
LECTURE 7-1

Vision-Language and Multimodal Models

Images and text in one model

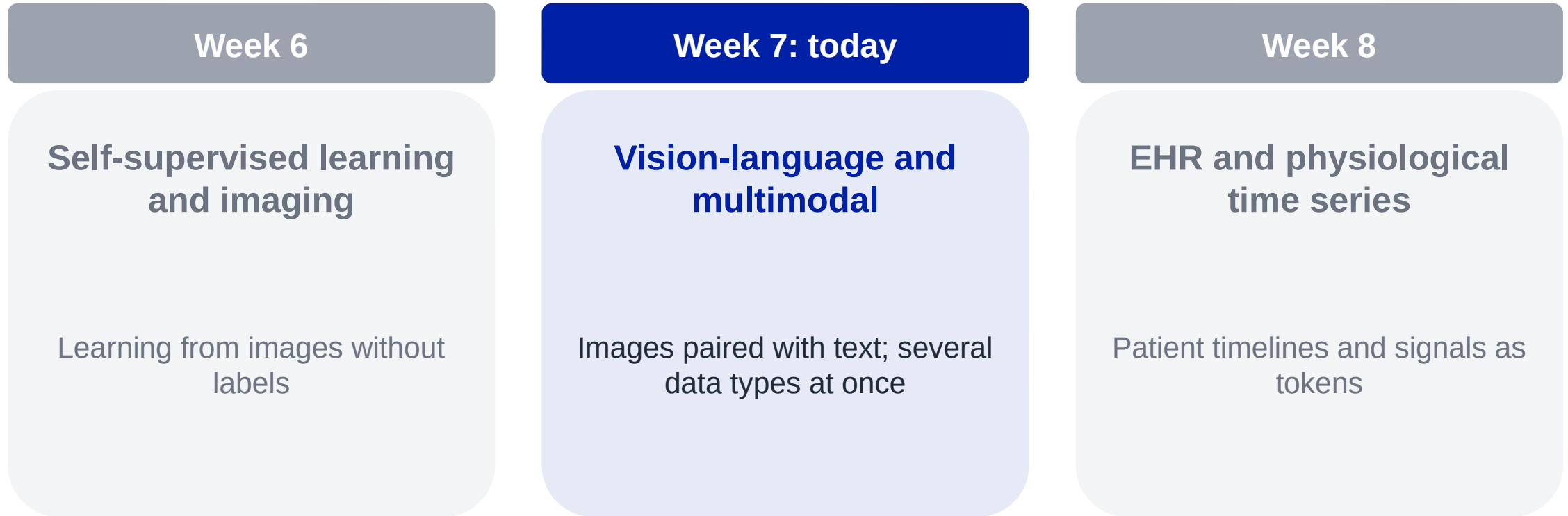


Parisa Rashidi · Herbert Wertheim College of Engineering, University of Florida



Where We Are

Week 6 learned from images alone. Today: images and text together.



Two Ways to Combine Images and Text

Match them in a shared space, or feed the image into a language model.

