Lesson 6: Connecting Text and Images

6.1 Components of a Multimodal Model

6.2 Vision-Language Understanding

6.3 Contrastive Language-Image Pretraining

6.4 Embedding Text and Images with CLIP

6.5 Zero-Shot Image Classification with CLIP

6.6 Semantic Image Search with CLIP

6.7 Conditional Generative Models

Lesson 6: Connecting Text and Images

6.8 Introduction to Latent Diffusion Models

6.9 The Latent Diffusion Model Architecture

6.10 Failure Modes and Additional Tools

6.11 Stable Diffusion Deconstructed

6.12 Writing Our Own Stable Diffusion Pipeline

6.13 Decoding Images from the Stable Diffusion Latent Space

6.14 Improving Generation with Guidance

6.15 Playing with Prompts

6.1

Components of a Multimodal Model

Multitudes of Media

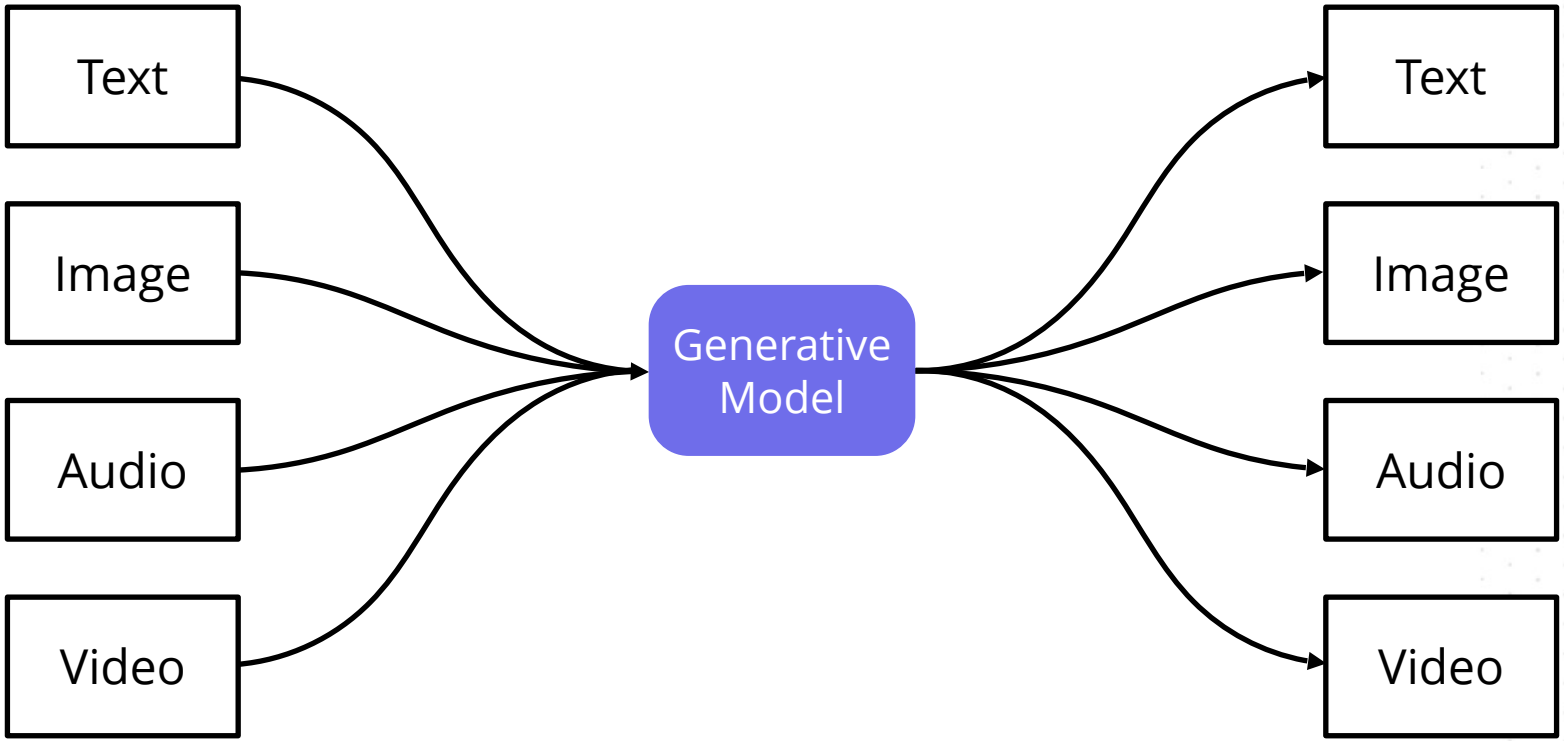
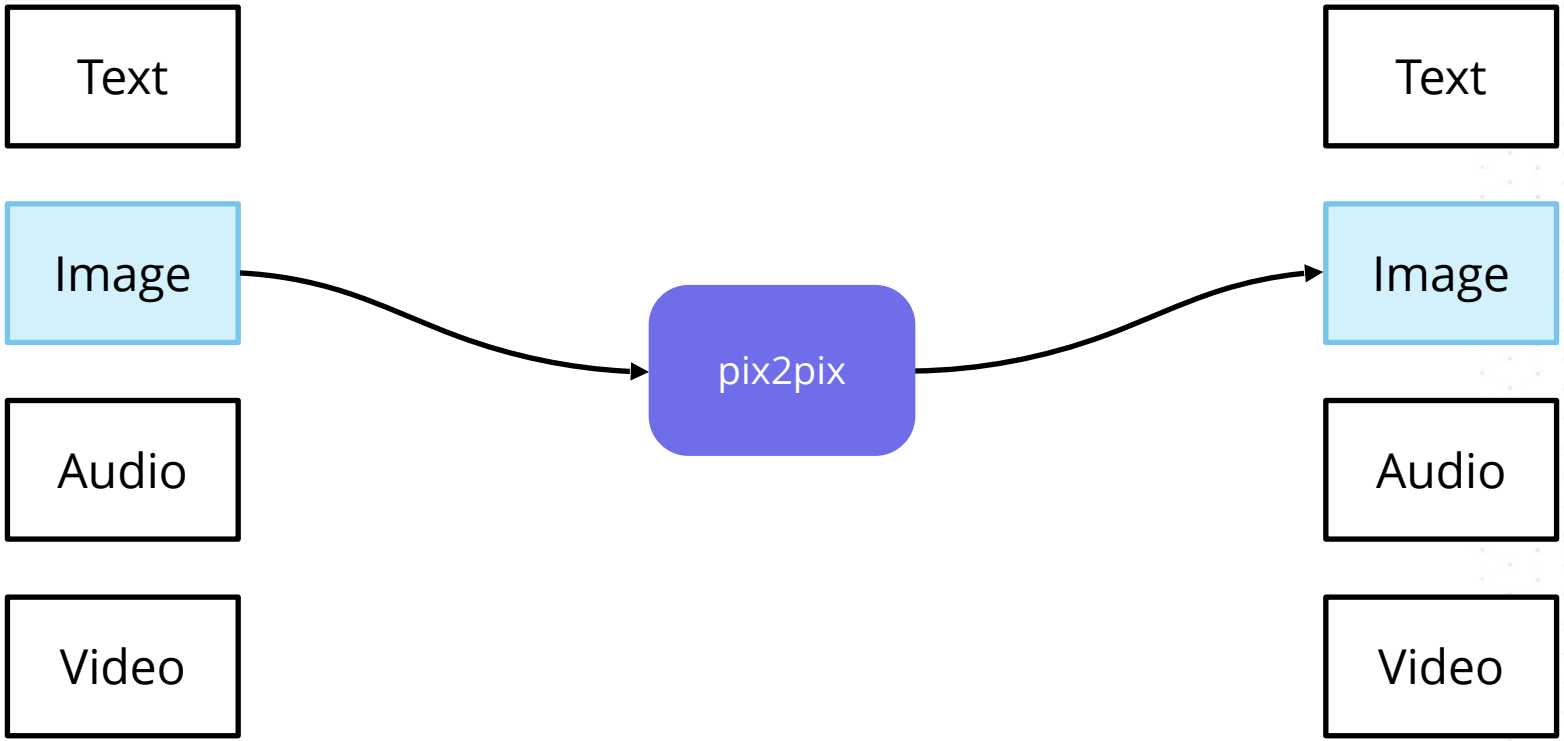


Image-to-Image (Upscaling, Restoration, Inpainting, Other)



