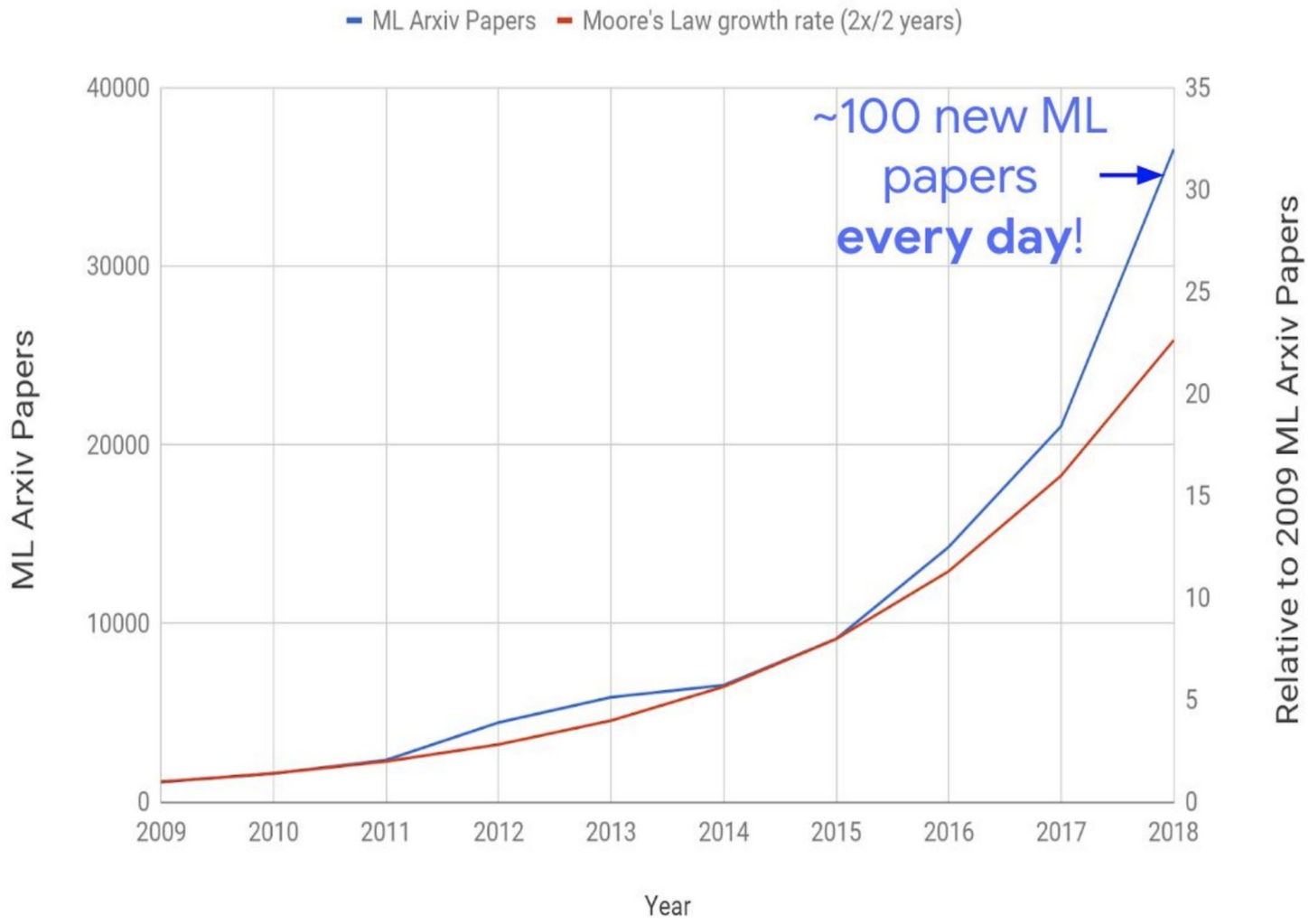
Papers With Jay

# eDiff-I: Text to Image Diffusion Models with Ensemble of Expert Denoisers




- Papers with Jay
  - Goals - Provide intuitive summary of state of the art papers in machine learning, and related fields.

Papers With Jay

# eDiff-I: Text to Image Diffusion Models with Ensemble of Expert Denoisers

Papers With Jay

# eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis,
Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, Ming-Yu Liu

NVIDIA Corporation

{ybalaji, snah, xunh, avahdat, jiamings, kkreis, maittala, taila, slaine, bcatanzaro, tkarras, mingyul}@nvidia.com

Archive Link

# Background - Diffusion Model

Papers With Jay

## Q: What is a diffusion model?



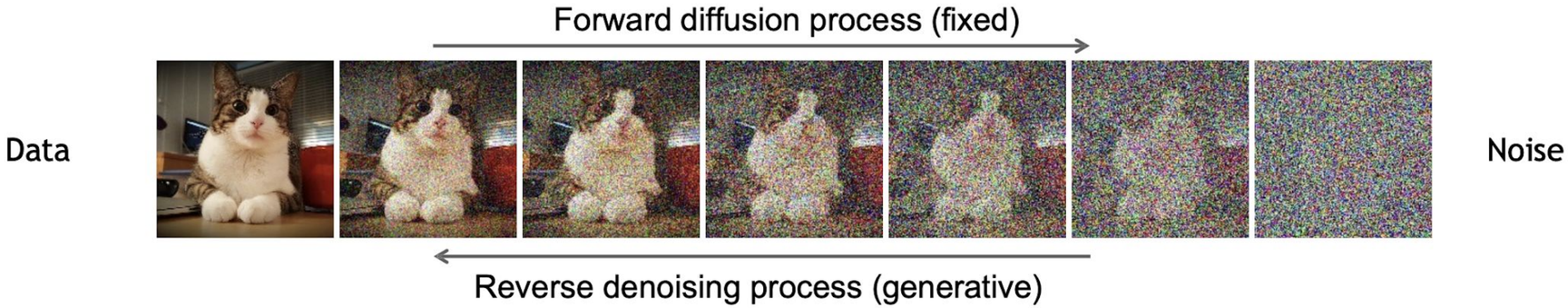

Slide Credits - Roni Sengupta Fall 2022 Lectures

Papers With Jay

# Background - Denoising Diffusion Models

Denoising diffusion models consist of two processes:

