DiffusionDemo

Activity Guide

Version of August 2025

Adithya Kameswara Rao
David S. Touretzky
Carnegie Mellon University

This work was funded by a grant from NEOM Company and by National Science Foundation award IIS-2112633.

Diffusion Demo

Overview

Experiment with the Stable Diffusion model to see how it generates an image from a text description.

Through the demo, the following goals are accomplished:

- Understand how an image is generated from noise.
- Understand the role of the *seed* in image generation.
- Develop an intuition about *latent spaces*.
- Understand the influence of the *guidance scale* in image generation.
- Understand how text influences image generation.
- Develop an intuition about the CLIP embedding space.
- Understand the role of a *negative prompt* in image generation.

Materials/Resources Required

- Any web browser
- Access to diffusiondemo.org (use http: not https:)
- Access to Cats in Latent Space video

Vocabulary Terms

These vocabulary words are also available as flash cards in the Vocabulary Flash Cards section.

- **Latent**: potential, hidden, or undeveloped
- **Latent space:** a compressed representation that captures the essential qualities of things
- **Latent image:** a compressed representation of an image as a point in a latent image space
- **Noise**: random values that obscure an image
- **Diffusion algorithm:** a process that incrementally removes noise from a latent image
- **Stable Diffusion:** a particular diffusion algorithm used to turn prompts into images
- **Manifold**: a collection of points forming a surface embedded in a high dimensional space
- **Denoising**:removing noise from a latent image, moving it closer to the manifold of noise-free images
- **Prompt**: a word or phrase describing the image to be generated
- **Inference step**: one step of noise removal; typically 8-20 steps are required
- **Seed**: a value used to kick off a sequence of random numbers forming an initial image that is pure noise.
- **CLIP**: Contrastive Language-Image Pretraining, a neural network model that learns embeddings for images and their captions
- **Semantic space**: a coordinate system in which each axis represents some aspect of meaning, such as "animate/inanimate" or "male/female".
- **Embedding**: a representation of meaning as a point in a semantic space.
- **Semantics**: meaning.
- **VAE**: Variational Auto-Encoder, a neural network that turns a point in a latent image space into an actual image
- **Interpolation**: estimating a value that lies between two known values.
- **Guidance scale**: parameter that controls how strictly a prompt should be followed.
- **Inpainting**: replacing some portion of an image with different content, such as adding sunglasses to a face

Getting Started

What is a Diffusion Model?

A *diffusion text-to-image model* generates an image based on a text description. It is a neural network that has been constructed by exposing it to lots of text and image pairs. Once trained, the model can be conditioned to perform a variety of tasks such as producing images based on scribbles, enhancing the quality of an image, and more.

Some of the best-known models are:

- DALL-E 3 (from OpenAI)
- Stable Diffusion 3 (from Stability AI)
- Midjourney (from Midjourney Inc.)
- Imagen (from Google)

Key Ideas, Activities, and Videos

 *Key Ideas* are marked with a lightbulb.

 *Activities* are marked with a ringing bell.

 Notes on *Video Clips* are marked with a camera.

Level: Beginner

Learning Objectives

Students will be able to:

- Perceive how images are generated from text
- Understand how images are generated from noise
- Understand the role seeds play in creating different images

Demo Investigations

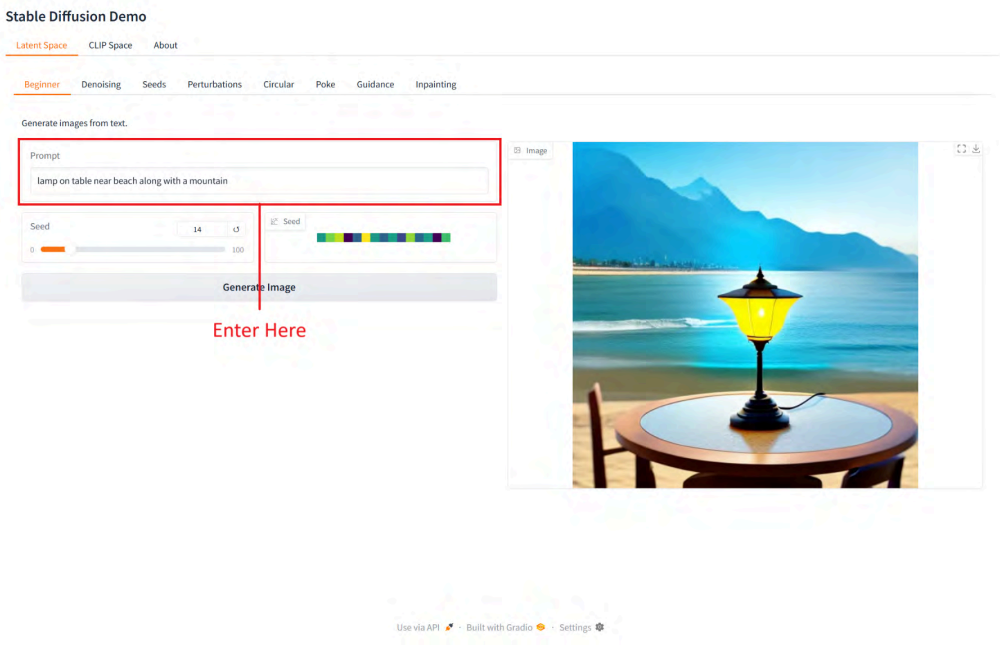
Importance of Prompts

Prompts are the instructions you give the model to create your desired image. Think of them as describing what you want to see in a picture.

Prompt - The text description that tells the diffusion model what kind of image to create. Like giving instructions to an artist about what to draw.

Open the demo at diffusiondemo.org.

Go to the **Beginner** tab under Latent Space (Figure 1).



Screenshot from demo

Figure 1 - Entering a prompt in the diffusion demo.

Simple vs Detailed Prompts

Compare these simple vs. detailed prompts:

Simple prompt: "lamp on table"
Detailed prompt: "lamp on table near beach along with a mountain"

	
lamp on table	**lamp on table near beach along with a mountain**

Figure 2 - The difference between a simple prompt and a detailed prompt.

Notice how adding location details (beach, mountain) gives the diffusion model more specific guidance (Figure 2).

Simple prompt: "man in a suit"

Detailed prompt: "photo of man in a suit in black and white using pencil"

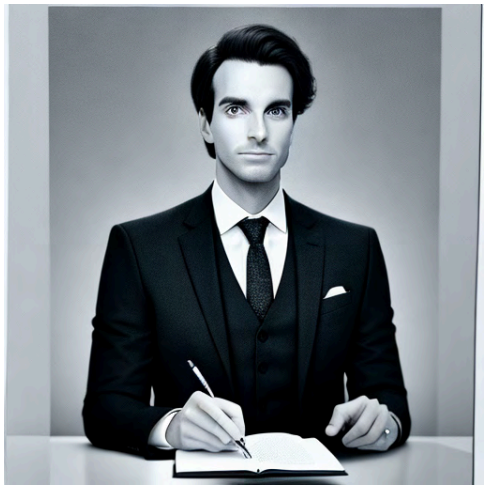
	
man in a suit	**photo of man in a suit in black and white using pencil**

Figure 3 - Results of prompts with different amounts of detail.

Observe how the detailed prompt specifies (Figure 3):

- Medium ("photo")
- Subject ("man in a suit")
- Style ("black and white")
- Technique ("using pencil")

How Artistic Styles Transform Images

Explore how artistic styles transform images:

Try this experiment:

1. Enter the base prompt "a woman holding an apple"
2. Add different artistic style terms to create each new prompt:
    - "a woman holding an apple, picasso cubist style"
    - "a woman holding an apple, salvador dali style"
    - "a woman holding an apple, reubens style"
    - "a woman holding an apple, disney cartoon style"
    - "a woman holding an apple, norman rockwell style"
    - "a woman holding an apple, zap comix style"
    - "a woman holding an apple, 19th century sepia photograph"

- "a woman holding an apple, michaelangelo statue style"
- "a woman holding an apple, illuminated manuscript style"

3. Click "Generate Images" for each style (Figure 5)

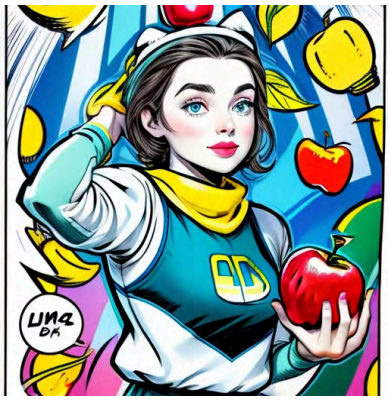
picasso cubist style

salvador dali style

reubens style

disney cartoon style

norman rockwell style



zap comix style

19th century sepia photograph

michaelangelo statue style

illuminated manuscript style

Figure 4 - A comparison of nine prompts generating different artistic styles.

Key takeaway: Vague prompts give the diffusion model freedom to fill in missing details, which might not match what you had in mind (Figure 4). Detailed prompts with specific elements, style, and context will produce images closer to your vision. Adding artistic styles provides even greater creative control over the aesthetic and mood of your generated images.

Importance of Noise

Diffusion models work by learning to transform noise into clear images.

Noise - Random patterns of pixels that look like TV static, with no recognizable features.

Navigation: Go to the "Denoising" tab under "Latent Space". Click on the arrow in the top left corner of the “About” box to close the box.

Screenshot from demo

Figure 5 - Generate an image in the demo with the **Generate Images** button.

From Noise to Image

How noise becomes an image:

During training:

- The model is shown clear images alongside noisy versions of the same images
- It learns to remove noise and recover the original picture
- The text prompt guides this denoising process, telling the model what to create

During generation (what you see in the demo):

- The model starts with 100% random noise
- Step by step, it removes this noise according to your prompt
- The random starting pattern gradually transforms into your requested image

Try this experiment in the Denoising tab:

1. Enter a simple prompt like "red apple"
2. Watch as random noise evolves into a recognizable apple
3. Notice how shapes emerge first, then colors and details

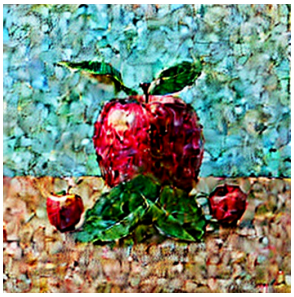
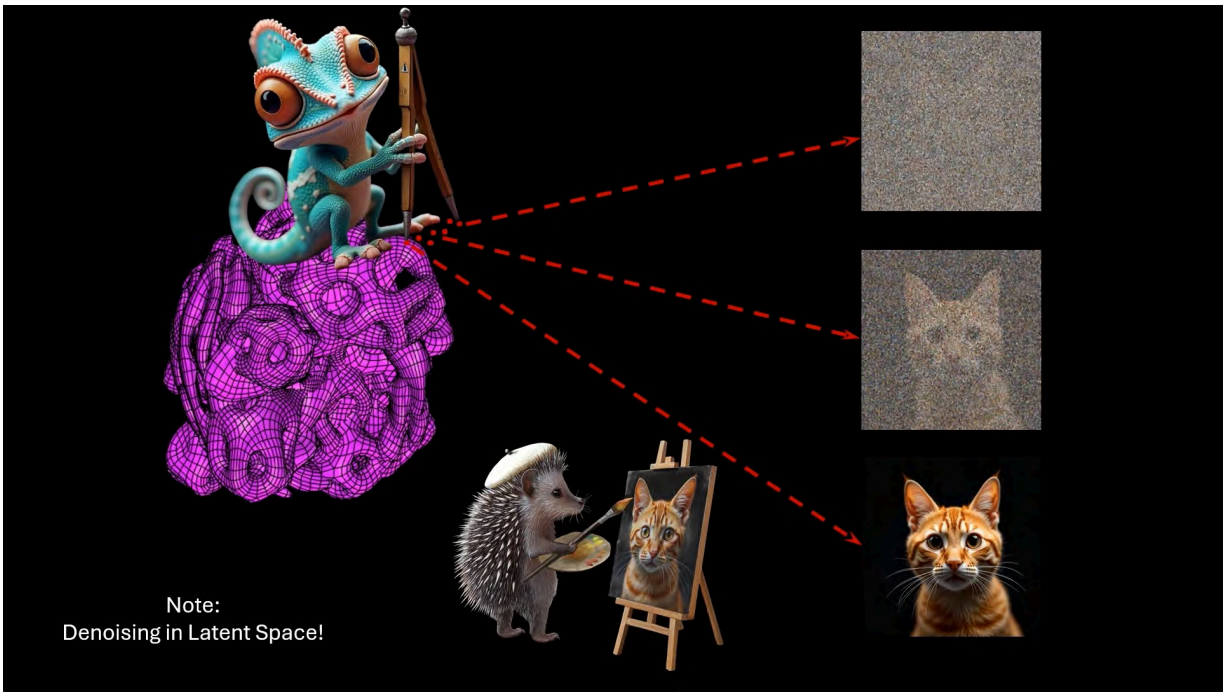
 Step: 1	 Step: 2	 Step: 3
 Step: 4	 Step: 5	 Step: 6
 Step: 7	 Step: 8	

Figure 6 - Sequence showing the progression of generating an image from noise.

