High-Resolution Image Synthesis with Latent Diffusion Models

Student: Itay Bachar

Student ID: 206948218

Course: ML2 – 097209

Date: 05/07/2024

Contents

Paper Summarization

a. Main Topic

This paper introduces Latent Diffusion Models (LDMs) also known as stable diffusion, a groundbreaking approach for synthesizing high-resolution images efficiently. LDMs operate by leveraging a lower-dimensional latent space created by pretrained autoencoders. This model significantly reduces computational costs while maintaining or even enhancing image quality.
By introducing cross-attention layers into the model architecture, latent diffusion models become powerful and flexible generators for general conditioning inputs such as text or bounding boxes. This approach demonstrates substantial improvements in scalability and versatility in image generation tasks, marking a notable advancement in the field of generative models.

b. Challenges and Techniques

1. Computational Inefficiency in Pixel-Space Diffusion Models

Typical diffusion models operate directly in pixel space, which makes both training and inference extremely computationally intensive. Training powerful DMs can take hundreds of GPU days, and inference is expensive due to the need for sequential evaluations, prolonging processing times.

Latent Diffusion Models (LDMs) address these challenges by applying diffusion processes in the latent space of pretrained autoencoders. Transitioning to a lower-dimensional latent space significantly reduces computational complexity and resource requirements, while optimizing the balance between reducing model complexity and detail preservation. By training in a perceptually equivalent latent space, LDMs achieve faster processing times while retaining fine details in generated images.

2. Model Flexibility and Generalization

Many existing generative models struggle to effectively incorporate various types of conditioning inputs, such as text or spatial masks due to fixed architectures and limited adaptability. Generating images from text, for example, requires models to translate textual information into visual features, a challenge not easily addressed by traditional models like GANs or autoregressive models that primarily operate in pixel space without mechanisms for processing text. Similarly, incorporating spatial masks, which denote areas for editing or inpainting within an image, demands the model to localize and adjust specific regions while maintaining global coherence, a capability often lacking in standard generative architectures.

LDMs overcome these limitations through a flexible UNet architecture enhanced with cross-attention mechanisms. This approach enables LDMs to effectively handle complex conditioning inputs, such as text or spatial masks, facilitating tasks like text-to-image synthesis and image inpainting with improved efficiency and quality. By utilizing a compressed latent space derived from pretrained autoencoders, LDMs offer improved versatility and performance across various generative tasks without the need for extensive retraining.

3. Training Stability and Performance Consistency:

While Generative Adversarial Networks (GANs) often suffer from training instability and limited output diversity, and traditional diffusion models require significant resources, LDMs handle those shortcomings.

LDMs leverage the stability of diffusion within a latent space learned by powerful autoencoders. Cross-attention mechanisms enable effective management of diverse conditioning inputs, ensuring consistent and high-quality image generation across tasks. This approach improves model reliability, making LDMs more robust and dependable for practical applications. Unlike typical DMs, LDMs are practical for a wide range of applications due to their efficiency and versatility.

c. Main Results

The experiments detailed in the paper highlight significant improvements in training and inference compared to pixel-based diffusion models:

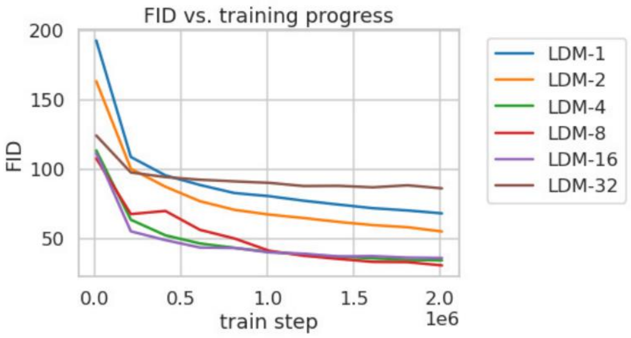
1. Perceptual Compression Tradeoffs

The behavior of LDMs was investigated with different downsampling factors $f \in \{1,2,4,8,16,32\}$, where LDM-1 corresponds to pixel-based DMs.

All models were trained on a single NVIDIA A100 for the same number of steps with the same number of parameters.