Text-to-Image

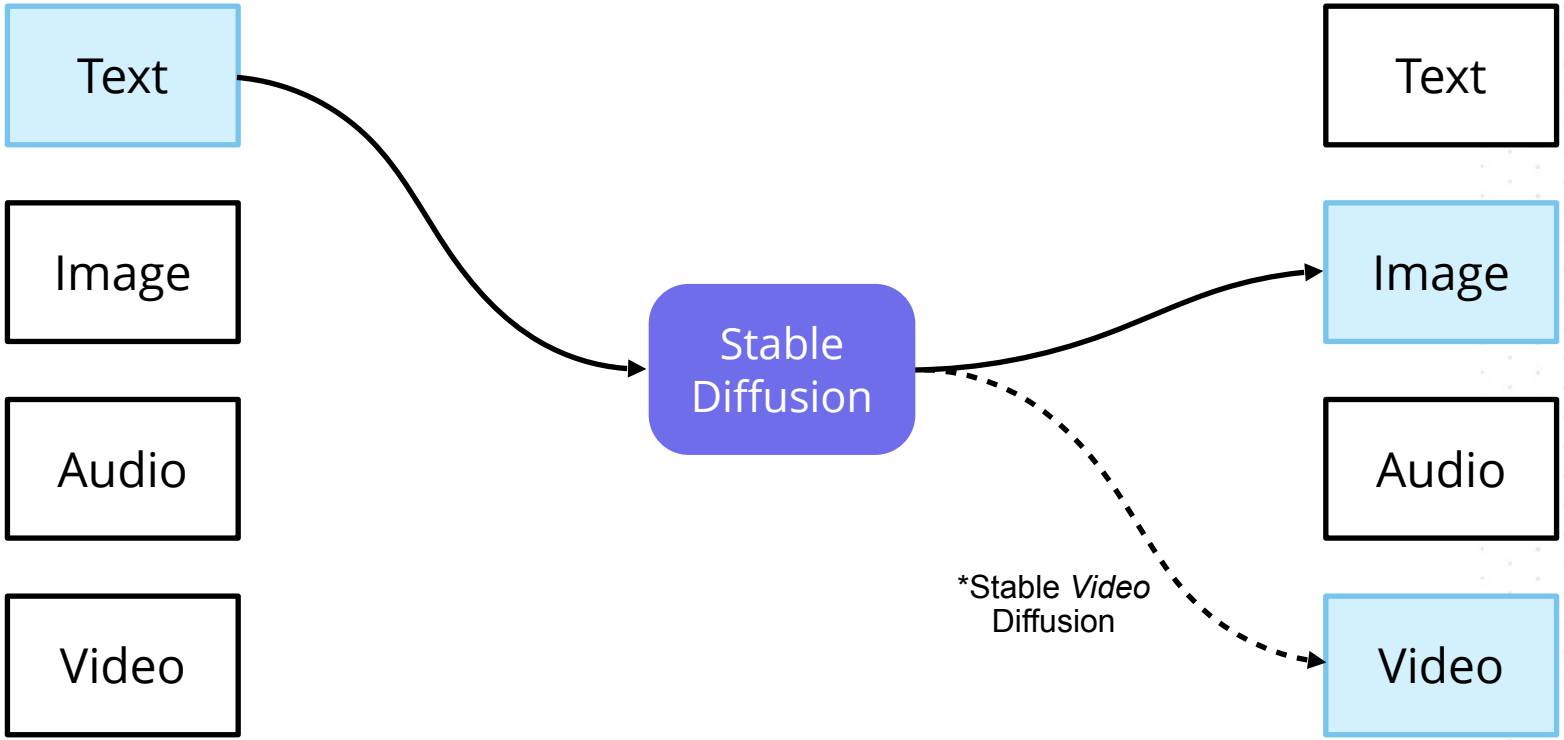
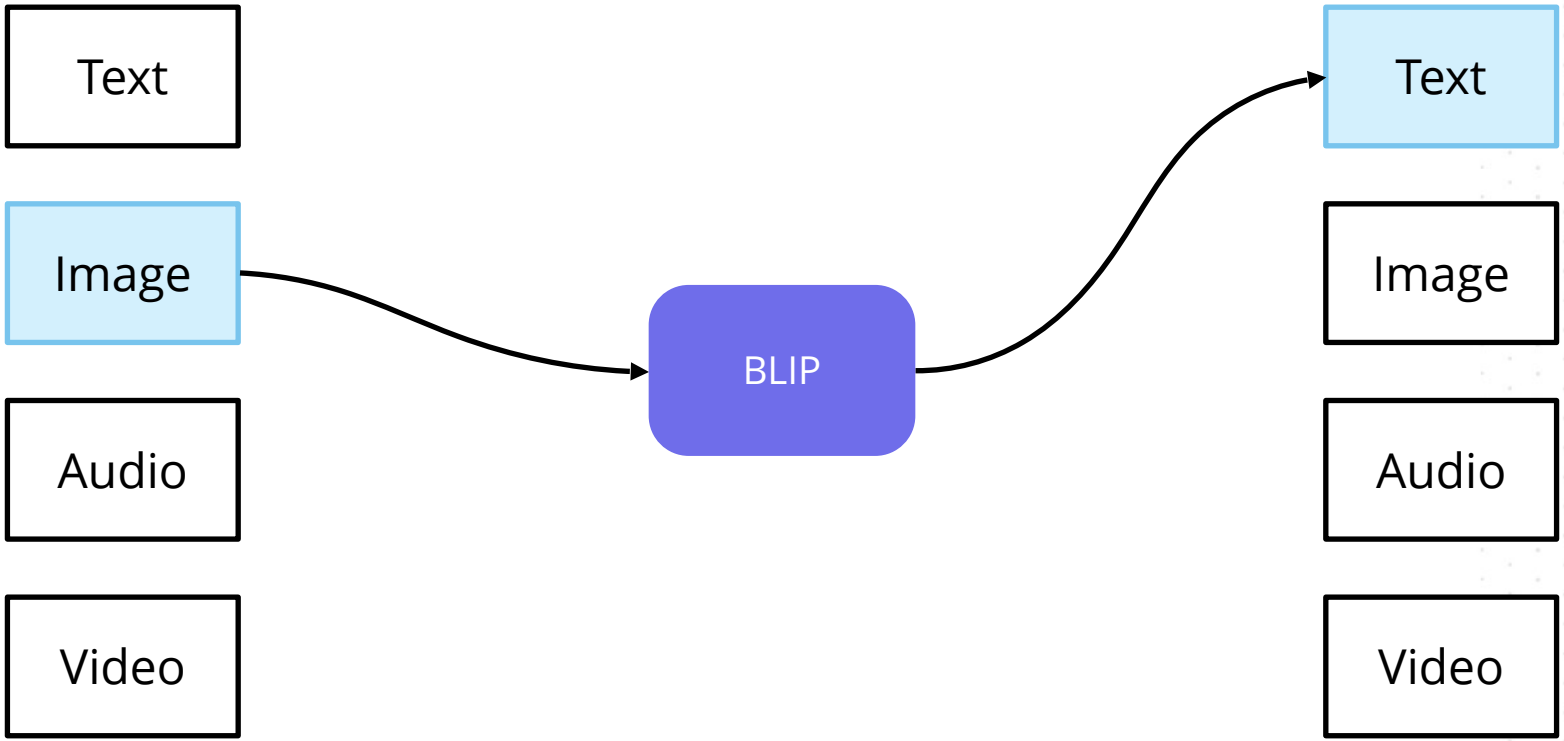
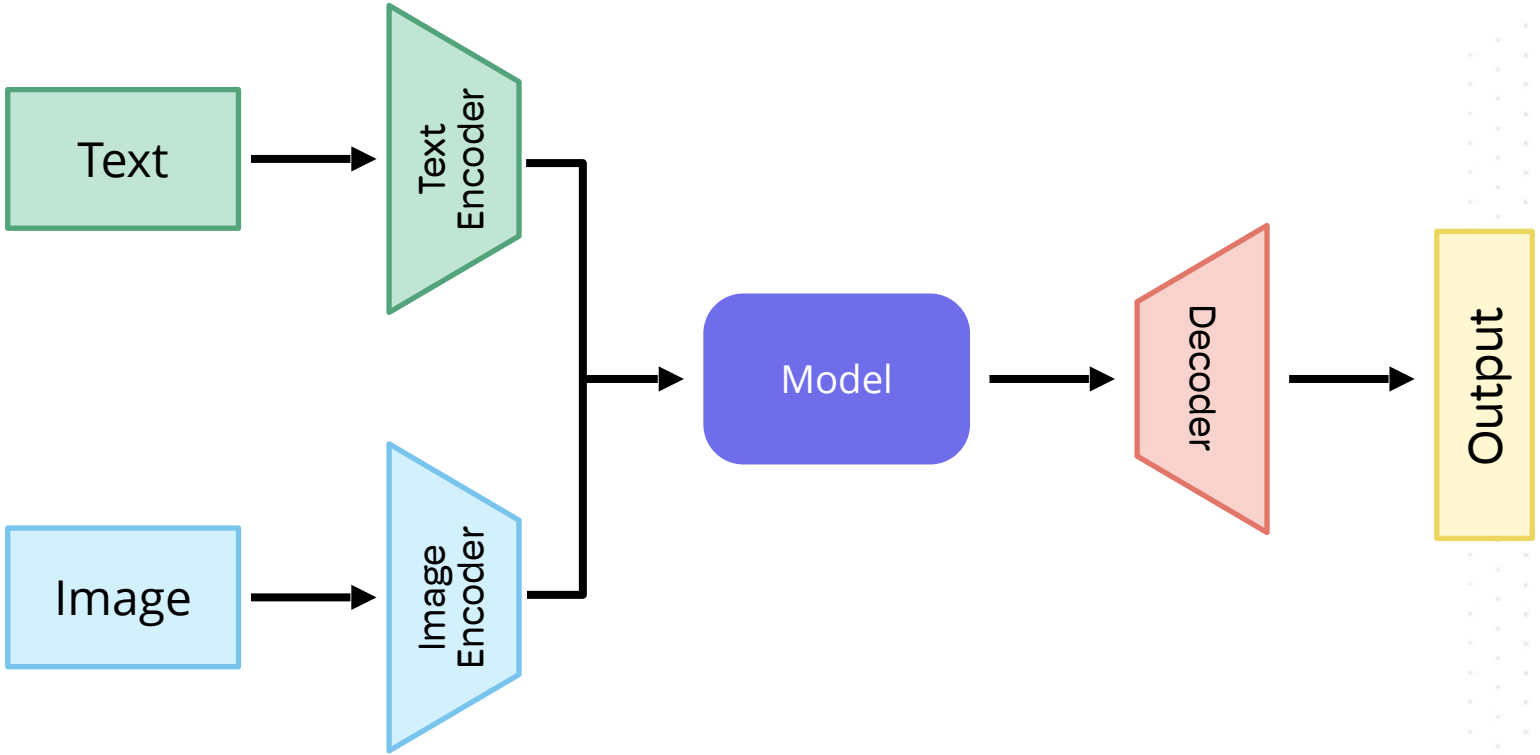


