140.800: How to AI (for Public Health)

Week 4: Multimodal Models

Yiqun T. Chen
Email: yiqunc@jhu.edu
Schedule office hours via email

Departments of Biostatistics and Computer Science
Data Science & AI Initiative and Malone Center for Engineering in Health

What Makes Data Multimodal?

Definition:

- Information from multiple modalities (text, images, audio, etc.)
- Each modality provides different types of information
- Combined analysis often more powerful than single modality

Think About How You Process Information:

MRI/ blood Biomarker

- Reading a paper: You see text *and* interpret figures
- Doctor's diagnosis: Patient symptoms (text) *and* X-ray images
- Drug discovery: Molecular structure (graph) *and* effect descriptions

The Key Insight:

Each modality tells part of the story. Together, they tell the complete story.

Biomedical Multimodal Examples

Variant Genes ↗ variant

parent A T C G

A T G G

Rich Multimodal Landscape:

- Text + Images: Research papers with figures and charts
- Molecular + Text: Chemical structures with property descriptions
- Genomics + Phenotype: DNA sequences with trait descriptions
- Medical Images + Reports: Scans with radiological findings
- Audio + Text: Voice biomarkers with clinical notes

Polygenic Risk Score

Why Multimodal Matters in Health:

- Redundancy: Cross-validate findings across modalities
- Completeness: Capture phenomena invisible to single modality
- Robustness: Handle missing or corrupted data
- Human-like: Matches how clinicians make decisions

The Multimodal Challenge: Why It's Hard

The Representation Gap:

- Images: High-dimensional pixel arrays, spatial patterns
- Text: Sequential tokens, semantic relationships
- Molecules: Graph structures, chemical bonds
- Audio: Time-frequency representations, temporal patterns

Fundamental Questions:

- How do we represent such different data types? (Image)
- How do we find connections between modalities? (Image, Text)
- When should we combine vs. analyze separately?
- How do we handle missing modalities?

How do we represent images?

red channel, Green, Blue

$\rightarrow \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 \\ 0 & 2 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$

How Do We Represent Images? From Pixels to CNNs

"Embeddings"

1024x1024

Start simple, then add structure:

- Raw pixels (vectorized): flatten an $H \times W \times C$ image into $\mathbb{R}^{HWC}$.
  - Pros: simplest numeric representation.
  - Cons: *destroys* spatial locality; no translation/scale invariance. → VERY High Dimension
- Patches / local descriptors: split into fixed-size patches (e.g., 16×16), summarize each patch.
  - Keeps some locality; can pool/aggregate over patches.
- Convolutions (CNNs): learn *local filters* with weight sharing to produce feature maps.
  - Hierarchy: edges $\rightarrow$ textures $\rightarrow$ parts $\rightarrow$ objects
  - Pooling provides translation tolerance; preserves spatial structure.

Key idea: respect image geometry (locality, stationarity) instead of treating pixels as an unordered vector.

label: Fish

Convolutional Neural Networks (CNNs)

Key idea: learn spatially-local filters shared across the image.

- A convolution applies a kernel $K \in \mathbb{R}^{k \times k}$ across the input:

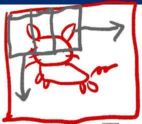
$$(F * K)(i,j) = \sum_{u=1}^{k} \sum_{v=1}^{k} K(u,v)\, F(i+u, j+v)$$

- Weight sharing: the same K is used everywhere $\rightarrow$ translation equivariance.
- Stacking layers builds a **hierarchy of features**:
  - Shallow: edges, corners, textures
  - Deeper: object parts, semantics
- Pooling layers reduce resolution, increase invariance.

Popular CNN backbones: AlexNet, VGG, ResNet (skip connections), EfficientNet (scaling).

can be used off the shelf from some hosting website!

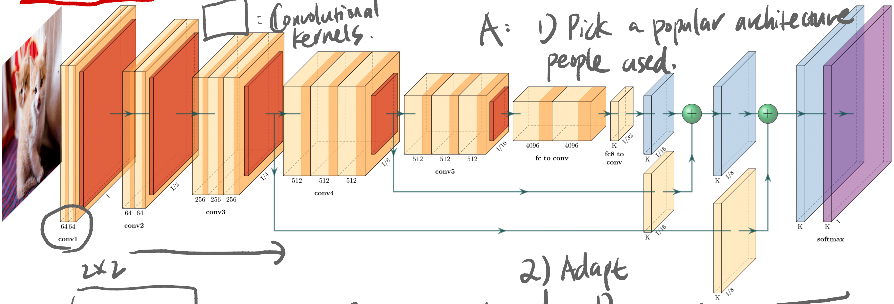
CNN Visualization (A great image feature extractor!)



