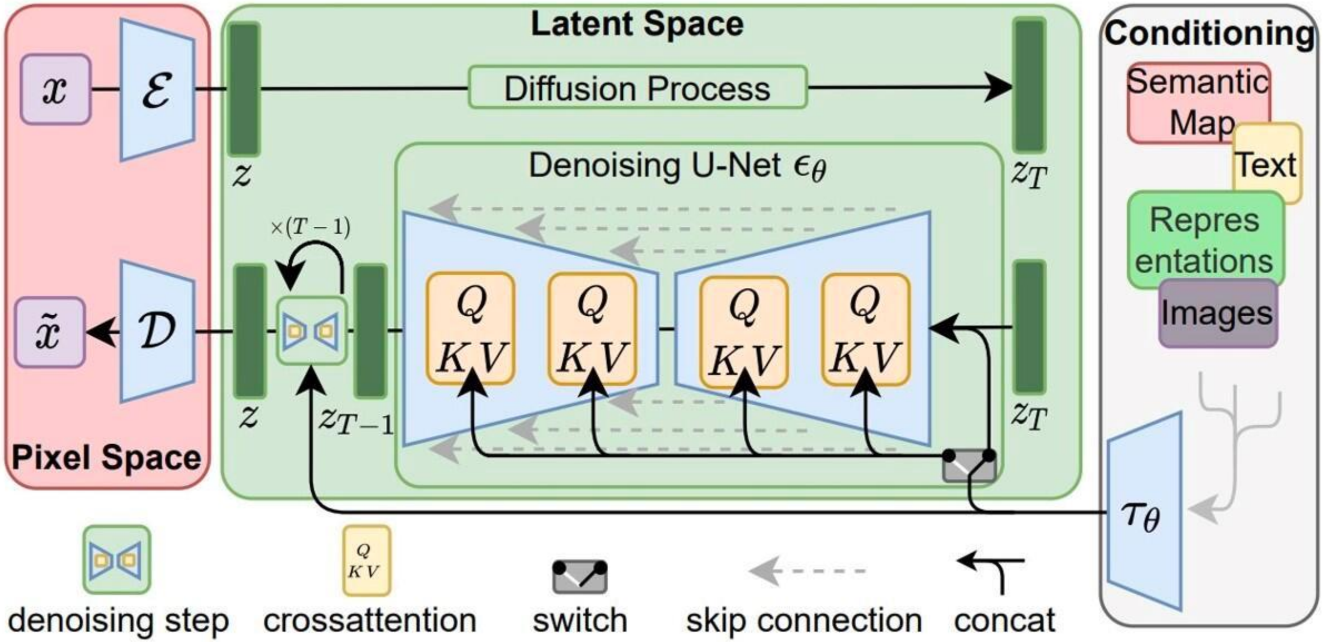
Conditional Diffusion

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)

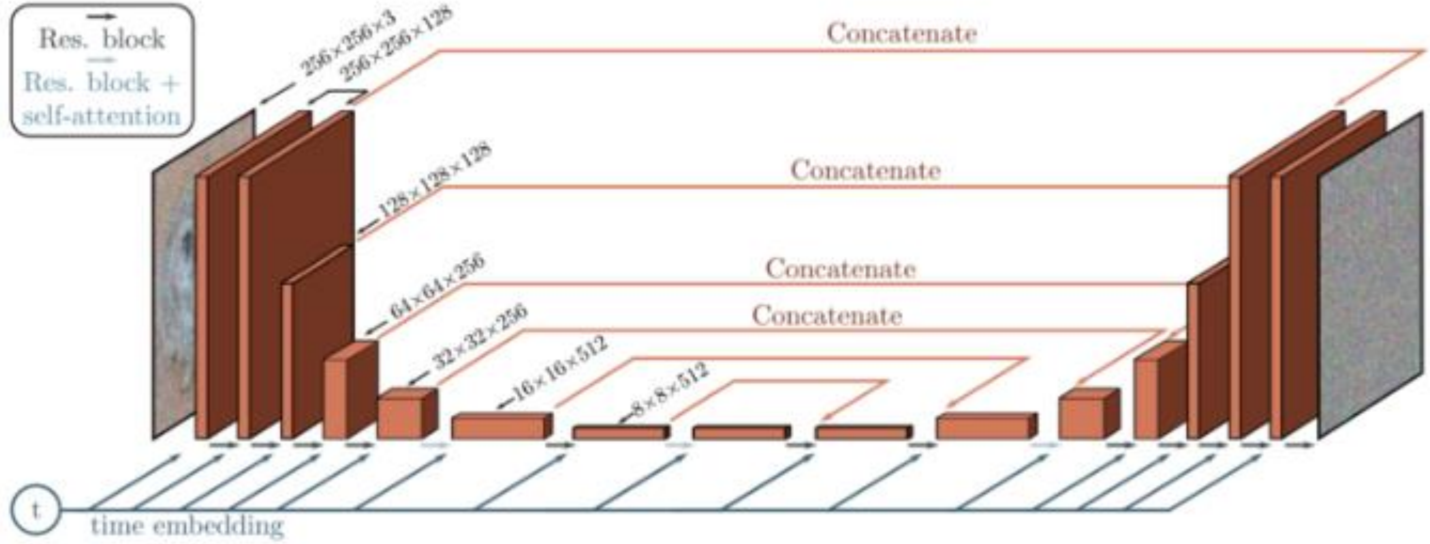
Also the CLIP model? That's a tool we could leverage

