## A. Dataset Collection Procedure

**Video Acquisition**: 413 English educational videos were downloaded from YouTube. From the initial list, we filtered and curated a smaller list of 10 speakers according to the following criteria: (1) the material must be presented in a slide-based style, (2) the slides must be stationary (i.e. external video clips cannot be played), and (3) the speaker makes use of their mouse to refer to specific figures on the slide. After filtering, 334 videos remained. Following prior work [1, 50, 53, 26], we adopt a strict protocol to mitigate ethical concerns of using publicly available Youtube data: (1) videos are internally checked to avoid offensive content, (2) the raw videos are not shared, but only the Youtube IDs and download scripts are shared, (3) creators have full control of the accessibility of their content and any personally identifiable data (only links to publicly available data are shared) and (4) all of the creators were individually contacted about the inclusion of their content in our dataset. [2]

**Slide Segmentation**: The quality of segmentation is crucial for our task of retrieval, therefore, we collected manual human annotations on MTurk. We presented the annotator with a lecture video and asked the annotator to use a slider to navigate to the end of each slide and mark its precise timestamp. A screenshot of this experiment can be found in Appendix C. In order to ensure the high quality of segmentations, we conduct the annotation process in multiple steps. (1) An internal team manually annotated 10 lecture videos for groundtruth annotations. (2) The experiment was made available to 100 MTurkers. (3) We evaluate their results, marking an annotation as correct if theirs matched ours within a 1-second interval. (4) Annotators who were able to perform above a 90% correctness threshold were assigned the full set.

**Figure Annotation and Labeling**: Our dataset is unique from previous datasets as our focus is centered around figure-level retrieval. In order to enable this task, our data must consist of precise bounding boxes and labels for each figure. Therefore, we design an MTurk experiment where annotators are shown slides and asked to create a bounding box around figure instances and label their classes. Our class labels are inspired from PRImA [2], a dataset that consists of layouts from scientific reports. We follow their taxonomy to find labels on figures, which consist of natural images, diagrams, table, and equations. In Appendix D, we provide details on figure class labels and a screenshot of the MTurk experiment. To obtain precise and accurate figure annotations, we follow a multiphase process. (1) An internal team manually annotated 10 lecture videos for groundtruth figure annotations and labels. (2) We make the experiment available to 100 MTurkers for 10 different slides. (3) We manually evaluate the annotations, marking an annotation as correct if the annotators had the same number of figures, equivalent types, and high overlap of bounding boxes. (4) Annotators who were able to perform above a 90% correctness threshold were assigned the full set. (5) To ensure the absolute highest quality of figure annotations, our internal team of annotators manually corrected all the annotations for any mislabeled bounding box annotations or incorrect regions.

**Text Extraction: ASR & OCR**: We use Google ASR [7] to extract spoken language from audio. We use the Video-Model, which has a reported WER of 16% (Amazon: 22%, Microsoft 24%, IBM Watson 29%, Google Speech-Model 37%). We manually verify 100 random segments in the dataset, and find that the WER is 17.1%. To extract OCR text from the images of slides, we use Tesseract [44]. We manually verify 100 random slides in the dataset, and find that the WER is 37.82%.

**Mouse Trace Location Extraction**: We extract the mouse trace location to be used as an additional grounding signal between visual objects and language. For each segmented slide, the background is static and the only object that is moving is the pointer. If there is any movement, we consider that as the pointer location. We manually verify 100 random mouse trace location in the dataset, and find that the percentage of correct keypoints (PCK) with a threshold of 50 pixels, is 77.1%.

## B. MTurk: Annotators

For each task, we approximate the time each takes with internal annotators to ensure a minimum payment of $8 per hour for MTurkers.

**Slide Segmentation:** The annotators are compensated at least $8 per hour. For the task of slide segmentation, as annotators are simply required to scroll through the video to find transition points (which takes around 10 minutes for an hour long video), we pay 50 cents for a 15-minute-long video (i.e $2 to scroll through an hour-long video). This approximates to roughly a payment of $10 per hour. We pay annotators a total amount of $856.95 for this task.

**Figure Annotation:** The annotators are compensated at least $8 per hour. For the task of figure annotations, we pay the annotators 5 cents per slide, where annotators are expected to spend around 20 seconds per slide (approximating around $9 per hour). As a result, we spent $451.55 for a total of 9031 slides.

[2] Implicit consent allows creators to be removed from the dataset at any time.