Q: How to choose convolutional features & layers?

: Convolutional kernels.

A: 1) Pick a popular architecture people used.



2) Adapt based on 1) and choose best adaptation using validation loss

2x2

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Filters

$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Vision Transformers (ViT)

Key idea: treat an image as a sequence of patches $\rightarrow$ use a Transformer encoder.

- Split image into N patches $\{x_1, \ldots, x_N\}$, flatten each to a vector.
- Project each patch with a linear map E:

$$z_i = E \cdot x_i + p_i, \quad i = 1, \ldots, N$$

  where p_i is a positional embedding.
- Add a special `[CLS]` token z_0; the Transformer encoder produces contextualized representations:

$$Z' = \mathrm{TransformerEncoder}([z_0, z_1, \ldots, z_N])$$

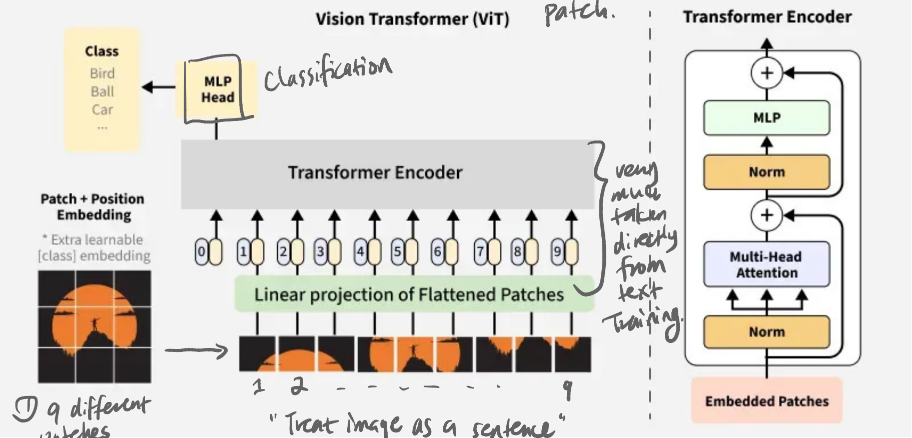
- Use the output of z_0 as a global **image embedding**.

Properties: scales well with data, captures long-range dependencies, now competitive or better than CNNs on many tasks.

ViT Visualization

"Pretrain"/self-supervised Training:

① with hold one of the pathces.
② use the Rest to reconstruct that patch.



Classification

very much taken directly from text Training.

1 2 - - - - - - 9

"Treat image as a sentence"

① 9 different patches (9 is a # that you can choose)

From Representation to Alignment

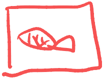


Image 1

Image2

A fish is swimming. Text 1

A cat is walking. Text 2

(Image 1, Text 1) ?
(Image 2, Text2) ?
(Image 1, text 2)
(Image 2, Text 1)

So far:

- We learned how to build strong image embeddings (CNNs, ViTs).
- We also know how to build text embeddings (LLMs, word embeddings).
  TF-IDF, Bag of words

Next question: How do we bring two modalities into the same space?

- Need a way to **compare** image and text representations.
- Requires a shared latent space for **cross-modal understanding**.

(Abstract!)

This motivates ***contrastive learning*** *for multimodal alignment.*

We need a way to measure "similarity" between text & image Representations

$\boxed{\text{Correlation !!}}$ ← Similarity (Image, Text)

Mathematical Foundation: Similarity Metrics

Core Question: How do we measure similarity between different modalities? Modalities ⇒ Vector / Embedding ⇒ similarity metric for alignment!