1. Match: CLIP

an image encoder and a text encoder
map both into one shared space

good for: search, zero-shot labels
cannot: write a report

2. Read: images as tokens

image vectors go into the language model
next to the text tokens

good for: reports, answers, dialogue
Med-PaLM M, MedGemma

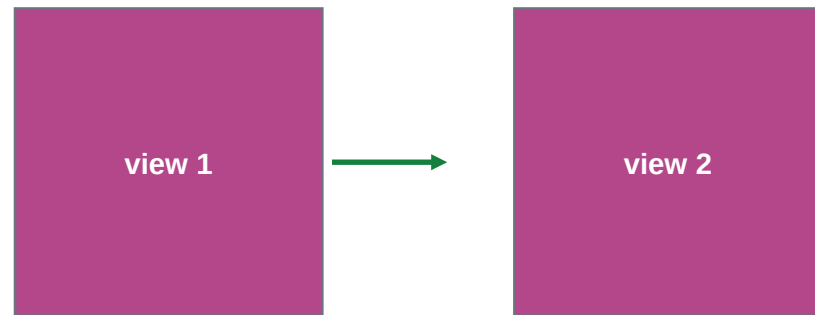


CLIP: Matching Images with Their Captions

From Two Views to an Image and Its Caption

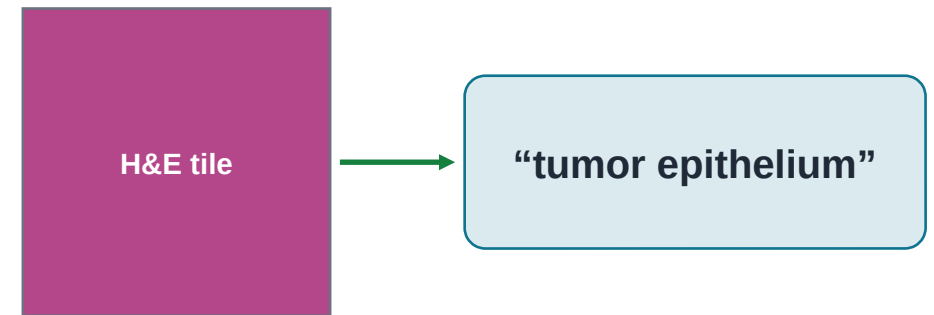
Week 6 paired two views of one image. CLIP pairs an image with its text.

Week 6: two views



pulled together

CLIP: image and caption

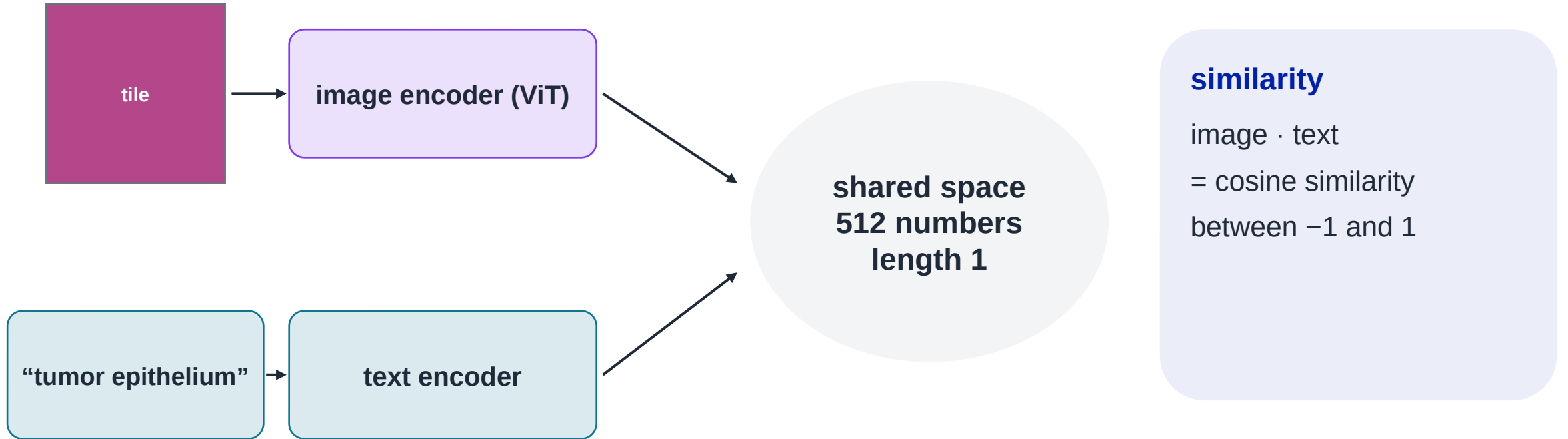


pulled together

Negatives: every other caption in the batch.