## C. MTurk: Slide Segmentation

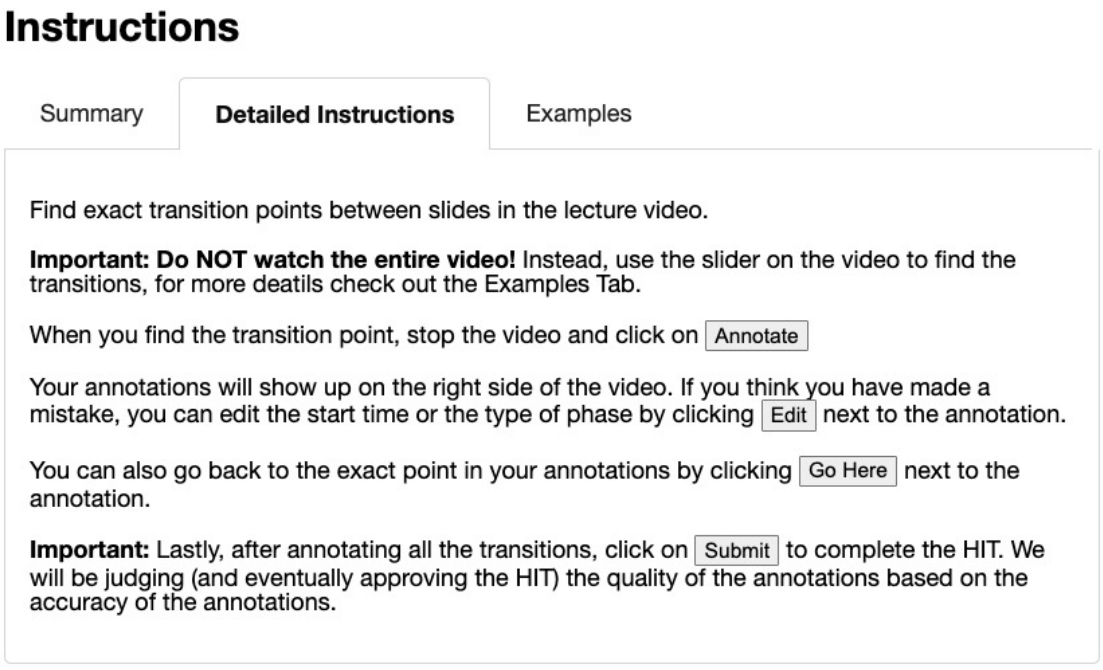
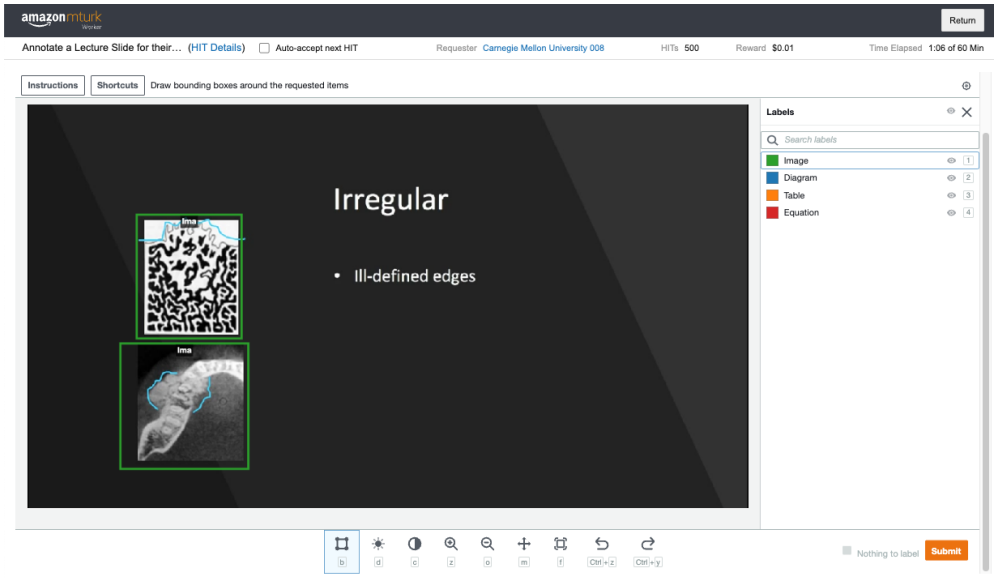




Figure 8: MTurk Screenshot and Instructions for Slide Segmentation

## D. MTurk: Figure Annotation




Annotation Instructions

1. To draw a bounding box click the correct label, and draw a rectangle using your mouse over each instance of the target.
2. If the target goes off the screen, label up to the edge of the image.
3. Capture each distinct image that is on the slide.
4. Do not mark logos or other images that do not add informative content to the topic of the slide.
5. Only capture slide images - do not mark speaker faces or videos stills that may have been played on a slide
6. If there is nothing to label, mark the checkbox "Nothing to Label" next to the submit button
7. You can delete an annotation by clicking the bounding box you just made and pressing delete on your keyboard.

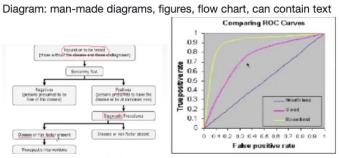
**Bounding Box Instructions**

Use the bounding box tool to draw bounding boxes over the requestsed regions, if they appear on the slide :

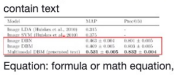
Shown below are the definitions for each requested region

1. Images: photographs of natural images, can contain text

2. Diagram: man-made diagrams, figures, flow chart, can contain text




3. Table: arrangement of information or data, typically in rows and columns, can contain text



4. Equation: formula or math equation, can contain text

Figure 9: MTurk Screenshot and Instructions for Figure Annotations

## E. Comprehensive Results for Each Speaker

| anat-1 | Figure-to-Text | | | Text-to-Figure | | |
|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| Random | 0.27 ± 0.38 | 3.18 ± 2.4 | 5.77 ± 1.73 | 0.26 ± 0.36 | 3.38 ± 1.42 | 8.07 ± 3.8 |
| CLIP | 0.53 ± 0.37 | 4.53 ± 1.21 | 9.11 ± 1.11 | 0.54 ± 0.38 | 3.19 ± 1.63 | 7.17 ± 3.15 |
| PVSE | 2.1 ± 0.4 | 7.33 ± 1.14 | 12.76 ± 2.07 | 1.83 ± 0.36 | 8.09 ± 0.92 | 11.73 ± 0.88 |
| PVSE (BERT) | 2.62 ± 0.43 | 5.49 ± 0.12 | 10.48 ± 1.71 | 1.04 ± 0.36 | 7.01 ± 2.21 | 10.68 ± 1.82 |
| PCME | 1.3 ± 0.71 | 4.18 ± 0.32 | 7.83 ± 0.16 | 1.3 ± 0.71 | 4.18 ± 0.32 | 7.83 ± 0.16 |
| PCME (BERT) | 1.3 ± 0.71 | 3.92 ± 0.63 | 8.09 ± 0.92 | 1.3 ± 0.71 | 3.92 ± 0.63 | 8.09 ± 0.92 |
| Ours | **11.23 ± 0.91** | 30.82 ± 6.1 | 42.31 ± 4.82 | **13.79 ± 2.34** | 34.34 ± 8.91 | 44.9 ± 6.53 |
| Ours w/ Trace | 9.64 ± 3.08 | **31.05 ± 6.71** | **46.49 ± 2.67** | 10.71 ± 0.54 | **36.86 ± 2.33** | **49.85 ± 1.14** |