The Only Open-Source Diffusion Model: Stable Diffusion



Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Architecture



Stable Diffusion Examples

A Dog In A Hat Looking
Like A Vintage Portrait

A Giant Panda In
Between A Celestial War

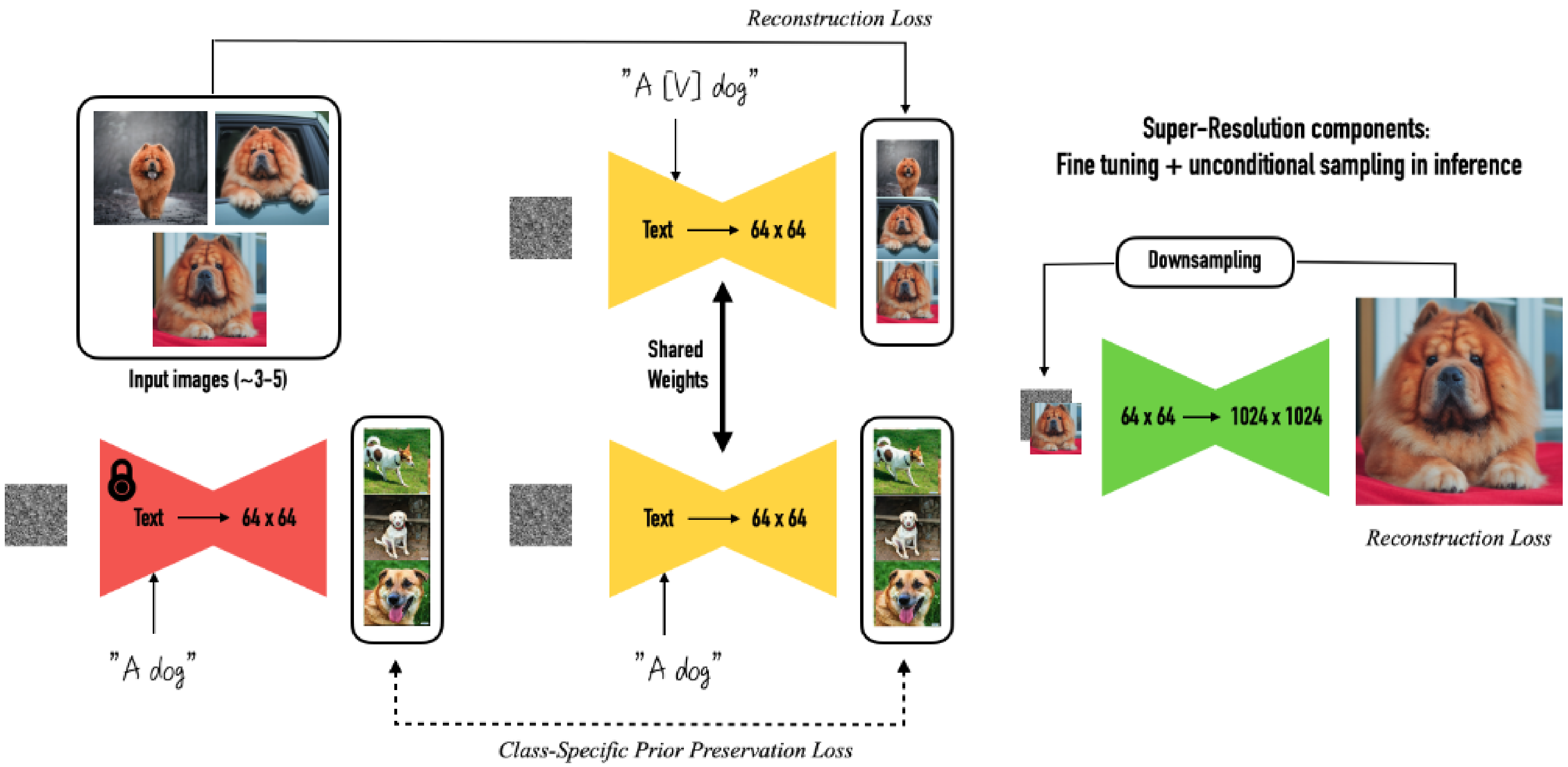
Some State of the Art Diffusion Applications

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

DreamBooth

Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

DreamBooth

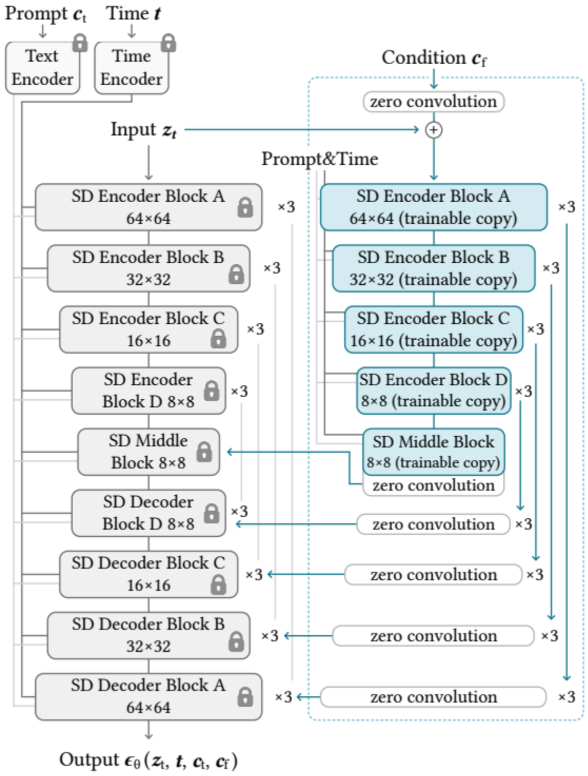


Reconstruction Loss
"A [V] dog"
Text → 64 x 64
Shared Weights
Input images (~3-5)
Text → 64 x 64
"A dog"
Text → 64 x 64
"A dog"
Class-Specific Prior Preservation Loss
Super-Resolution components:
Fine tuning + unconditional sampling in inference
Downsampling
64 x 64 → 1024 x 1024
Reconstruction Loss

ControlNet

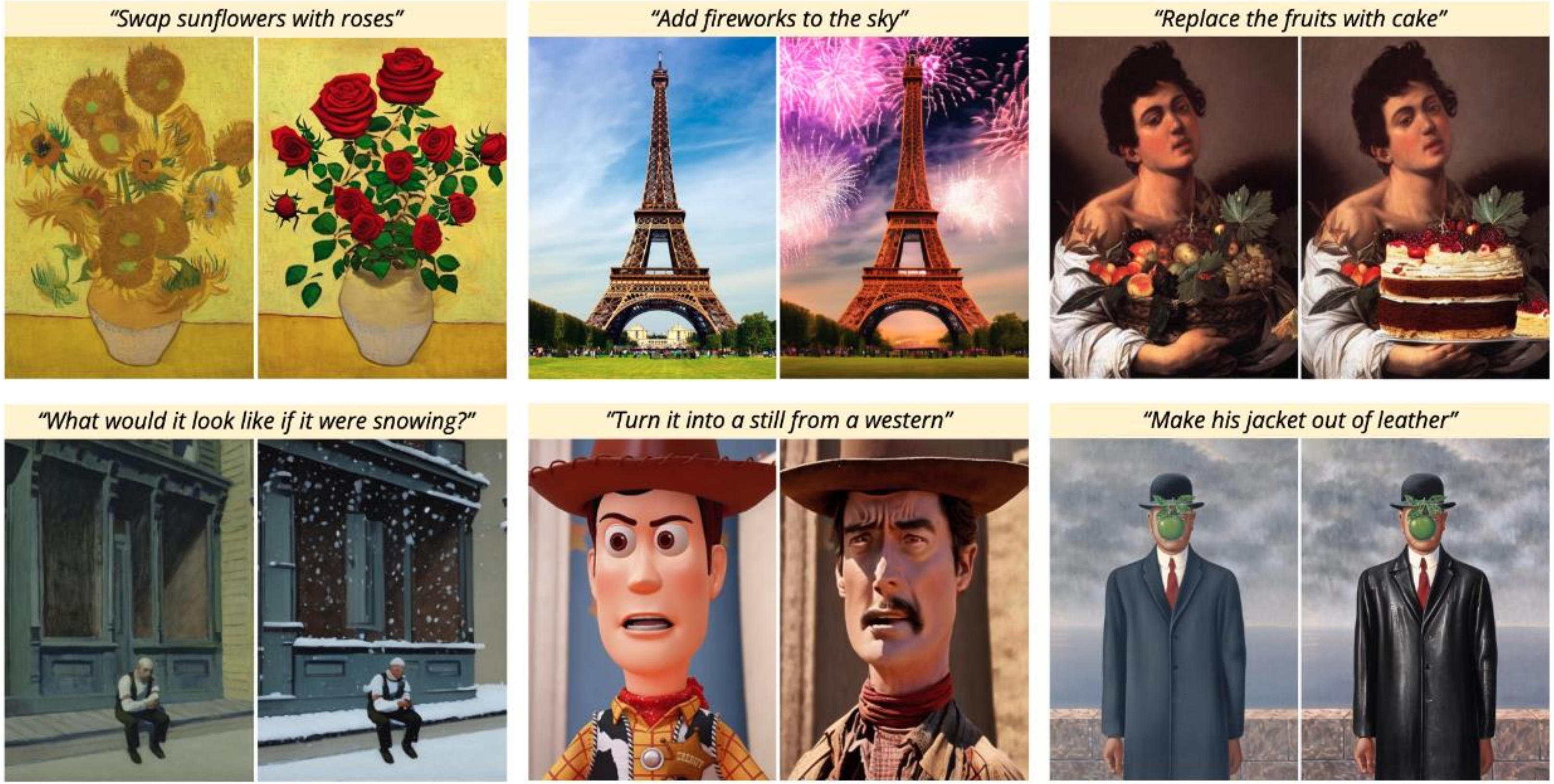
Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

