Object Customization with Textual Inversion

STABLE DIFFUSION MODEL

Stable Diffusion is a popular text-to-image diffusion model developed by Stability AI. It generates images from text prompts by gradually removing noise from a random starting point. The model works by:

1. Starting with pure noise (random pixels)
2. Progressively "denoising" the image in steps
3. Using text guidance to steer the denoising process toward images that match the description

TEXTUAL INVERSION

Textual Inversion is a technique that allows you to "teach" Stable Diffusion new concepts using only a few example images. Here's how it relates to Stable Diffusion:

- Stable Diffusion has a text encoder that converts text prompts into embeddings (numerical representations) that guide the image generation
- Textual Inversion creates a new "pseudo-word" (often written like <my-concept>) with a custom embedding that represents your specific concept
- This custom embedding is learned by optimizing it to reproduce your reference images when used in prompts
- Once trained, you can use this pseudo-word in prompts as if it were part of the model's original vocabulary

OBJECT CUSTOMIZATION WITH TEXTUAL INVERSION

Collect reference images: Gather 3-5 images of the specific object we want to customize (e.g., your pet, a unique piece of furniture, etc.)

Training process:

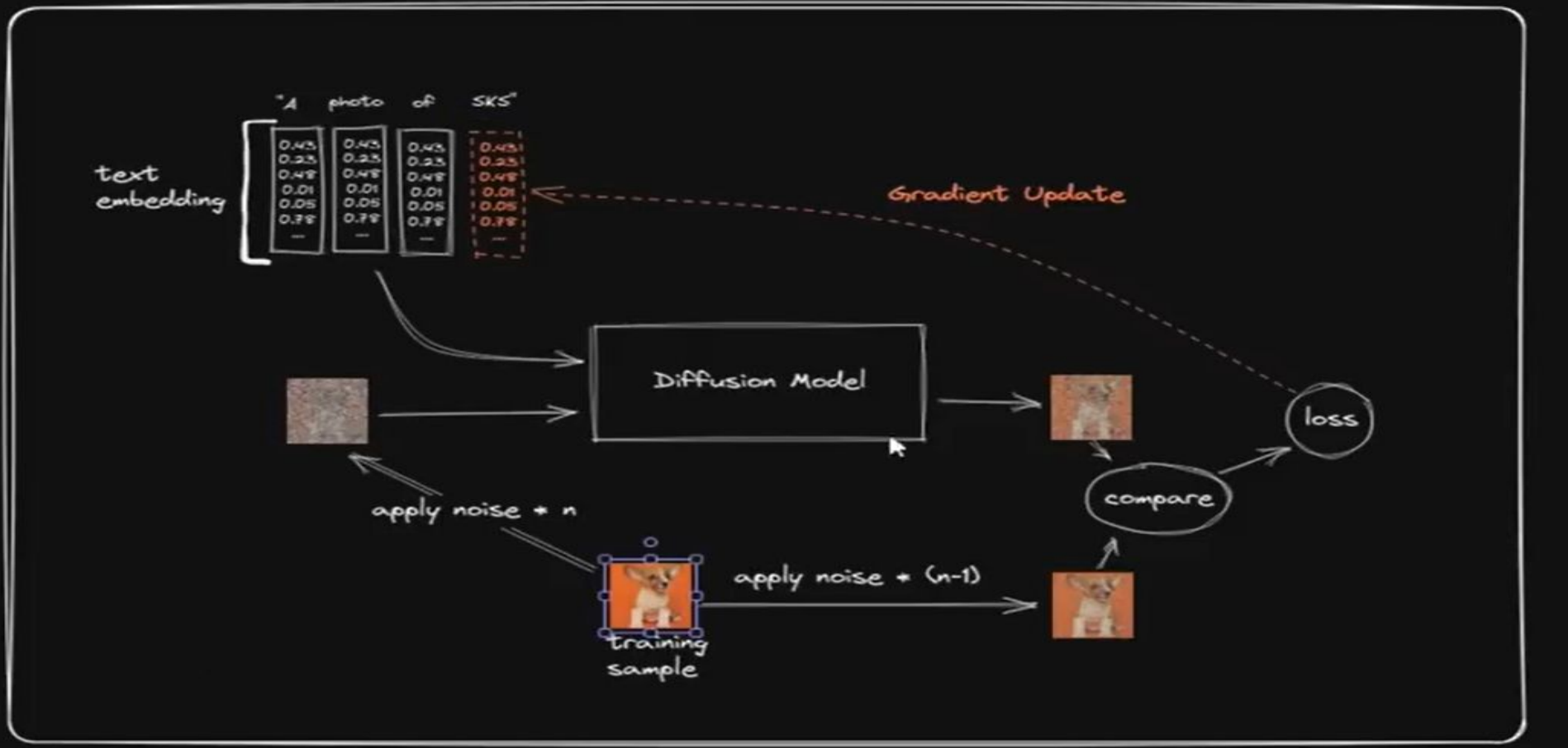
- Choose a placeholder token (e.g., <my-cat>)
- Initialize this token's embedding randomly or from a similar concept
- Train only this embedding while keeping the rest of the model frozen
- The training process optimizes the embedding to reconstruct your reference images when prompted