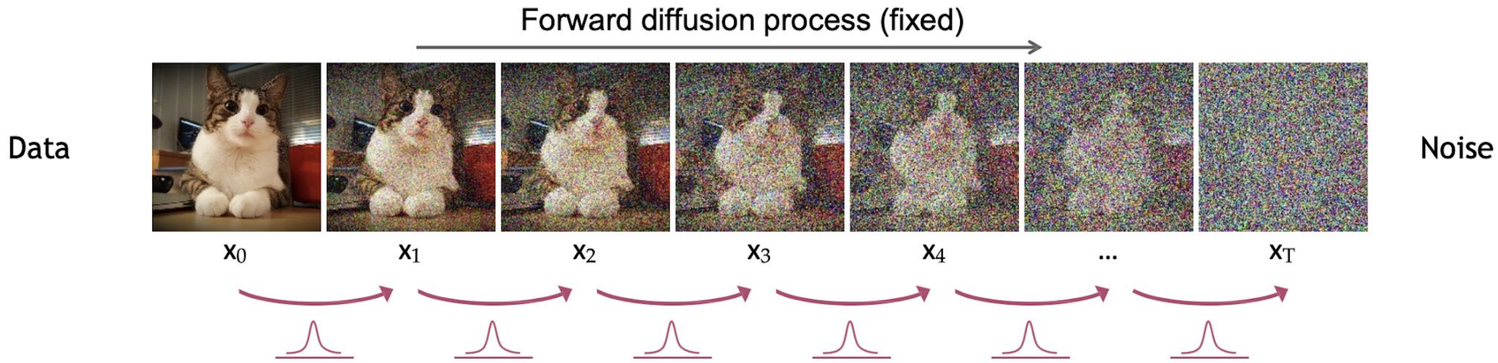
- Forward diffusion process that gradually adds noise to input
- Reverse denoising process that learns to generate data by denoising



# Background - Denoising Diffusion Model

## Forward Diffusion Process

The formal definition of the forward process in T steps:




$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x_t}; \sqrt{1-\beta_t}\mathbf{x_{t-1}}, \beta_t\mathbf{I}) \quad \Rightarrow \quad \text{Sample: } x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_{t-1}$$

mean variance

where, $\epsilon_{t-1} \sim \mathcal{N}(0, \mathbf{I})$

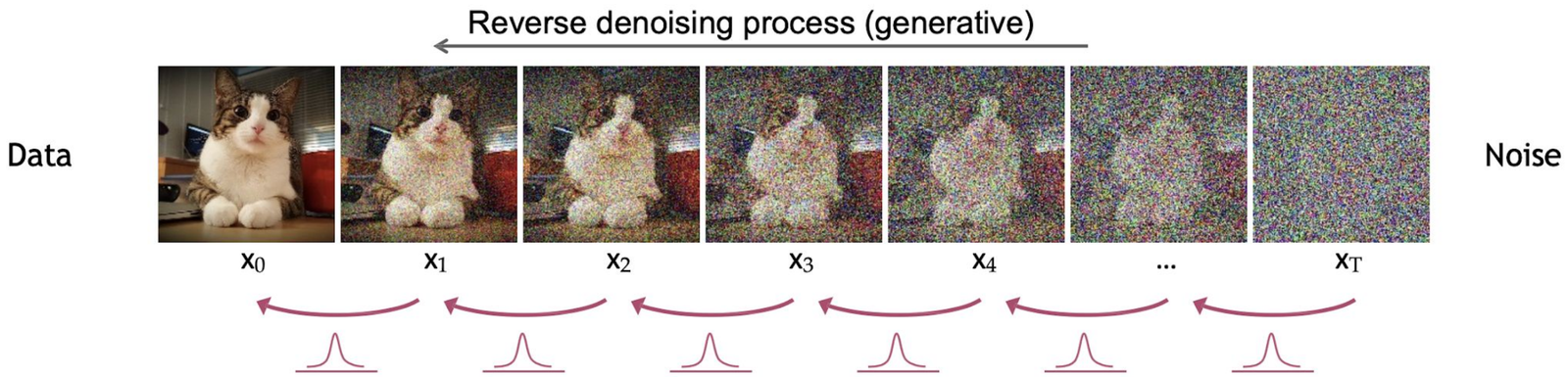
$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\,\epsilon \qquad \alpha_t := 1-\beta_t \text{ and } \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$$

Noisy image $\mathbf{x}_t$ can be represented in terms of $\mathbf{x}_0$

# Background - Diffusion Model

## Reverse Denoising Process

Formal definition of forward and reverse processes in T steps:




$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\mu_\theta(\mathbf{x}_t, t)}, \sigma_t^2\mathbf{I})$$

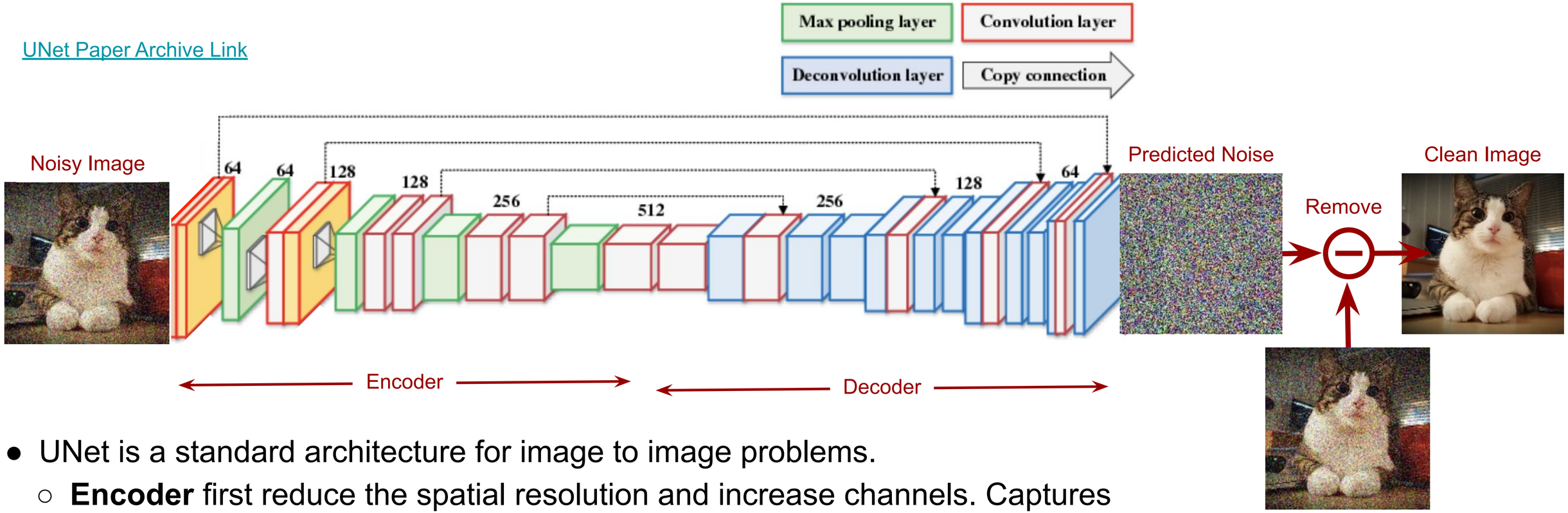
Trainable network
(U-net, Denoising Autoencoder)

Given noisy images, train a UNet style model to recover the noiseless image.