ControlNet

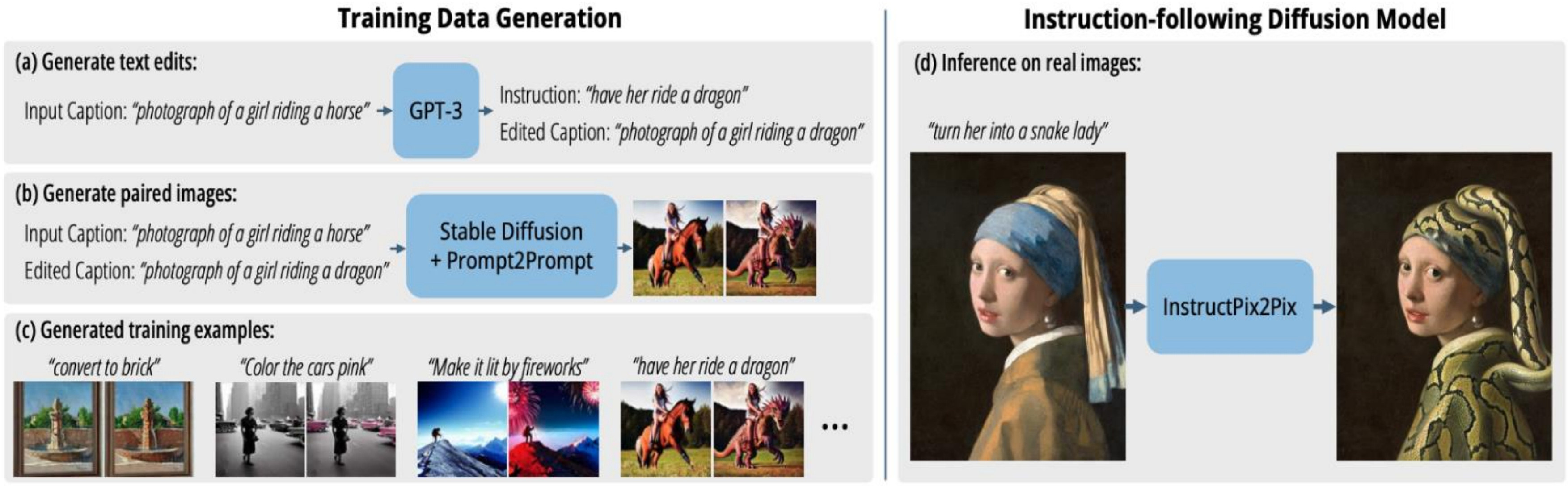


(a) Stable Diffusion (b) ControlNet

InstructPix2Pix

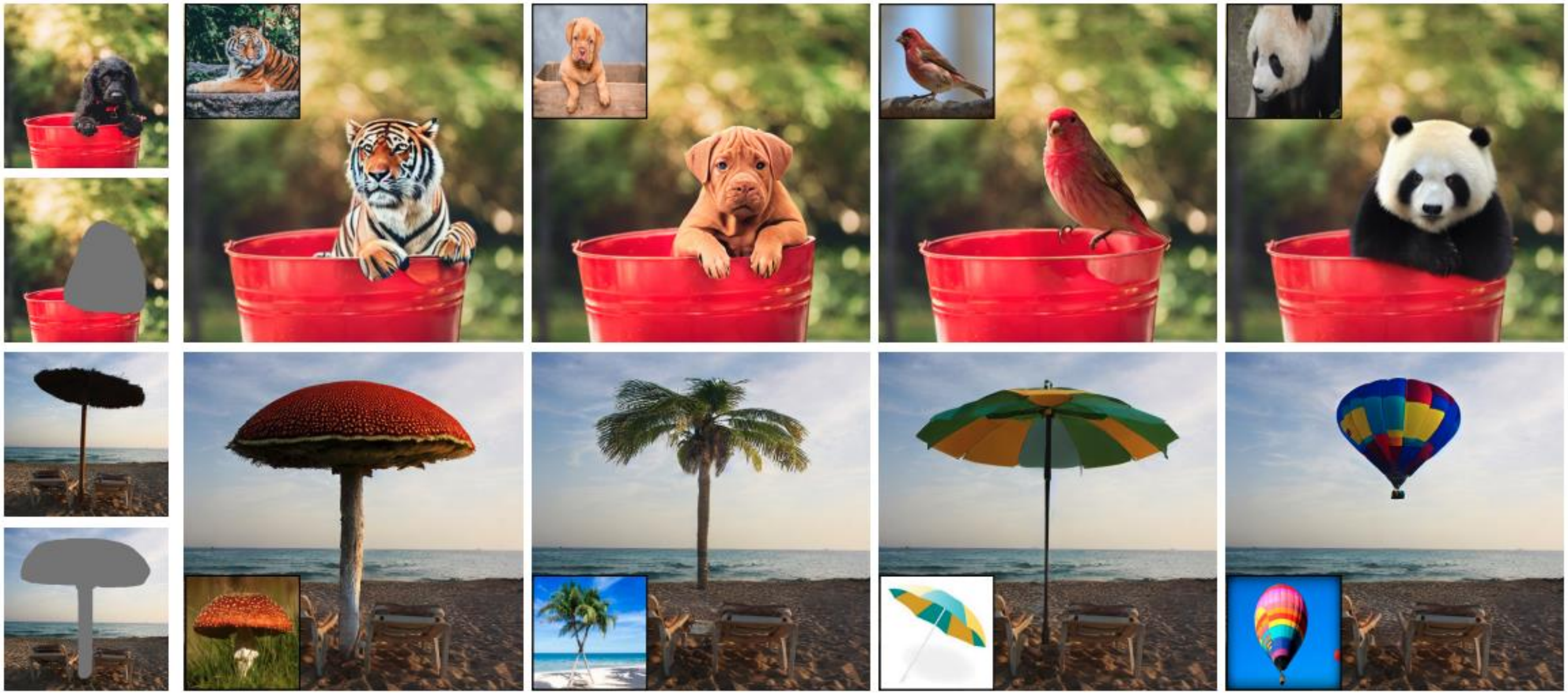


InstructPix2Pix



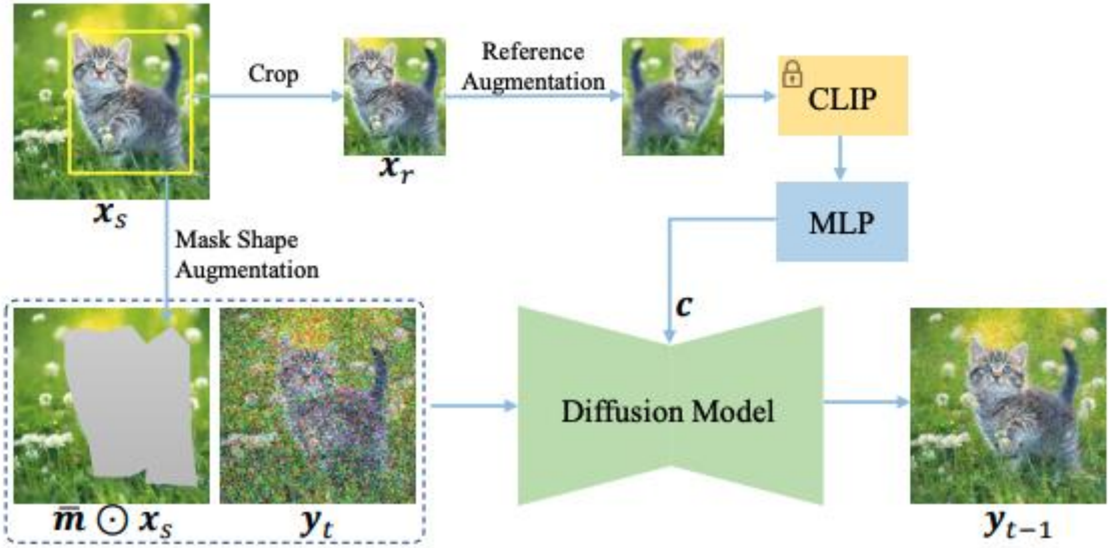
Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