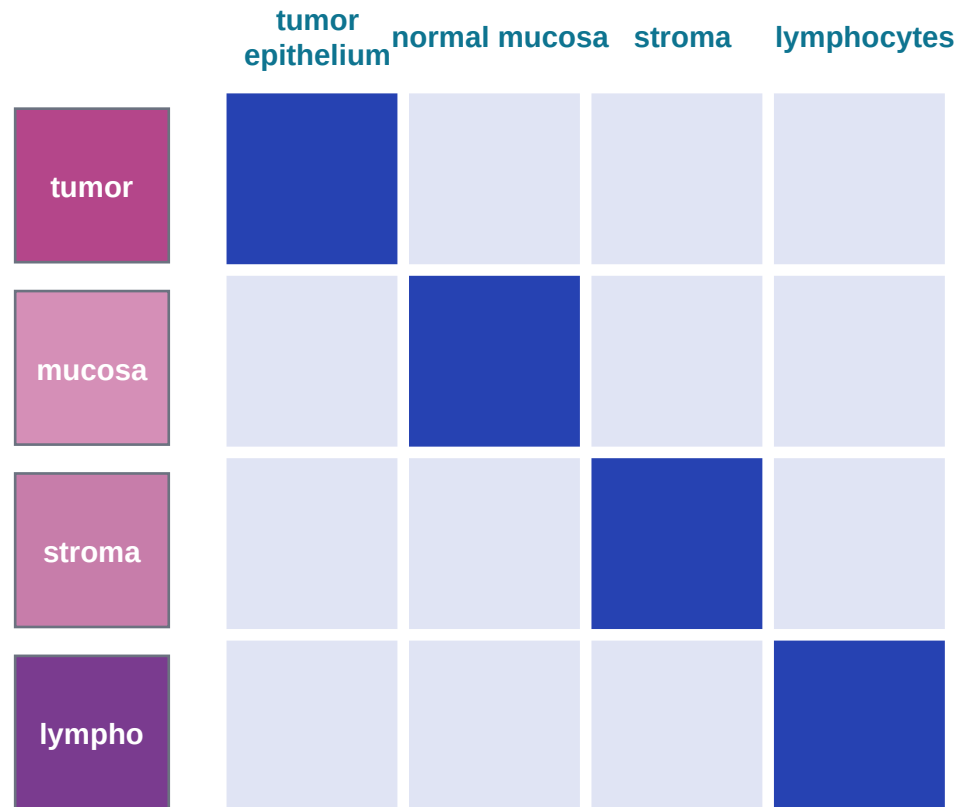
Two Encoders, One Shared Space

Each encoder maps its input to a vector of the same size, normalized to length 1.



The Similarity Grid

A batch of image and caption pairs gives a grid; matching pairs lie on the diagonal.



diagonal: matching pairs
(pushed up)

off-diagonal: wrong pairs
(pushed down)

The CLIP Loss

Cross-entropy along every row and every column, averaged.

$$\text{loss} = \frac{1}{2} [\text{CE}(\text{rows: image} \rightarrow \text{text}) + \text{CE}(\text{columns: text} \rightarrow \text{image})]$$

similarity $\div \tau$

τ is learned, starting at 0.07

softmax per row

which caption matches this image?

softmax per column

which image matches this caption?

target

the diagonal

Scale: Web Pairs and Biomedical Pairs

CLIP learned from 400 million web pairs; BiomedCLIP from 15 million figure captions.

CLIP (2021)

400 million web image-text pairs
batch of 32,768 pairs
temperature learned, from 0.07
little medical imagery in the data

BiomedCLIP

15 million figure-caption pairs
from 4.4 million PubMed Central articles
PubMedBERT text encoder
same contrastive recipe

Animation: CLIP: Matching Images with Their Captions

Animation clip: week07-clip1.mp4

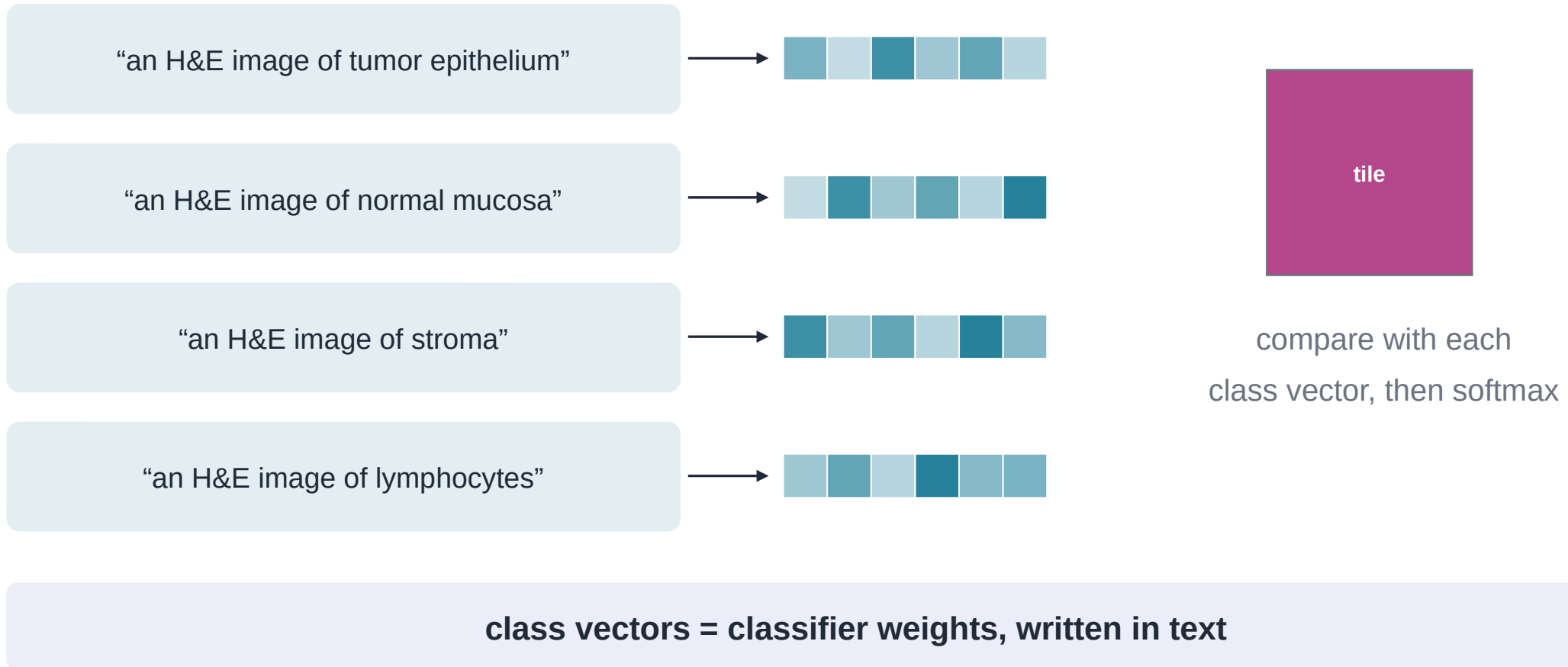
Silent version embedded (click to play). The narrated version is in the course repository.