| anat-2 | Figure-to-Text | | | Text-to-Figure | | |
|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| Random | 1.75 ± 2.48 | 21.05 ± 8.59 | 50.87 ± 9.92 | 8.77 ± 2.48 | 31.58 ± 8.6 | 56.14 ± 8.94 |
| CLIP | 7.12 ± 2.42 | 23.2 ± 6.48 | 57.21 ± 7.01 | 3.51 ± 4.96 | 16.08 ± 8.61 | 37.43 ± 3.61 |
| PVSE | 7.02 ± 2.48 | 38.6 ± 2.48 | 66.67 ± 2.48 | 7.02 ± 2.48 | 38.6 ± 6.56 | 68.42 ± 0.0 |
| PVSE (BERT) | 5.26 ± 4.3 | 33.33 ± 4.96 | 61.4 ± 8.94 | 5.26 ± 0.0 | 33.34 ± 6.56 | 52.63 ± 8.59 |
| PCME | 5.26 ± 0.0 | 28.07 ± 4.96 | 56.14 ± 2.48 | 5.26 ± 0.0 | 28.07 ± 4.96 | 56.14 ± 2.48 |
| PCME (BERT) | 5.26 ± 0.0 | 28.07 ± 2.48 | 52.63 ± 0.0 | 5.26 ± 0.0 | 28.07 ± 2.48 | 52.63 ± 0.0 |
| Ours | **7.02 ± 2.48** | **54.39 ± 6.56** | **78.95 ± 4.3** | **8.77 ± 2.48** | **57.89 ± 4.3** | 75.44 ± 6.56 |
| Ours w/ Trace | **8.77 ± 4.96** | **49.12 ± 6.56** | **73.68 ± 7.44** | 7.02 ± 6.56 | **49.12 ± 8.94** | **77.19 ± 6.56** |

| bio-1 | Figure-to-Text | | | Text-to-Figure | | |
|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| Random | 0.79 ± 0.43 | 3.13 ± 1.19 | 4.7 ± 0.89 | 8.77 ± 2.48 | 31.58 ± 8.6 | 56.14 ± 8.94 |
| CLIP | 0.51 ± 0.03 | 3.41 ± 1.53 | 5.7 ± 2.4 | 3.51 ± 4.96 | 16.08 ± 8.61 | 37.43 ± 3.61 |
| PVSE | 0.97 ± 0.07 | 4.05 ± 0.54 | 6.0 ± 1.06 | 7.02 ± 2.48 | 38.6 ± 6.56 | 68.42 ± 0.0 |
| PVSE (BERT) | 0.79 ± 0.43 | 5.07 ± 1.15 | 8.55 ± 1.52 | 5.26 ± 0.0 | 33.34 ± 6.56 | 52.63 ± 8.59 |
| PCME | 0.66 ± 0.29 | 2.11 ± 0.4 | 4.68 ± 0.49 | 5.26 ± 0.0 | 28.07 ± 4.96 | 56.14 ± 2.48 |
| PCME (BERT) | 0.48 ± 0.03 | 2.39 ± 0.23 | 4.68 ± 0.49 | 5.26 ± 0.0 | 28.07 ± 2.48 | 52.63 ± 0.0 |
| Ours | **4.23 ± 0.9** | **12.53 ± 2.6** | **19.15 ± 1.29** | **8.77 ± 2.48** | **57.89 ± 4.3** | 75.44 ± 6.56 |
| Ours w/ Trace | **2.91 ± 0.82** | **7.14 ± 1.08** | **12.65 ± 2.73** | 7.02 ± 6.56 | **49.12 ± 8.94** | **77.19 ± 6.56** |

| **bio-3** | Figure-to-Text | | | Text-to-Figure | | |
|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| Random | 0.57 ± 0.4 | 3.68 ± 1.63 | 6.58 ± 1.56 | 1.16 ± 1.11 | 4.07 ± 1.72 | 8.35 ± 1.33 |
| CLIP | 0.0 ± 0.0 | 5.4 ± 1.52 | 11.35 ± 2.29 | 0.85 ± 0.02 | 3.66 ± 1.04 | 7.09 ± 1.22 |
| PVSE | 1.14 ± 0.81 | 6.61 ± 0.93 | 15.8 ± 1.33 | 1.16 ± 0.43 | 6.34 ± 1.25 | 12.35 ± 1.28 |
| PVSE (BERT) | 2.87 ± 1.11 | 7.47 ± 0.94 | 12.93 ± 1.45 | 1.43 ± 0.39 | 5.12 ± 1.02 | 9.19 ± 0.7 |
| PCME | 1.7 ± 0.65 | 5.15 ± 0.56 | 8.28 ± 2.13 | 1.7 ± 0.65 | 5.15 ± 0.56 | 8.28 ± 2.13 |
| PCME (BERT) | 1.7 ± 0.65 | 5.44 ± 0.76 | 9.16 ± 1.55 | 1.7 ± 0.65 | 5.44 ± 0.76 | 9.16 ± 1.55 |
| Ours | **1.74 ± 0.75** | **12.03 ± 0.37** | **19.14 ± 1.56** | **4.57 ± 1.45** | **13.11 ± 3.02** | **20.04 ± 2.47** |
| Ours w/ Trace | **1.17 ± 0.83** | **8.85 ± 1.9** | **14.85 ± 3.03** | 3.42 ± 0.6 | **10.03 ± 0.35** | **16.36 ± 2.08** |

