

Multimodal Models

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 16 of "The NeuroAI Handbook"

How machines fuse vision, language, and sound, and generate new images from noise, echoing the brain's own multisensory fusion.

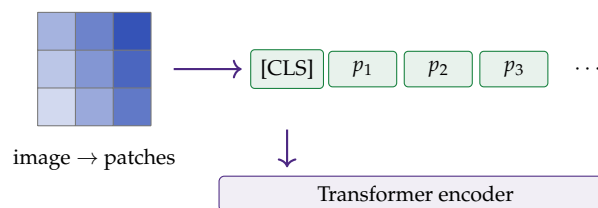
One-sentence thesis. Multimodal AI reaches human-like understanding by doing two brain-like things, fusing separate senses into one shared representation, and generating rich data by learning to reverse the destruction of it, with cross-attention and diffusion as the two mechanisms that make each possible.

PART 1: Fusing the Senses

One Concept, Many Streams

You see a dog, hear it bark, and feel its fur, and none of that arrives as a single signal, yet the brain fuses the streams into one concept, "dog." Multimodal AI chases the same goal: models that reason across text, images, and audio rather than being trapped in one. The first requirement is a way to turn each modality into the same kind of object, a sequence of vectors a Transformer can chew on.

For images, the breakthrough was the **Vision Trans-**



The Vision Transformer treats an image as a sequence of patch tokens, plus a [CLS] token, so the same architecture that reads text can read pictures.

A Shared Space for Words and Pixels

Having encoders is not enough; vision and language must live in the *same* space to be compared. CLIP built that shared space from 400 million image-text pairs, using a *contrastive* objective: pull each true image-caption pair together while pushing all mismatched pairs apart. The model learns that the pixels of a cat and the letters "cat" land near each other, which grants *zero-shot* classification, name a new category in words and CLIP can recognize it.

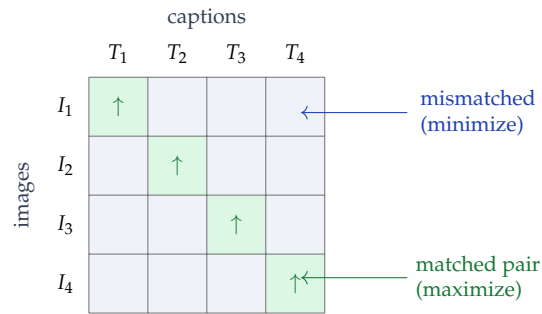
former (ViT). Rather than sliding convolutional filters, ViT chops a 224x224 image into a grid of 16x16 patches, flattens each into a vector, adds a position embedding, and feeds the whole sequence to a standard Transformer, exactly as if the patches were words.

The trade-off. Dropping the CNN's built-in assumptions (locality, translation equivariance) means ViT must *learn* everything from data: it loses to CNNs on small datasets but wins decisively past roughly 100 million images, which is why it anchors large-scale models like CLIP.

ANALOGY

CLIP is Hebbian learning at web scale, "neurons that fire together, wire together." When you see an apple and hear the word "apple" together, the visual and auditory populations co-activate and their connection strengthens, so later one can evoke the other. CLIP's contrastive loss does the same thing to image and text embeddings.

In symbols, training maximizes the similarity of matched pairs against all others in the batch, the bright diagonal of a similarity matrix.



CLIP's contrastive objective. Across a batch, the correct image-caption pairs on the green diagonal are pulled together while every off-diagonal mismatch is pushed apart.

Anatomy of a Multimodal Model

Zoom out and nearly every multimodal system has the same three parts. A **specialized encoder** for each modality (a ViT for images, a Transformer for text) turns raw input into vectors. A small **projection head** maps each encoder's output into the shared

embedding space so the vectors are directly comparable. And a **fusion mechanism** combines them, which can be as light as CLIP's contrastive loss or as rich as the cross-attention layers we turn to next, where one modality actively interrogates another instead of merely sitting beside it.

Letting One Sense Question Another

A shared space lets you *compare* modalities, but real fusion means letting one *interrogate* the other: given the question “what animal is this?”, which image patches matter? **Cross-attention** is the mechanism. One modality supplies the *query* (the text question); the other supplies the *keys* and *values* (the image patches). The query is matched against every key, the softmax turns those matches into weights, and the output is a weighted blend of the values, high weight on the “dog” patch, low weight on grass and sky.

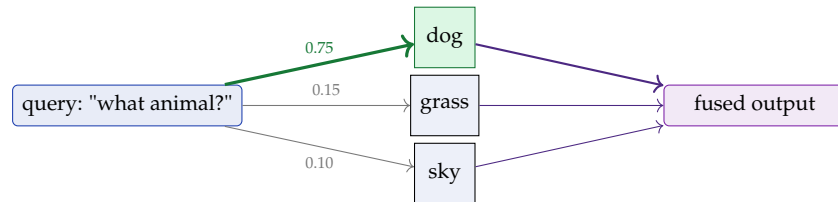
What makes cross-attention powerful is that it is *asymmetric*: text can ask a specific question of an image instead of both being averaged together blindly.

In symbols, with query Q from one modality and keys/values K, V from another:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

The Brain Does This Too

Cross-attention is a close cousin of how association cortex binds the senses. Sensory streams converge in regions like the superior temporal sulcus, where single neurons respond to both a face and its matching voice, and the integration is *selective*, weighted by temporal synchrony and spatial coincidence rather than summed blindly. The most vivid proof that vision can *override* audition is the **McGurk effect**: lips forming “ga” dubbed over the sound “ba” are heard as a third syllable, “da.” Perception is not passive reception but active inference, the brain’s best guess from all evidence, exactly what a fusion model computes.