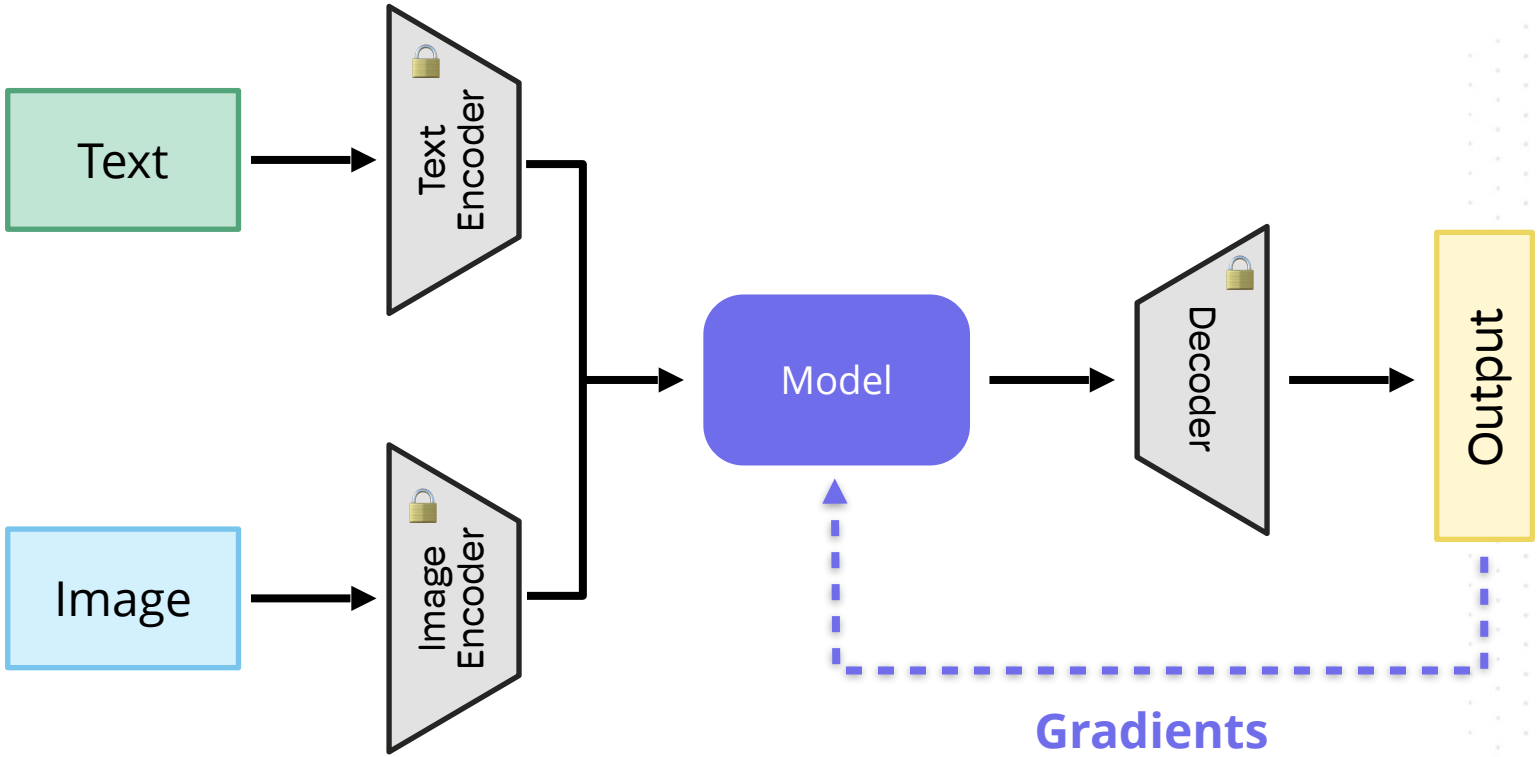
Image-to-Text (Image Captioning)