Cosine Similarity (Most Common):

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}$$

Toy Example:

- Image embedding: $\mathbf{u} = [0.8, 0.6, 0.0]$
- Text embedding: $\mathbf{v} = [0.6, 0.8, 0.0]$
- Cosine similarity: $\frac{0.8 \times 0.6 + 0.6 \times 0.8 + 0.0 \times 0.0}{\sqrt{0.8^2+0.6^2+0.0^2}\sqrt{0.6^2+0.8^2+0.0^2}} = \frac{0.96}{1.0 \times 1.0} = 0.96$

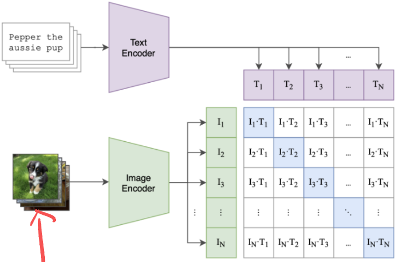
Why Cosine?

① $[-1, 1]$ where "similarity" ↑ as the metric ↑

② It's input magnitude agnostic ← $\text{sim}(2u, 2v) = \text{sim}(u, v)$

Patient: pathology slide, EHR Data, genomics data, ...

Key intuition: If you show a model images and their captions together, it can learn to connect visual and textual concepts.



"A brown dog running in the grass"

cosine similarity ↑

- Positive pair: (dog image, matching caption) should be close in embedding space.
- Negative pairs: (dog image, caption about a car) should be far apart.

cosine similarity ↓

Big picture: The model **aligns** images and text into a shared semantic space by performing "classification".

$$(\text{Image } i, \text{Caption } j) = \begin{cases} 1 & \text{correct pairing} \\ 0 & \text{incorrect pairing} \end{cases}$$

Contrastive Learning: The Core Idea

$\langle a, b \rangle \iff \|a-b\|_2^2$
if $\|a\|_2^2 = \|b\|_2^2 = 1$

Dog, Fish $\Rightarrow$ optimized $\langle , \rangle$ closer - 1
Dog, Fish, Plants, Views $\Rightarrow$ $\langle , \rangle \uparrow$

A classification loss in disguise!

Goal: Bring matched pairs close, push mismatched pairs apart.

- Given a batch of image–text pairs $\{(I_i, T_i)\}_{i=1}^{N}$:

(I_1, T_1) correct
(I_1, T_2) ✗

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log \frac{\exp(\langle f(I_i), g(T_i)\rangle/\tau)}{\sum_j \exp(\langle f(I_i), g(T_j)\rangle/\tau)}\right]$$

cosine similarity. $= \langle \underbrace{f(I_i)}_{\text{image embedding of dog 1}}, \underbrace{g(T_i)}_{\text{Text embedding of dog 1.}} \rangle$

- Where:
  - $f(I)$ = image encoder (CNN/ViT)
  - $g(T)$ = text encoder (Transformer/LLM)
  - τ = temperature parameter controlling sharpness of similarity: lower τ $\Rightarrow$ more peaky distributions (hard negatives matter more).
- Encourages **alignment** between true pairs and **separation** otherwise.

$\langle f(I_i), g(T_i)\rangle$ to be high
$\langle f(I_i), g(T_j)\rangle$ for $j \neq i$ to be low.

$\tau \iff$ how "black & white" the pos & neg are.

Alignment works whenever we have encoders f and g:

caption

A fish to the left

- - - Right

- - vertical

Fish in the middle.

Image Encoder   Text

$$f : \mathcal{X} \to \mathbb{R}^d, \quad g : \mathcal{Y} \to \mathbb{R}^d$$

$$s(x, y) = \frac{\langle f(x), g(y) \rangle}{\|f(x)\| \|g(y)\|}$$

① Fix f and g

↓

Alignment

Advantage:

① super efficient to learn

② Focus on alignment, also data-efficient

Disadvantage:

① limited expression

Examples:

- Image ↔ Text (CLIP) (d = 512)
- Audio ↔ Text (speech + transcripts)
- Molecule ↔ Text (chemistry + descriptions)

Use cases: zero-shot classification, retrieval, embedding search.

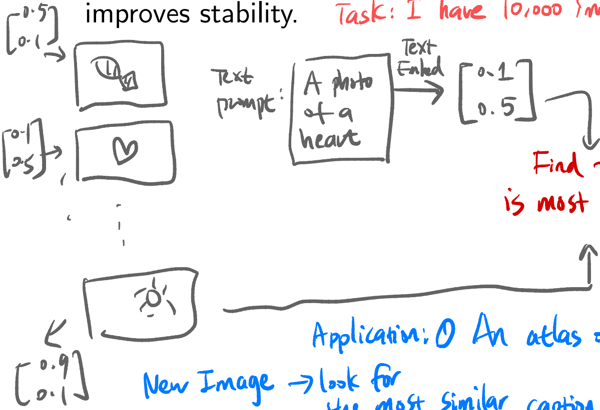
When alignment isn't enough: We need fusion (cross-attention, hybrid models) for fine-grained grounding and reasoning – more coming.