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

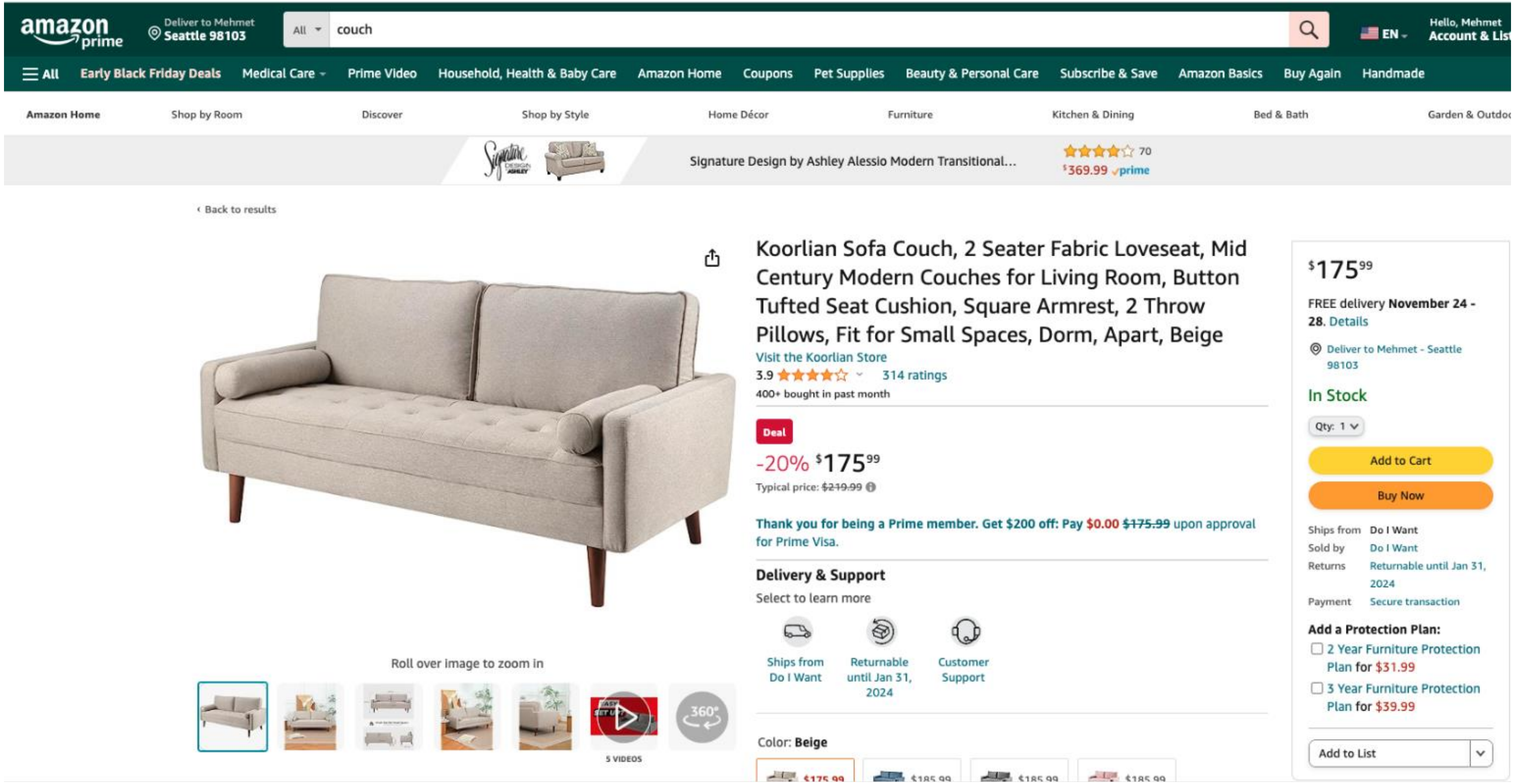


My Diffusion Research

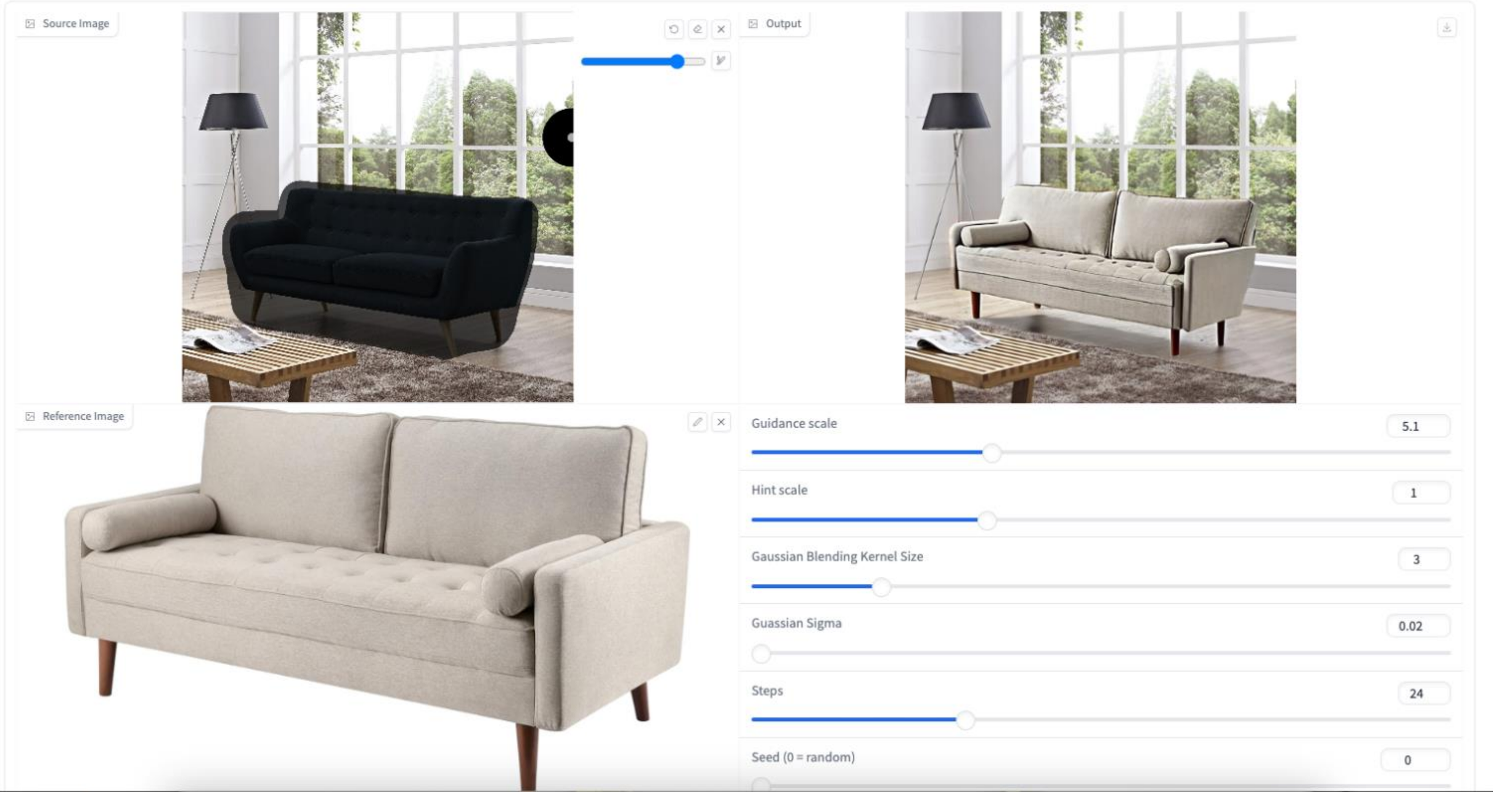
- I try to achieve Virtual Try-All

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

Virtual Try-All cont’d



Virtual Try-All cont'd



How?

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