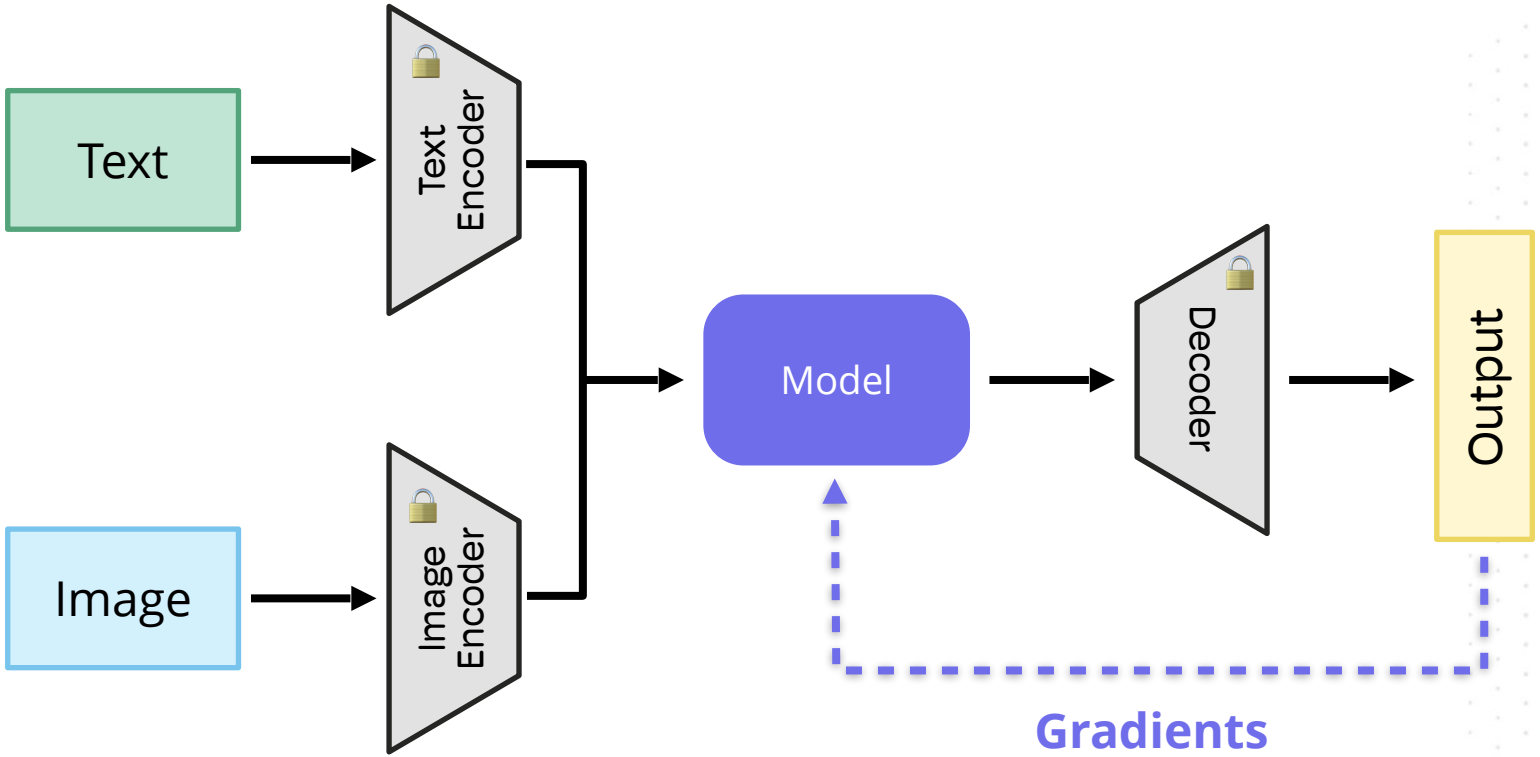
Multimodal Models



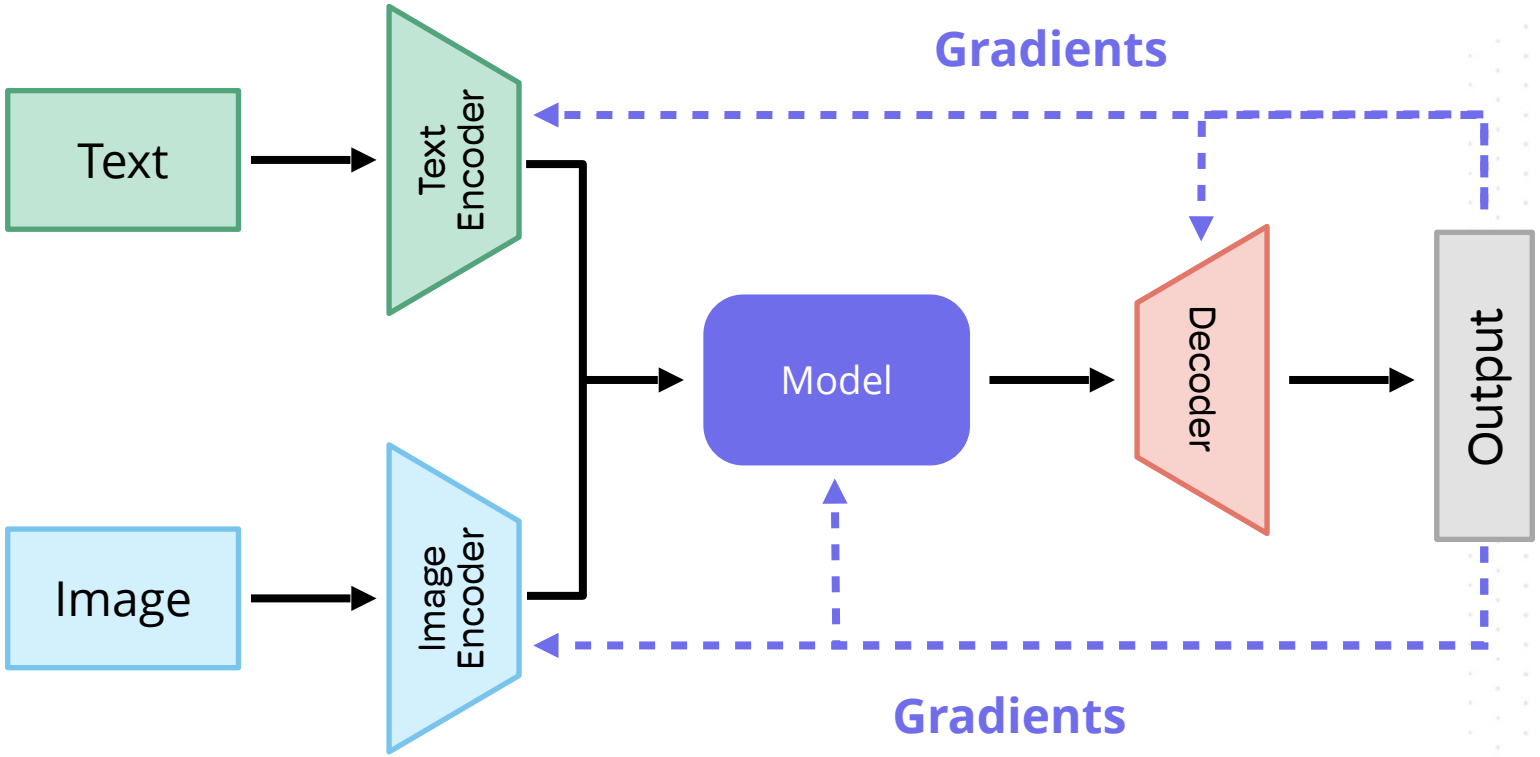
Training Multimodal Models



Training Multimodal Models



Training Multimodal Models



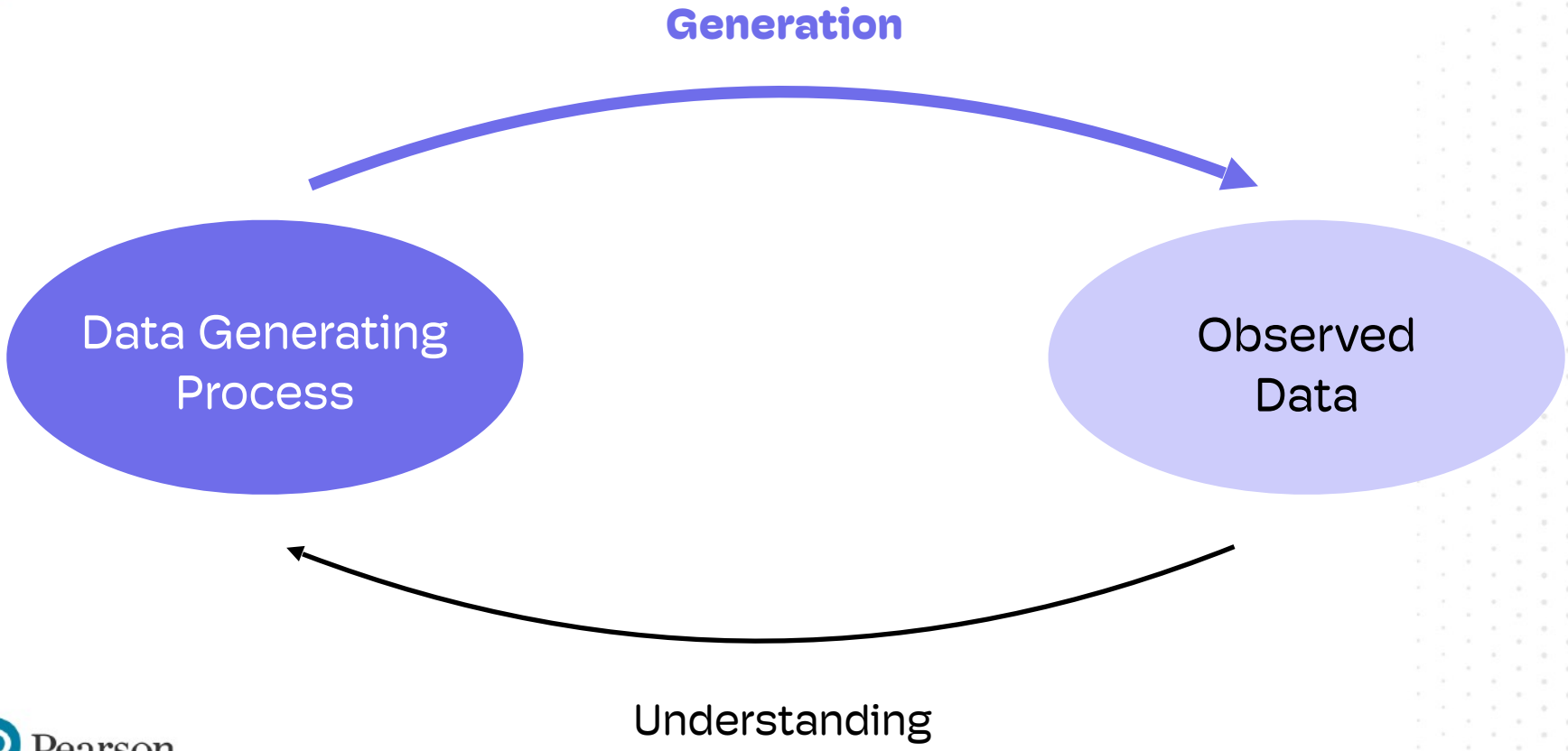
6.2

Vision-Language Understanding

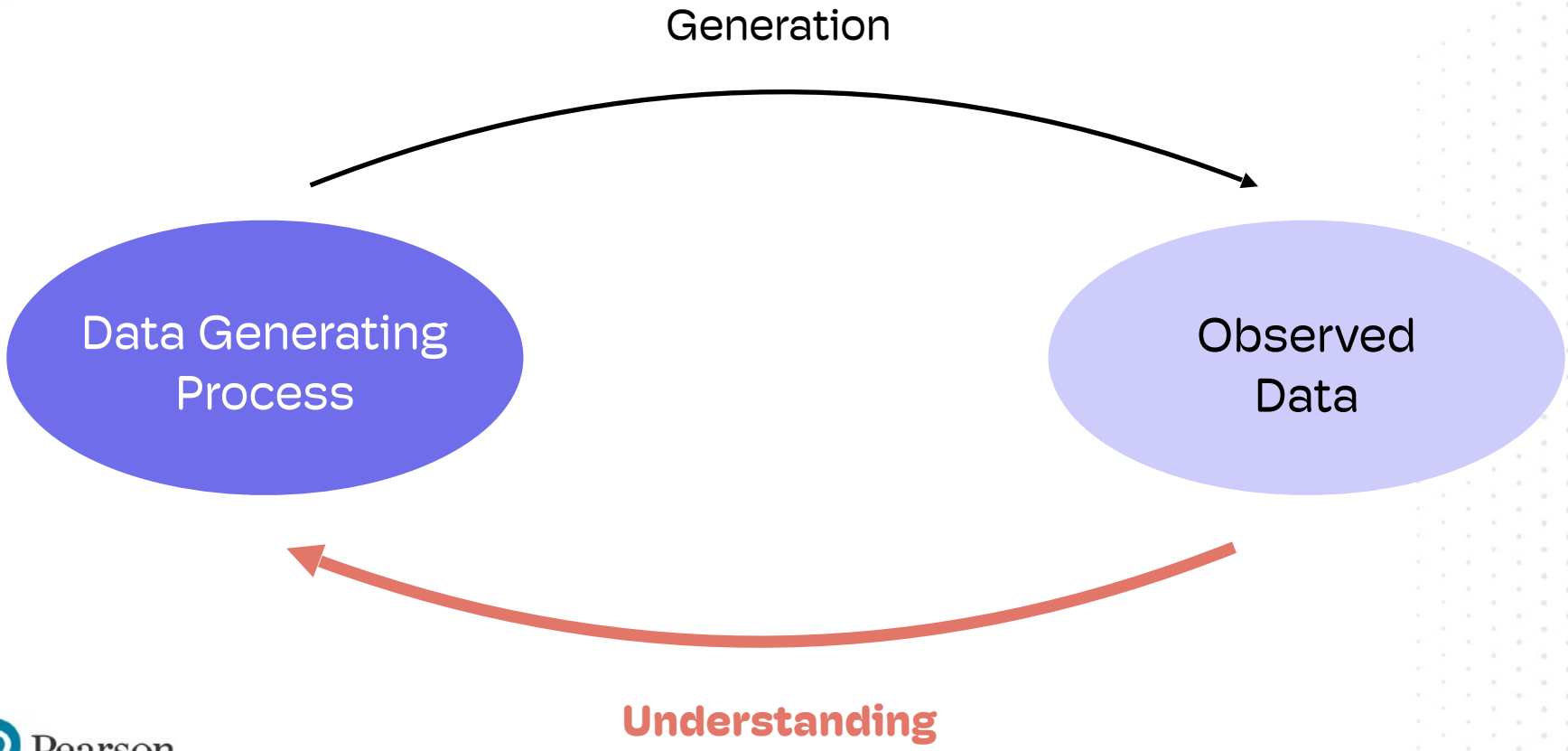
A Taxonomy of Multimodal Architectures

- **Vision-Language Model:** Understanding
- **Latent Diffusion Model:** Generation

Generative Processes



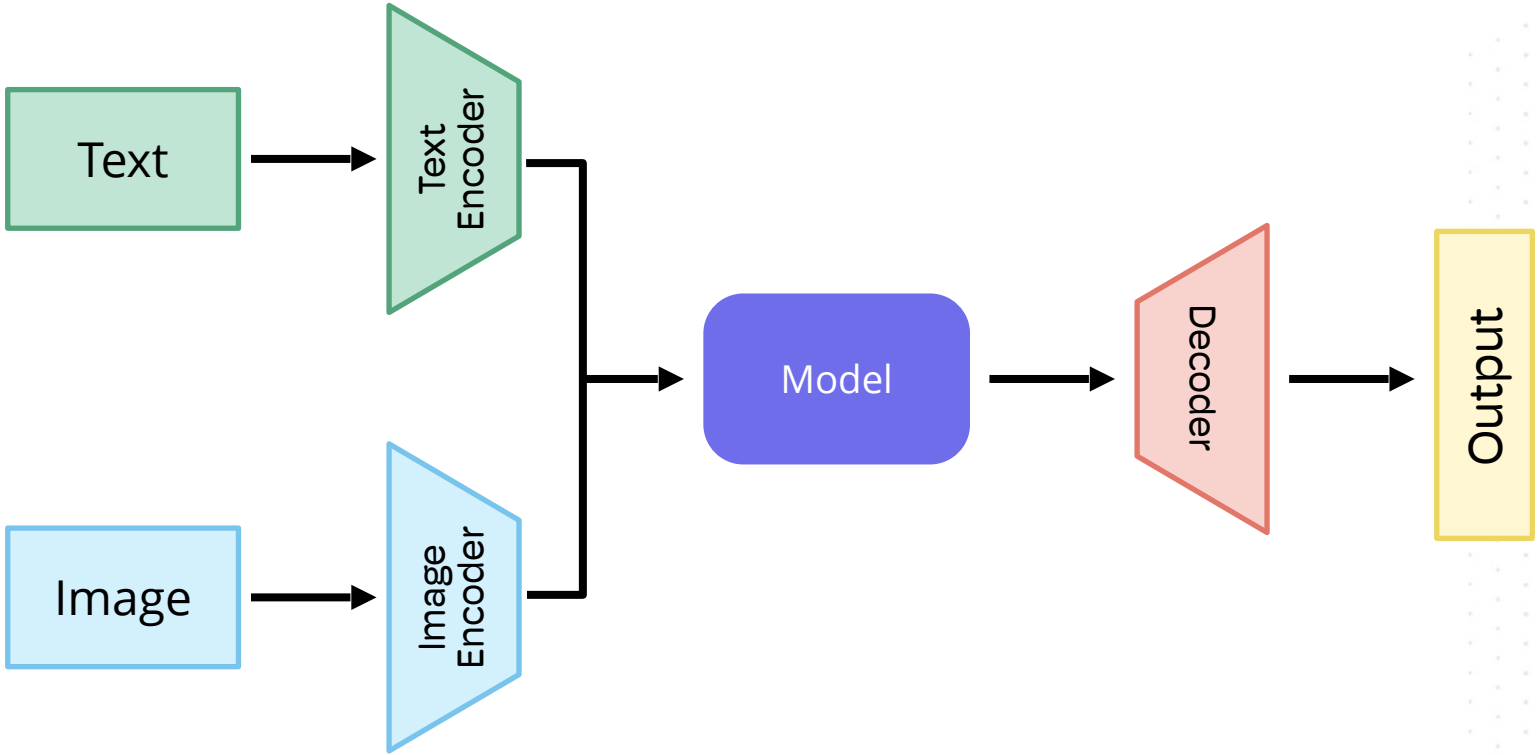
Generative Processes



Vision-Language Models

> Is it an image model that can take text conditioning (i.e., stable diffusion) or a text model (LLM) than can take image inputs (LLaVa/ChatGPT4)?