# UNet Architecture

Papers With Jay

[UNet Paper Archive Link]




- UNet is a standard architecture for image to image problems.
  - **Encoder** first reduce the spatial resolution and increase channels. Captures the useful information in the images.
  - Now **decoder** creates image from encoder output.
  - **Skip connections** pass information from encoder to decoder at the same resolution.
- Diffusion models use UNet to predict noise in noisy image.
  - Removing this noise from noisy image gives the clear image.

# Training Diffusion Models

What is the loss function? (Ho et al. NeurIPS 2020 )

$$L_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}(1,T)} \left[ ||\epsilon - \epsilon_\theta(\underbrace{\sqrt{\bar{\alpha}_t}\, \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon}_{\mathbf{x}_t}, t)||^2 \right]$$

1. Pick clean image (x_0) from batch.
2. Decide number of steps t of noise to apply.
3. Compute noisy image (x_t) using closed form expression.
4. Calculate loss above, compute gradients and do gradient descend to calculate UNet weights.

**Algorithm 1** Training

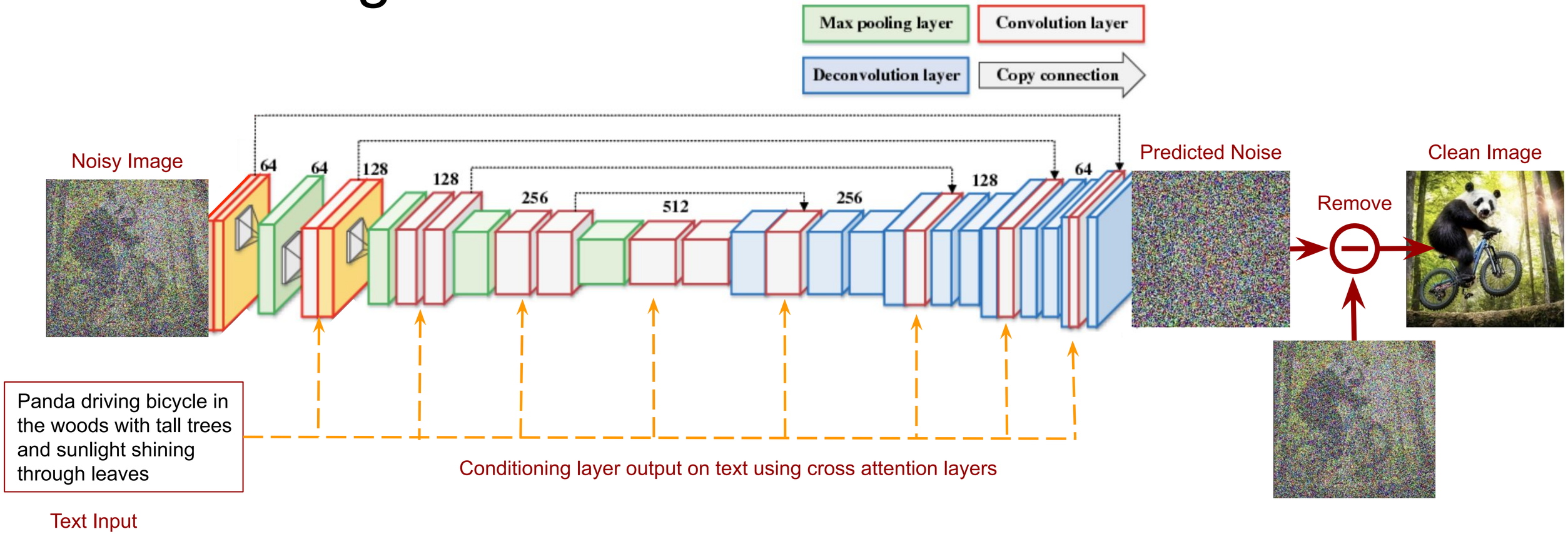
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\underbrace{\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}}_{\mathbf{x}_t}, t) \right\|^2$$
6: **until** converged

# Inference

1. Goal generate noise less image.
2. First generate noisy image from Gaussian distribution.
3. Apply the trained UNet to predict the noise in the image,
   a. Subtract predicted noise from noisy image.
4. Repeat step 3 T times.

$$
\begin{aligned}
&1: \ \mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
&2: \ \textbf{for } t = T, \ldots, 1 \textbf{ do} \\
&3: \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
&4: \quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z} \\
&5: \ \textbf{end for} \\
&6: \ \textbf{return } \mathbf{x}_0
\end{aligned}
$$

# Text to Image Models




- Condition diffusion models on text input to predict images depicting the text.
- Conditioning done using cross attention layers applied between UNet output and input text embedding.

# Back to the paper - What can it do?

## Complex text to photo realistic images


*A highly detailed digital painting of a portal in a mystic forest with many beautiful trees. A person is standing in front of the portal.*


*A highly detailed zoomed-in digital painting of a cat dressed as a witch wearing a wizard hat in a haunted house, artstation.*


*An image of a beautiful landscape of an ocean. There is a huge rock in the middle of the ocean. There is a mountain in the background. Sun is setting.*

# What can it do?

## Complex text and style image to photo realistic images

Style reference

*A photo of a duckling wearing a medieval soldier helmet and riding a skateboard.*

# What can it do?

## Complex text and layout image to photo realistic images


*A digital painting of a half-frozen lake near mountains under a full moon and aurora. A boat is in the middle of the lake. Highly detailed.*

# Paper Contributions (1/3) - Ensemble of Expert DMs

Papers With Jay

- Key observation - Diffusion model behaves differently at different noise levels.
  - At initial stage (high noise in input visual data), diffusion model uses text prompt for denoising.
  - At later stages, diffusion model uses visual features for denoising, and ignores text prompt.
- Hence authors trains different diffusion models each specializing for  a different noise levels.
- No additional computational cost during inference.