<u>Training Progress</u>: Small downsampling factors (LDM- {1,2}) result in slow training progress, while overly large downsampling factors cause image fidelity to stop after fewer training steps due to excessive perceptual compression.

<u>Optimal Balance</u>: LDM- {4-16} strike a good balance between efficiency and perceptually faithful results, with LDM-8 showing a significant FID gap of 38 compared to pixel-based diffusion (LDM-1) after 2M training steps.



FID (Fréchet Inception Distance) assesses image quality by comparing the feature distributions of generated and real images using Inception v3 features. A lower FID score indicates greater similarity between generated and real images, implying higher quality and realism.

<u>Sample Throughput</u>: LDM- {4-8} outperform models with unsuitable compression ratios, achieving lower FID scores and increased sample throughput. Complex datasets like ImageNet require reduced compression rates to maintain quality.

2. Image Generation with Latent Diffusion

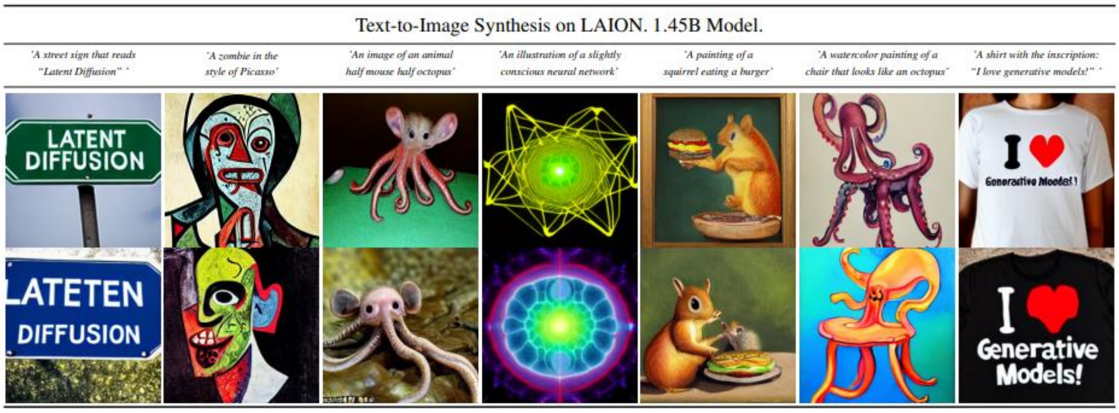
Unconditional models of 256^2 images were trained on CelebA-HQ, FFHQ, LSUN-Churches, and LSUN-Bedrooms. The evaluation was based on FID and Precision-and-Recall metrics.

LDMs consistently improved upon GAN-based methods in Precision and Recall, confirming the advantages of their mode-covering likelihood-based training objectives.

3. Conditional Latent Diffusion

Using cross-attention-based conditioning, LDMs effectively manage diverse input modalities:

Text-to-Image: A 1.45B parameter KL-reg LDM trained on LAION-400M effectively translates complex language prompts into high-quality images, surpassing AR and GAN-based methods.



Text-to-Image Synthesis on LAION. 1.45B Model.

'A street sign that reads "Latent Diffusion" '
'A zombie in the style of Picasso'
'An image of an animal half mouse half octopus'
'An illustration of a slightly conscious neural network'
'A painting of a squirrel eating a burger'
'A watercolor painting of a chair that looks like an octopus'
'A shirt with the inscription: "I love generative models!" '
LATENT DIFFUSION
LATETEN DIFFUSION
I Generative Models!

Flexibility: LDMs adapt well to synthesizing images from semantic layouts on OpenImages dataset, showcasing their versatile conditioning capabilities.

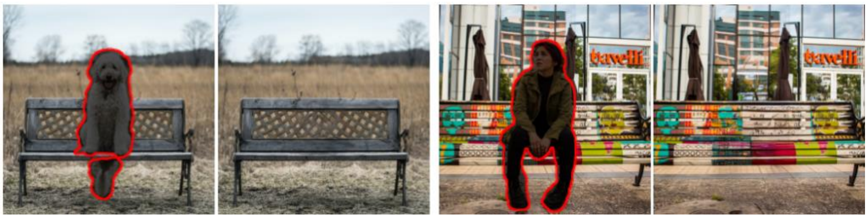
LDMs serve as efficient image-to-image translators:

Semantic Synthesis: Trained on landscapes with semantic maps, LDMs generalize to high resolutions, producing images up to megapixel quality.