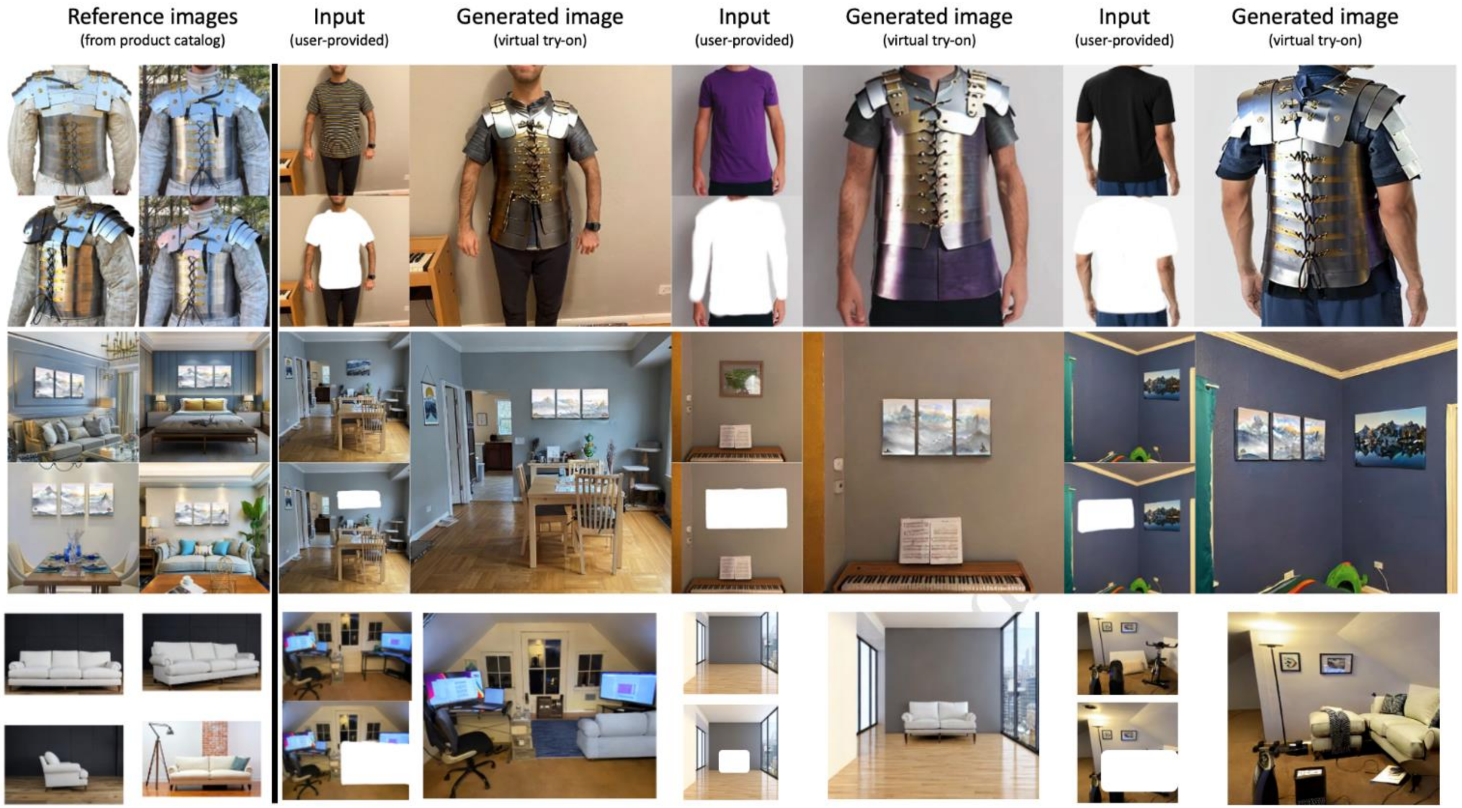
1. Operate in any 'in-the-wild' user image, and reference image,
2. Integrate the reference product harmoniously with the surrounding context while maintaining the product's identity
3. Perform fast inference to facilitate real-time usage across billions of products and millions of users.