# Empirical Proof 1

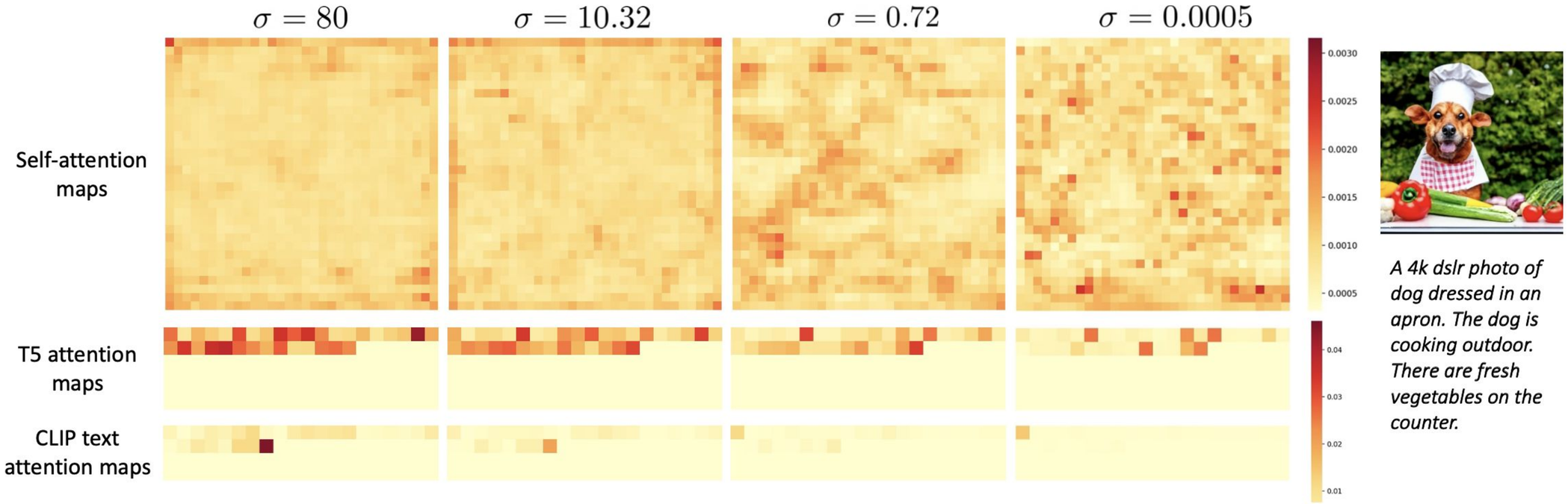



Figure 3. Visualization of attention maps at different noise levels. We plot the self-attention heat maps for visual features (top). Each value indicates how often the visual feature at this location is used and is the result of averaging over all attention queries. We also plot the cross-attention heat maps (bottom), where each value indicates how often the corresponding text token is used. We plot the cross-attention heat maps for both the T5 text tokens and the CLIP text tokens. Note that our cross-attention layer also includes a null token, which is not shown in the figure. When the attention values for all text tokens are low, the null token is mostly attended. The figure shows that the text attention value is strong at higher noise levels where the core image formation occurs. In these regions, the model relies on the text to generate a rough layout consistent with the text description. On the other hand, at lower noise levels, the text attention value is weak because the images are mostly formed at this point. The models need not rely on text information for denoising. Conversely, in the case of self-attention maps, the attention values are evenly distributed at higher noise levels, as the image content in these regions is not informative. At lower noise levels, the attention maps exhibit patterns better correlated with the image content.

# Empirical Proof 2 - Switching text at different noise levels

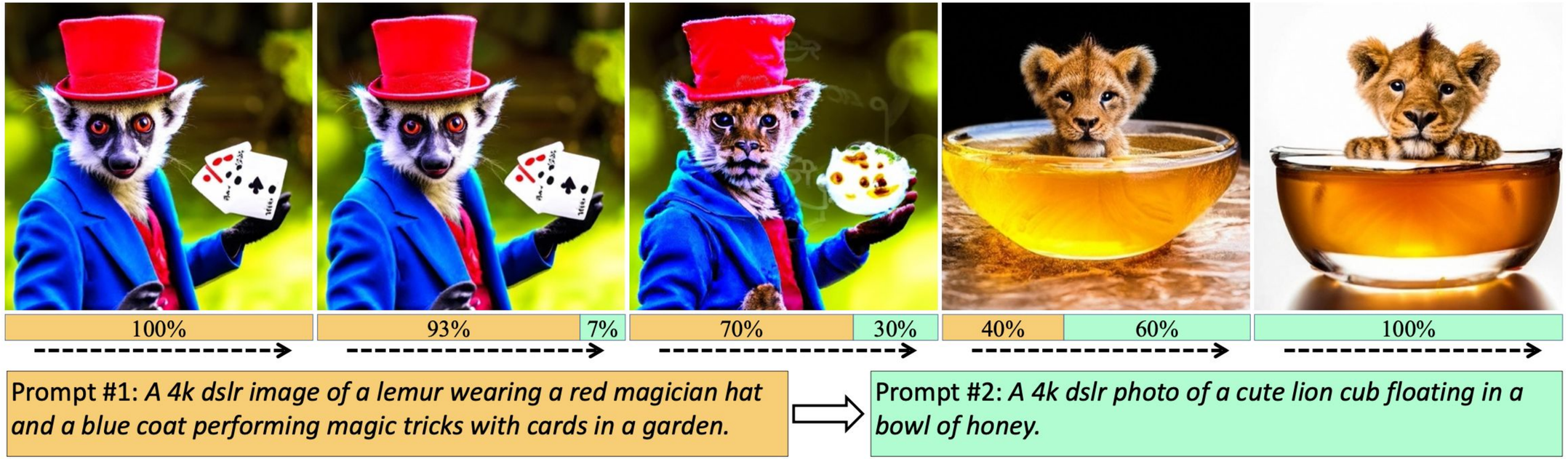



Figure 4. Impact of prompt switching during iterative denoising. We change the input text to Prompt #2 after a fixed percentage of denoising steps have been performed using Prompt #1. From left to right, the 5 images are produced with different transition percentages, which are 0%, 7%, 30%, 60%, and 100%, as visualized in the figure. Comparing the first and second outputs, we note that the text inputs have no visible impact on the output when used in the last 7% of denoising, which suggests text prompt is not used at the end of iterative denoising. The third output shows influences from both prompts, where the lemur and cards are replaced with lion and honey, respectively. The fourth output suggests the text input in the first 40% of denoising is overridden by the text input in the remaining 60%. From these results, we find the denoiser utilizes text input differently at different noise levels.

# Paper Contributions (2/3) - Ensemble of Encoders

Papers With Jay

- Authors use ensemble of encoders to provide information to the diffusion model
  - T5 encoder on text prompt.
  - CLIP text encoder.
  - CLIP image encoder on style and layout images.
- Key observations - different encoders capture different information.
  - CLIP text encoder captures the global layout of image.
  - T5 text encoder captures fine-grained text details.
  - CLIP image encoder captures style and layout information in images.