| **bio-4** | Figure-to-Text | | | Text-to-Figure | | |
|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| Random | 0.77 ± 0.44 | 2.15 ± 0.45 | 4.62 ± 1.04 | 0.77 ± 0.21 | 2.93 ± 0.98 | 5.84 ± 1.13 |
| CLIP | 0.32 ± 0.46 | 2.48 ± 0.16 | 4.95 ± 0.83 | 0.16 ± 0.23 | 1.88 ± 0.81 | 5.16 ± 1.5 |
| PVSE | 1.7 ± 1.44 | 4.75 ± 1.41 | 7.82 ± 1.96 | 2.17 ± 1.18 | 4.31 ± 1.25 | 6.45 ± 1.62 |
| PVSE (BERT) | 1.08 ± 0.58 | 2.61 ± 0.57 | 5.07 ± 0.09 | 1.22 ± 0.75 | 3.52 ± 0.72 | 5.66 ± 1.61 |
| PCME | 1.86 ± 1.98 | 3.39 ± 1.55 | 4.92 ± 1.54 | 1.86 ± 1.98 | 3.39 ± 1.55 | 4.92 ± 1.54 |
| PCME (BERT) | 0.62 ± 0.22 | 1.23 ± 0.57 | 3.99 ± 1.17 | 0.62 ± 0.22 | 1.23 ± 0.57 | 3.99 ± 1.17 |
| Ours | **4.28 ± 2.04** | **12.22 ± 6.66** | **18.5 ± 9.12** | **2.9 ± 1.72** | **11.79 ± 5.52** | **20.1 ± 6.1** |
| Ours w/ Trace | **3.67 ± 1.82** | **12.07 ± 5.46** | **19.9 ± 6.82** | 2.29 ± 0.97 | **12.68 ± 5.64** | **21.88 ± 7.02** |

| **dental** | Figure-to-Text | | | Text-to-Figure | | |
|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| Random | 0.17 ± 0.14 | 0.87 ± 0.25 | 1.91 ± 0.28 | 0.29 ± 0.21 | 0.92 ± 0.29 | 1.67 ± 0.28 |
| CLIP | 0.06 ± 0.08 | 1.09 ± 0.35 | 2.14 ± 0.25 | 0.23 ± 0.08 | 0.98 ± 0.23 | 1.85 ± 0.32 |
| PVSE | 0.4 ± 0.21 | 1.73 ± 0.13 | 2.48 ± 0.42 | 0.29 ± 0.08 | 1.44 ± 0.19 | 2.65 ± 0.26 |
| PVSE (BERT) | 0.34 ± 0.0 | 1.84 ± 0.57 | 2.65 ± 0.33 | 0.64 ± 0.31 | 1.57 ± 0.65 | 2.6 ± 0.32 |
| PCME | 0.23 ± 0.08 | 0.86 ± 0.13 | 1.73 ± 0.41 | 0.23 ± 0.08 | 0.86 ± 0.13 | 1.73 ± 0.41 |
| PCME (BERT) | 0.23 ± 0.08 | 0.86 ± 0.23 | 1.67 ± 0.34 | 0.23 ± 0.08 | 0.86 ± 0.23 | 1.67 ± 0.34 |
| Ours | **0.63 ± 0.29** | **2.72 ± 0.23** | **6.18 ± 0.53** | **1.15 ± 0.16** | **5.31 ± 0.25** | **8.36 ± 1.45** |
| Ours w/ Trace | **0.69 ± 0.23** | **3.28 ± 1.04** | **6.16 ± 1.1** | 0.8 ± 0.39 | **3.28 ± 0.69** | **5.88 ± 0.58** |

| **ml-1** | Figure-to-Text | | | Text-to-Figure | | |
|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| Random | 0.28 ± 0.2 | 1.88 ± 0.28 | 3.5 ± 0.48 | 0.29 ± 0.21 | 0.92 ± 0.29 | 1.67 ± 0.28 |
| CLIP | 0.43 ± 0.34 | 1.69 ± 0.36 | 4.83 ± 1.94 | 0.23 ± 0.08 | 0.98 ± 0.23 | 1.85 ± 0.32 |
| PVSE | 1.48 ± 0.47 | 5.65 ± 0.66 | 7.51 ± 0.96 | 0.29 ± 0.08 | 1.44 ± 0.19 | 2.65 ± 0.26 |
| PVSE (BERT) | 0.54 ± 0.16 | 3.78 ± 1.36 | 6.44 ± 1.18 | 0.64 ± 0.31 | 1.57 ± 0.65 | 2.6 ± 0.32 |
| PCME | 0.66 ± 0.34 | 2.43 ± 0.19 | 4.49 ± 0.61 | 0.23 ± 0.08 | 0.86 ± 0.13 | 1.73 ± 0.41 |
| PCME (BERT) | 0.54 ± 0.16 | 2.68 ± 0.53 | 4.61 ± 0.48 | 0.23 ± 0.08 | 0.86 ± 0.23 | 1.67 ± 0.34 |
| Ours | **0.82 ± 0.05** | **4.76 ± 1.93** | **7.89 ± 1.87** | **1.15 ± 0.16** | **5.31 ± 0.25** | **8.36 ± 1.45** |
| Ours w/ Trace | **1.22 ± 0.3** | **3.71 ± 0.89** | **6.2 ± 1.84** | 0.8 ± 0.39 | **3.28 ± 0.69** | **5.88 ± 0.58** |

| **psy-1** | Figure-to-Text | | | Text-to-Figure | | |
|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| Random | 4.48 ± 4.78 | 11.1 ± 6.21 | 21.22 ± 2.19 | 4.15 ± 1.35 | 15.36 ± 1.05 | 26.73 ± 1.25 |
| CLIP | 4.05 ± 3.89 | 13.22 ± 3.03 | 30.03 ± 5.88 | 2.68 ± 2.34 | 13.38 ± 2.66 | 22.35 ± 2.92 |
| PVSE | 3.99 ± 0.86 | 16.71 ± 4.75 | 29.65 ± 4.62 | 5.07 ± 2.48 | 17.98 ± 1.51 | 35.8 ± 1.29 |
| PVSE (BERT) | **5.71 ± 2.87** | 18.87 ± 3.56 | 27.79 ± 3.23 | 4.16 ± 1.39 | 20.46 ± 3.25 | 34.84 ± 2.7 |
| PCME | 5.08 ± 2.49 | 14.75 ± 1.36 | 26.29 ± 3.24 | 4.16 ± 1.39 | 15.68 ± 2.66 | 30.92 ± 9.63 |
| PCME (BERT) | 4.16 ± 1.39 | 13.54 ± 6.14 | 26.93 ± 10.49 | 4.16 ± 1.39 | 14.46 ± 7.45 | 26.93 ± 10.49 |
| Ours | **2.27 ± 3.21** | **19.5 ± 0.76** | **38.23 ± 2.51** | **9.38 ± 5.52** | **32.4 ± 1.45** | **43.52 ± 5.59** |
| Ours w/ Trace | **3.39 ± 1.54** | **19.18 ± 3.81** | **33.78 ± 4.26** | 7.51 ± 3.75 | **20.83 ± 6.17** | **37.8 ± 5.6** |