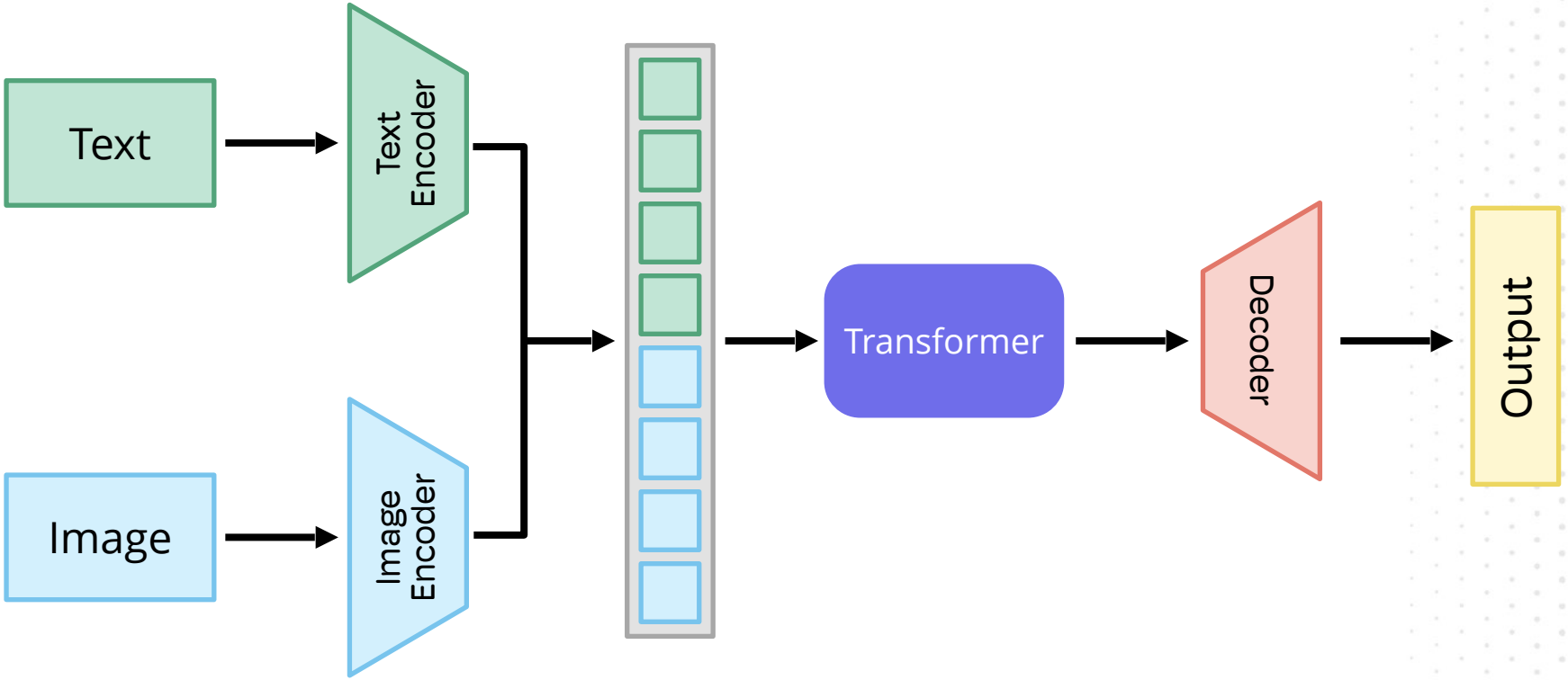
Multimodal Models



Text-to-Image

Image-to-Text (Image Captioning)

Foundation Models



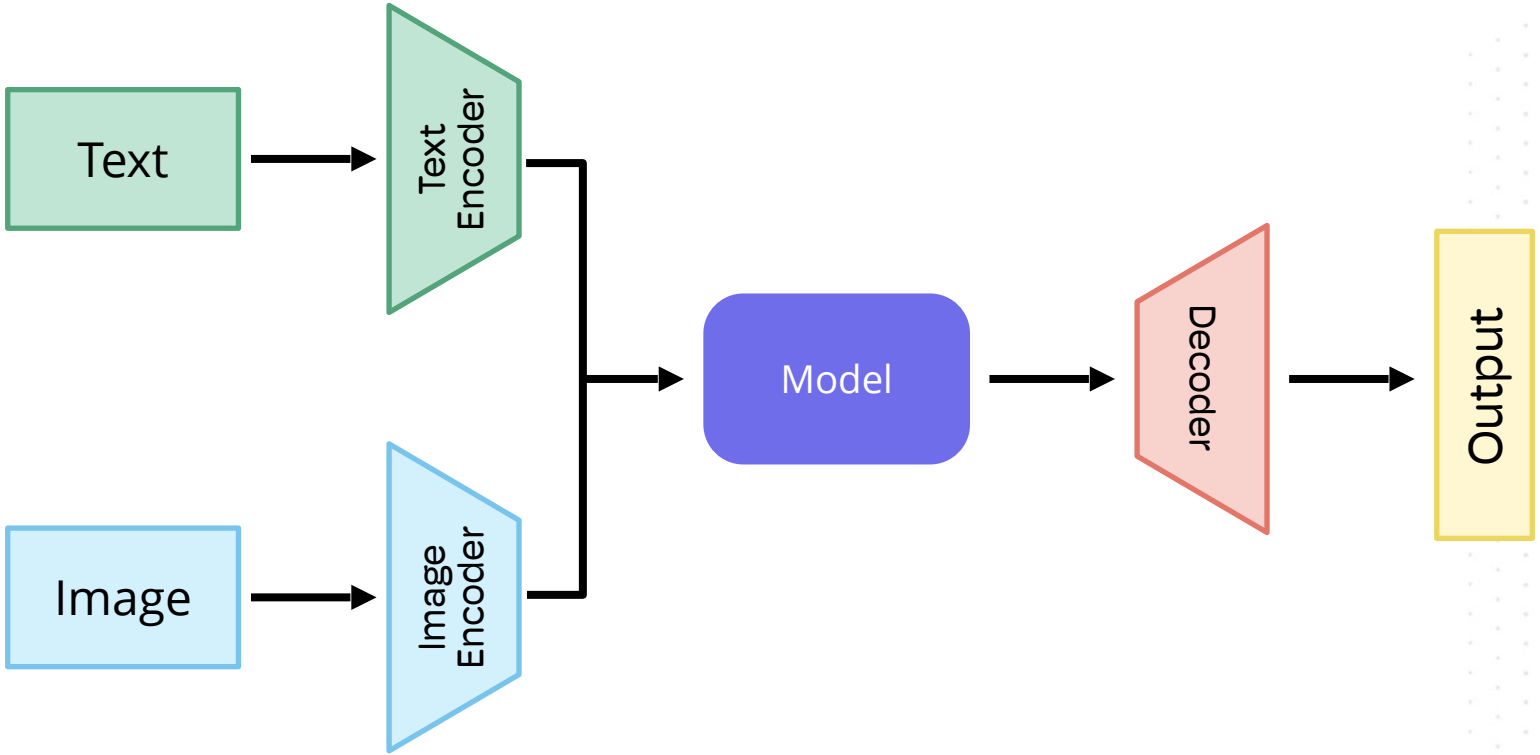
6.3

Contrastive Language–Image Pre-training

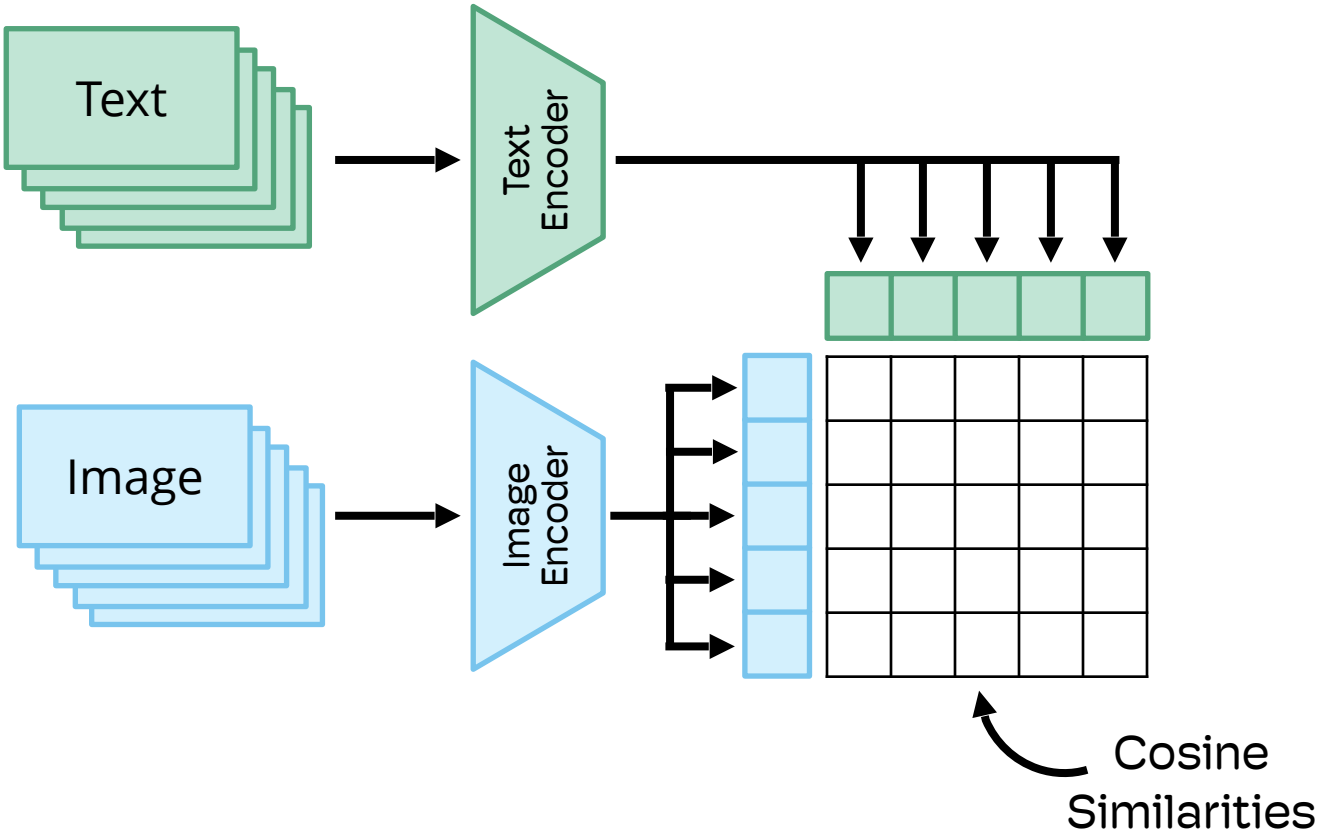
CLIP Innovations

- **Multiple encoder modalities**
- **Contrastive loss for a shared latent space**
- **Web scale (image, text) pair dataset**