Super-Resolution: LDM-SR achieves superior FID performance compared to SR3, supported by competitive Inception Scores (IS) and validated through user studies.

Inpainting: LDMs demonstrate significant speed and quality improvements over specialized methods, making them preferred for image inpainting tasks (the task of filling in missing or damaged parts of an image).

Inpainting



d. Theoretical Framework and Proofs

1. Perceptual Image Compression

Objective: The paper introduces an autoencoder-based perceptual compression model designed to encode images x into a latent representation $z = \mathcal{E}(x)$ using an encoder $\mathcal{E}$ and decode them back into reconstructed images $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$.

Loss Functions: It employs a combination of perceptual loss and patch-based adversarial objectives to ensure that reconstructed images are realistic and locally consistent, avoiding the blurriness associated with pixel-wise losses like L2 or L1 norms.

Regularization: Two regularization strategies are explored: KL- reg and vector quantization VQ- reg within the decoder, which stabilize the learning process and preserve structural details in z .

More on that later in d.5.

2. Diffusion Models

Objective: DMs are probabilistic models that learn the data distribution $p(x)$ by denoising a normally distributed variable through a sequence of steps T.
This denoising process can be interpreted as learning the reverse process of a fixed Markov Chain of length T.
This process effectively reconstructs the input x from its noisy versions x_t.

Training: The models utilize denoising autoencoders $\theta(x_t, t)$ to predict denoised versions of x_t.

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right], \qquad (1)$$

In equation (1), $\epsilon_\theta(x_t, t)$ is the denoised variant predicted by the neural network ϵ_θ given the noisy input x_t and the current time step t.

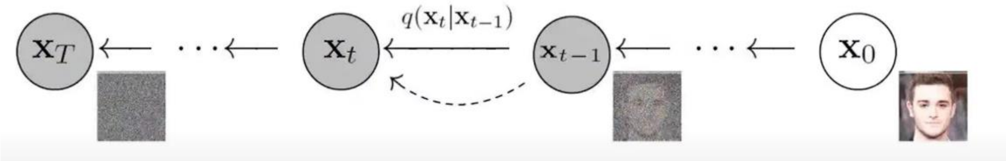
So, diffusion model is trained to minimize the difference between the actual noise added to the input (denoted as ϵ) and the noise predicted by the network.

At each step t of the diffusion process, noise ϵ sampled from a Gaussian distribution $\mathcal{N}(0,1)$ is added to the input x to produce the noisy version x_t .

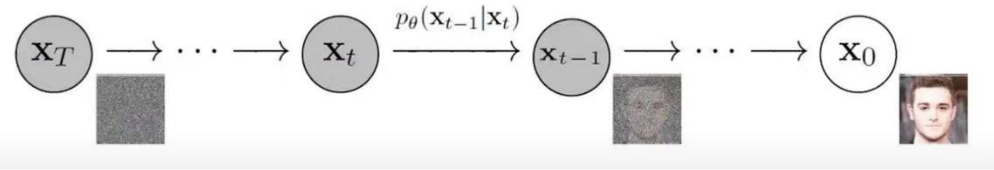
$x_\mathrm{t} = \frac{\mathrm{T\text{-}t}}{\mathrm{T}} \cdot x_0 + (1\text{-}\frac{\mathrm{T\text{-}t}}{\mathrm{T}}) \cdot \epsilon$

So, L_{DM} can be rewritten by: $L_{DM} = \mathrm{MSE}(\epsilon, \epsilon_\theta[\frac{\mathrm{T\text{-}t}}{\mathrm{T}} \cdot x_0 + (1\text{-}\frac{\mathrm{T\text{-}t}}{\mathrm{T}}) \cdot \epsilon])$

Diffusion process
Forward:



Backward:



3. Generative Modeling of Latent Representations:

Utilization of Compressed Latent Space: LDMs leverage the efficient, low-dimensional latent space z from the perceptual compression model $\mathcal{E}$ and $\mathcal{D}$, focusing on modeling high-frequency details essential for generative tasks.