DreamPaint

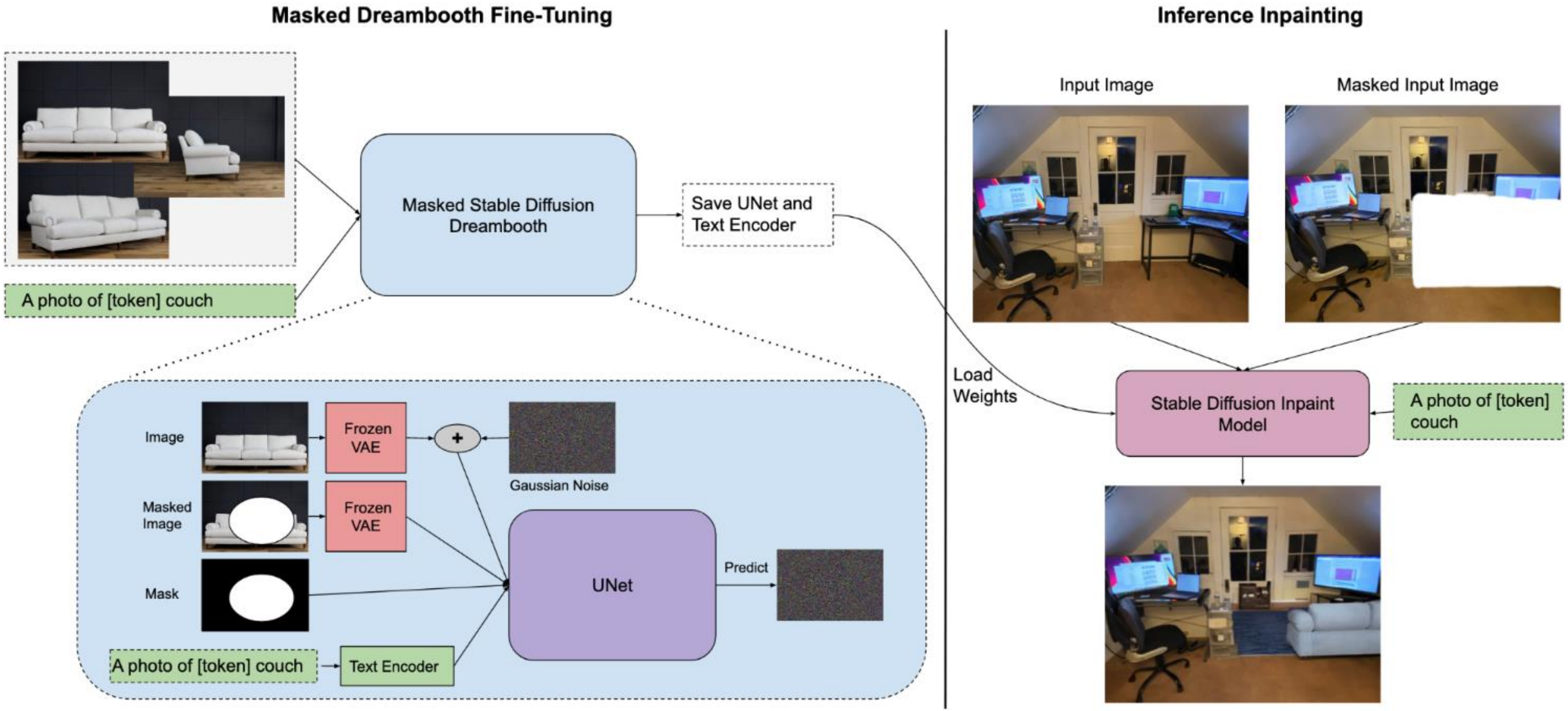
Previously, we implemented DreamPaint [1] (Dreambooth-Inpaint), which is a framework to intelligently inpaint any e-commerce product on any user-provided context image without requiring any expensive 3D AR/VR inputs.