# Paper Contributions (3/3) - Paint with words input

Papers With Jay

- Allows users to provide a layout image with words written in it to condition the denoising process using a cross attention layer.


*A digital painting of a half-frozen lake near mountains under a full noon and aurora. A boat is in the middle of the lake. Highly detailed.*

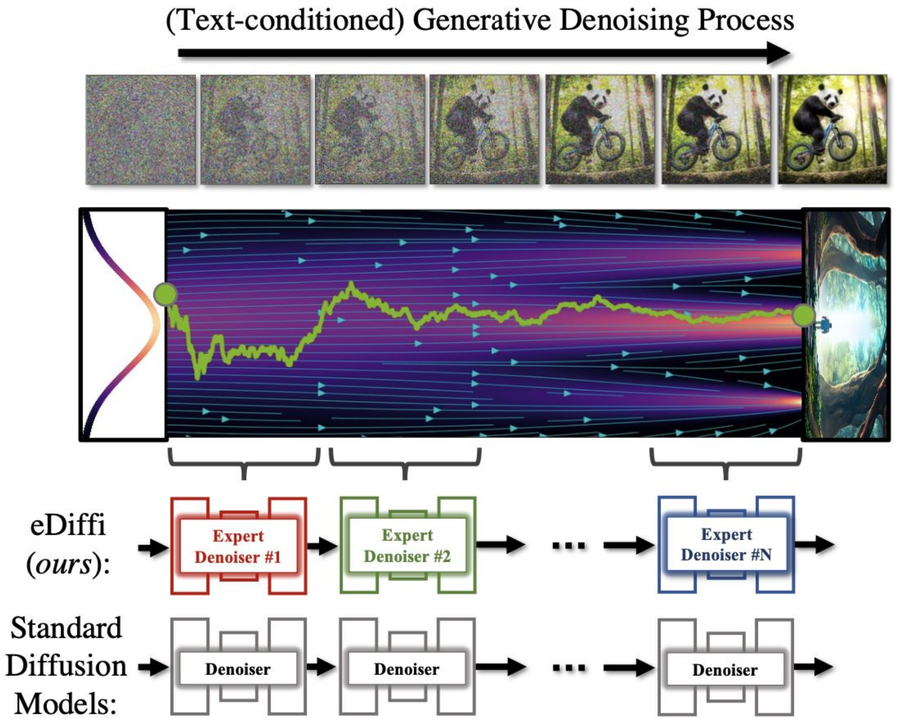
# Ensemble of Expert Denoisers

Papers With Jay

- Neural network D progressively de noise a Gaussian noise image to generate the output image.
- Denoiser at each noise level uses the input noisy image and the text prompt to generate output.
- Key observation - Diffusion model behaves differently at different noise levels.
  - At initial stage (high noise in input visual data), diffusion model uses text prompt for denoising.
  - At later stages, diffusion model uses visual features for denoising, and ignores text prompt.

# Ensemble of Expert Denoisers

- Existing papers learn a single diffusion model for all noise levels.
- Authors argue that this leads to a low capacity model as it has to learn the behavior at all noise levels.
- Instead authors propose learning different DM for different noise ranges.