Zero-Shot Classification

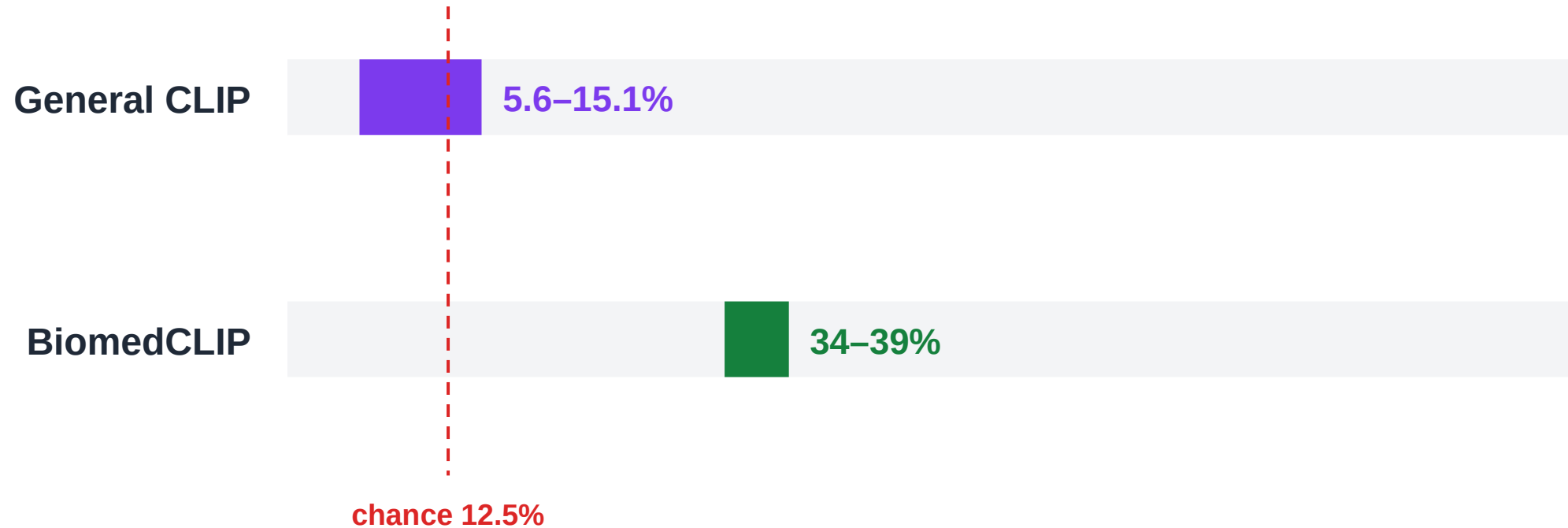
Prompts as Classifier Weights

Encode one text prompt per class; the class vectors act as a linear classifier.



How Well Does It Work on Pathology?