Key takeaway: The diffusion model starts with pure random noise, and then gradually removes the noise in a way that reveals an image matching your text description (Figure 6).

How this experiment connects to the "Cats in Latent Space" Video:

When you watch the steps unfold in our apple experiment, you're seeing exactly what the video describes (Figure 7) - the UNET (Oona) progressively removing noise in latent space, while the VAE decoder (Vi) translates each denoised state into increasingly clear images. The early steps show "latent noise with maybe a tiny bit of apple flavor," and as denoising continues, the apple becomes more defined - demonstrating how the point in latent space gradually approaches the manifold of realistic images.



Screenshot from video at timestamp - 6:24

Figure 7 - A denoising example from the *Cats in Latent Space* video.

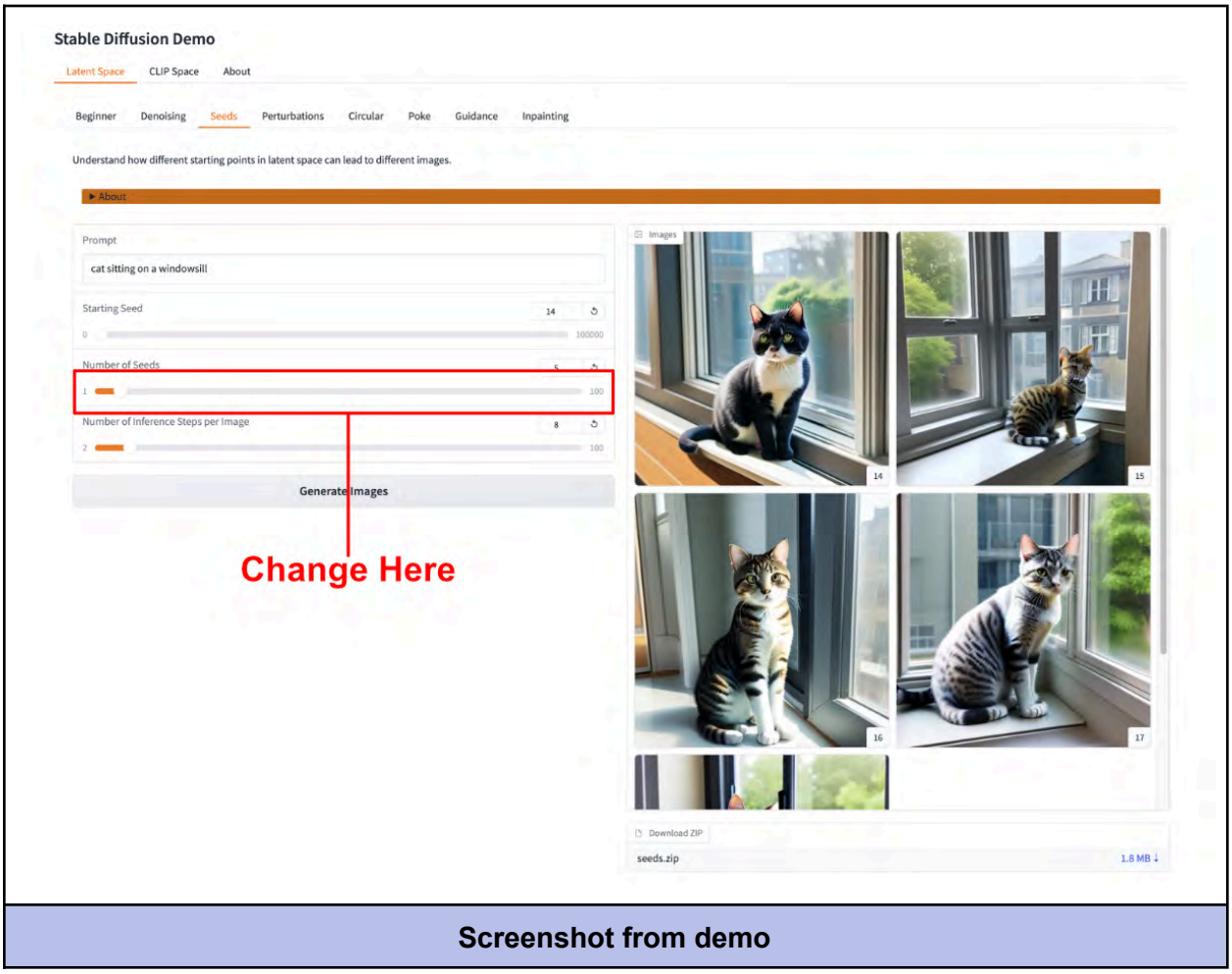
Importance of Seeds

Seeds are the starting point for your image creation journey. They determine which unique version of your requested image will be generated.

Seed - Think of a seed like your starting location on a city map when following directions. Imagine you have instructions to "go 2 blocks north, then 2 blocks east" to reach a specific restaurant. If you and your friend start at different locations (different seeds) but follow the exact same directions (same prompt),

you'll both end up at different restaurants. Similarly, different seeds create different starting points in the noise pattern, leading to different final images even when following the exact same prompt.

Navigation: Go to the "Seeds" tab under “Latent Space” (Figure 8)



Screenshot from demo

Figure 8 - Change the number of seeds with the **Number of Seeds** slider.

See how different seeds create different images:

Try this experiment:

1. Enter the prompt "cat sitting on a windowsill"
2. Set **Number of Seeds** to 6
3. Click **Generate Images**

Observe how the same prompt produces six completely different images (Fig. 9):

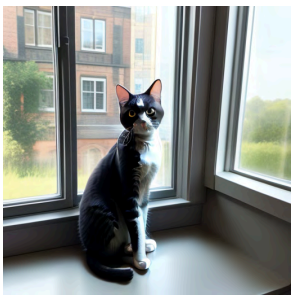
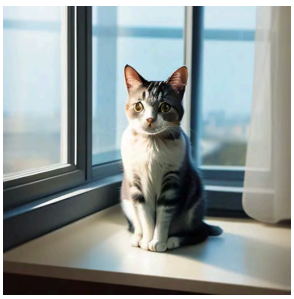
 Seed: 1	 Seed: 2	 Seed: 3
 Seed: 4	 Seed: 5	 Seed: 6

Figure 9 - Six differently seeded images generated from the same prompt.

Key takeaway: Seeds let you explore different possibilities without changing your prompt. When you find an image you like, note its seed value so you can recreate exactly the same image later. This is especially useful when generating multiple variations to choose from.

Level: Intermediate

Learning Objectives

Students will be able to:

- Explain how increasing inference steps improves image quality and detail
- Demonstrate how guidance scale values control the balance between creativity and prompt adherence
- Utilize negative prompts to exclude unwanted elements from generated images

Demo Investigations

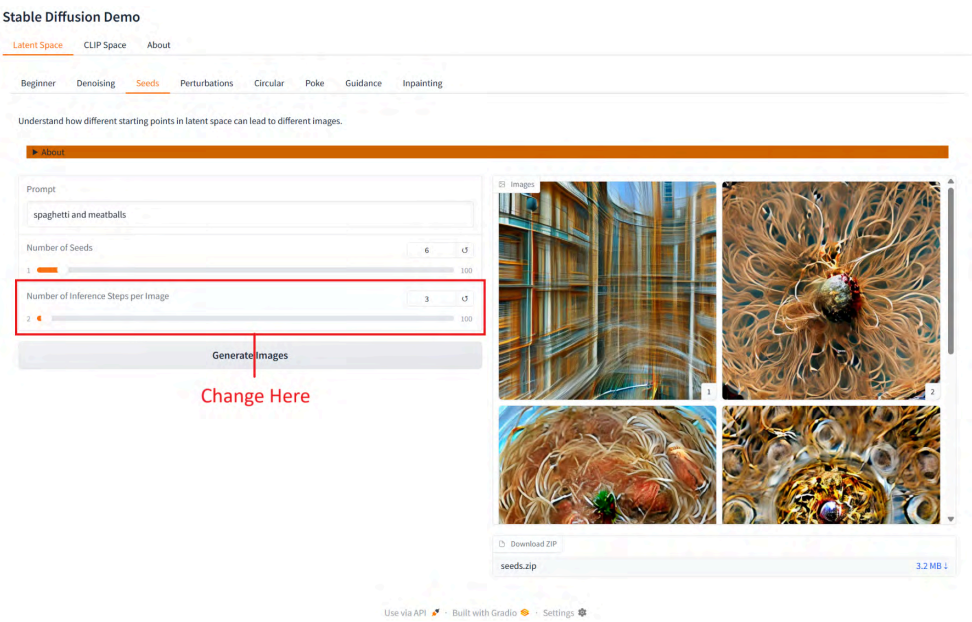
Importance Of Inference Steps

The number of inference steps determines how thoroughly the model refines your image. Think of it as controlling the amount of time the diffusion model spends creating your picture.

 *Inference* - When the model uses the prompt and creates a new picture, it is called inference.

 *Denoising* - Removing all those random colors and shapes to transform noise into a clear image.

Navigation: Go to the "Seeds" tab under "Latent Space" (Figure 10)



Screenshot from demo

Figure 10 - Setting the number of inference steps per image with the **Number of Inference Steps per Image** slider.

See how inference steps affects image quality:

Try this experiment:

1. Enter the prompt "spaghetti meatballs"
2. Set the **Number of Seeds** to 3
3. Set **Number of Inference Steps per Image** to 3
4. Click **Generate Images** button
5. Then, try it with 4 inference steps and 8 inference steps (Figure 11)