Using your custom concept:

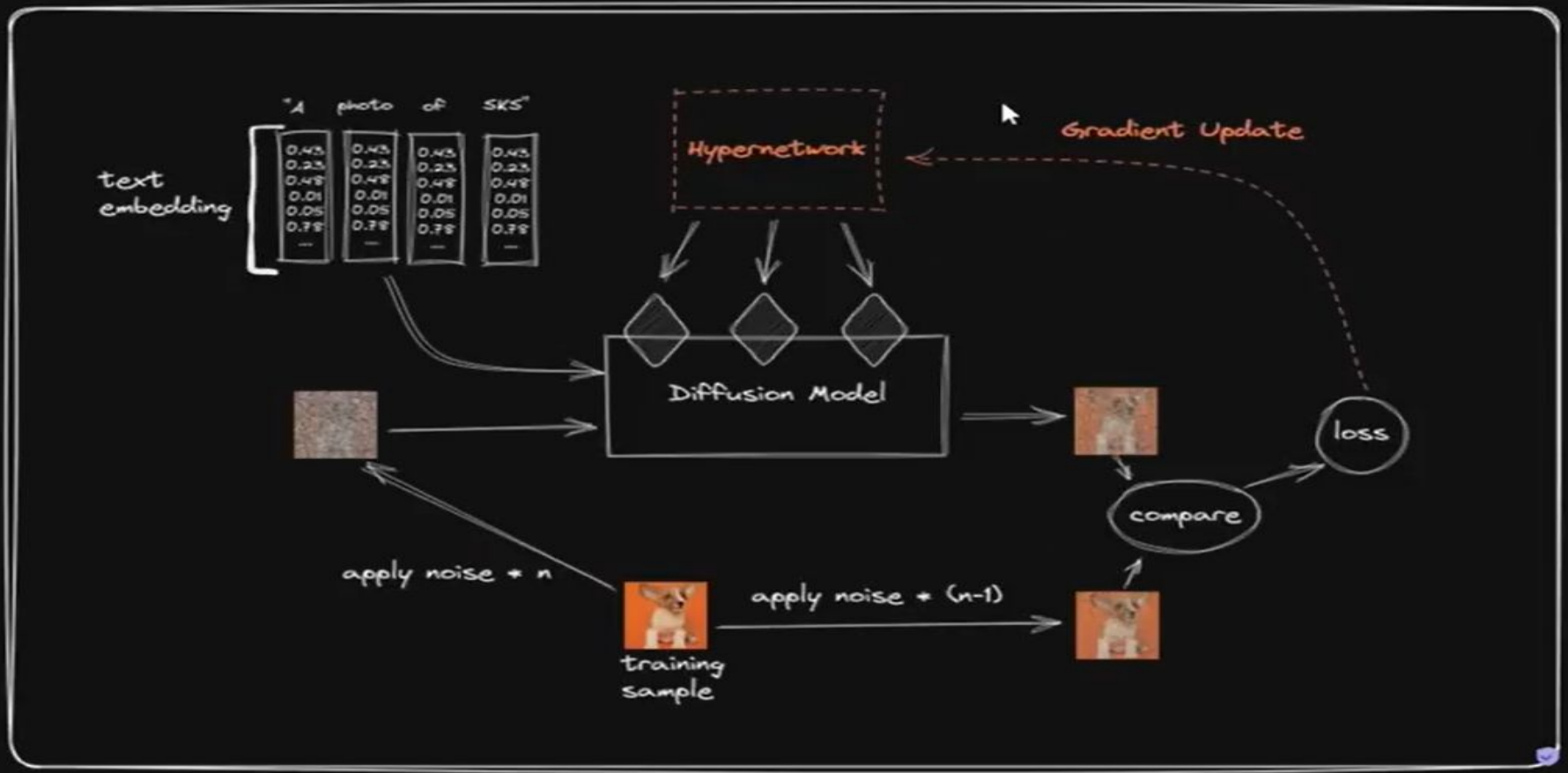
- Once trained, you can use the token in prompts: "A painting of <my-cat> in a renaissance style"
- You can combine it with other styles, settings, and concepts: "<my-cat> on the moon", "A cartoon version of <my-cat>"
- The model will generate images that maintain the essence of your specific object while applying the new context