② Train f & g and alignment: Fusion

Zero-shot via prompt templates:

- Build text prompts like: "a photo of a {label}" , "an X-ray showing {condition}" , "a histology slide of {tissue}."
- **Prompt ensembling:** average embeddings over multiple templates; improves stability.

Task: I have 10,000 images, I want to find an image based on my verbal description.



Find the image whose embedding is most similar to my text embedding !

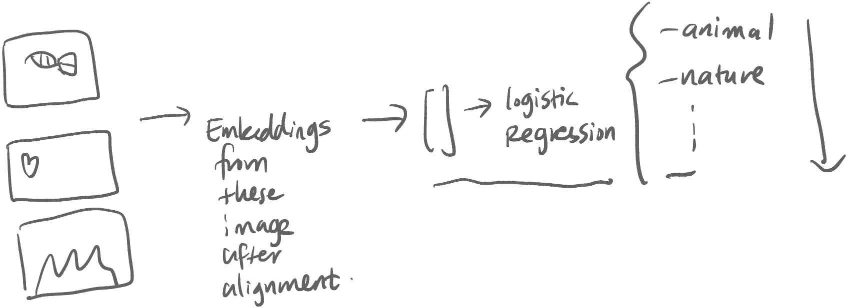
Application: ① An atlas of patients / samples

New Image → look for the most similar caption.

Using CLIP: Linear Probe (Few-Shot)

Linear probe:

- Freeze image encoder; extract embeddings
- Fit a light classifier (logistic regression, tree models) on a small labeled set
- Strong baseline, avoids full fine-tuning cost



Domain Adaptation: From CLIP to BiomedCLIP

Analogous to BERT → PubMEDBERT

Gene BRCA1 → BR/CA/1 in BERT
→ 1 token in PubMED

Challenge: Web captions ≠ biomedical language/images.

- BiomedCLIP: pretrain on biomedical image–text pairs (PMC, etc.).
- **Adapters/LoRA**: parameter-efficient fine-tuning on limited domain data. Fusions

extract all image + captions in published papers.

Practical knobs:

- *Freeze* vision encoder; fine-tune text prompts (prompt tuning) or small adapters.
- *Curriculum* prompts: generic → domain-specific (radiology → finding-level).
- **Regularize**: weight decay, mixup on embeddings, moderate τ. Temperature

Beyond Alignment: Cross-Attention Fusion

Why fuse? Some tasks need **fine-grained grounding** (token $\leftrightarrow$ patch).

- **Cross-attention** lets text tokens query image patches (and/or vice versa).
- Used in BLIP-2/LLaVA/Flamingo-style models for VQA, captioning, grounding.

Minimal math:

$$\mathrm{Attn}(\mathbf{Q}_{\mathsf{text}}, \mathbf{K}_{\mathsf{image}}, \mathbf{V}_{\mathsf{image}}) = \mathrm{softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_i^\top}{\sqrt{d}}\right) \mathbf{V}_i$$

Text asks; image answers. Great for localization, step-by-step reasoning.

Cross-Modal Attention: The Bridge Between Modalities

Standard Self-Attention: Q, K, V all from same modality
Cross-Modal Attention: Mix queries and keys from different modalities

Example Configurations:

1. **Image** $\rightarrow$ **Text**: Q from text, K, V from image

$$\text{Attention}(\mathbf{Q}_{\text{text}}, \mathbf{K}_{\text{image}}, \mathbf{V}_{\text{image}}) \tag{1}$$

2. **Text** $\rightarrow$ **Image**: Q from image, K, V from text

$$\text{Attention}(\mathbf{Q}_{\text{image}}, \mathbf{K}_{\text{text}}, \mathbf{V}_{\text{text}}) \tag{2}$$

Intuition:

- Text asks questions (Q), image provides answers (K, V)
- "What does this medical image show?" $\rightarrow$ Relevant image patches

Vision–Language Models (VLMs)

GPT-4oV , Qwen-VL
GPT5

Definition: Models that **jointly process images and text** to produce shared understanding.

image ⇒ "text"   ↳ LLM

- **Inputs:** multimodal pairs (image + caption, scan + report, diagram + text)
- **Architecture:** often two encoders (vision, text) with **fusion layers** (cross-attention)
  ("f", "g")
- **Outputs:** can be text (caption, answer), labels (classification), or embeddings (retrieval)

Visual Question Answering (VQA)

Task: Answer a natural language question about an image.