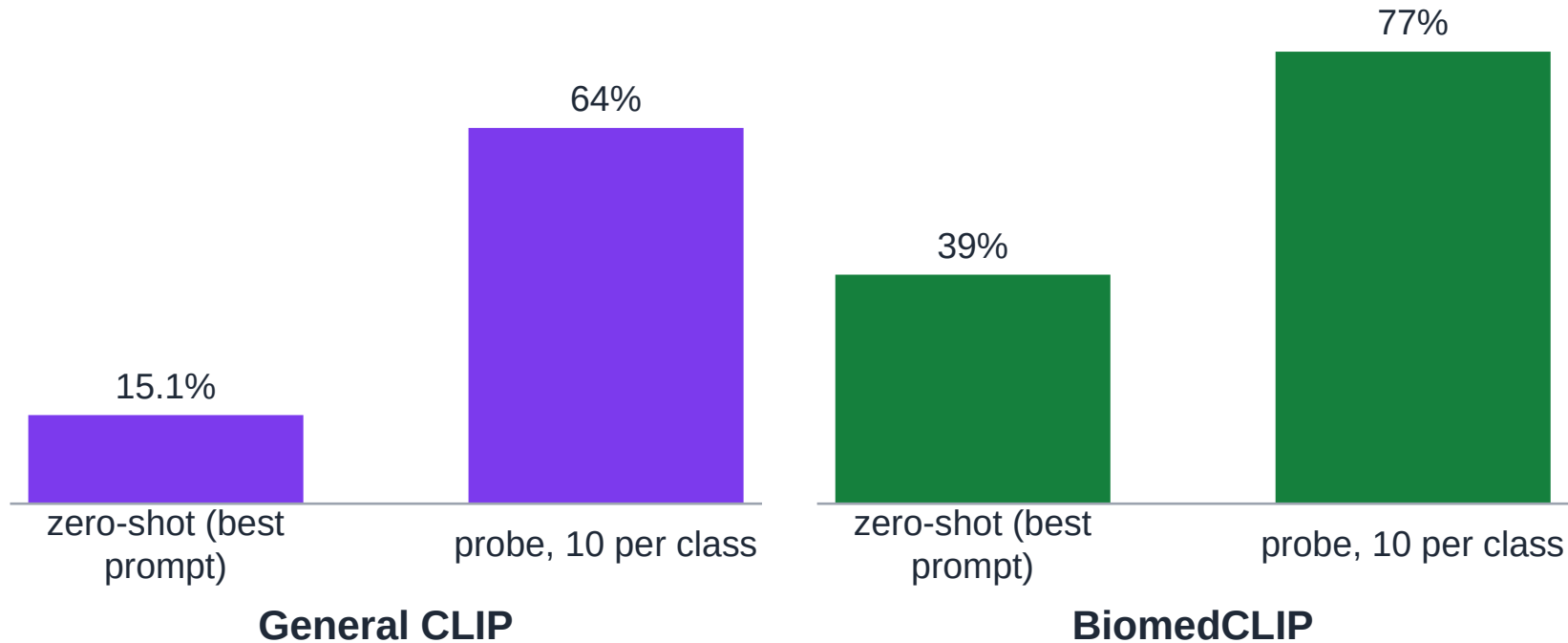
General CLIP is near or below chance on H&E tiles; BiomedCLIP does better.



range = different prompt wordings; 5,000 Kather tiles, 8 classes

A Few Labels Beat a Clever Prompt

A linear probe with 10 labeled tiles per class outperforms every prompt, for both models.



For projects

try zero-shot first;
then label a few
examples and probe

Animation: Zero-Shot Classification, and Where It Fails

Animation clip: week07-clip2.mp4

Silent version embedded (click to play). The narrated version is in the course repository.



Images as Tokens: Med- PaLM M

Matching Is Not Writing

CLIP can score a report against an image, but it cannot write one.