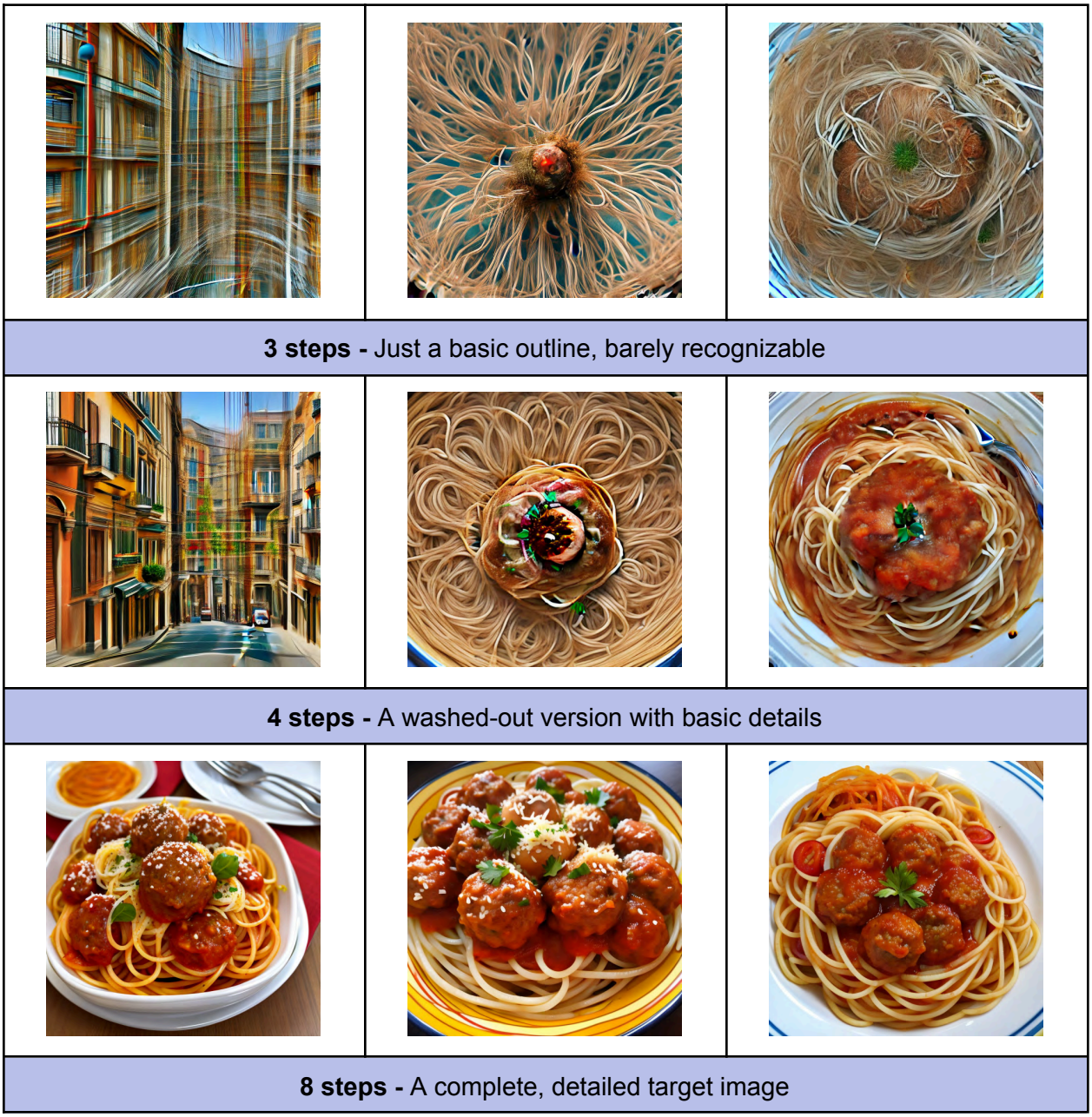


Figure 11 - Output at varying numbers of inference steps.

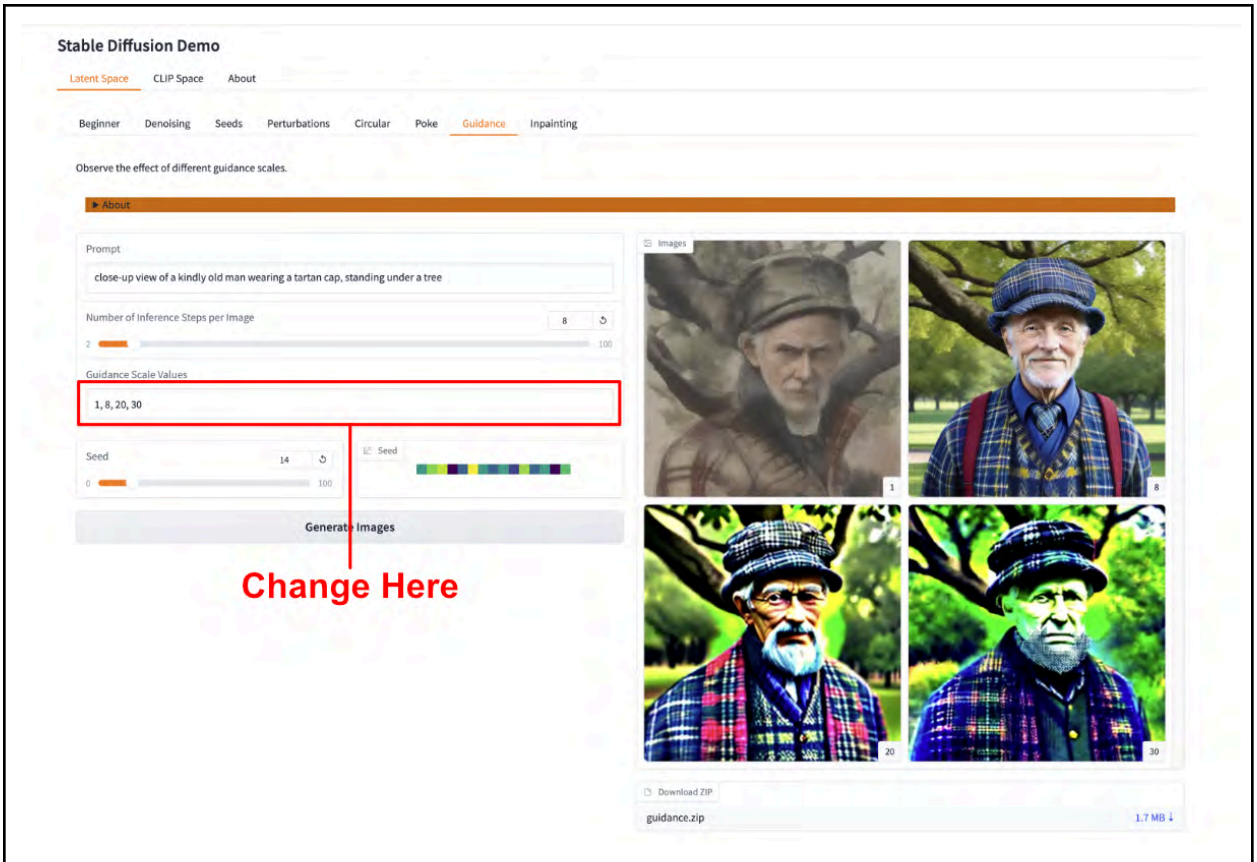
Key takeaway: More steps allow the model to refine details gradually. Fewer steps save time but produce rougher images. For quick drafts, use fewer steps; for polished results, use more steps.

Importance Of Guidance Scale Values

Guidance scale controls how strictly the model follows your prompt versus how much creative freedom it takes.

Guidance Scale - A control knob that tells the model how closely it should follow your idea. If you turn it up, the model will create a picture very close to your prompt. If you turn it down, the model will be more creative and make a picture that might be different from your idea.

Navigation: Go to the "Guidance" tab under "Latent Space" (Figure 12)



Screenshot from demo

Figure 12 - Adjusting the guidance scale values with the **Guidance Scale Values** slider.

See how guidance values affect your image:

Try this experiment:

1. Enter a prompt of your choice
2. Try the guidance scale values: 8, 20, and 30 (Figure 13)

		
Guidance: 8 Normal-looking image	**Guidance: 20** Oversaturated image	**Guidance: 30** Distorted oversaturated image

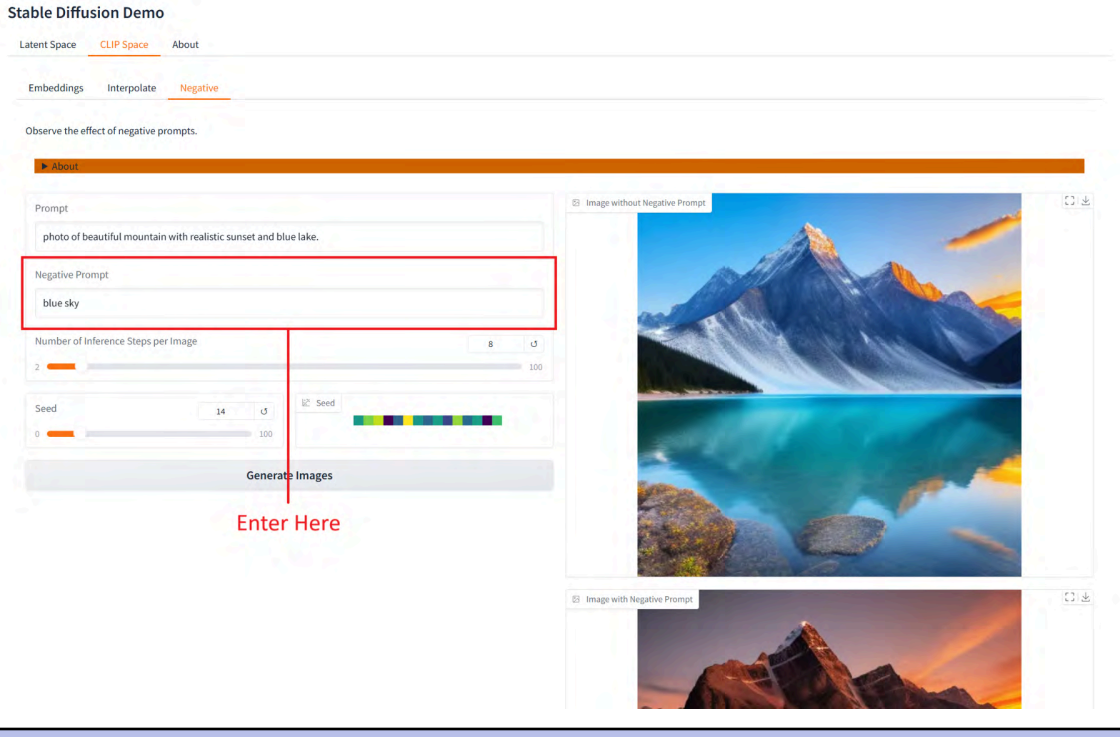
Figure 13 - Effect of adding too much guidance.

Key takeaway: Think of guidance like steering a car: too little guidance (low values) and the diffusion model wanders off course; too much guidance (high values) and you overcompensate, creating distorted results.

Importance Of Negative Prompts

Negative prompts tell the model what NOT to include in your image.

Navigation: Go to the "Negative" tab under "CLIP Space" (Figure 14)



Screenshot from demo

Figure 14 - Adding a negative prompt.

See how negative prompts influence your image:

Try this experiment:

1. Enter the prompt "photo of beautiful mountain with realistic sunset and blue lake"
2. Now add the negative prompt "blue sky" and generate

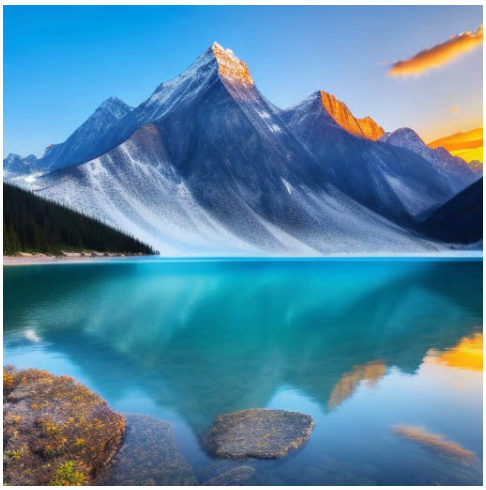
	
No negative prompt	**With negative prompt**

Figure 15 - Two images generated with the same prompt, one with a negative prompt.

Compare the two images (Figure 15) - notice how the model avoids including a blue sky in the second image. The colors of the sky show sunset colors instead.

Key takeaway: Negative prompts are powerful for avoiding unwanted elements in your image. Use them to steer the model away from specific colors, objects, or styles that you don't want to appear in the final result.