| **psy-2** | Figure-to-Text | | | Text-to-Figure | | |
|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| Random | 4.48 ± 4.78 | 11.1 ± 6.21 | 21.22 ± 2.19 | 0.44 ± 0.62 | 5.96 ± 2.95 | 12.74 ± 4.36 |
| CLIP | 4.05 ± 3.89 | 13.22 ± 3.03 | 30.03 ± 5.88 | 1.62 ± 1.46 | 5.77 ± 1.85 | 14.12 ± 0.88 |
| PVSE | 3.99 ± 0.86 | 16.71 ± 4.75 | 29.65 ± 4.62 | 3.47 ± 2.08 | 11.2 ± 2.34 | 19.22 ± 1.49 |
| PVSE (BERT) | **5.71 ± 2.87** | 18.87 ± 3.56 | 27.79 ± 3.23 | **4.47 ± 0.6** | 11.56 ± 1.24 | 18.26 ± 2.92 |
| PCME | 5.08 ± 2.49 | 14.75 ± 1.36 | 26.29 ± 3.24 | 2.18 ± 2.24 | 8.91 ± 3.03 | 17.98 ± 2.88 |
| PCME (BERT) | 4.16 ± 1.39 | 13.54 ± 6.14 | 26.93 ± 10.49 | 1.83 ± 0.76 | 8.63 ± 3.26 | 15.94 ± 6.21 |
| Ours | **2.27 ± 3.21** | **19.5 ± 0.76** | **38.23 ± 2.51** | **1.36 ± 1.08** | **14.89 ± 7.3** | **26.92 ± 5.74** |
| Ours w/ Trace | **3.39 ± 1.54** | **19.18 ± 3.81** | **33.78 ± 4.26** | 2.72 ± 2.15 | **16.06 ± 0.29** | **27.66 ± 2.79** |

| **speaking** | Figure-to-Text | | | Text-to-Figure | | |
|---|---|---|---|---|---|---|
| | Recall@1 | Recall@5 | Recall@10 | Recall@1 | Recall@5 | Recall@10 |
| Random | 3.26 ± 2.72 | 21.46 ± 6.17 | 44.79 ± 7.36 | 0.44 ± 0.62 | 5.96 ± 2.95 | 12.74 ± 4.36 |
| CLIP | 4.34 ± 1.65 | 14.16 ± 7.03 | 38.77 ± 3.21 | 1.62 ± 1.46 | 5.77 ± 1.85 | 14.12 ± 0.88 |
| PVSE | 8.54 ± 1.64 | 27.64 ± 5.15 | 51.18 ± 5.45 | 3.47 ± 2.08 | 11.2 ± 2.34 | 19.22 ± 1.49 |
| PVSE (BERT) | **7.43 ± 1.39** | 24.44 ± 2.67 | 38.4 ± 3.71 | **4.47 ± 0.6** | 11.56 ± 1.24 | 18.26 ± 2.92 |
| PCME | 3.19 ± 0.1 | 16.04 ± 3.08 | 32.01 ± 3.53 | 2.18 ± 2.24 | 8.91 ± 3.03 | 17.98 ± 2.88 |
| PCME (BERT) | 3.19 ± 0.1 | 15.97 ± 0.49 | 30.83 ± 0.59 | 1.83 ± 0.76 | 8.63 ± 3.26 | 15.94 ± 6.21 |
| Ours | **13.75 ± 3.68** | **29.58 ± 7.24** | **53.06 ± 4.52** | **1.36 ± 1.08** | **14.89 ± 7.3** | **26.92 ± 5.74** |
| Ours w/ Trace | **5.28 ± 2.9** | **32.71 ± 9.07** | **52.92 ± 9.48** | **2.72 ± 2.15** | **16.06 ± 0.29** | **27.66 ± 2.79** |

Table 4: Speaker-wise results for Figure-to-Text and Text-to-Figure retrieval. PolyViLT consistently outperforms previous baselines.

## F. Held-out Testing Results

| | Text-to-Figure | | | Figure-to-Text | | |
|---|---|---|---|---|---|---|
| | r@1 | r@5 | r@10 | r@1 | r@5 | r@10 |
| Random | 0.47 ± 0.12 | 3.17 ± 0.76 | 6.27 ± 0.26 | 0.77 ± 0.17 | 2.7 ± 0.57 | 5.15 ± 1.4 |
| PCME | 0.52 ± 0.0 | 2.91 ± 0.0 | 4.95 ± 0.0 | 0.52 ± 0.0 | 2.91 ± 0.0 | 4.95 ± 0.0 |
| PVSE | 0.91 ± 0.34 | 3.69 ± 0.39 | 7.12 ± 0.56 | 0.85 ± 0.1 | 3.58 ± 0.39 | 6.34 ± 0.32 |
| PolyViLT (Ours) | **1.25 ± 0.33** | **5.6 ± 1.0** | **9.54 ± 0.98** | **0.99 ± 0.43** | **5.12 ± 0.67** | **8.0 ± 1.01** |

Table 5: Held-out testing results, we train on a source speaker and test on held-out target speaker for 4 speaker pairs (bio-1 $\rightarrow$ bio-3, anat-1 $\rightarrow$ anat-2, psy-1 $\rightarrow$ psy-2, bio-1 $\rightarrow$ psy-1)

## G. Slide Explanation Generation

| | Rouge-1 | Rouge-2 | Rouge-L |
|---|---|---|---|
| anat-1 | 0.153 | 0.048 | 0.123 |
| bio-1 | 0.092 | 0.033 | 0.076 |
| psy-1 | 0.066 | 0.015 | 0.053 |