CLIP

image + candidate text
→ a similarity score

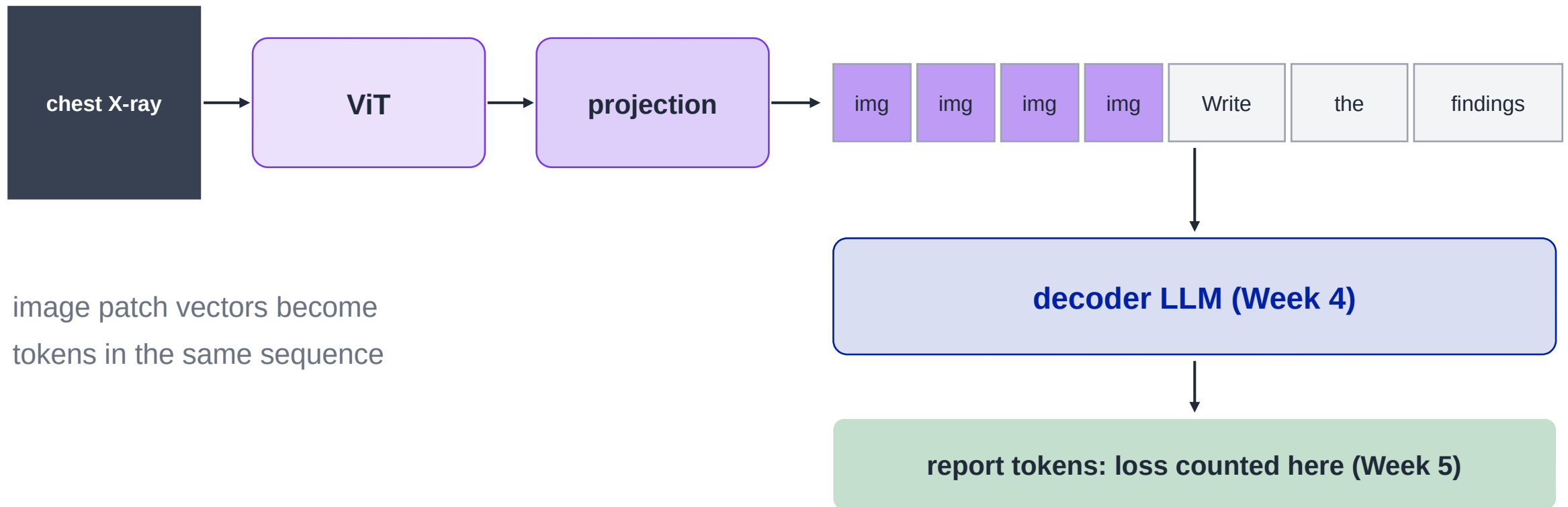
vision-language model

image + instruction
→ a written report

Writing needs the decoder from Weeks 3 to 5, with the image as part of its input.

Images as Tokens in the Language Model

Image vectors are projected and interleaved with text tokens; the rest is Weeks 4 and 5.



Med-PaLM M: One Model, Many Biomedical Tasks

PaLM-E (PaLM plus a ViT), fine-tuned on 14 biomedical tasks with one set of weights.

Model

PaLM-E: PaLM + ViT
sizes: 12B, 84B, 562B
one set of weights
for every task

MultiMedBench: 14 tasks

medical question answering (MedQA and others)
radiology report generation and summarization
visual QA on radiology and pathology images
mammography and dermatology classification
genomic variant calling

What the Evaluation Shows, and What It Does Not

A generalist model is competitive, but radiologists' reports still win most comparisons.

Reports (246 chest X-rays)

Med-PaLM M preferred over the radiologist's report in up to 40.5% of cases (side-by-side ranking)

Zero-shot tuberculosis

87.7% accuracy with no task training, vs 92.6% for a specialized ensemble trained on 138 images

**“Up to 40.5%” means the radiologist was preferred most of the time.
Retrospective rankings are not outcomes (Week 12).**

Animation: Images as Tokens: Med-PaLM M

Animation clip: week07-clip3.mp4

Silent version embedded (click to play). The narrated version is in the course repository.

Fusion and Missing Modalities

Where the modalities meet decides what happens when one is missing.

Early fusion

all modalities as tokens in one sequence

Intermediate

combine learned representations midway

Late fusion

one model per modality;
combine predictions

**Clinical reality: a modality is often missing.
Suggested project: does fusion still help when one is missing?**

Recap: Images and Text

Match

Two encoders, one space; contrastive loss along rows and columns of the grid.

Zero-shot

Prompts become classifier weights; wording matters; a few labels beat a clever prompt.

Read

Images as tokens in a decoder; the same next-token loss on report tokens.

Evaluate

Up to 40.5% preferred is still a minority; retrospective is not outcomes.

Before Thursday

Read

Radford et al., CLIP (ICML 2021) and Tu et al., Med-PaLM M (NEJM AI 2024).

Write

Reading responses on Canvas before 1:00 PM Thursday.

Explore

Week 7 reference notebook (ungraded): the CLIP loss from scratch, zero-shot versus a linear probe on Kather tiles.

Next week

Foundation models for EHR and physiological time series.