Level: Advanced

Learning Objectives

Students will be able to:

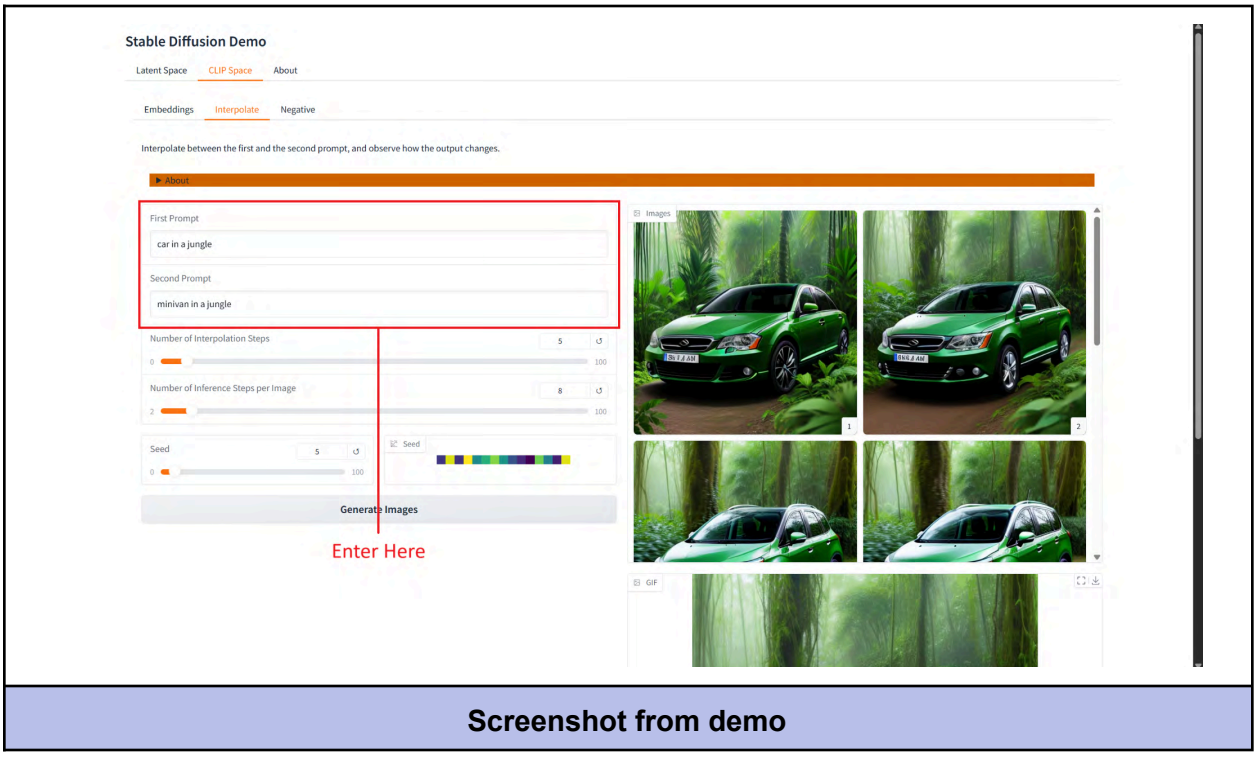
- Visualize how images transition between prompts through interpolation
- Explore how small changes in latent space affect images
- Understand how prompts are represented in the CLIP embedding space

Exploring prompt interpolations

Interpolation shows a smooth transition between two different prompts, blending their concepts together.

Interpolation - Imagine you mix two different colors of paint to get a new color.

Navigation: Go to the "Interpolate" tab under "CLIP Space" (Figure 16)



Screenshot from demo

Figure 16 - Setting two prompts for interpolation.

Watch concepts blend together:

Try this experiment:

1. Enter "king" as the first prompt
2. Enter "queen" as the second prompt
3. Set the "Number of Interpolation Steps" to 4
4. Click on "Generate Images"

 Initial Image	 Interpolation Step: 1	 Interpolation Step: 2
 Interpolation Step: 3	 Interpolation Step: 4	 Final Image

Figure 17 - A succession of images generated from two prompts.

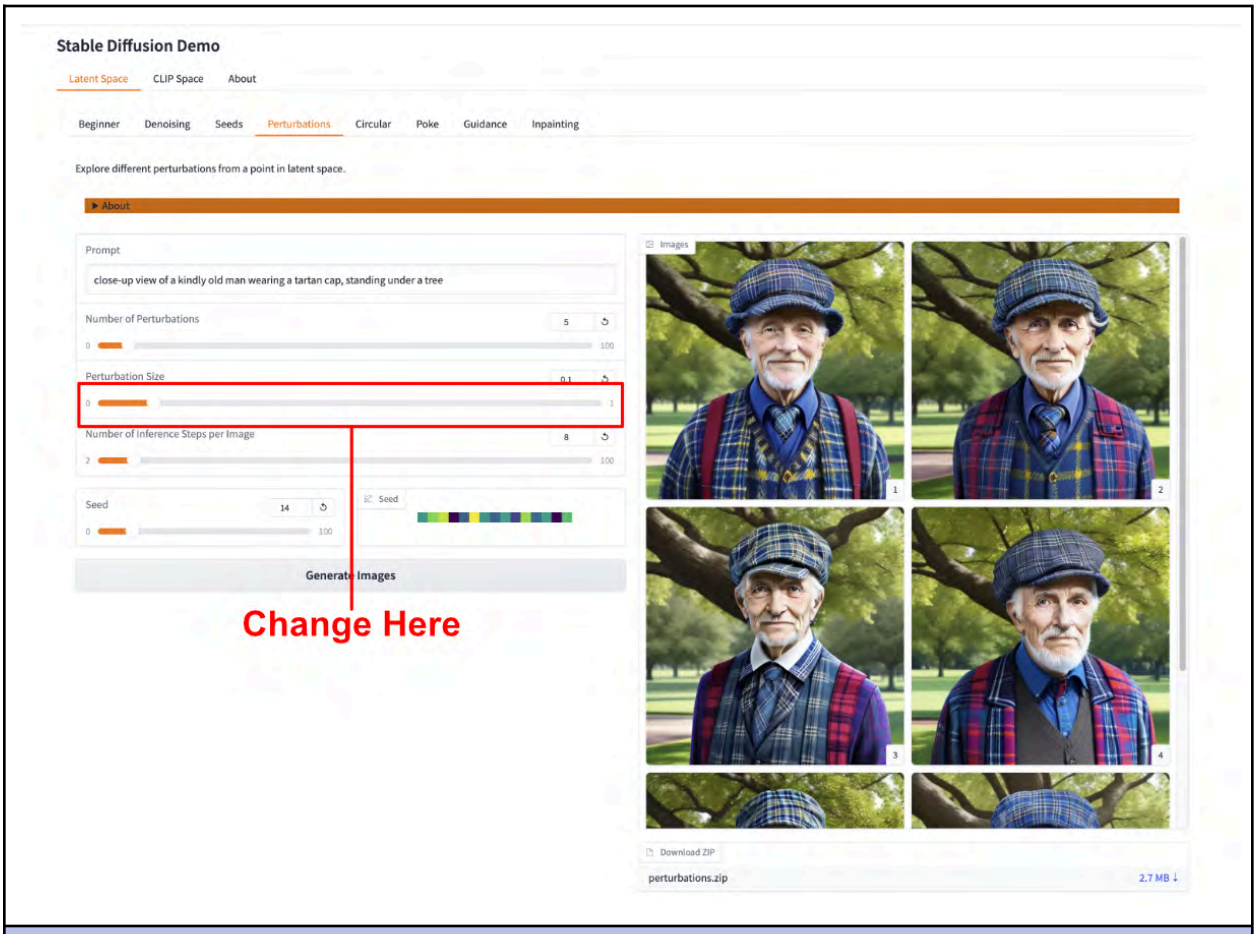
Observe how the images gradually change from a king to a queen (Figure 17). Notice how some images in the middle show features of both

Key takeaway: Prompt interpolation reveals how the diffusion model understands relationships between concepts. The smooth transition shows the diffusion model doesn't just switch between concepts but can blend them naturally.

Exploring latent space

The latent space is where the diffusion model creates its "mental image" of what you're asking for. Points that are close to each other in this space will produce similar-looking images.

Navigation: Go to the "Perturbations" tab under "Latent Space" (Figure 18)



Screenshot from demo

Figure 18 - Setting the perturbation size with the **Perturbation Size** slider.

See how distance in latent space affects similarity:

Try this experiment:

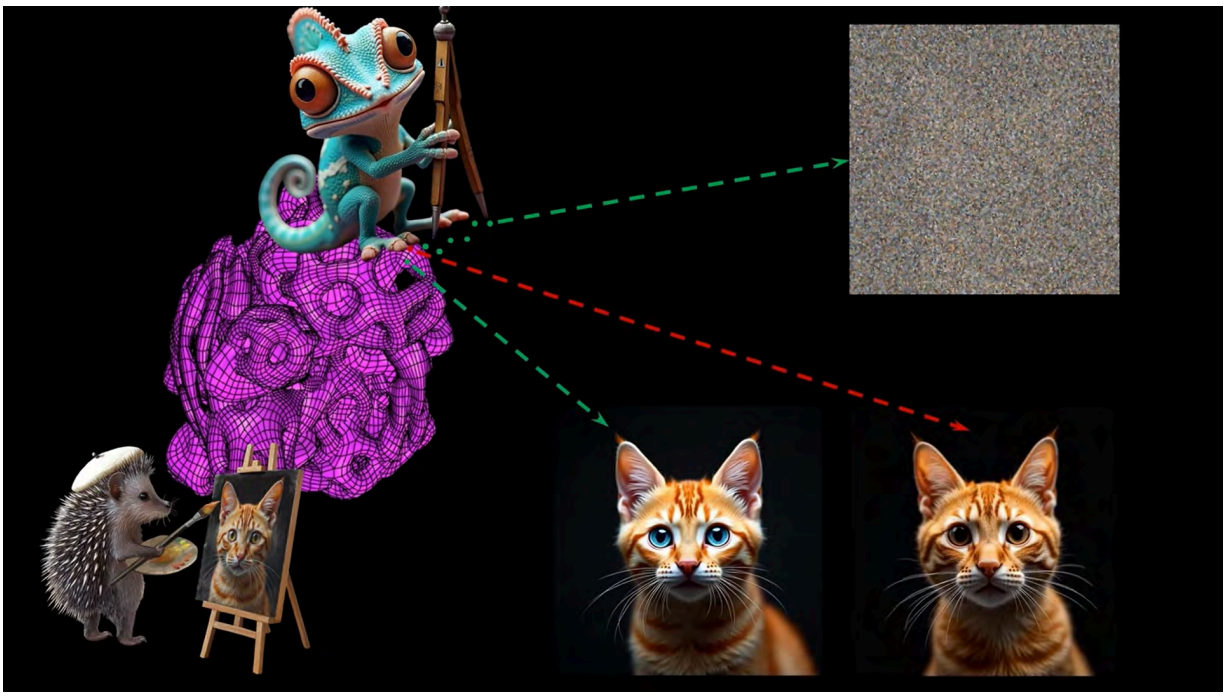
1. Enter the prompt “cat face”
2. Set the “Number of Perturbations” to 3
3. Set "Perturbation Size" to 0.05 and generate
4. Note how all images look very similar (small changes)
5. Now set "Perturbation Size" to 1.0 and generate again
6. Observe how the images are now more different from each other (bigger changes)

Figure 19 - The effect of perturbation values on multiple images generated from the same prompt.

Key takeaway: Small movements in latent space (0.1) create subtle variations like changing lighting or minor details (Figure 19). Larger movements (1.0) can change major elements while keeping the overall concept. This shows how the diffusion model organizes its internal representation of images - like a map where similar concepts are neighbors.

How this experiment connects to the *Cats in Latent Space* Video:

When you set the "Perturbation Size" to 0.05 (small shifts in the starting noise), Oona (UNET) guides these slightly different starting points to nearby locations on the manifold, resulting in cat faces that Vi (VAE) renders as very similar with only subtle variations in details (Figure 20). But when you increase to 1.0 (larger shifts in starting point), Oona (UNET) takes these more widely separated starting points to more distinct locations on the manifold, producing cat faces that Vi (VAE) renders with more significant differences while still maintaining the core concept. This demonstrates how the latent space organizes similar concepts as neighbors, just as the video explains: "Oona will take the shifted point to a slightly different spot on the manifold than before, and this causes Vi to produce a different cat image.