- Input: (Image, Question) e.g., Chest X-ray + "What abnormality is visible?" Answer in {"healthy", "lung cancer"}.
- Model: process image patches + text tokens; combine with cross-attention
- Output: Answer text or categorical label

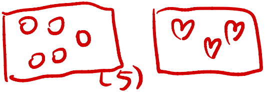
Infer the safety level of traveling?
"Bias"

Why it matters:

- Brings interaction: models respond to specific queries
- Biomedical use: "Where is the tumor?", "What stage is this?", "Has pneumonia improved?"
- VQA = the canonical benchmark for multimodal reasoning

Geographic Benchmark

Can VLM like GPT-4 Count?



(5) (3) ⇒ Ask Models how many objects there are!

city 1

city 2

⇒ Infer the city?

Representative VLMs

(Open-weight: Download the actual model & Deploy it yourself).

General-domain:

Given an image ⇒ Generates its caption

- **BLIP / BLIP-2:** pretrain on image–text pairs, instruction-tune for captioning & QA
- **Flamingo:** frozen LLM + cross-attention adapters
- **LLaVA:** connect CLIP vision encoder to LLaMA (open-weight LLM from meta) for multimodal dialogue

Biomedical adaptations: Train the general model on a Biomedical Dataset

- **MedVQA:** chest X-ray VQA datasets (e.g., VQA-RAD)
- **PMC-VQA:** millions of figure–caption QA pairs from PubMed Central
- **BioLLaVA:** LLaVA variants tuned on biomedical images + text

Three Key Fusion Strategies

Early Fusion (Feature Level; what we just covered):

- Combine raw features from different modalities
- Pro: Rich interaction between modalities
- Con: High dimensionality, potential for overfitting

Late Fusion (Decision Level):

- Process each modality separately, combine final outputs
- Pro: Modular, interpretable, handles missing data
- Con: Limited cross-modal interaction

Hybrid Fusion (Intermediate):

- Combine at multiple stages of processing
- Pro: Balance between interaction and modularity
- Con: More complex architecture and training

Text ⇒ Trying the find the "Right" Images

Retrieval (Image $\leftrightarrow$ Text)

- Task: Given a query (e.g., *"Which image shows a doctor examining an X-ray?"*), retrieve the correct image or caption.
- Key Metrics:

  100 Text Questions ⇒ 85 of the Top 10 similar images which contain at least one correct answer.

  - **Recall@K**: Fraction of queries where the correct match appears in the top-K results.

    Recall@1, 5, ..., 10

    $$\text{Recall@K} = \frac{\#\{\text{queries with correct in top-K}\}}{\#\{\text{queries}\}}$$

    k=10

    Example: Recall@10 = 0.85 means 85% of queries had the correct answer within the first 10 retrieved results.

    $\frac{10}{10{,}000}$!

  - **Mean Reciprocal Rank (MRR)**: Average of reciprocal ranks of the first relevant item. Rewards higher ranking of the correct match.

    $$\left[\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}\right]$$

    Small Ranks ⇓ higher Rewards

    Top 1 ↓ ✓

  - Other: Precision@K, nDCG (normalized discounted cumulative gain) for graded relevance.

VLM Tasks: Captioning

① Caption/generate a medical Report based on MRI/ X-ray/etc.

Captioning (includes some classification)

Generative.

- Task: Given an image, generate a natural language description (e.g., *"A nurse holding a newborn baby in the delivery room."*)
- Most Common Metric: $\boxed{\textbf{BLEU}}$ / Human Evaluation RLHF (Ranking reward).
  - Measures n-gram precision between generated caption and references.
  - Formula (BLEU-N):

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where:

- p_n = modified n-gram precision
- w_n = weight (often uniform, $1/N$)
- $\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$ is the brevity penalty, with c = candidate length, r = reference length

  - Other metrics: CIDEr (consensus-based), SPICE (semantic graph).