Cross-attention. A text query weights the image patches by relevance (softmax attention) and returns their weighted blend, letting language selectively pull the visual information it needs.

Learning to Undo Destruction

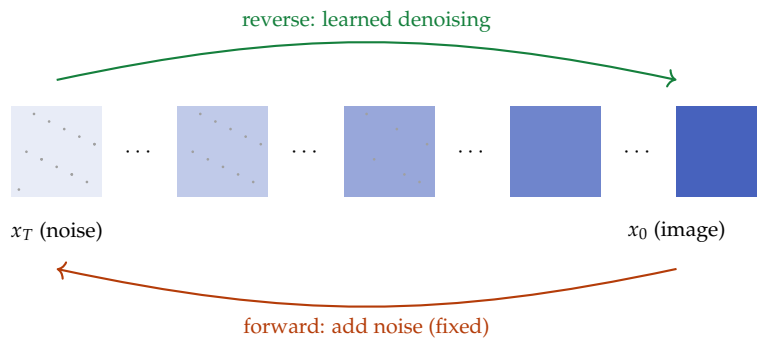
Understanding data is one thing; *creating* it is another, and diffusion models are the engine behind DALL-E, Stable Diffusion, and Midjourney. The trick is to split generation into a hard problem and an easy one. The easy, *forward* process just adds a little Gaussian noise to an image over many timesteps until nothing remains but static, that direction is trivial and needs no learning. The hard, *reverse* process is what the model learns: given a noisy image at step t , predict the noise that was added so you can subtract it and step back toward a clean image.

ANALOGY

Diffusion is a sandcastle and the tide. The forward process is erosion: wave after wave washes the castle down to a formless mound, easy and predictable. The model learns to run the film in reverse, not by memorizing one recording but by learning the *rule* for how to un-erode one step at a time, until it can sculpt a brand-new castle from any random mound of sand.

The Time-Conditioned U-Net

The network doing the denoising is a **U-Net**, an encoder-decoder with skip connections that preserve fine detail. Critically it is *time-conditioned*: it takes both the noisy image and the timestep t , so it can learn coarse structure at high noise and fine detail at low noise. Run it iteratively and structured images emerge from pure noise.



Diffusion. The forward process destroys an image with noise (fixed, easy); the model learns only the reverse process, denoising step by step to generate new images from static.

Understanding Meets Generation

Stable Diffusion fuses the two halves of this chapter: a CLIP-style text encoder for *understanding* and a diffusion U-Net for *generation*. Two ideas make it practical. First, **latent diffusion**: instead of denoising $512 \times 512 \times 3 = 786,432$ pixels, a VAE compresses the image to a $64 \times 64 \times 4$ latent (about **49x** fewer values) and the whole diffusion runs there, cutting cost dramatically for the same quality, much as the brain imagines an apple from a compressed cortical code, not a photoreceptor-by-photoreceptor reconstruction.

Second, **text conditioning via cross-attention**: at every denoising step the prompt’s embedding is fed into the U-Net through cross-attention, so “a red apple” steers the noise toward an apple shape filled with red.

Turning the Prompt Up

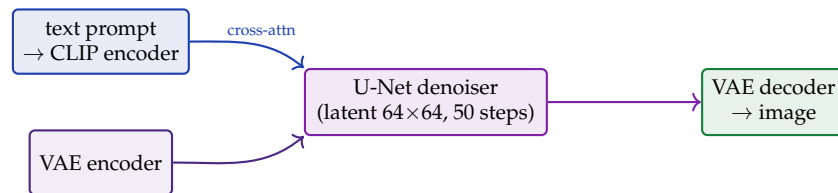
To make the model obey the prompt more strongly, **classifier-free guidance** computes two noise predic-

tions each step, one conditioned on the text and one unconditioned, then pushes further in the conditioned direction by a *guidance scale* w .

In symbols:

$$\tilde{\epsilon} = \epsilon_{\text{uncond}} + w(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}})$$

Watch the knob. Low w gives creative but loose outputs; $w \approx 7\text{--}8$ is the sweet spot; push w too high and images turn oversaturated and unnatural. The same trick generalizes past images: speech models (Whisper for recognition, VALL-E for synthesis) tokenize audio and treat it as one more sequence, and multimodal medical models fuse MRI, clinical notes, and labs to beat any single modality.



A latent diffusion model. Denoising runs in the VAE’s compressed latent space (49x cheaper), guided at every step by the text prompt through cross-attention, then decoded to a full image.

Quick Self-Test

1. How does a Vision Transformer turn an image into something a Transformer can process, and why does it need far more training data than a CNN?
2. Describe CLIP’s contrastive objective. In what sense is it a large-scale implementation of Hebbian learning?
3. In cross-attention between text and an image, which modality provides the query and which provides the keys and values, and why does that asymmetry matter?
4. Diffusion splits generation into a forward and a reverse process. Which one is learned, which is fixed, and what does the network actually predict at each step?
5. What two innovations let Stable Diffusion generate high-quality images cheaply, and what does the guidance scale w control?

If You Remember One Thing

Multimodal AI stands on two mechanisms, each with a brain analog. Cross-attention lets one modality selectively query another, the same selective binding association cortex performs and the McGurk effect exposes. Diffusion generates by learning to reverse the easy destruction of data, one denoising step at a time. Stable Diffusion is simply both at once: a text encoder that understands, a latent diffusion model that creates, and cross-attention stitching them together.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552