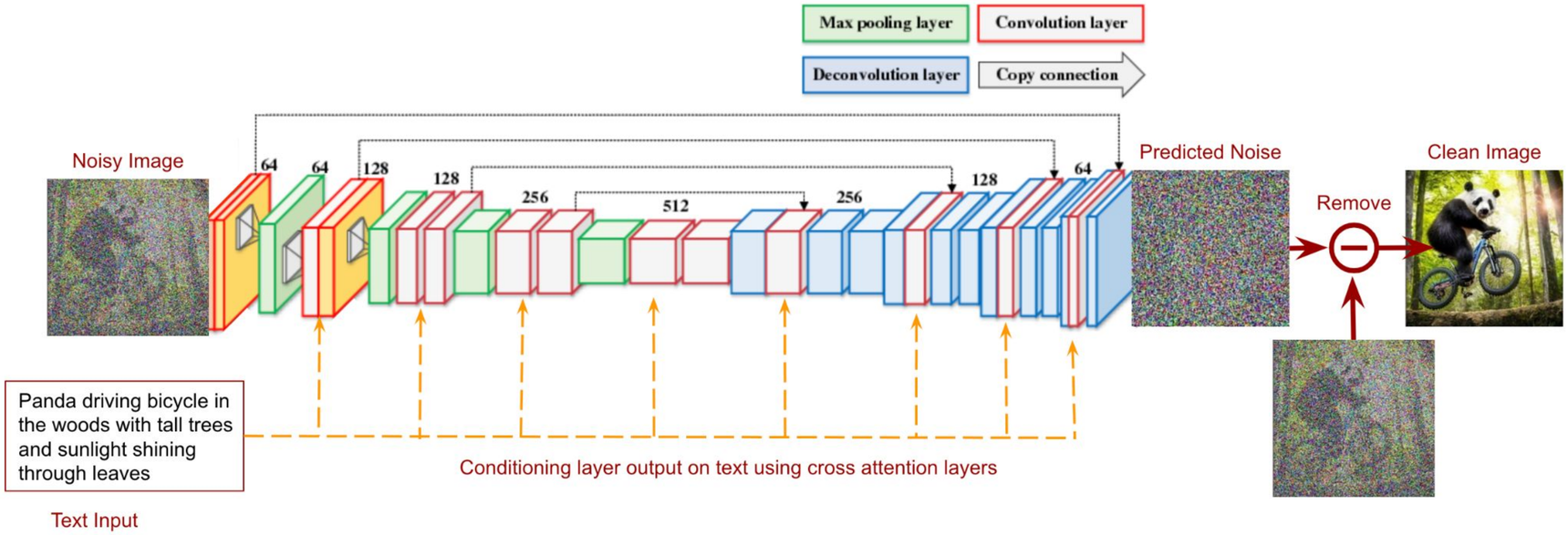


# Ensemble of Expert Denoisers

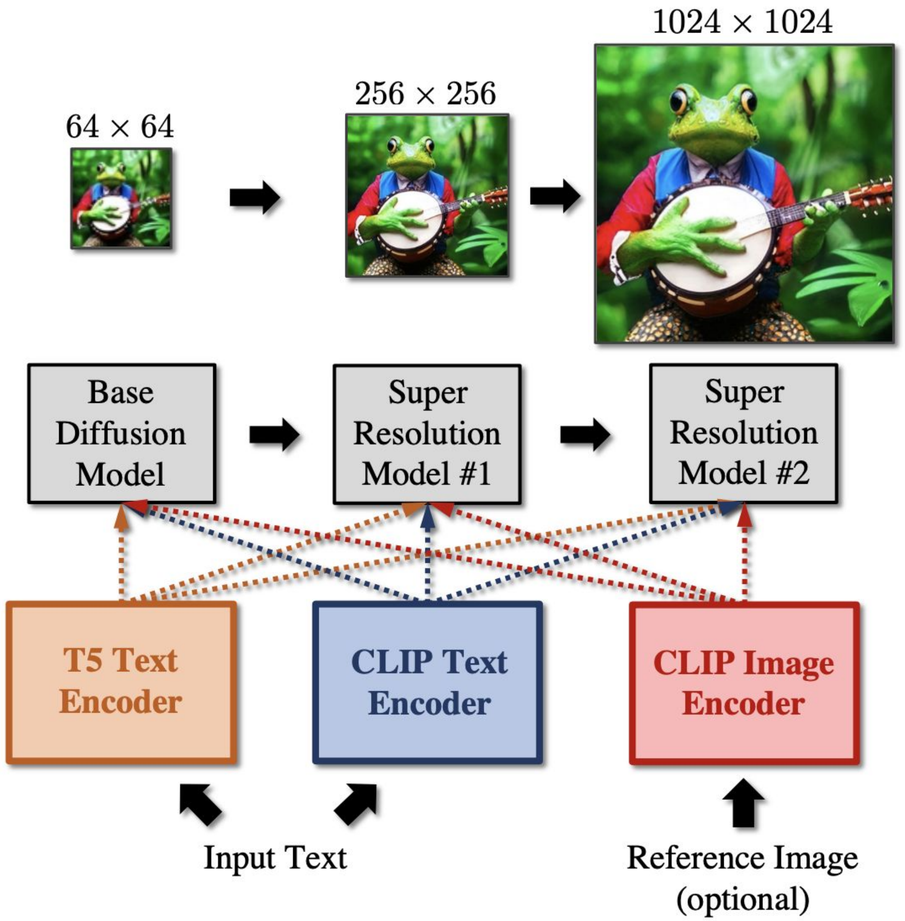
- No increase in inference complexity.
- To reduce training complexity
  - First train a shared model across all noise levels.
  - Now fine-tune different expert models on different noise ranges.
- Do this in a binary tree fashion
  - After models of level l are trained, fine-tune models are level l+1.
- Authors use 3 experts in the end
  - One DM at low noise level, another at high noise level
  - Third DM for intermediate noise levels.

# Multiple Conditional Inputs

- Use multiple embeddings to condition
  - T5 text embedding, CLIP text embedding and CLIP image embedding.
- Precompute embedding on dataset to speed up training.
- Cross attention is computed between the conditioning signal and the output of the UNet Diffusion model.



# Generating high resolution images




- Use cascade of Diffusion Models with UNet architecture

# Paint with words


*A digital painting of a half-frozen lake near mountains under a full moon and aurora. A boat is in the middle of the lake. Highly detailed.*

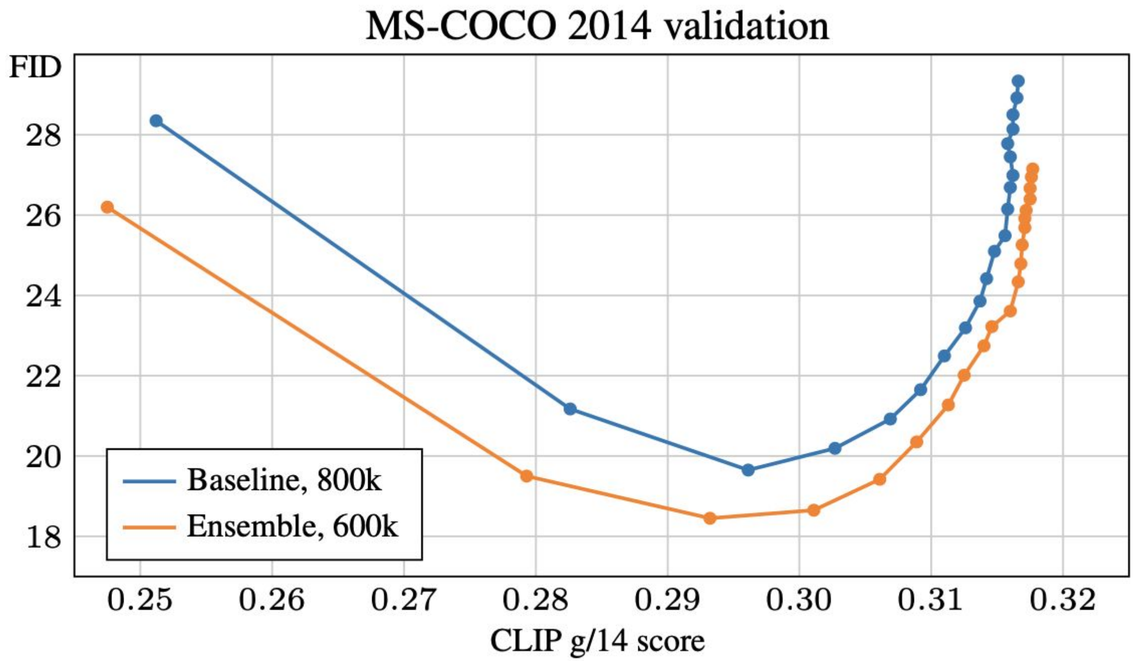
- Compute cross attention between the layout image and text tokens.
- Encourage higher attention between matching text and layout image regions.