VLM Tasks: Visual Question Answering (VQA)

Visual Question Answering

- Task: Answer natural language questions about an image. Example: Q: *"What color is the traffic light?"* → A: "Green"
- Key Metrics:
  - **Accuracy**: Proportion of questions answered correctly.

    (✓)

$$\text{Accuracy} = \frac{\#\{\text{correct answers}\}}{\#\{\text{total questions}\}}$$

Alternative: Whether "Green" is in the predicted answer

  - **Exact Match (EM)**: 1 if predicted string exactly matches reference, 0 otherwise.
  - **F1 score**: Harmonic mean of precision and recall at the token level.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

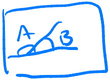
A: Green
B: The light is Green
C: Green light

VLM Tasks: Reasoning Benchmarks

→ VLM is much less developed than LLM

Multimodal Reasoning

- Task: Answer complex questions requiring reasoning or external knowledge. Examples:
  - OK-VQA: *"Which company manufactures the phone in the image?"*
  - ScienceQA: *"Which planet is shown in the diagram?"*
  - MathVista: *"What is the angle at point A in the figure?"*
- Metrics:
  - **Exact Match / F1** (string-level correctness).
  - **Chain-of-thought consistency**: fraction of steps logically valid.
  - **Human evaluation**: correctness and faithfulness of reasoning traces.



→ Q: Give me the angle of A .. → A: 120°

VLM Tasks: Multimodal Classification

Classification

- Task: Predict categorical label from fused modalities. Example: Input = meme image + caption. Output = "Hateful" or "Not hateful."
- Key Metrics:
  - **Accuracy**: fraction of correct predictions.
  - **Macro-F1**: average F1 across all classes, treating them equally.

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} \text{F1}_c$$

  - **AUROC**: area under ROC curve; probability that positive is ranked above negative.

Structured Perception (Detection & Segmentation)

- Task: Parse structured elements from multimodal inputs. Example: Given a scanned document, identify table regions and read cell values.
- Key Metrics:
  - **Mean Average Precision (mAP)** for detection:

$$\text{AP} = \int_0^1 p(r)\,dr \quad \text{and} \quad \text{mAP} = \frac{1}{C}\sum_{c=1}^{C} AP_c$$

    where $p(r)$ = precision as a function of recall.
  - **Mean Intersection over Union (mIoU)** for segmentation:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|}, \quad \text{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \text{IoU}_i$$

    where P = predicted region, G = ground truth.

Popular Pretrained Multimodal Models

- **CLIP / Open-CLIP**: Contrastive alignment of image/text embeddings → strong retrieval and zero-shot classification.
- **BLIP / BLIP-2**: Generative pretraining (captioning) + adaptation to VQA and reasoning tasks.
- **Flamingo (DeepMind)**: Few-shot multimodal QA by adding cross-attention layers to frozen LLMs.
- **LLaVA, GPT-4V**: Visual encoder + LLM backbone; strong reasoning and general-purpose QA.

Key multimodal pretraining strategies:

- Contrastive (aligning vision/text in shared space)
- Generative (predicting captions or answers)
- Instruction tuning (aligning multimodal input with human queries)

Adapting Pretrained Models to Tasks

Zero-Shot / Prompting

- CLIP: zero-shot classification with natural prompts
- GPT-4V: direct domain-specific reasoning (e.g., chart analysis)

Lightweight Finetuning

- LoRA / Adapters: update small parameter subsets ($W \approx W_0 + AB^\top$)
- Instruction-tuning with limited task-specific data

Task-Specific Heads

- Detection: bounding box regression head
- VQA: cross-attention head into LLM decoder

LLM vs VLM Finetuning

- **LLM finetuning**: focuses on text-only alignment (e.g., domain language adaptation).
- **VLM finetuning**: requires jointly adapting vision + language modules (e.g., aligning image features to textual concepts).

Applications of VLMs

- **Healthcare**: Radiology reports, pathology slides, multimodal risk prediction (Metrics: AUROC, sensitivity/specificity)
- **Science & Education**: Diagram QA, chart interpretation, multimodal tutoring (Metrics: Exact Match, F1, reasoning consistency)
- **Web & Industry**: Content moderation, e-commerce retrieval, multimodal search (Metrics: F1, Recall@K, CTR = clicks / impressions)

What is a potential application in **your research area**?