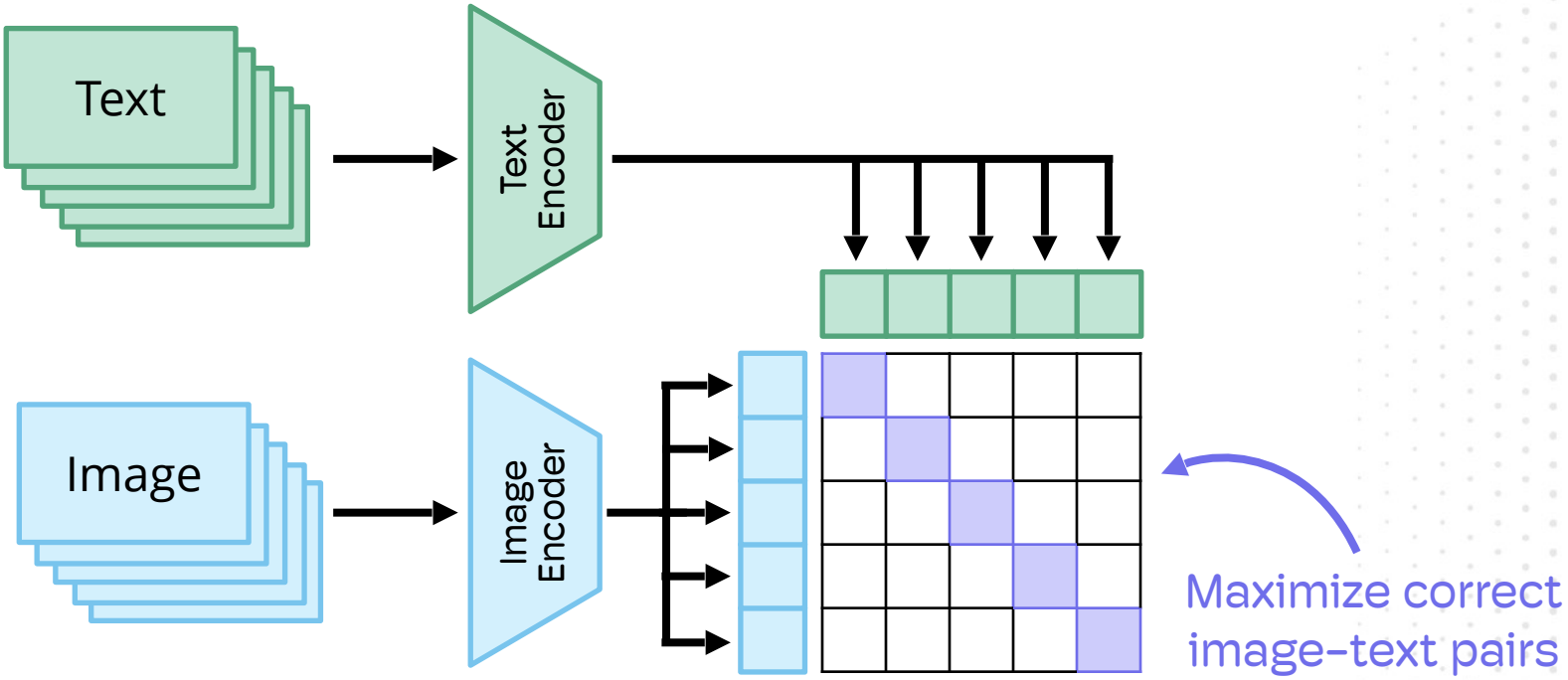
Multimodal Models



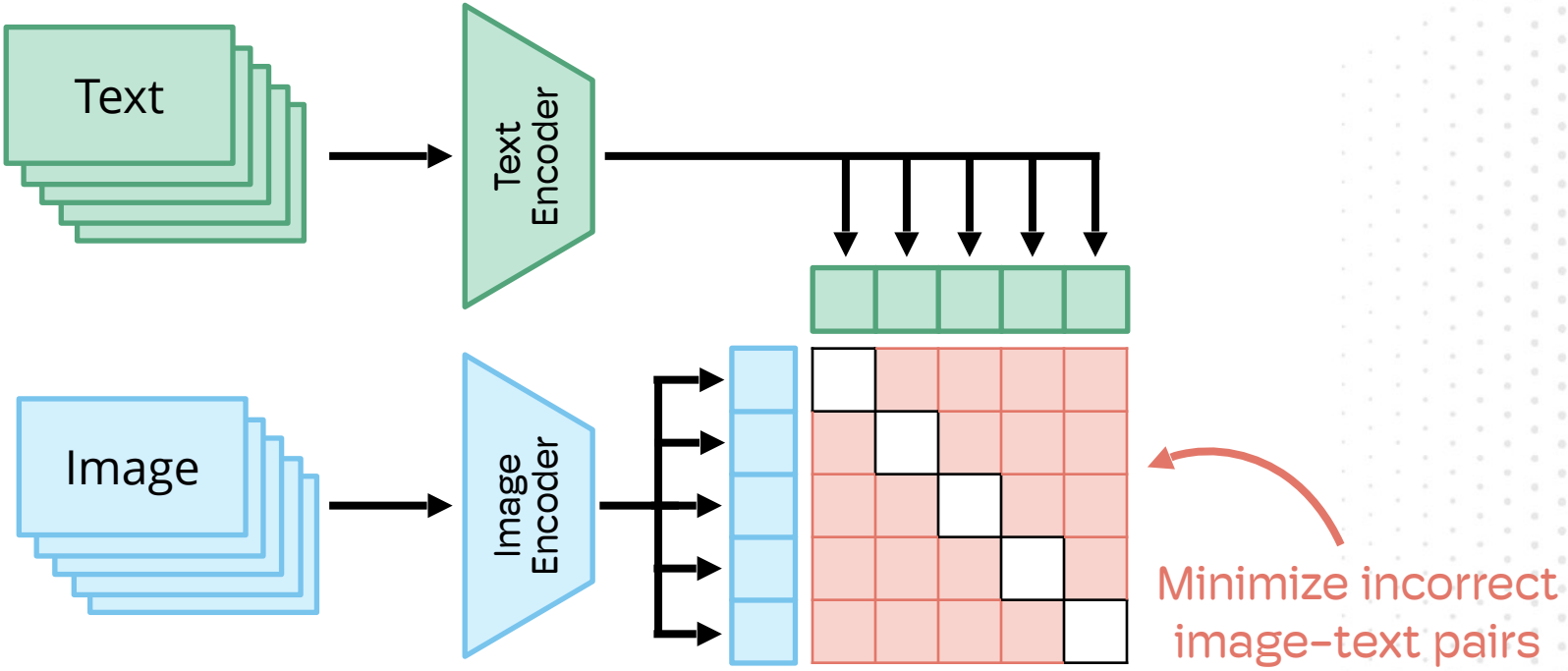
CLIP


Text
Text Encoder
Image
Image Encoder
Cosine Similarities

Contrastive Loss



Contrastive Loss



6.4

Embedding Text and Images with CLIP

Live Coding

6.5

Zero-Shot Image Classification with CLIP

Live Coding

6.6

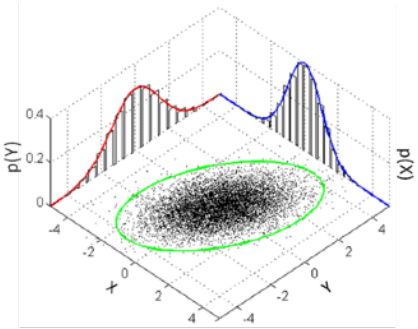
Semantic Image Search with CLIP

Live Coding

6.7

Conditional Generative Models

Generative Model

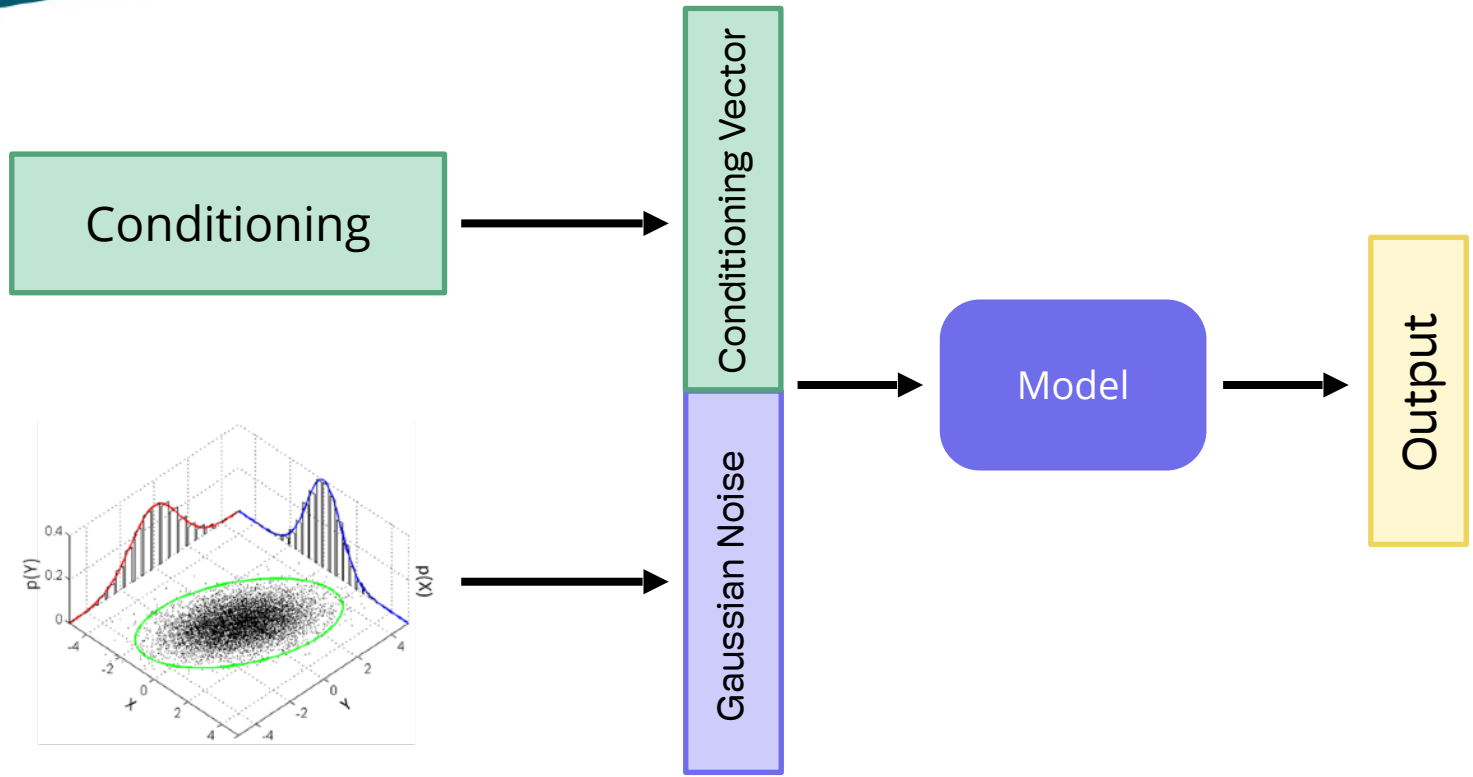


Gaussian Noise

Model

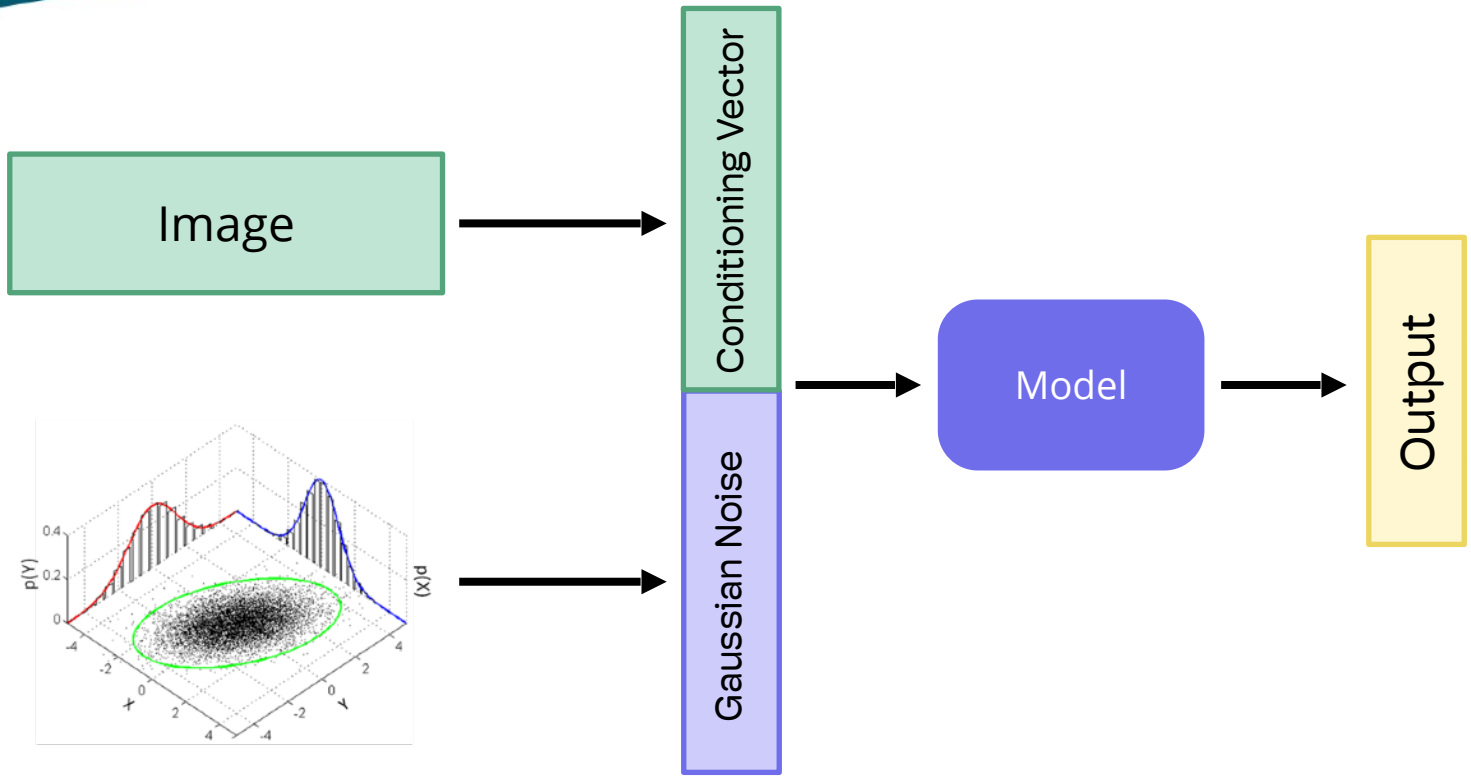
Output

Conditional Generative Model

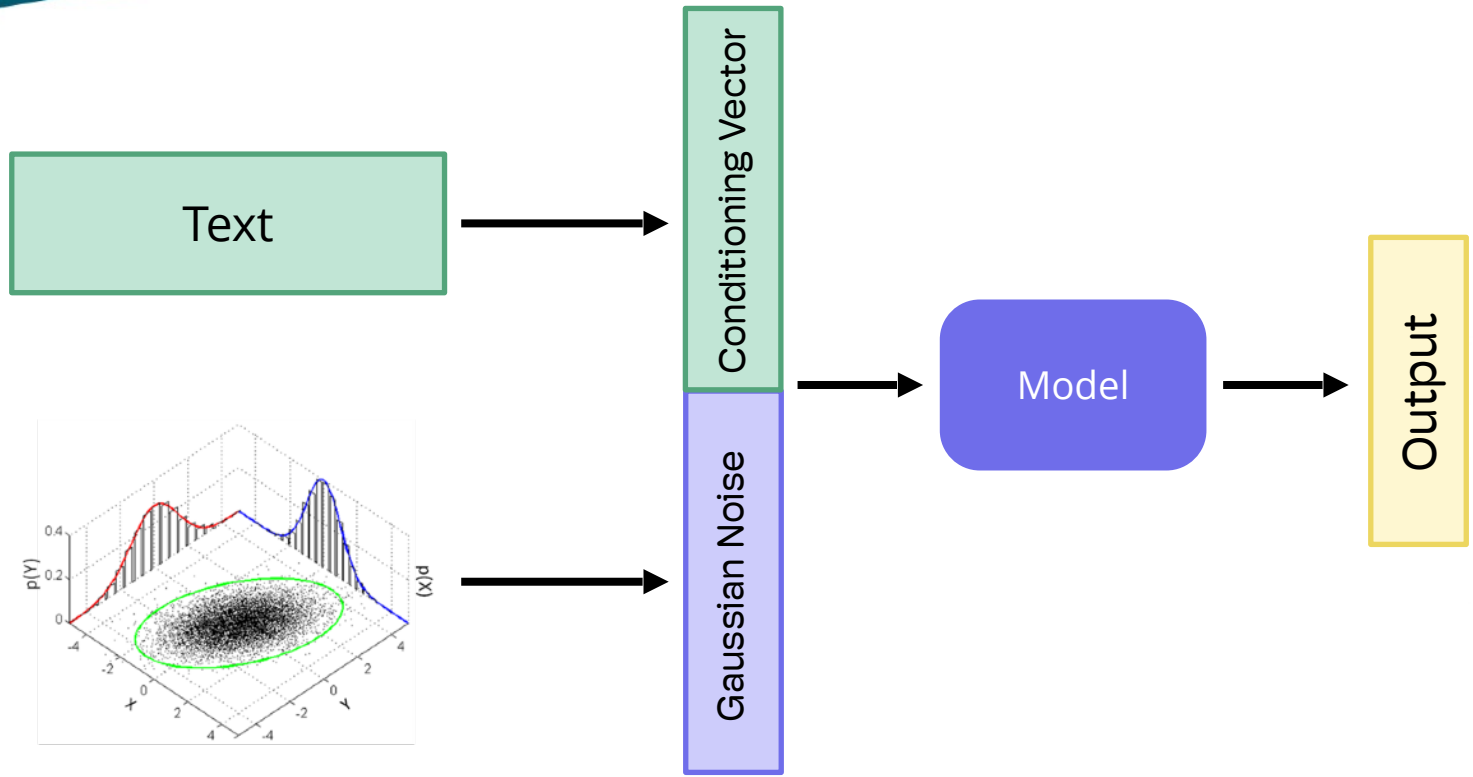


Pearson

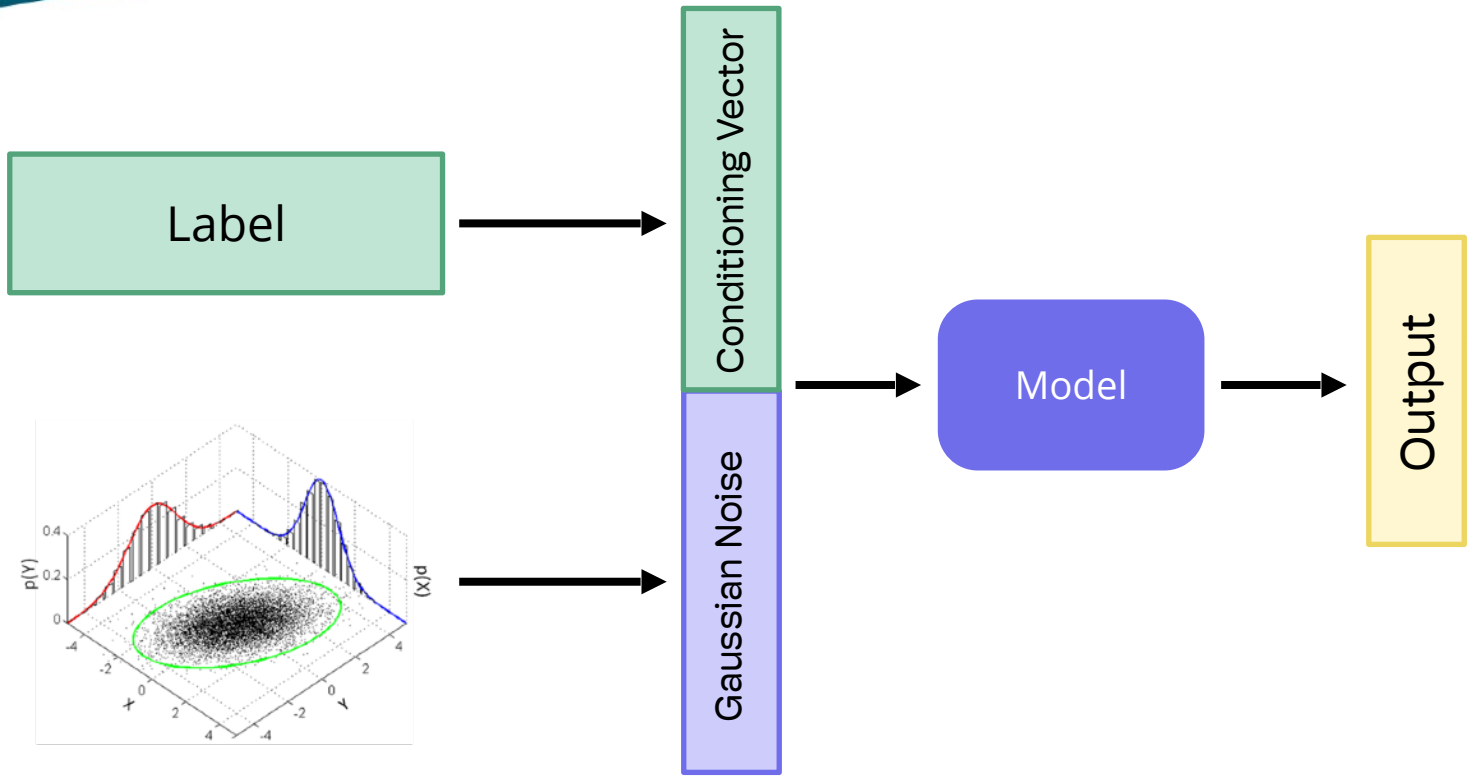
Conditional Generative Model



Conditional Generative Model

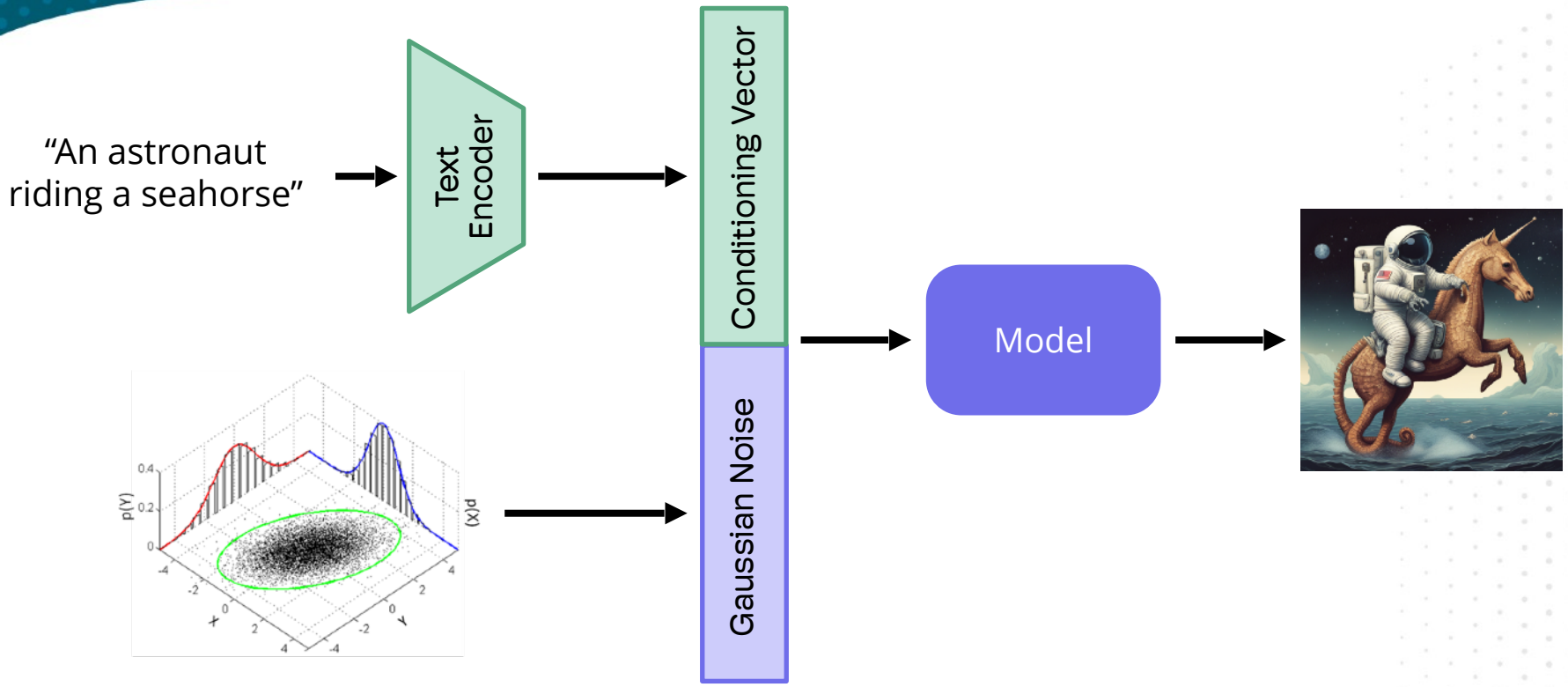