Table 6: Finetuned slide explanation generation scores

## H. Fine-tuning CLIP

| | Text-to-Figure | | | Figure-to-Text | | |
|---|---|---|---|---|---|---|
| | r@1 | r@5 | r@10 | r@1 | r@5 | r@10 |
| CLIP | 2.22 | 8.45 | 14.45 | 1.63 | 7.59 | 19.0 |
| CLIP - Finetuned | 3.56 | 11.76 | 20.66 | 1.95 | 8.59 | 16.27 |

Table 7: Results on 10 speaker average of zero-shot pretrained CLIP vs fine-tuned CLIP on our crossmodal retrieval objectives

## I. Keyword Identifiability

| **PolyViLT r@10** | tf-idf rank | | | |
|---|---|---|---|---|
| | <5 | 5 -10 | 10 - 30 | 30 - 50 |
| Text-to-Figure | **0.236** | 0.2 | 0.122 | 0.132 |
| Figure-to-Text | **0.249** | 0.22 | 0.066 | 0.124 |

Table 8: Recall@10 scores for Keyword Identifiability measured by TF-IDF ranks

Specifically for figures which contain text, which consists of 54.9% of our dataset, there are many cases where the pairing between text and figures can be easily found by identifying the keyword and finding its existence in the figure or the spoken language. Naively finding the existence of identical words in two instances is trivial and could lead to incorrect retrievals. The core challenge lies in correctly identifying the keyword that defines the slide segment.

In order to understand the importance of identifying the keyword and how our model performs for text-inclusive figures, we measure the term frequency–inverse document frequency (or tf-idf) of each word in the spoken language, except stopwords which are filtered out. The words are then ranked according to their tf-idf values. We iterate through each word, find the words that also exist in the ocr output of the figure, and extract the word with the lowest tf-idf rank. Under this condition, if the tf-idf rank for a word is 5, this can be intuitively seen as the fifth most important keyword that defines the slide. Simply stated, if the tf-idf rank of a word is low, the keyword can be easily detected in the slide and the spoken language. On the other hand, if the tf-idf rank of a word is high, this implies that the keyword is hard to detect. [3]

In Table 8 We measure the recall@10 score conditioned on tf-idf ranks, which indicates how well PolyViLT does under varying levels of difficulty of identifying the keyword. PolyViLT's struggles for cases with easier keyword identifiability and suffers even more with harder cases. This calls for a need for PolyViLT to effectively address easier cases, via using tf-idf directly as a feature, and relying more on the vision when the keyword is not easily identifiable.

[3] Note that this method of retrieval is intractable with more number of words and documents

## J. Long Range Sequence and OOV Tokens

| CLIP | (a) Length of Spoken Language | | | | | (b) Number of Subwords | | | |
|---|---|---|---|---|---|---|---|---|---|
| **r@10** | <100 | 100 - 200 | 200 - 400 | 400 - 600 | 600+ | <10 | 10 - 20 | 20-30 | 30 - 50 |
| Figure-to-Text | 0.0447 | 0.0465 | 0.0567 | 0.0676 | 0.175 | 0.065 | 0.062 | 0.0543 | 0.0619 |
| Text-to-Figure | 0.0793 | 0.0662 | 0.0599 | 0.0571 | 0.14 | 0.0704 | 0.055 | 0.0498 | 0.0473 |

| PVSE | (a) Length of Spoken Language | | | | | (b) Number of Subwords | | | |
|---|---|---|---|---|---|---|---|---|---|
| **r@10** | <100 | 100 - 200 | 200 - 400 | 400 - 600 | 600+ | <10 | 10 - 20 | 20-30 | 30 - 50 |
| Figure-to-Text | 0.0779 | 0.0777 | 0.0644 | 0.0901 | 0.0928 | 0.0973 | 0.0842 | 0.0656 | 0.0667 |
| Text-to-Figure | 0.0901 | 0.116 | 0.1063 | 0.1013 | 0.0814 | 0.108 | 0.0839 | 0.0602 | 0.084 |

| PCME | (a) Length of Spoken Language | | | | | (b) Number of Subwords | | | |
|---|---|---|---|---|---|---|---|---|---|
| **r@10** | <100 | 100 - 200 | 200 - 400 | 400 - 600 | 600+ | <10 | 10 - 20 | 20-30 | 30 - 50 |
| Figure-to-Text | 0.0342 | 0.1301 | 0.082 | 0.0733 | 0.053 | 0.0744 | 0.0556 | 0.0617 | 0.0271 |
| Text-to-Figure | 0.0342 | 0.1301 | 0.082 | 0.0752 | 0.0536 | 0.076 | 0.0518 | 0.0603 | 0.0309 |

Table 9: For all competitive baselines, performance (a) peaks at 100–200 words then drops with increasing length of spoken language and (b) drops with increasing number of subwords

## K. Qualitative Cases of Failure

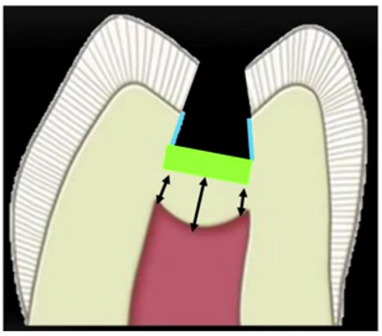
| Source Image | Ground Truth Text | Retrieved Text |
|---|---|---|
|   | alright so another thing we like to do in both surgical extraction and after a simple extraction is irrigation so here is a mono jet syringe and we can load this up with sterile saline in order to clean out the extraction site so we use a steady stream of sterile saline or water during bone removal for a surgical extraction and it prevents heat generation from the spinning bird that can damage bone it also increases the efficiency of the surgical bone of the surgical birth and is because it washes away the chips of bone and provides lubrication for the surgical bird so like i mentioned before you want to irrigate during surgical removal of bone and at the completion of any extraction to flush the socket of any debris a infectious or inflamed tissue you can also give a patient this to take home with them for gentle warm salt water rinses and flushing out of the healing socket at home i think that can be a very useful tool to use at for home care | all right and so those were the hand instruments we talked about sickle scalars and curettes and now we can talk about ultrasonic scalars which are used for tenacious calculus is calculus that can be harder to remove now you can also use hand instruments to get to get tenacious calculus off that's for sure but for the board exam just remember that they like to test that ultrasonic scalars can specifically be used for some tough to get calculus there are contraindicated for patients with pacemakers infectious diseases spread by aerosol and at risk for respiratory disease that's because they spit out a lot of water and in terms of the the pacemaker here electronic dental instruments like ultrasonic scalars and also apex locator has used in endodontics could potentially interfere with pacemakers because they use electrical impulses to maintain proper heart rhythm so that's just an important thing to note there are two different main types of ultrasonics there is a magnetic stricte vulture sonic |
|  | liner and then a base can be used for metal restorations and when liner is used so it goes over the liner in order to protect it from being resorbed and washed out the base also provides thermal protection and can distribute local stress across all the underlying dentin resin-modified glass ionomer is what this stands for is frequently used as a base pitch or bond is our example here | could result in congenitally missing or supernumerary teeth if it occurs early on but more likely you're going to see a cyst oh don toma gemination or fusion or dens and dente depending on the amount of cell differentiation and that has occurred |