Model Architecture: The generative process is facilitated by a time-conditional UNet backbone $\epsilon_\theta(\circ, t)$ which effectively translates latent representations z_t back into high-resolution images using a single pass through the decoder $\mathcal{D}$.

Reweighted Bound: The objective function L_{LDM} emphasizes the importance of perceptually relevant features in z, ensuring that the model focuses on the most significant details for accurate image generation.

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\Big] \,. \qquad (2)$$

4. Conditioning Mechanisms

Conditional Image Synthesis: LDMs are extended to handle conditional distributions $\mathrm{p}(z|\mathrm{y})$, where y represents additional input modalities such as text prompts or semantic maps.

Integration of Cross-Attention: To incorporate y effectively, a domain-specific encoder τ_θ projects y into an intermediate representation $\tau_\theta(\mathrm{y})$, which is then integrated into the UNet layers via cross-attention mechanisms. This allows the model to selectively attend to relevant features of y during the image synthesis process.

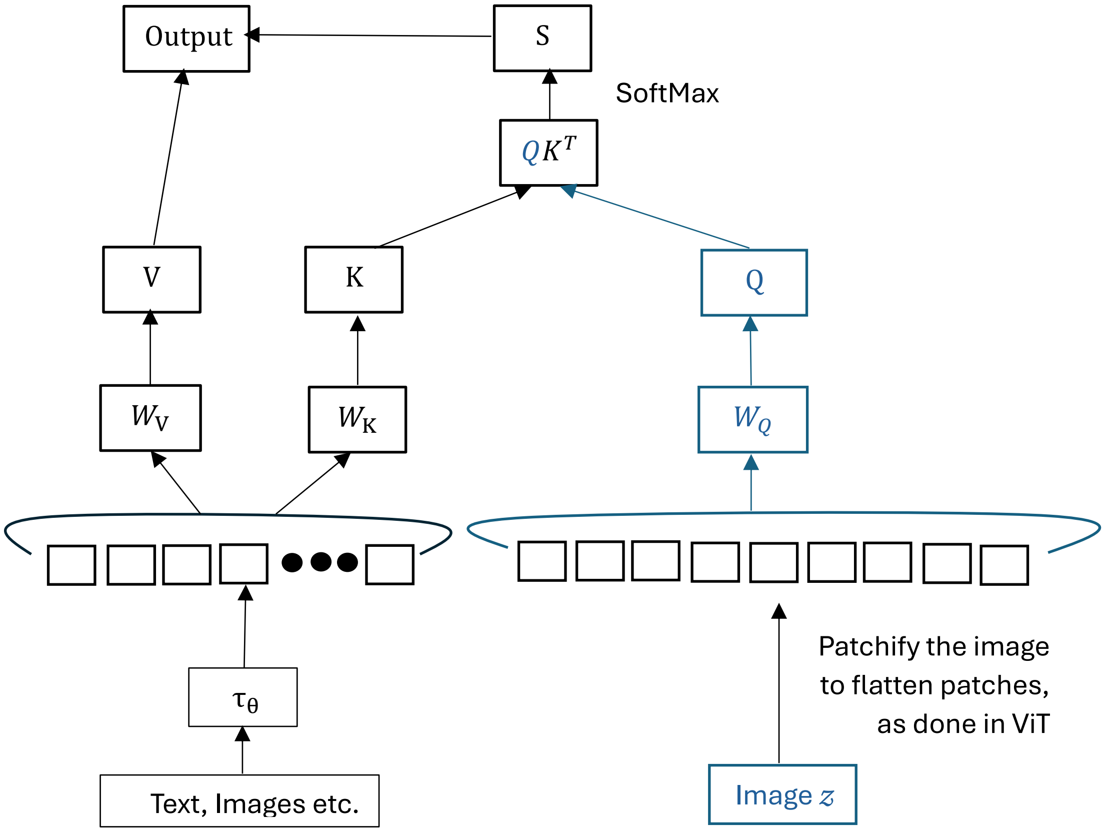
The cross-attention mechanism is implemented as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$$