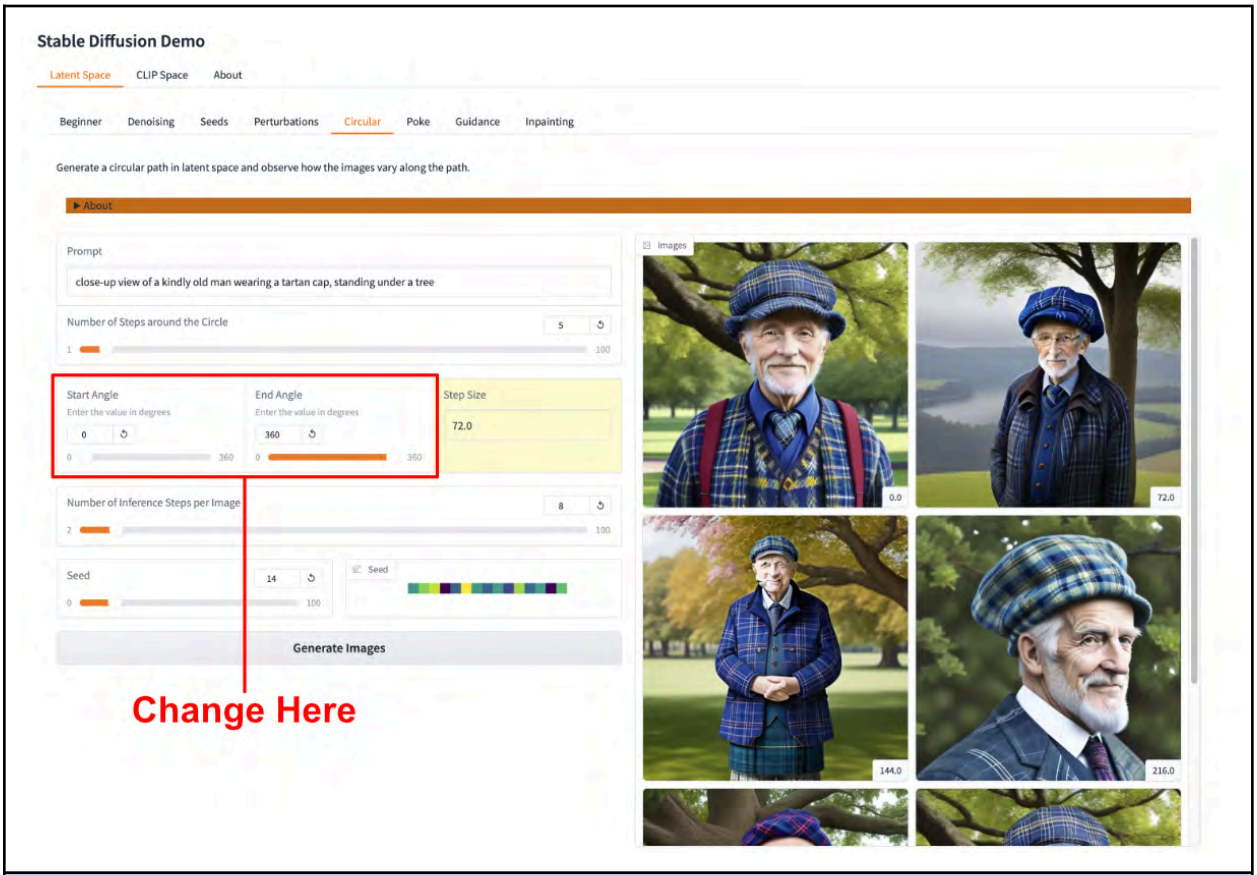
Screenshot from video at timestamp - 6:57

Figure 20 - A screen capture from *Cats in Latent Space* demonstrating small shifts in latent space.

Circular paths through latent space

This visualization creates a loop of images by tracing a circular path through latent space.

Navigation: Go to the "Circular" tab under "Latent Space" (Figure 21)



Screenshot from demo

Figure 21 - Setting the start and end angles for a circular latent space traversal.

See gradual transformation in a circle:

Try this experiment:

1. Enter the prompt "sleeping cat"
2. Change the "Number of Step around the Circle" to 9
3. Press the "Generate Images" button (Figure 22)
4. Then, decrease the "End Angle" to 40°
5. Change the "Number of Step around the Circle" to 10
6. Press the "Generate Images" button again (Figure 23)

 Angle: 0	 Angle: 40	 Angle: 80
 Angle: 120	 Angle: 160	 Angle: 200
 Angle: 240	 Angle: 280	 Angle: 320
	 Angle: 360	

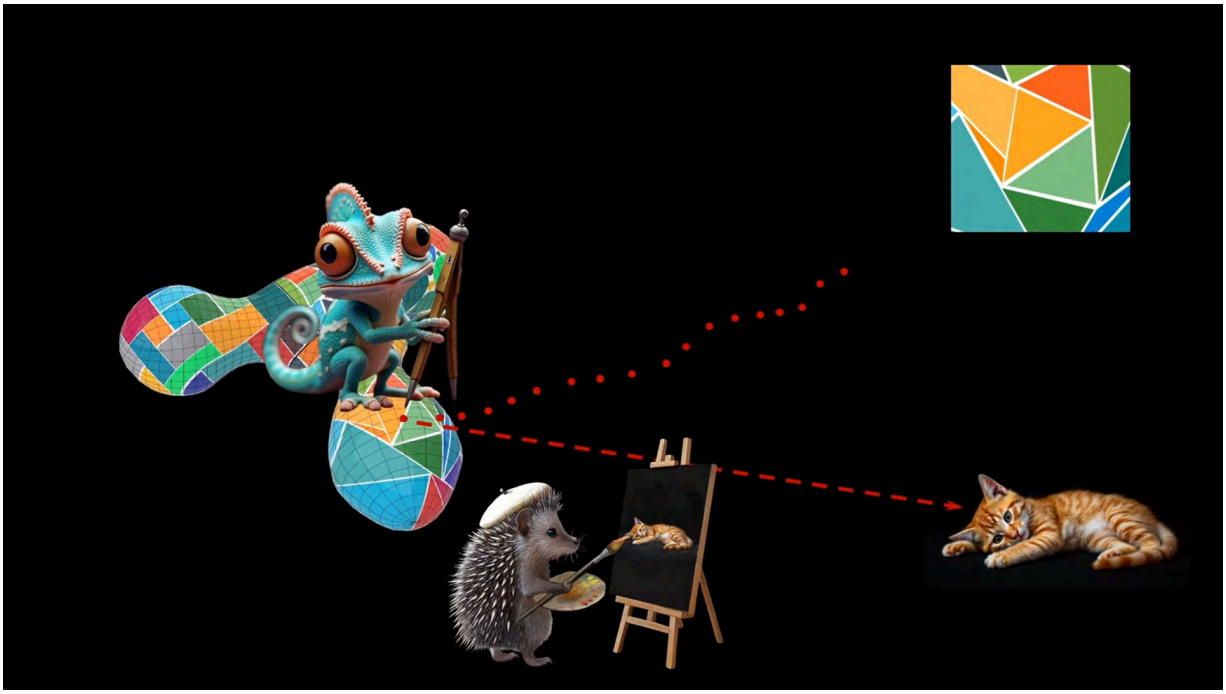
Figure 22 - A succession of images generated at increasing angular values, ending with the same image as the first.

 Angle: 0	 Angle: 4	 Angle: 8
 Angle: 12	 Angle: 16	 Angle: 20
 Angle: 24	 Angle: 28	 Angle: 32
 Angle: 36	 Angle: 40	

Figure 23 - A succession of images generated at increasing angular values in a partial traversal of circular space.

How this experiment connects to the "Cats in Latent Space" Video:

The images around the circle represent cats within the "lobe" of the manifold where Oona (UNET) is currently positioned (Figure 24). This particular lobe corresponds to cats lying down or on their side - distinctly different from other lobes. The diffusion model organizes its "thoughts" (representations) such that all similar cat images cluster together in a specific lobe. When we traverse the lobe in a circle from 0 to 40 degrees, the images remain very similar, confirming that points close together in latent space produce similar images when decoded by Vi (VAE). This demonstrates how the manifold is structured into specialized regions, with each lobe containing variations of a specific concept (in this case, cats lying on their side).



Screenshot from video at timestamp - 10:50

Figure 24 - A screen capture from *Cats in Latent Space* demonstrating lobes in latent space.

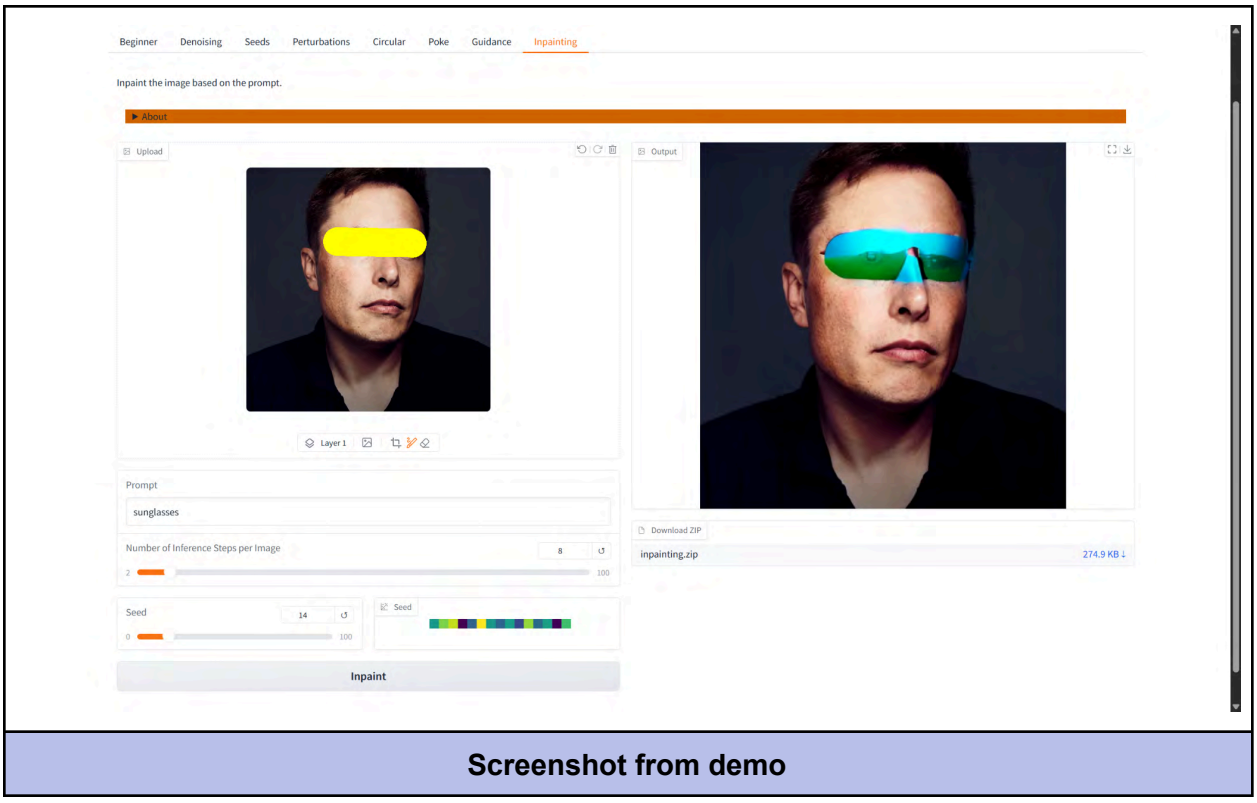
Key takeaway: The diffusion model's latent space is continuous and organized into specialized lobes. When we traverse this particular lobe (containing cats lying down), we see a smooth transition between similar images that gradually change as we move further around the circle.

Exploring inpainting

Inpainting lets you selectively change parts of an image while keeping the rest untouched.

Inpainting - Imagine you have a picture with a small piece missing, like a puzzle. Inpainting is when the model fills in that missing piece based on the rest of the picture.

Navigation: Go to the "Inpainting" tab under the "Latent Space" tab. (Figure 25)



Screenshot from demo

Figure 25 - The inpainting screen.

See how you can modify specific parts of an image:

Try this experiment:

1. Download the sample image from the provided link
2. Upload it to the "Inpainting" tab
3. Use the drawing tool to create a mask over the eyes area
4. Type "sunglasses" in the prompt field
5. Click the "Inpaint" button

	
Image before inpainting	**Image after inpainting**

Figure 26 - Using inpainting to modify a portion of an image.

Watch how the diffusion model adds sunglasses only to the masked area while keeping the rest of the image unchanged (Figure 26).