Figure 10: Figure-to-Text: Failure Case for dental (top-1 retrieval result shown on right)

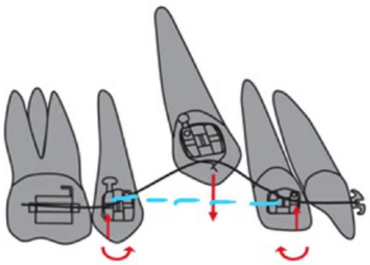
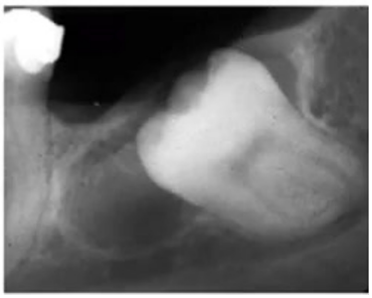
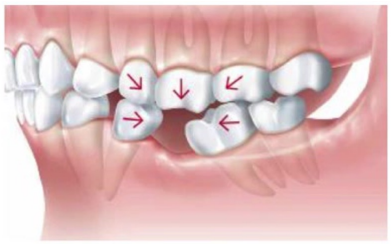
| Source Text | Ground Truth Image | Retrieved Image |
| --- | --- | --- |
| so there are really two phases of the orthodontic wire and that's activation which is also known as loading and that refers to the amount of force applied to engage the wire into the brackets lat so it's engaging the wire into the brackets lat and then tying it into place deactivation or unloading is letting the wire return back to its original shape and it's the amount of force that the wire applies to the tooth to get back to its original shape so based on how the wire is deflected when it's tied into these brackets it will apply a force to them because the wire wants to return back to its original shape because of its inherent elasticity so in this example the wire originally started out as a straight line a straight horizontal line and so when we deflected up here we can't quite get it into the brackets |  |  |
| apex ford integer assist now we have the radial lucency which is attached to the cej or the cemento and will junction where the enamel meets the cementum of the root and you can see how this radiolucency comes neatly attached to that point of the tooth now it's most common with canines and third molars and here this looks like it could be a second molar or third molar let's just say this is in fact a wisdom tooth and it's an accumulation of fluid between the crown and the reduced enamel epithelium which if you notice |  |  |

Figure 11: Text-to-Figure: Failure Cases for dental (top-1 retrieval result shown on right)

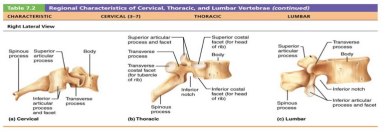
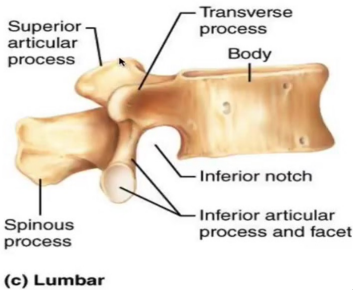
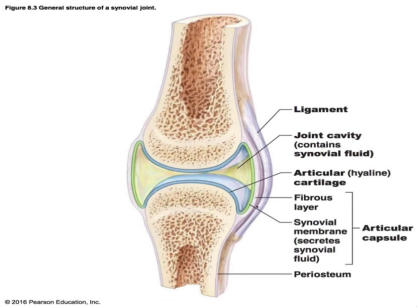
| Source Figure | Ground Truth Text | Retrieved Text |
|---|---|---|
|  | types again here's a side view notice the spine and saw the spinous processes it's much larger thicker inner thigh just for a lot of these muscles these larger muscles to come and test you same thing with the transverse process very different than what we had over here there's no ribs over here over here so this is the inferior articular process infested over here that you see here is a superior articular process over there you can see the faceting fortunately | so here you go everything you see in the midline of the body this is your axial skeleton so all your limbs they make up your appendicular skeleton so if you look over here you have your skull okay and then you have your your ribcage right this is your thoracic cage also is called and the thoracic cage again is made up of your customers this is ribs the sternum and the backbones all right they also cause posteriorly the castles the articulate with the um t1 through t12 your vertebral column the bones of the vertebral column so yeah now we're going to be looking at all this stuff in detail as we move through the course to this chapter actually |
|  | types again here's a side view notice the spine and saw the spinous processes it's much larger thicker inner thigh just for a lot of these muscles these larger muscles to come and test you same thing with the transverse process very different than what we had over here there's no ribs over here over here so this is the inferior articular process infested over here that you see here is a superior articular process over there you can see the faceting | let's move further you go you can see again in this picture then anterior longitudinal ligament very clearly notice it's a very broad it's a large strong ligament over here and comparatively the posterior longitudinal ligament much smaller so it's much weaker this is strong one this is the weak one and yes you can see the this is the body of the vertebrae the transverse processes the vertebrae over here |
|  | fluid so in this illustration we see a synovial joint so here are the two articulating bones and you can see that the articular bones there are lined with this articular cartilage at the articulating surface now we see this fibrous layer here that the outer fibrous layer and then we have this inner synovial membrane that forms the articular capsule the spaces between within this capsule it constitutes the joint cavity and is within this joint cavity we will find the synovial fluid we also see this outer get over here not this would be an example of this capsular ligaments and this is a it thickens the the fibrous layer to reinforce and straighten this joint | women so when you look over here this is scoliosis notice how you have this axis you have a lateral rotation of your thoracic spine over here too let's push more towards this side to ensure that the left side and it is towards the right and then you look over here this is this kyphosis this is over here so this the hunchback then if you look over here in lordosis again you have this access anterior curvature of the your lumbar spine so again over here you seeing this in the lumbar spine kyphosis generally speaking you'll tend to see in the the thoracic region thoracic spine in this of scoliosis also again commonly again you can see |