[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

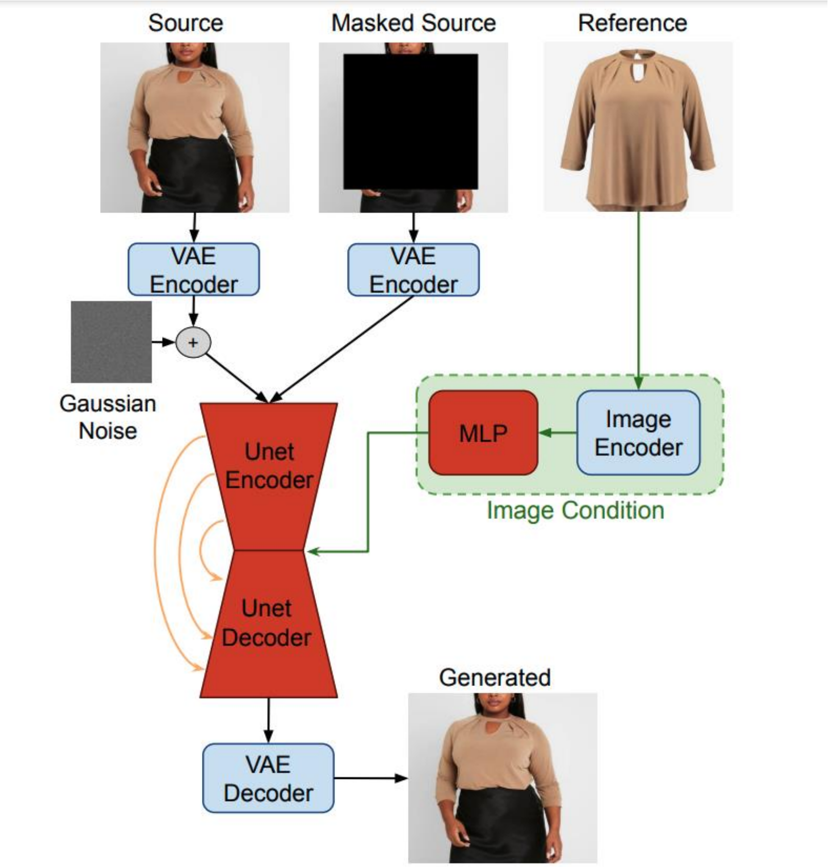


DreamPaint

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)

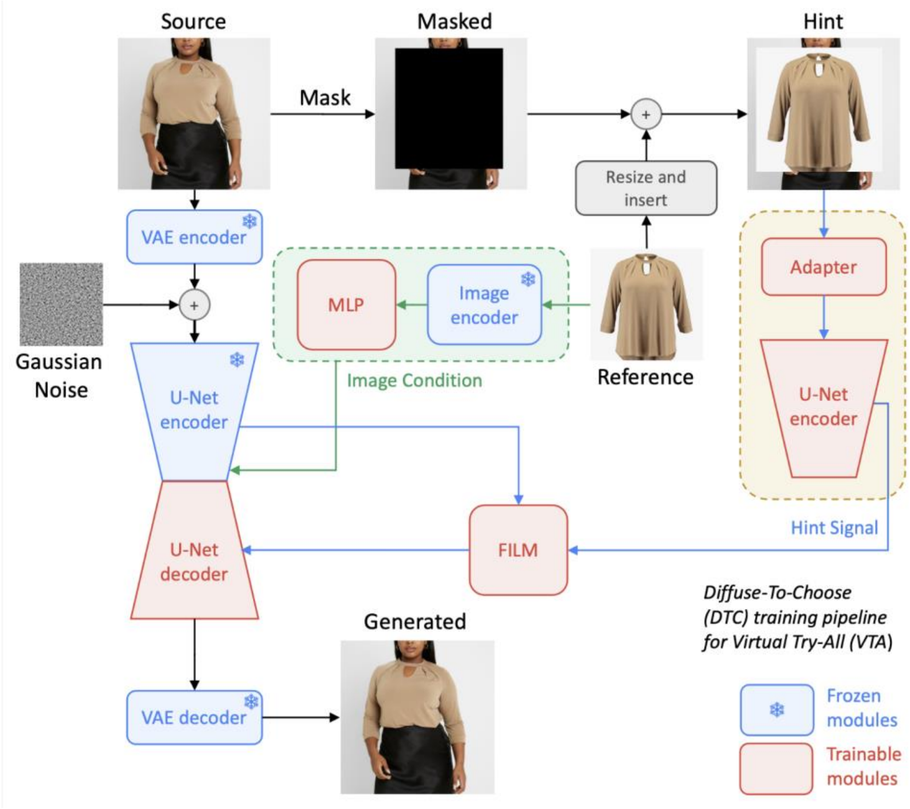
- For catalog items, we don't have to constraint ourselves with self-referencing.
  - Thus no need for the information bottleneck and aggressive augmentations.
- How far can we go with this approach?