Conditional Generative Model



Pearson

Text-to-Image



6.8

Introduction to Latent Diffusion Models

Making It Multimodal

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.

Paper: https://arxiv.org/abs/2112.10752

Making It Multimodal

Stable Diffusion = Three Generative Models

Making It Multimodal

Stable Diffusion = Diffusion + Transformer + VAE

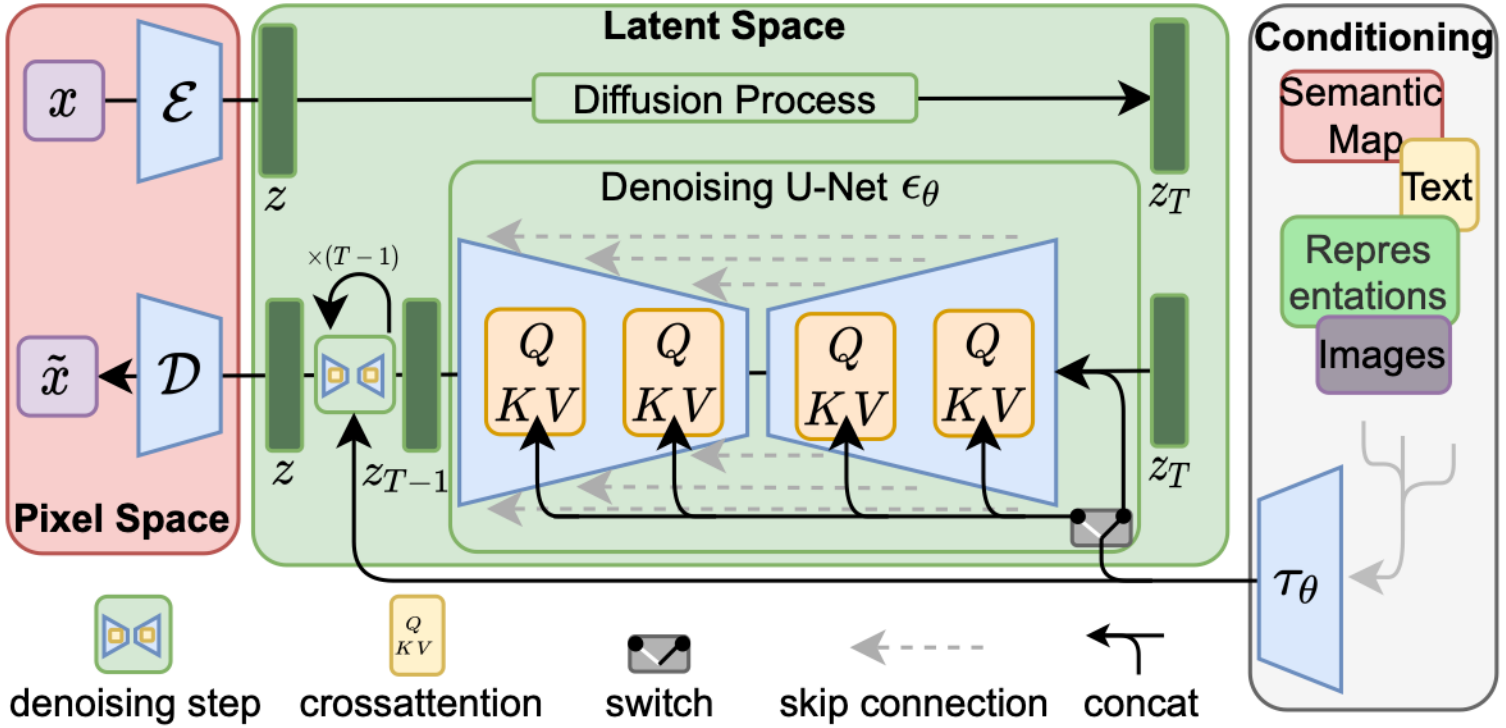
Latent Diffusion Model (LDM) Innovations

- **Semantic Compression:** Scales to higher dimensions
- **Latent Space Diffusion Process:** Much more efficient computationally (training and inference)
- **End-to-end training:** More practical due to ease of use
- **Cross-attention conditioning:** Task flexibility

6.9

The Latent Diffusion Model Architecture

LDM Architecture



Source: High-resolution image synthesis with latent diffusion models (2022)

Live Lecture

6.10

Failure Modes and Additional Tools

Rough Edges

- Realistic/legible text in images
- Faces and identities
- Proper anatomy (teeth, fingers, toes, poses, other)
- Geometric/spatial reasoning
- Logical and deductive reasoning

Some GUI Tools

- https://github.com/AUTOMATIC1111/stable-diffusion-webui
- https://github.com/comfyanonymous/ComfyUI
- Companies/Products
  - https://dreamstudio.ai/
  - https://runwayml.com/
  - https://www.midjourney.com
  - https://openai.com/dall-e-2

6.11

Stable Diffusion Deconstructed

Live Coding

6.12

Writing Our Own Stable Diffusion Pipeline

Live Coding

6.13

Decoding Images from the Stable Diffusion Latent Space

Live Coding

6.14

Improving Generation with Guidance

Live Coding

6.15

Playing with Prompts

Live Coding