Where,

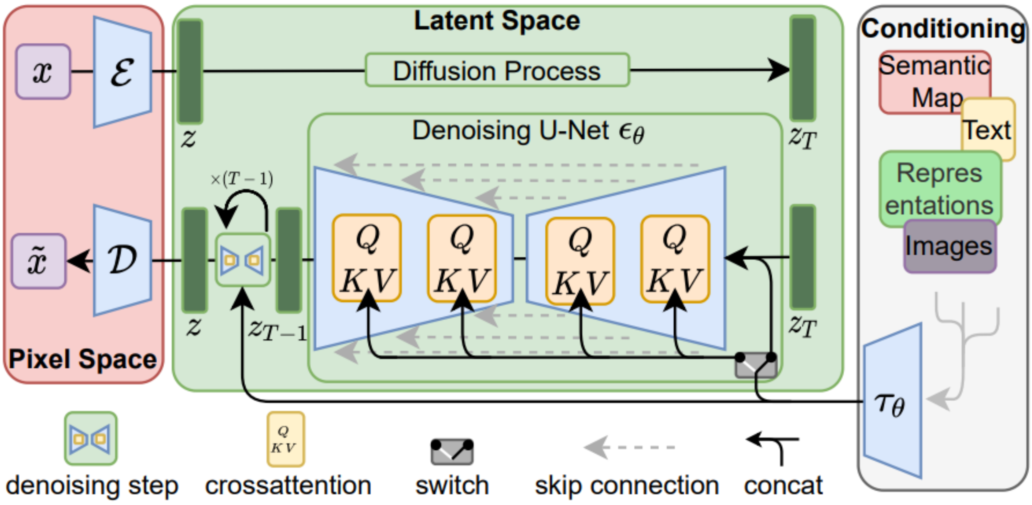
$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), \; K = W_K^{(i)} \cdot \tau_\theta(y), \; V = W_V^{(i)} \cdot \tau_\theta(y).$$

$\varphi(z_t)$ denotes intermediate representation of the UNet backbone
$W_v^{(i)}, W_{\mathrm{Q}}^{(i)}, W_{\mathrm{K}}^{(i)}$ are learnable projection matrices



Flexible Conditioning: The conditional L_{LDM} enables controlled image synthesis by optimizing both τ_θ and ϵ_θ jointly, facilitating diverse applications including text-guided image synthesis.

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right], \quad (3)$$



This image illustrates a block diagram depicting the stages of a latent diffusion model process. It outlines each phase involved in the model's operation.

The "switch part" refers to the option of either concatenating additional conditioning information directly to the z_t image or use the cross-attention mechanism.

5. Autoencoder Loss Function

The overall objective function $L_{Autoencoder}$ combines reconstruction loss, adversarial loss, log likelihood of the discriminator, and regularization term L_{reg}.

Where,

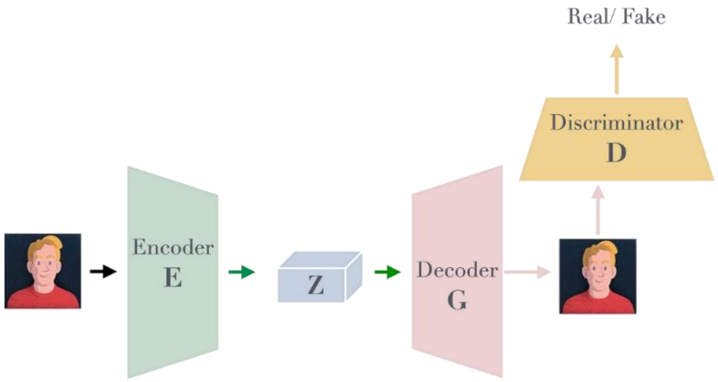
$$L_{rec} = MSE(x, \mathcal{D}(\mathcal{E}(x))$$

$$L_{reg} = KL\big(\mathcal{E}(x) = L, \mathcal{N}(0,1)\big)$$

$$L_{adv} = log[D_\psi(\mathcal{D}(\mathcal{E}(x) \approx x)]$$

$$L_{dis} = log\big(D_\psi\big)$$

$$L_{\text{Autoencoder}} = \min_{\mathcal{E},\mathcal{D}} \max_{\psi} \Big(L_{rec}(x, \mathcal{D}(\mathcal{E}(x))) - L_{adv}(\mathcal{D}(\mathcal{E}(x))) + \log D_\psi(x) + L_{reg}(x; \mathcal{E}, \mathcal{D})\Big)$$