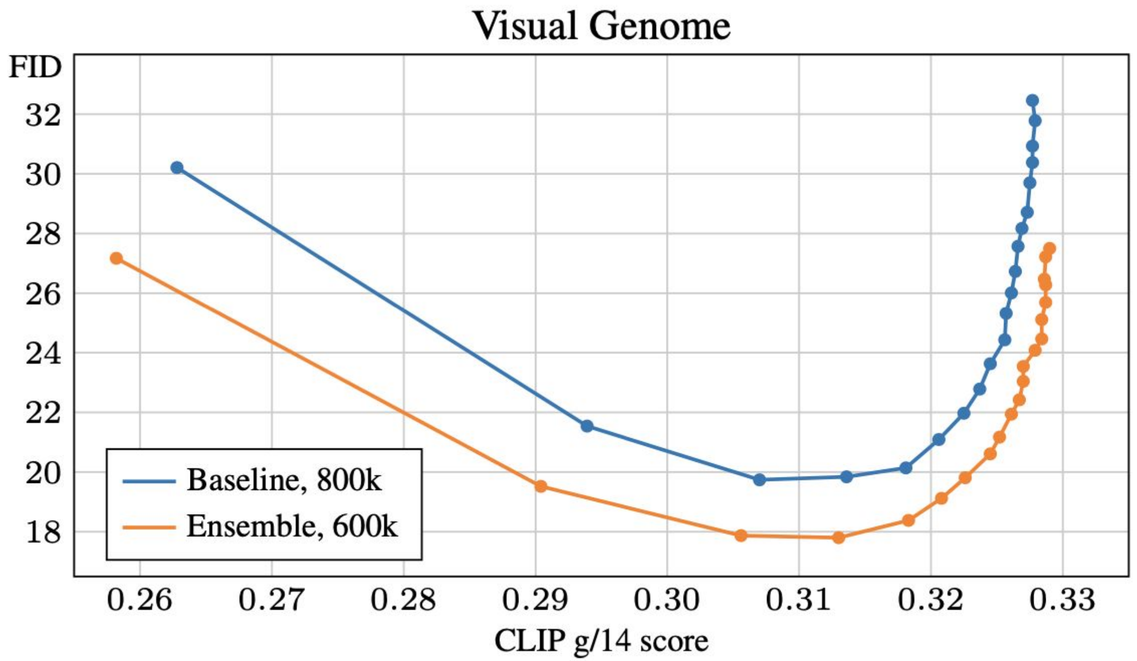
# Experiments



- Improved qualitative results compared to existing methods.
- Ablation studies.
- Demonstrate that both embeddings are better than either one of them.
- Visual results for style transfer and paint with words.

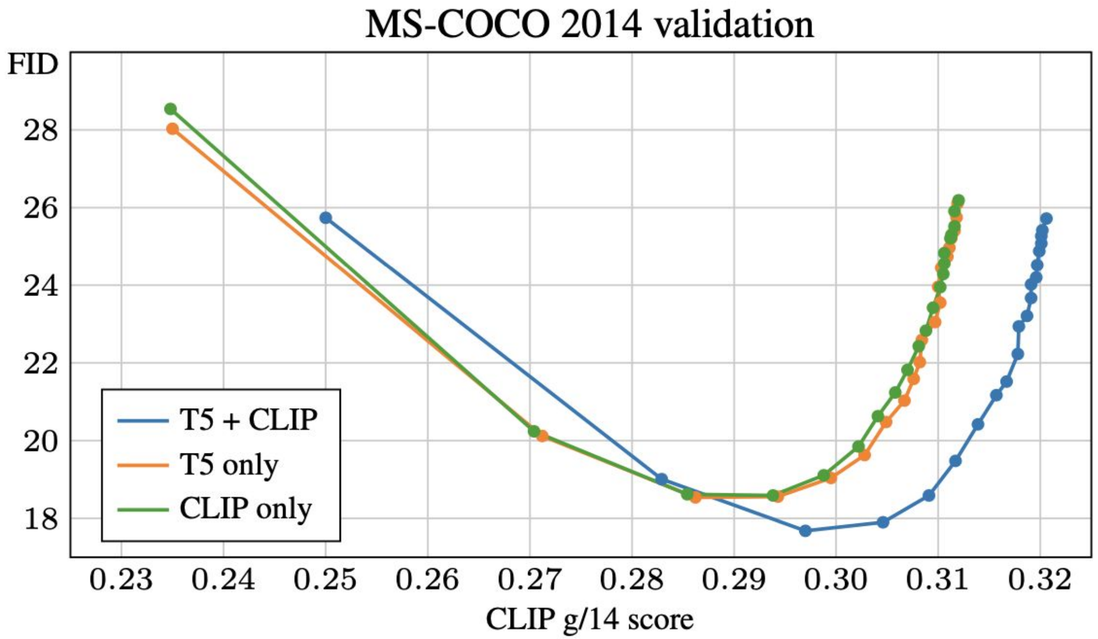
# Comparison with Baseline

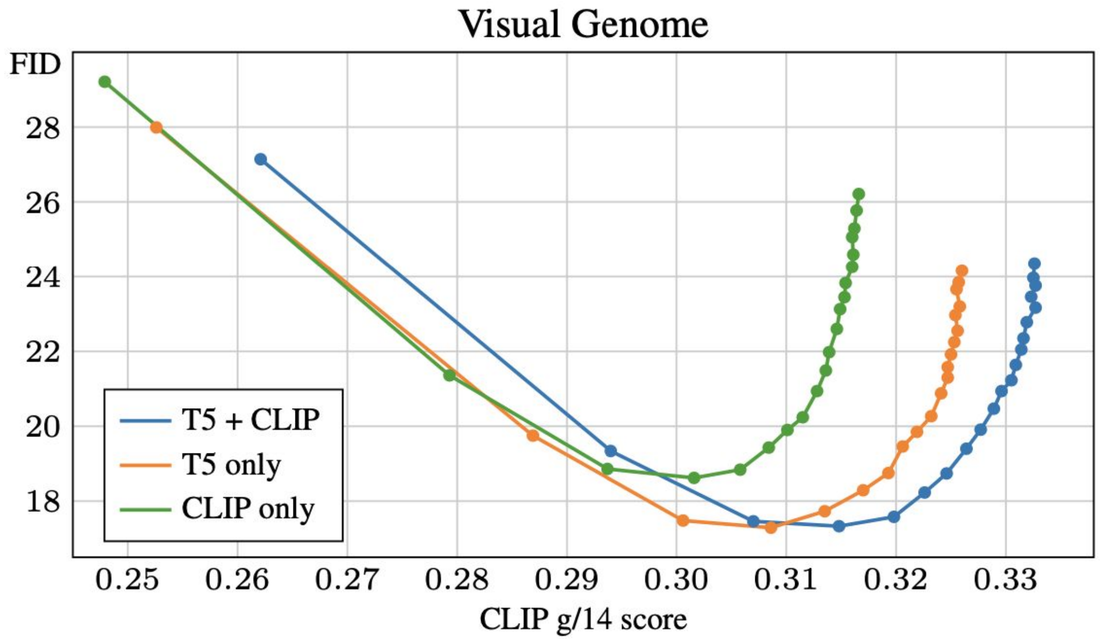



Higher CLIP score and lower FID score desirable.

# Comparison on conditional inputs




Higher CLIP score and lower FID score desirable.