Textual Inversion

+ output is a tiny embedding

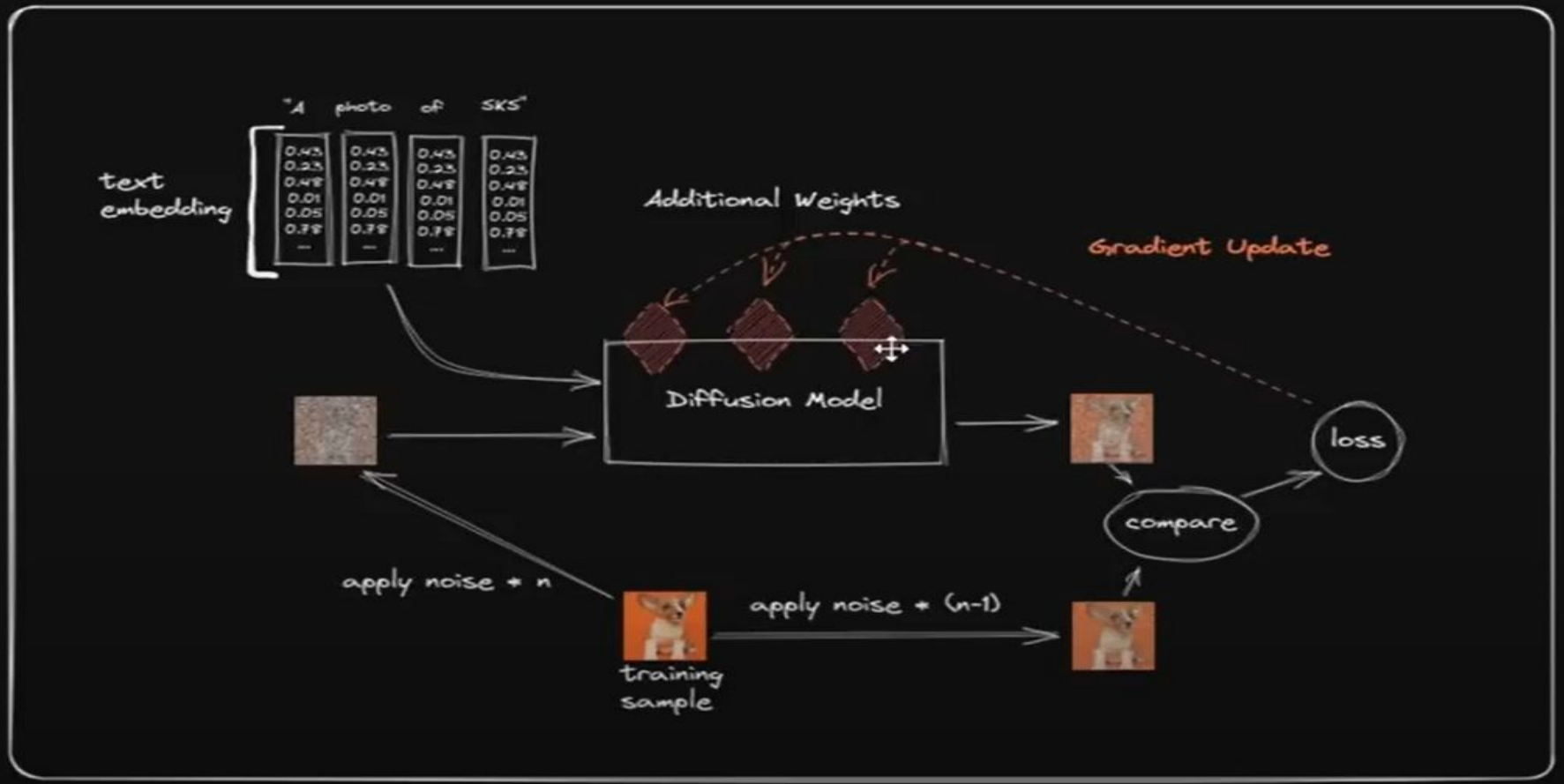


Hypernetworks



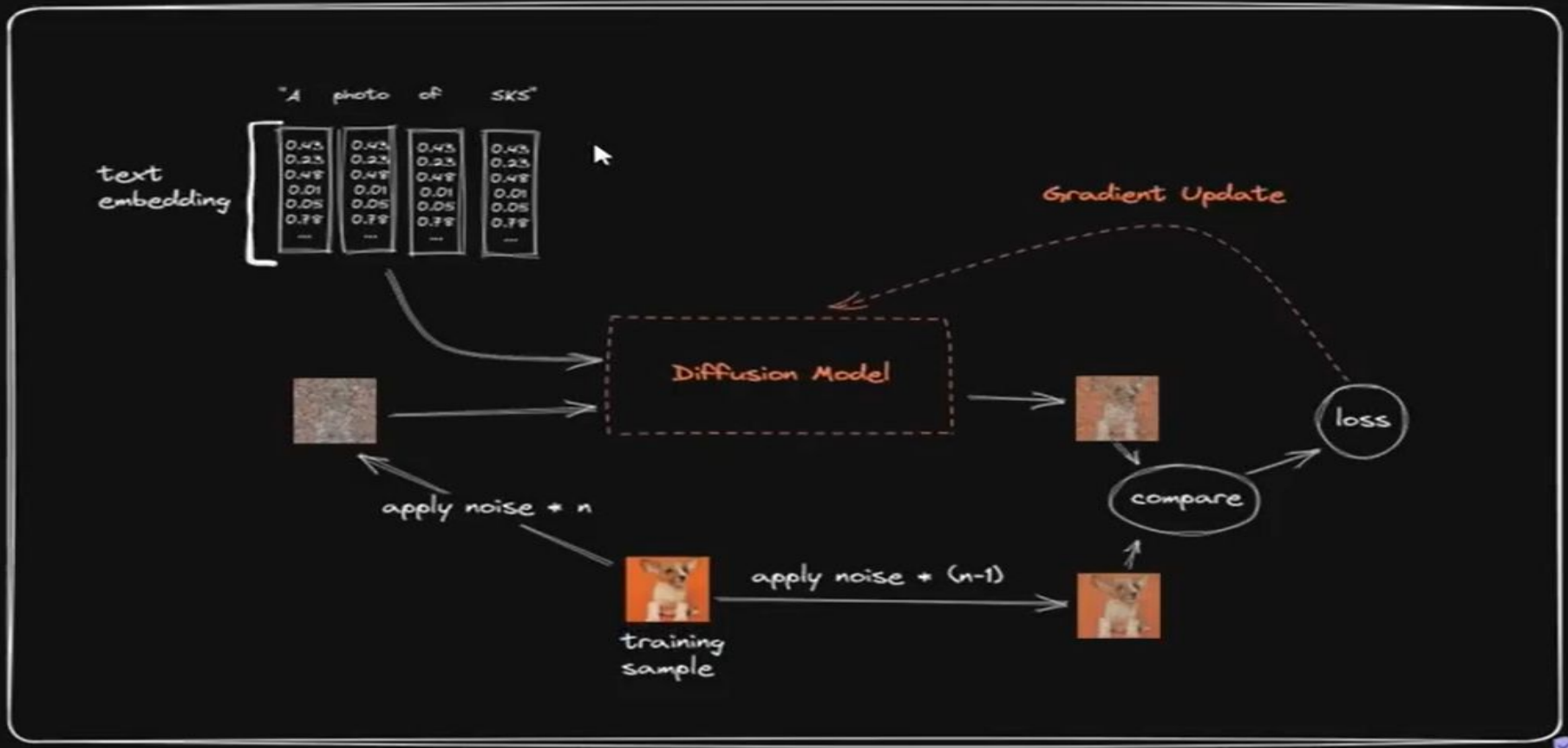
LoRA

+ quick to train



Dreambooth

+ probably the most effective
- storage inefficient (whole new model to deal with)



SHIFU

SHIFU

Riding a Bicycle

Surfing

in a Bathtub

Printed on a Surfboard

in a Nest on a Tree

Doing Yoga

CODE IMPLEMENTATION

https://colab.research.google.com/drive/1myVKmZtVDIp8M6o7xhGes2Uc3Tc_aAZ6?usp=sharing

Generated Output:

THANK YOU