Second Part – Theoretical

i. Novelty and methods

The novelty and methods of LDMs have been detailed in earlier sections. These discussions highlighted the significant advancements in computational efficiency, model flexibility, and the innovative use of cross-attention mechanisms. The methods outlined include the utilization of pretrained autoencoders for latent space representation, the implementation of latent diffusion processes, and the usage of a UNet structure for generative modeling.

ii. Limitations

1. Dependence on Pretrained Autoencoders

LDMs rely on pretrained autoencoders to create the latent space in which the diffusion process operates. The quality and performance of LDMs are inherently tied to the effectiveness of these autoencoders. If the autoencoders fail to capture all necessary details or introduce artifacts, it can affect the overall image quality produced by LDMs.

2. Compression Trade-offs

While operating in a lower-dimensional latent space reduces computational costs, it also introduces the challenge of balancing compression with image fidelity. Too much compression can lead to loss of detail and image quality, while too little compression may not yield significant computational savings.

3. Complexity of Training

The integration of cross-attention mechanisms and the necessity to handle various conditioning inputs add to the complexity of training LDMs. Training such models requires careful tuning and substantial computational resources, especially for tasks involving large datasets and high-resolution images.

4. Resource Requirements:

Despite improvements in efficiency, training LDMs still demands significant computational power and memory, particularly for large-scale or high-resolution tasks. This can be a barrier for applications with limited resources.

ii. Comparisons to Other Techniques

1. GANs vs LDMs

GANs are known for efficiently sampling high-resolution images with good perceptual quality. However, they are challenging to optimize and often struggle to capture the full data distribution. In contrast, LDMs offer more stable training and better handling of diverse conditioning inputs through cross-attention mechanisms. While GANs excel in producing sharp and photorealistic images when extensively fine-tuned, they can be complex to implement for tasks with less complex conditional inputs.

2. VAEs and flow-based models vs LDMs

These methods enable efficient synthesis of high-resolution images but typically do not achieve sample quality on par with GANs. They focus on good density estimation but may not capture the complex features of images as effectively as LDMs.

3. ARMs vs LDMs

ARMs achieve strong performance in density estimation but are computationally demanding and limited to low-resolution images due to their sequential sampling process. Two-stage approaches combining ARMs with latent image spaces have been explored to scale to higher resolutions but suffer from high computational costs. LDMs address these drawbacks by operating in a compressed latent space, making training computationally cheaper without significant reduction in synthesis quality.