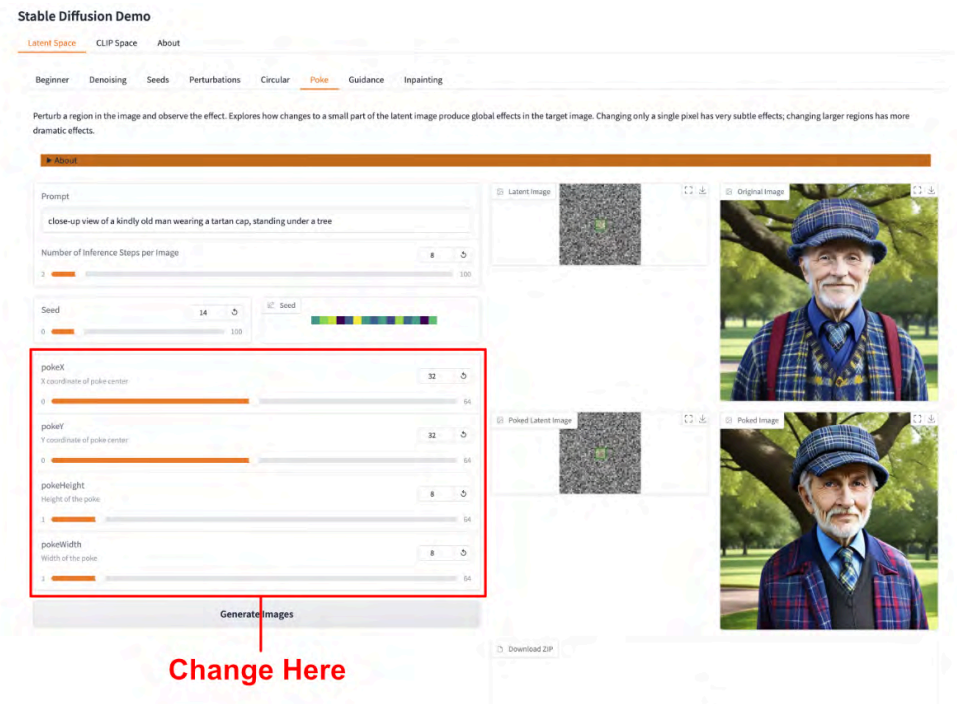
Key takeaway: Unlike other techniques that generate completely new images, inpainting allows for targeted edits. This is useful for:

- Adding or removing objects from photos
- Changing specific features without altering the whole image
- Fixing unwanted elements in an otherwise good image

Latent space manipulation

Small changes to the initial noise pattern can significantly alter the final image.

Navigation: Go to the "Poke" tab under the "Latent Space" tab. (Figure 27)



Screenshot from demo

Figure 27 - Altering poke values.

See how "poking" the latent space changes images:

Try this experiment:

1. Use the prompt "women holding apple" and generate images
2. Observe the original and poked images
3. Set the white box to a very small size (1×1 pixel) using "pokeHeight=1" and "pokeWidth=1"
4. Notice how even this tiny change creates subtle but meaningful differences throughout the entire image (like slight changes in facial features)

5. Move the tiny white box just one pixel to the right using "pokeX"
6. Observe how this minimal shift creates entirely different subtle alterations
7. Now increase the box size (try pokeHeight=32, pokeWidth=32)
8. Note how larger changes essentially create a completely different image, similar to using a new seed

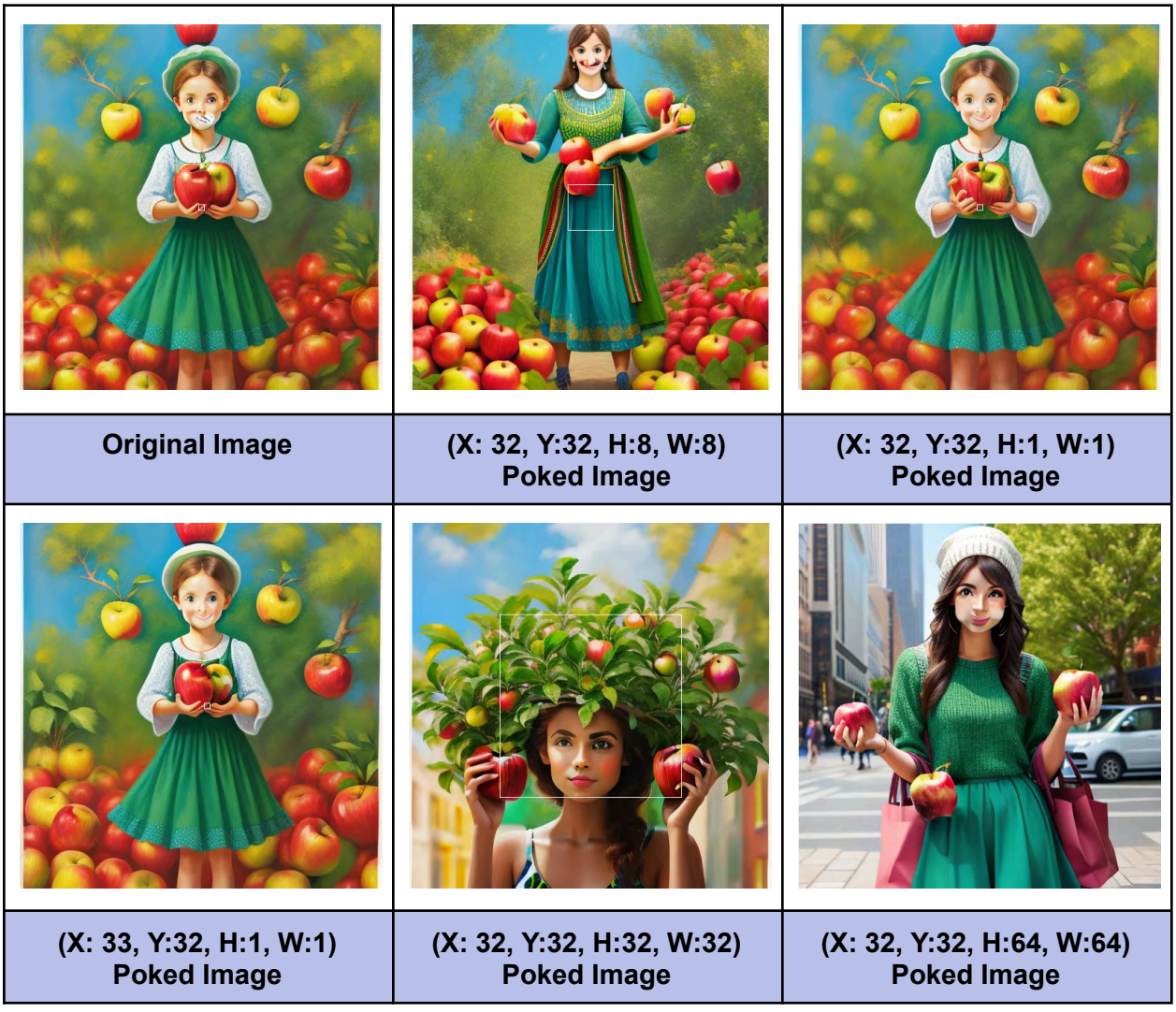
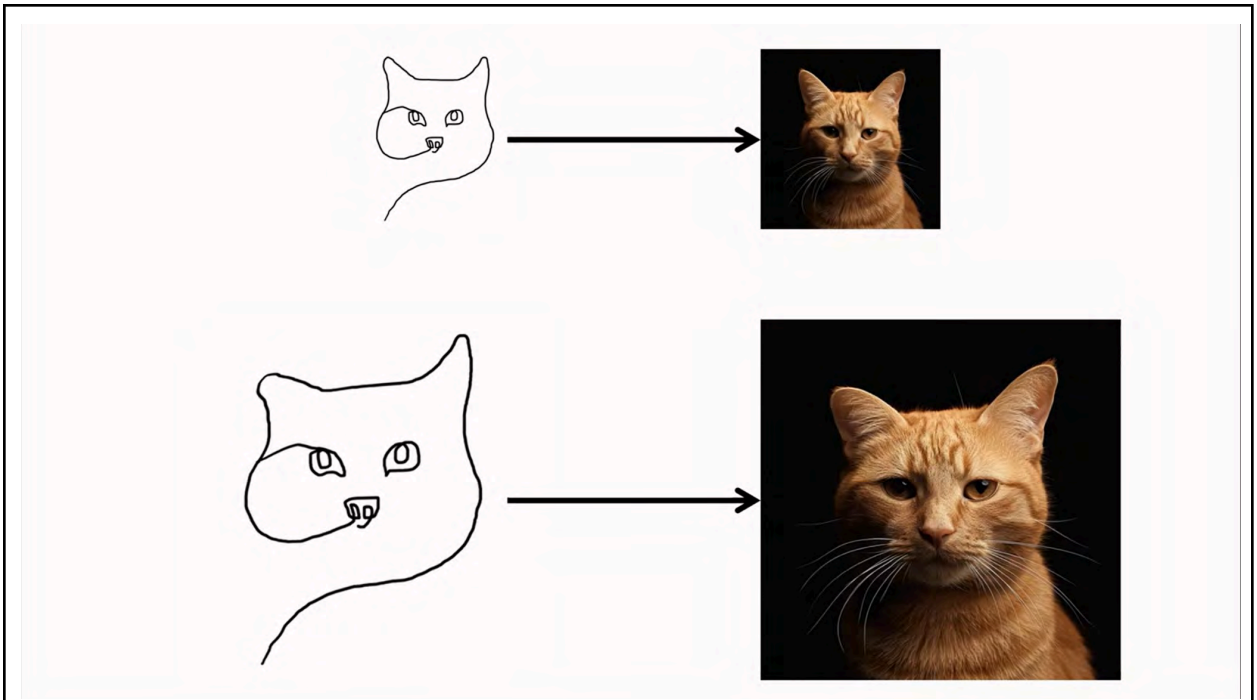


Figure 28 - Small changes in the latent space can result in very different output.

Key takeaway: Unlike inpainting (which modifies a specific part of a finished image), poking the latent space affects the entire image (Figure 28). This demonstrates how interconnected all parts of the latent representation are - changing one area can affect the entire image.

How this experiment connects to the "Cats in Latent Space" Video:

Just as the video shows how simple scribbles contain the "essence" of a cat that Vi can transform into detailed images, the latent representation in diffusion models works the same way (Figure 29). When you "poke" the latent space in the experiment, you're essentially modifying these "scribbles" - the fundamental patterns that guide image creation. As explained in the video: "Any small change to a line results in some change to the image." This is exactly what you're seeing when you poke the latent space - small modifications to the underlying representation cause Vi (VAE) to produce a different final image based on Oona's (UNET) interpretation of these modified "scribbles." The latent representation captures the potential or hidden elements needed to create the complete image, just like how simple lines can suggest a cat that an artist can elaborate into a detailed drawing.



Screenshot from video at timestamp - 1:30

Figure 29 - A screen capture from *Cats in Latent Space* showing a mapping from line art to generated images.

Exploring CLIP space

CLIP is responsible for making the diffusion model understand the essence of text and convert it into visuals.

CLIP - A dictionary that connects words to visual concepts, helping the model translate text into images.

Navigation: Go to the "Embeddings" tab under the "CLIP Space" tab (Figure 30).

Embeddings - Special numerical codes that represent words in a way the model can understand.

Screenshot from demo

Figure 30 - A view of the CLIP space.

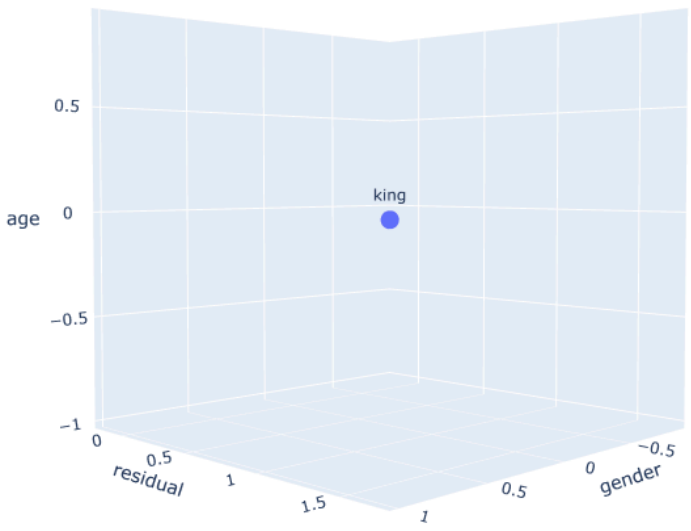
See how words are represented in 3D semantic space:

Semantics - The meaning behind words and images.