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

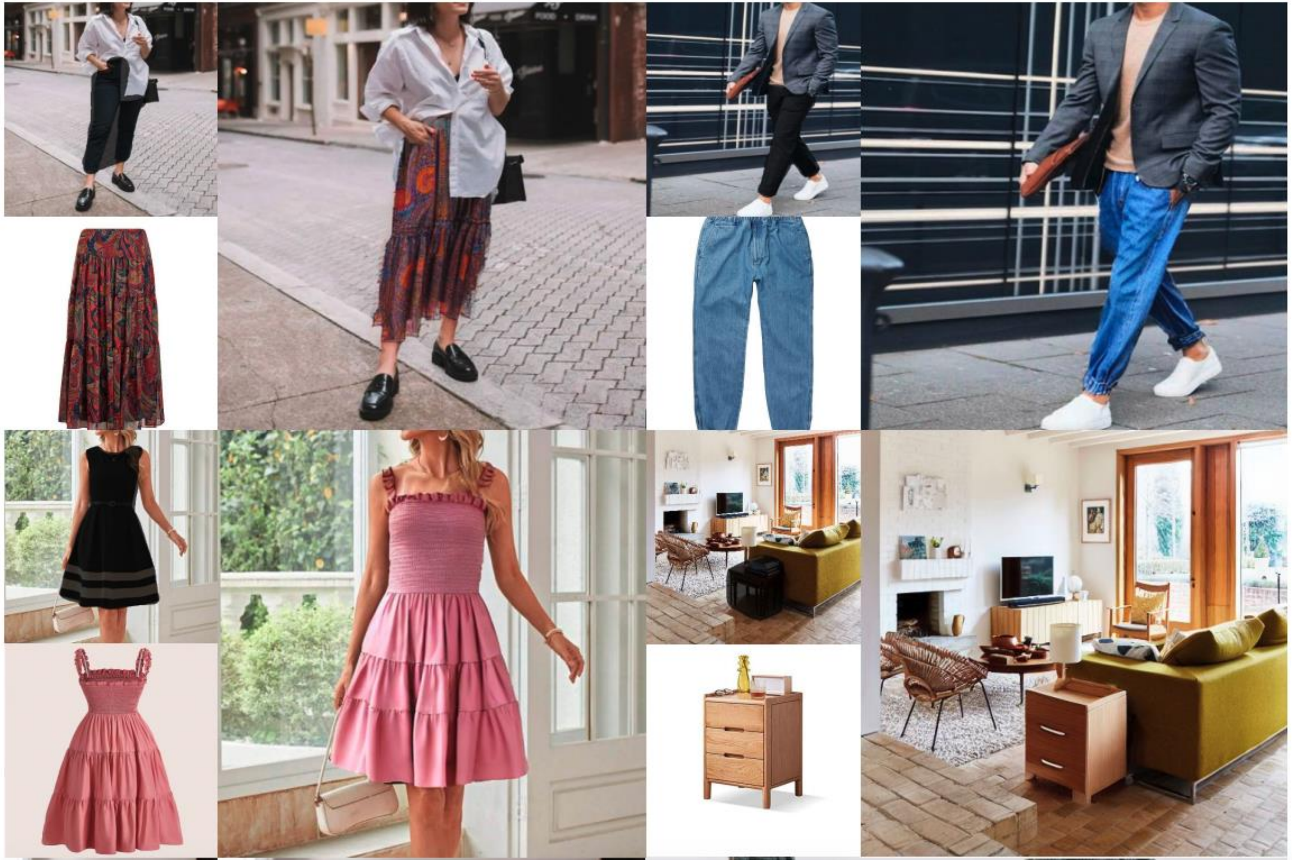


Diffuse to Choose

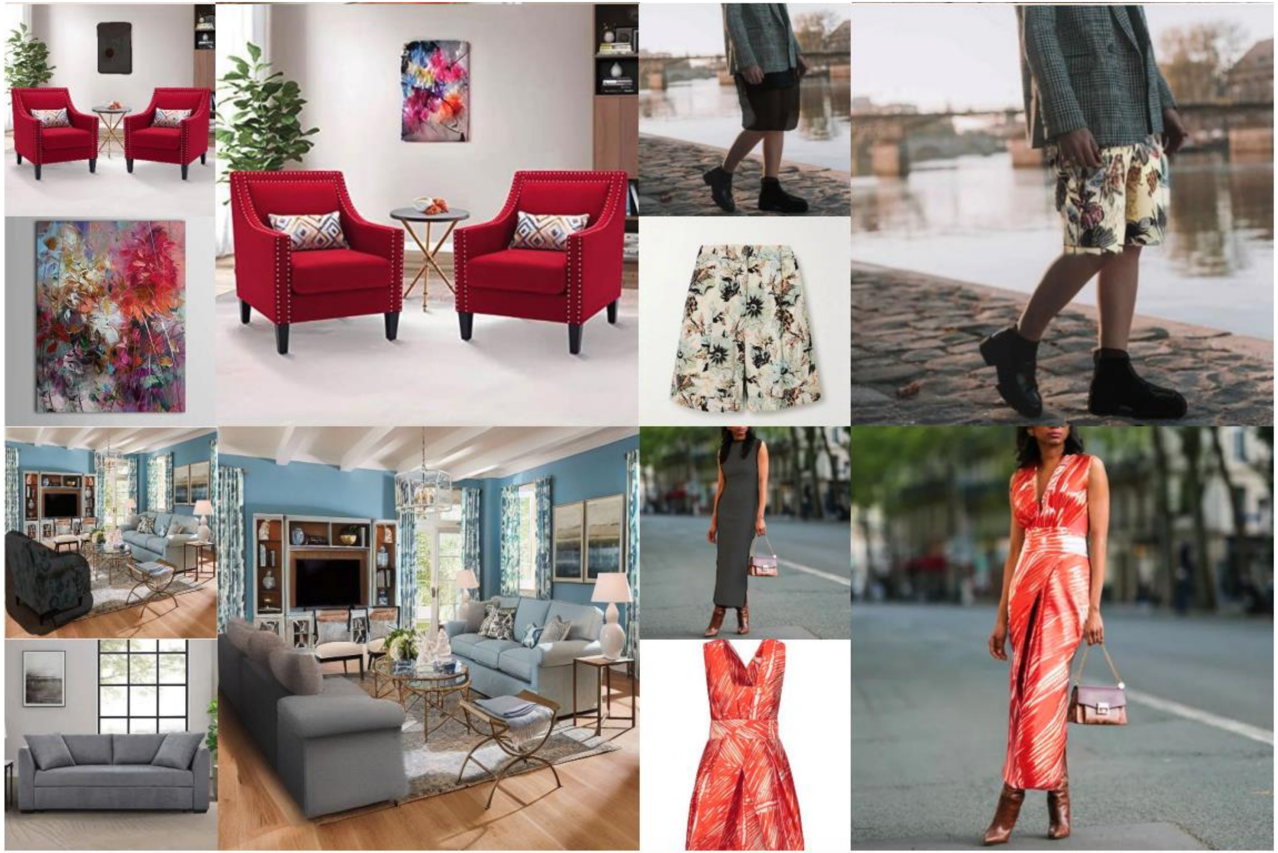
Demo

Visual Results

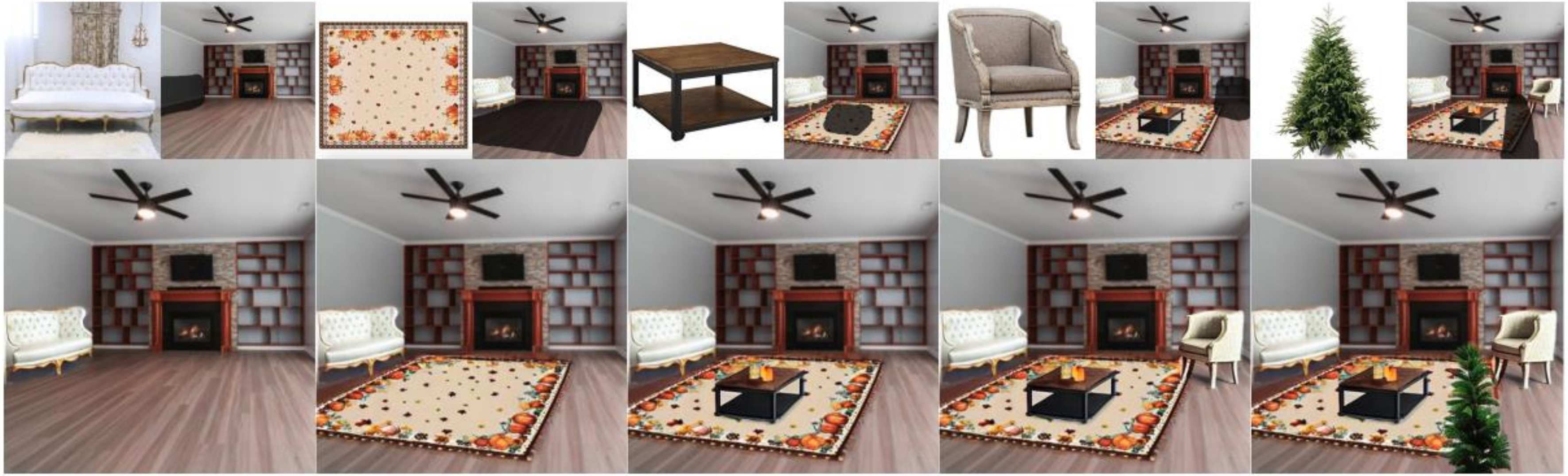
Visual Results



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 (↑)	FID (↓)
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 (↑)	FID (↓)
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

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 (↓)	Semantic Blending (↓)
PBE_{best}	3.7	3.13
DreamPaint [25]	**2.83**	2.53
Ours	2.9	**2.5**