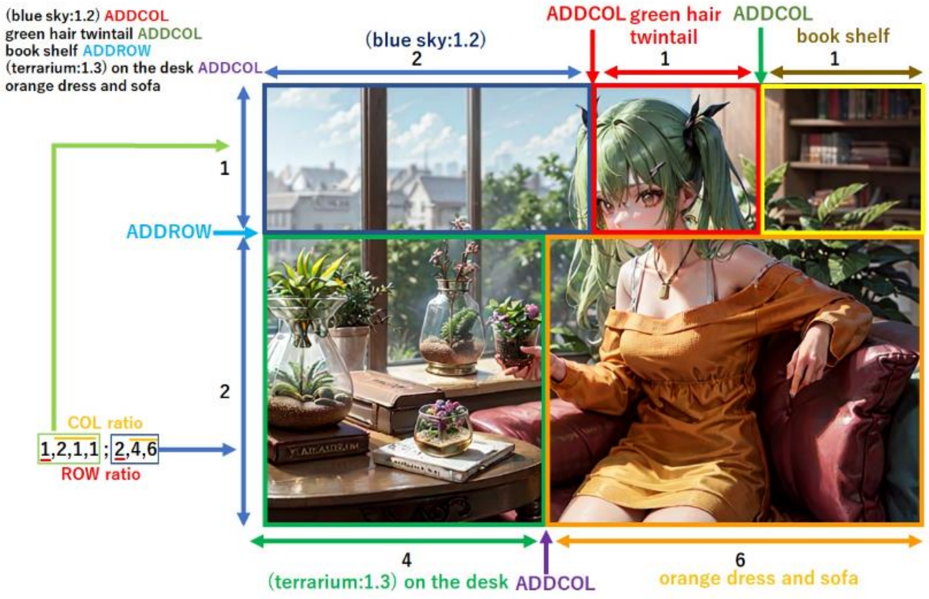
4. DMs vs LDMs

DMs have achieved state-of-the-art results in density estimation and sample quality. They leverage the natural biases of image-like data but evaluate and optimize models in pixel space, leading to low inference speed and high training costs. In contrast, LDMs operate in a compressed latent space, maintaining synthesis quality while speeding up inference and reducing training costs compared to DMs.

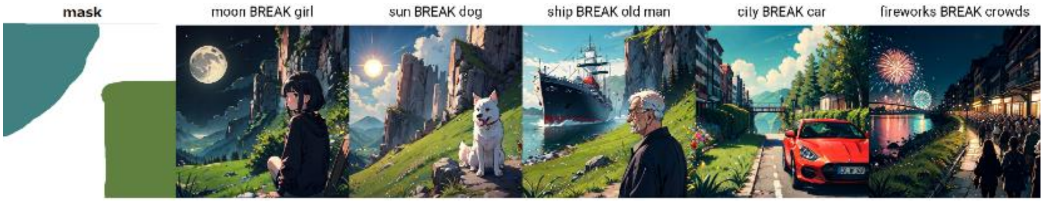
iii. Extensions

1. Regional Prompter

The Regional Prompter extension was developed by Hako Mikan, a Japanese innovator known for enhancing the capabilities of Latent Diffusion Models. It introduces a novel method to specify and manipulate specific regions within generated images using localized prompts or masks. This extension enables users to delineate regions of interest through hand-drawn masks or uploaded images containing predefined masks. By integrating these masks into the conditioning of LDMs, users can direct the model's attention to specific areas while preserving global coherence in the generated images. specific regions or features of interest within an image.



One notable contributor to this extension is Itay Irmay, with whom I served in reserve duty during the war, and he has significantly expanded its functionality. An experimental feature, termed "Mask Regions aka Inpaint+," allows users to specify multiple regions using detailed hand-drawn masks. This feature facilitates precise control over image synthesis by assigning different prompts or effects to designated areas within the image. Moreover, the extension supports the creation, saving, and loading of presets for masks, enhancing usability across different projects and datasets.



Link to this extension HERE

2. Extension Proposal: iLDMs

Interactive Latent Diffusion Models (iLDMs) represent an advancement in generative models that integrates real-time user interaction into the image generation process. Unlike traditional LDMs that rely on fixed inputs like text descriptions or spatial masks, iLDMs enable users to actively guide and refine the image synthesis process during generation. Key features of this extension could include:

Real-time Conditioning Updates: Users can dynamically adjust conditioning inputs such as text descriptions or spatial masks while the model is generating images. This capability requires the latent space representation to adapt in real-time, ensuring responsive adjustments to user inputs.

Feedback Loops: Introducing iterative feedback mechanisms where users can refine generated images based on immediate visual feedback. This allows users to annotate or modify specific areas of the image, guiding the model towards desired outcomes with each iteration.

Interactive Interface: Designing an intuitive interface that facilitates user interaction with iLDMs, making it accessible for non-experts to engage in the creative process of image synthesis. The interface would support seamless interaction and control over the generative process.
This could be built using frameworks like Tkinter, PyQt, or web-based tools.

Implementing iLDMs would expand the functionality of traditional LDMs by offering a dynamic and interactive approach to generative image synthesis. This extension holds promise for applications in creative industries, interactive design tools, and personalized content generation, paving the way for innovative uses in user-driven generative modeling.

Nowadays, even with the release of Stable Diffusion 3, one of the most challenging aspects of this field remains generating text within images accurately due to the complexity of maintaining readability and consistency with various fonts and sizes.

All the Stable Diffusion WebUI extensions can be found HERE
For explanations of most of them HERE

Thanks for reading,
Itay Bachar.