Try these experiments:

Experiment 1: Visualizing individual concepts

1. Clear any existing words from the visualization
2. Add the word "king" to see its position in 3D space (Figure 31)



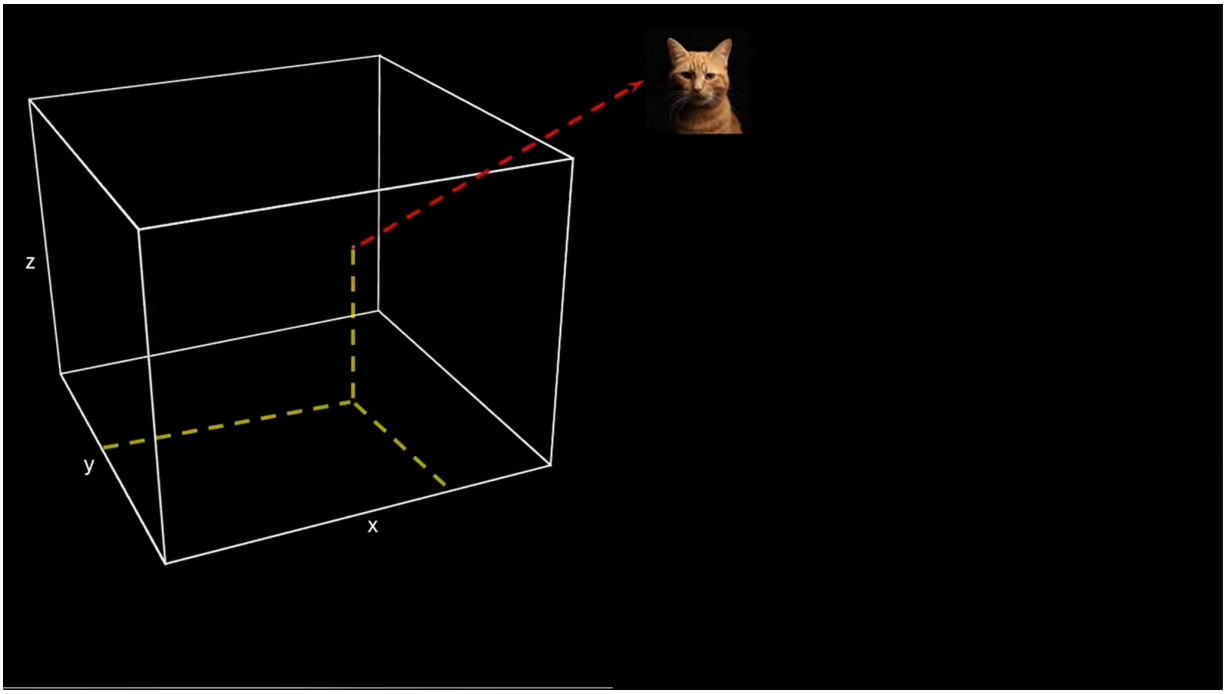
Screenshot of 3D space

Figure 31 - Adding the word “king” to the CLIP space.

Notice how a single concept is represented by coordinates in latent space.

How this experiment connects to the "Cats in Latent Space" Video:

This experiment directly illustrates the video's explanation that "we could describe a cat by writing down three numbers: x, y, and z." (Figure 32) Just as the video explains that "every latent cat would be a point at those coordinates in a three-dimensional latent cat space," you can see how the word "king" exists as a specific point in the CLIP embedding space. While the actual embedding space has many more dimensions, this 3D visualization helps us understand how the diffusion model encodes concepts as precise numerical coordinates. This mathematical representation is how the model "understands" what you're asking for - each concept has its own unique address in semantic space, just like every cat has a specific location in the latent cat space described in the video.

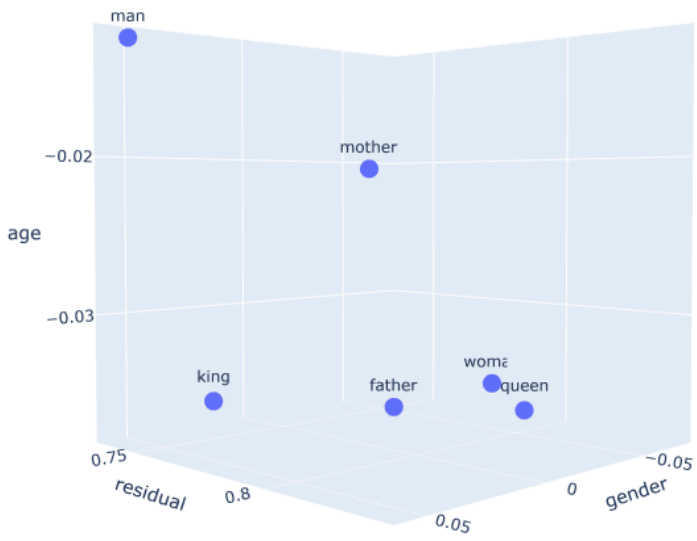


Screenshot from video at timestamp - 2:15

Figure 32 - A screen capture from *Cats in Latent Space*, demonstrating how a particular cat might be represented in CLIP space.

Experiment 2: Discovering manifolds

Clear words and add: "king queen man woman father mother"



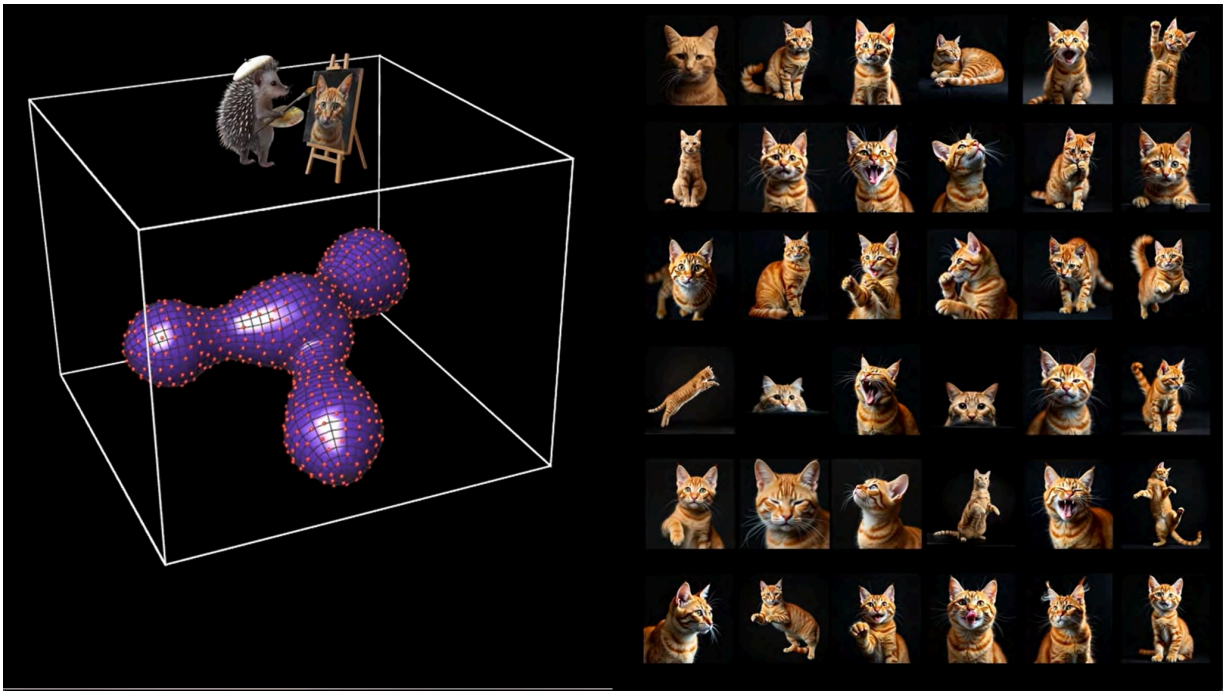
Screenshot of 3D space

Figure 33 - Multiple words in CLIP space.

Notice how words form a complex surface, or manifold (Figure 33). Rotate the view to see how these words relate to each other. Observe how words with similar meanings cluster together

How this experiment connects to the "Cats in Latent Space" Video:

This experiment directly illustrates the video's explanation that "after sampling enough points, they saw that the points lie on a complex surface, called a manifold." When you add related words like "king," "queen," "man," "woman," "father," and "mother," you can observe this in action. The words form a structure in the embedding space. This is the manifold described in the video - a surface where related concepts exist in relationship to each other. (Figure 34)

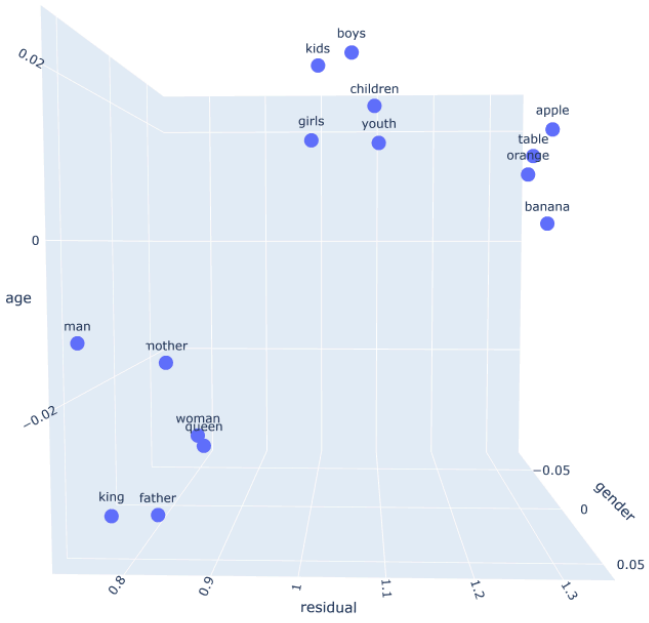


Screenshot from video at timestamp - 2:34

Figure 34 - A screen capture from *Cats in Latent Space*, showing multiple cats mapped to CLIP space.

Experiment 3: Exploring multiple manifolds

1. Clear words and add: "king queen man woman father mother boys girls kids children youth apple orange banana table"
2. Rotate the view so the residual dimension faces toward you



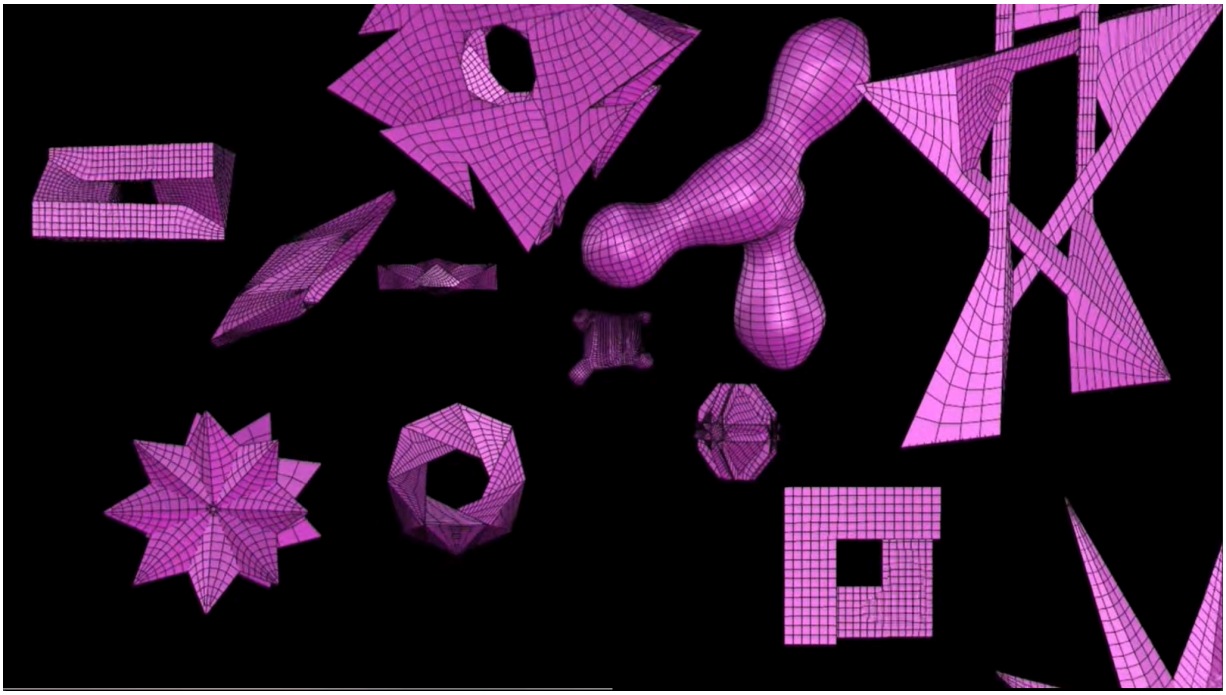
Screenshot of 3D space

Figure 35 - Organization of multiple words in CLIP space.