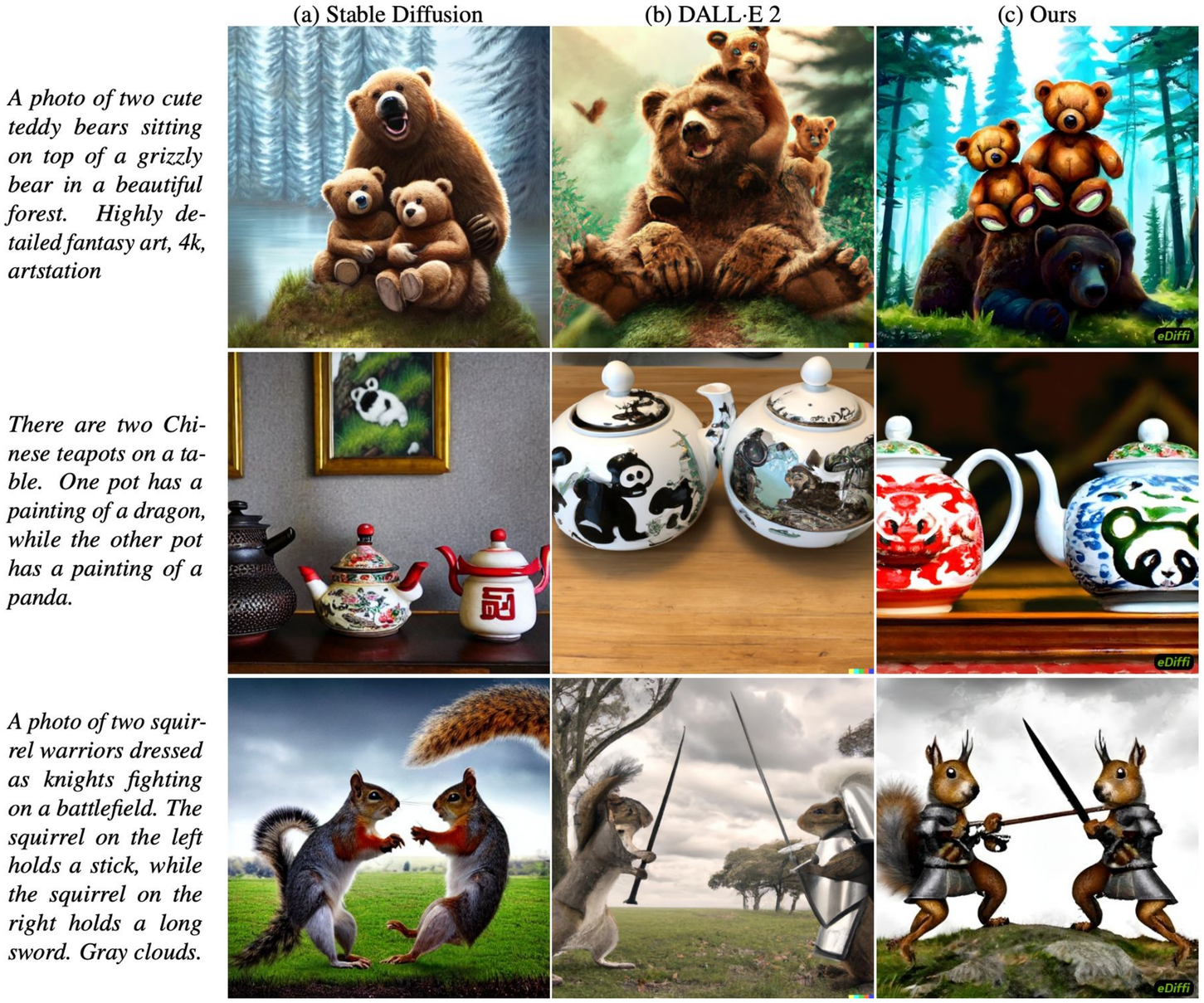
# Quantitative Evaluation

Papers With Jay

Table 1. Zero-shot FID comparison with recent state-of-the-art methods on the COCO 2014 validation dataset. We include the text encoder size in our model parameter size calculation.

| Model | # of params | Zero-shot FID ↓ |
|---|---|---|
| GLIDE [49] | 5B | 12.24 |
| Make-A-Scene [15] | 4B | 11.84 |
| DALL·E 2 [55] | 6.5B | 10.39 |
| Stable Diffusion [57] | 1.4B | 8.59 |
| Imagen [61] | 7.9B | 7.27 |
| Parti [83] | 20B | 7.23 |
| eDiffi-Config-A | 6.8B | 7.35 |
| eDiffi-Config-B | 7.1B | 7.26 |
| eDiffi-Config-C | 8.1B | 7.11 |
| eDiffi-Config-D | 9.1B | **7.04** |

# Qualitative Results



# Style Transfer Results

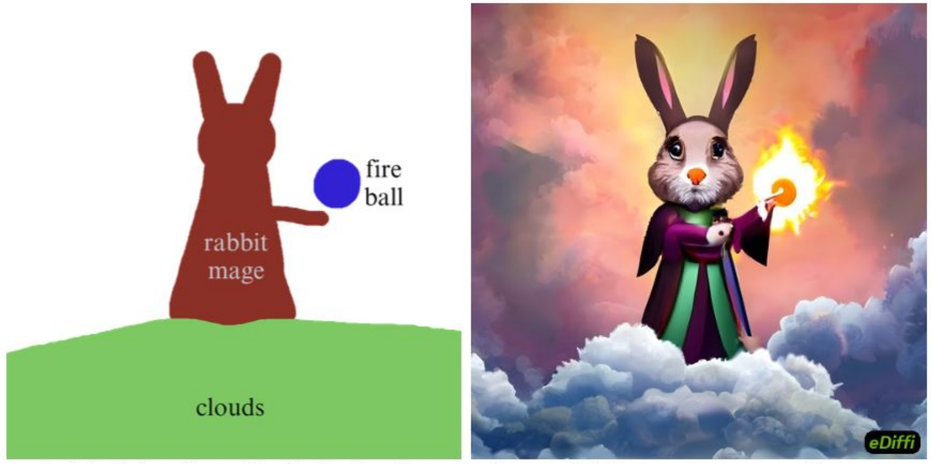


# Paint with words results


*A dramatic oil painting of a road from a magical portal to an abandoned city with purple trees and grass in a starry night.*


*A Halloween scene of an evil pumpkin. A large red moon in the sky. Bats are flying and zombies are walking out of tombs. Highly detailed fantasy art.*


*A monster and a teddy bear playing dungeons and dragons around a table in a dark cellar. High quality fantasy art.*




*A highly detailed digital art of a rabbit mage standing on clouds casting a fire ball.*

# Conclusion

- eDiffi - State of the art text to image diffusion model.
- Uses ensemble of diffusion models each specializing in a different noise range.
- Denoising process conditioned on multiple embeddings
  - T5 text, CLIP text, CLIP image.
- Introduces paint with words capability.