Notice how different concept categories form separate manifolds (Figure 35). Observe how the model organizes concepts into semantic groups.

How this experiment connects to the "Cats in Latent Space" Video:

This experiment directly illustrates the video's statement that "if Vi knows how to draw many different things, they have learned many manifolds." (Figure 36) As you add diverse words from different categories (people, fruits, furniture), you can see how these form distinct manifolds in separate regions of the space, just as the video explains that "trees form another manifold in a different part of the latent space" from cats. By rotating the view to see the residual dimension, you're essentially looking at how the model organizes its knowledge of different categories, keeping related concepts together while separating unrelated ones.



Screenshot from video at timestamp - 3:10

Figure 36 - A screen capture from *Cats in Latent Space*, showing manifolds for various concepts.

Not Safe For Work (NSFW) Filtering in DiffusionDemo

The Diffusion Demo incorporates an important safety feature—the Stable Diffusion Safety Checker—which automatically filters potentially inappropriate content. This system helps ensure the demo remains appropriate for educational settings.

How NSFW Filtering Works: When you enter potentially problematic prompts (like "sexy teenage girl"), the safety filter activates, returning a black image instead of potentially inappropriate content (Figure 37).

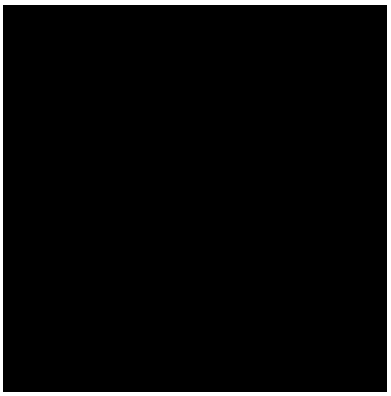
	
Without NSFW Filtering	**With NSFW Filtering**

Figure 37 - Effect of the NSFW filter.

Behind the Scenes: The safety checker uses a CLIP-based neural network model that analyzes the generated image by comparing it against known patterns of inappropriate content. If the system detects too much similarity to problematic reference patterns, it automatically blocks the image from being displayed.

Filtering Generated Images by Concept Similarity

This tab demonstrates how you can automatically filter out generated images based on their similarity to a specified concept. In other words, it lets you set a content rule *before* generation: any image that is too similar to the “forbidden” concept will be flagged and removed. This is useful for avoiding certain subjects in your results (for example, filtering out images containing cats if you’re allergic to them!) by using a filter threshold to decide what gets through.

Filter Concept – The keyword or phrase for the content you want to screen out. The demo will check each generated image for visual similarity to this concept. For example, if the filter concept is “cat,” the system will look for cat-like features in every image.

💡 **Filter Threshold** – A sensitivity setting (a small number, typically between -1 and 1) that determines how much similarity is “too much.” If an image’s similarity score to the filter concept exceeds this threshold, that image is filtered (flagged and not allowed through). Lowering the threshold makes the filter stricter (even a slight hint of the concept will be flagged); raising it makes the filter more lenient.

Navigation: Go to the “Filter” tab under the “CLIP Space”

Try this experiment: Follow these steps to see the filter in action.

1. In the Prompt field, enter “A child with their pet.”
2. In the Filter Concept field, type “cat.”
3. Set the Filter Threshold to 0.03
4. Click the Generate Images button.

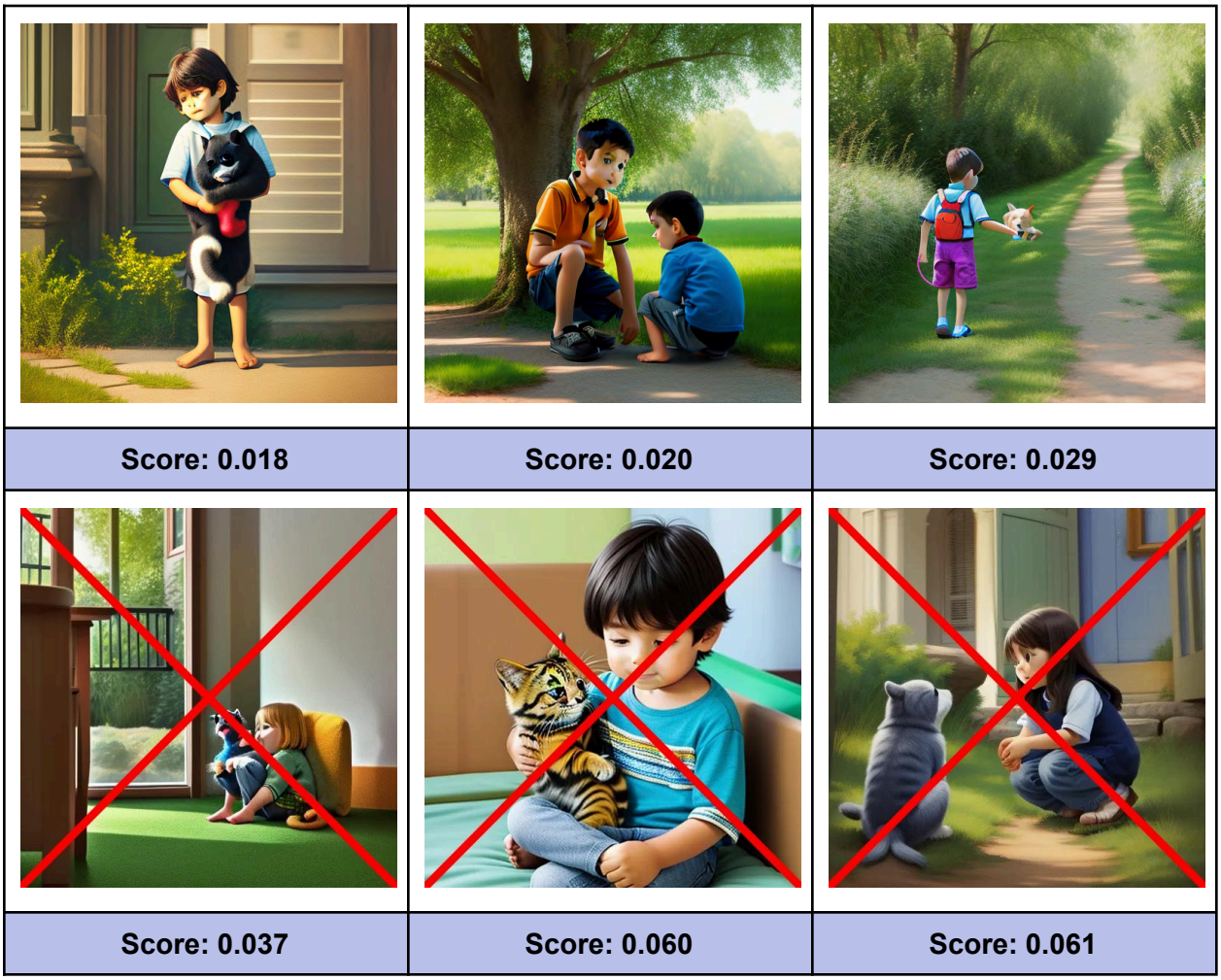


Figure 38 - Filtering images that include a cat.

Each generated image will display a score (for how “cat-like” it is). Any image with a score above 0.03 will be flagged and filtered out (Figure 38, second row). Images with a score below 0.03 remain unfiltered (Figure 38, first row). You should see that images where the child’s pet is not a cat (e.g. a dog or another animal) have low scores (under the threshold) and appear normally, while any image where the child’s pet is a cat will have a higher score (exceeding 0.03) and thus be marked with a red X (filtered).

Key takeaway: Content filtering in diffusion models is a powerful tool for controlling image outputs. By understanding and tuning the filter concept and threshold, you can decide which concepts are allowed or blocked in the generated images. This experiment shows how a diffusion model, assisted by a similarity-checking mechanism (much like the built-in NSFW safety filter), can automatically screen out unwanted content. Adjusting the threshold is crucial – it lets you find a balance between catching all instances of an undesired concept and avoiding false alarms. In summary, the Filter tab helps you explore how to enforce content guidelines on AI-generated images, giving you greater control over the model’s creativity.

Other Resources

- Diffusion Explainer: Stable Diffusion Explained with Visualization
- The Illustrated Stable Diffusion
- UMAP Zoo
- Word Embedding Demo

Collection:

- Stable diffusion: resources and discussion - Part 2 2022/23 - fast.ai Course Forums
    - https://scorebasedgenerativemodeling.github.io/
    - https://huggingface.co/blog/stable_diffusion
    - https://github.com/huggingface/notebooks/tree/main/diffusers
    - https://huggingface.co/blog/annotated-diffusion
    - https://pharmapsychotic.com/tools.html

- https://sdtools.org/
- https://jalammar.github.io/illustrated-stable-diffusion/
- Sebastian Kamph - YouTube
- enigmatic_e - YouTube
- How to Train Your Series
- parrot zone | Notion
- https://course.fast.ai/Lessons/part2.html

- https://stable-diffusion-art.com/
- Beginner's Guide to Getting Started With Stable Diffusion
- https://prompthero.com/resources
- https://huggingface.co/blog/stable_diffusion

- GitHub - alen-smajic/Stable-Diffusion-Latent-Space-Explorer
- Stable Diffusion: Latent Exploration | Jerome Swannack
- Uncover the Magic of Stable Diffusion: Exploring the Latent Space
- Cat Creator

Vocabulary Flash Cards

LATENT

Potential, hidden, or undeveloped.

LATENT SPACE

A compressed representation that captures the essential qualities of things.

LATENT IMAGE

A compressed representation of an image as a point in a latent image space.

NOISE

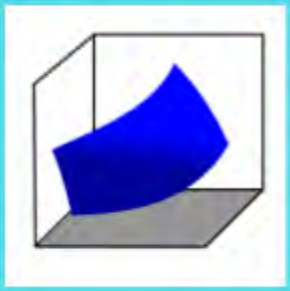
Random values that obscure an image.

DIFFUSION ALGORITHM

A process that incrementally removes noise from a latent image.

STABLE DIFFUSION

A particular diffusion algorithm used to turn prompts into images.



MANIFOLD

A collection of points forming a surface embedded in a high dimensional space.

DENOISING

Removing noise from a latent image, moving it closer to the manifold of noise-free images.

PROMPT

A word or phrase describing the image to be generated.

INFERENCE STEP

One step of noise removal; typically 8-20 steps are required.

SEED

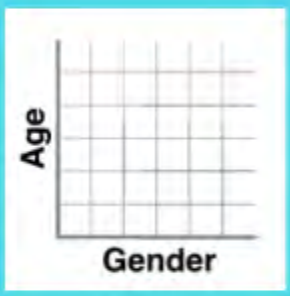
A value used to kick off a sequence of random numbers forming an initial image that is pure noise.

CLIP

Contrastive Language-Image Pretraining, a neural network model that learns embeddings for images and their captions.

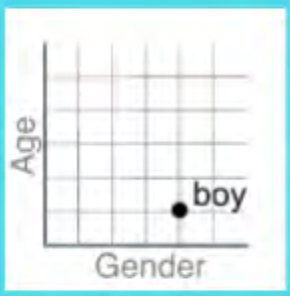
VAE

Variational Auto-Encoder, a neural network that turns a point in a latent image space into an actual image.



SEMANTIC SPACE

A coordinate system which each axis represents some aspect of meaning, such as "young/old" or "male/female".

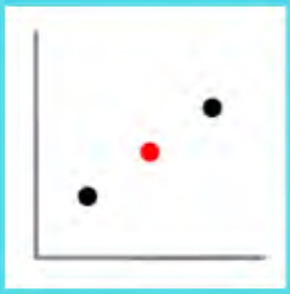


EMBEDDING

A representation of meaning as a point in a semantic space.

SEMANTICS

Meaning carried by a word, a phrase, or any piece of language.



INTERPOLATION

Estimating a value that lies between two known values.

GUIDANCE SCALE

Parameter that controls how strictly a prompt should be followed.

INPAINTING

Replacing some portion of an image with different content, such as adding sunglasses to a face.

This work was funded by a grant from NEOM Company and by National Science Foundation Award IIS-2112633.

DiffusionDemo uses Jetstream2 at Indiana University through allocation CIS240773 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.