Figure 12: Figure-to-Text: Failure Case for Anat-1 (top-1 retrieval result shown on right)

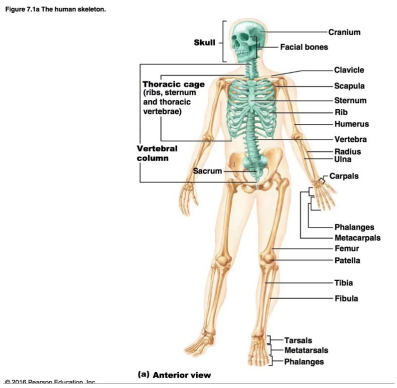
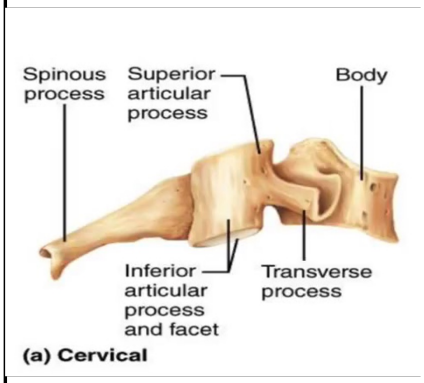
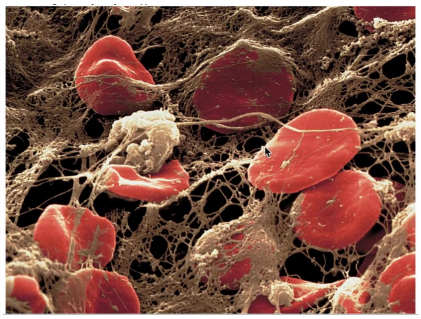
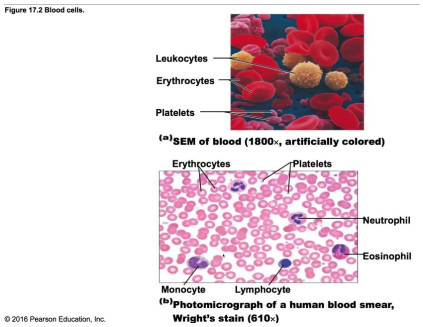
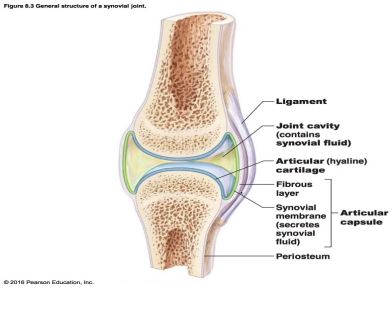
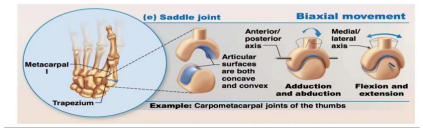
| Source Text | Ground Truth Image | Retrieved Image |
| --- | --- | --- |
| so here you go everything you see in the midline of the body this is your axial skeleton so all your limbs they make up your appendicular skeleton so if you look over here you have your skull okay and then you have your your ribcage right this is your thoracic cage also is called and the thoracic cage again is made up of your customers this is ribs the sternum and the backbones all right they also cause posteriorly the castles the articulate with the um t1 through t12 your vertebral column the bones of the vertebral column so yeah now we're going to be looking at all this stuff in detail as we move through the course to this chapter actually |  |  |
| formed the side here we're looking at a scanning electron micrograph and you can see this fibermesh that's over here and within this fiber mesh you can see all the red blood cells that are trapped |  |  |
| fluid so in this illustration we see a synovial joint so here are the two articulating bones and you can see that the articular bones there are lined with this articular cartilage at the articulating surface now we see this fibrous layer here that the outer fibrous layer and then we have this inner synovial membrane that forms the articular capsule the spaces between within this capsule it constitutes the joint cavity and is within this joint cavity we will find the synovial fluid we also see this outer get over here not this would be an example of this capsular ligaments and this is a it thickens the the fibrous layer to reinforce and straighten this joint |  |  |

Figure 13: Text-to-Figure: Failure Cases for Anat-1 (top-1 retrieval result shown on right)

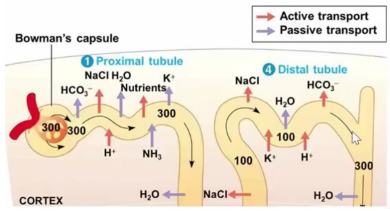
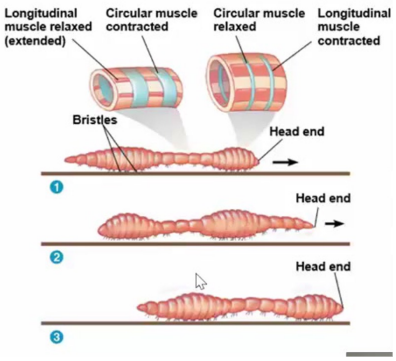
| Source Figure | Ground Truth Text | Retrieved Text |
|---|---|---|
|  | process now we get up to the distal tube which is again next to the proximal tube and now you see more salt being pulled out more water being pulled out and then you get potassium hydrogen and proton ions in this case and again concentrations regulates those concentrations and then the ions contributes to the ph regulation in this case so that it stays neutral ph | that that an issue so like i said angiotensin raises the blood pressure and decreases the blood flow in the capillaries in the kidneys and then like i said if you have chronic high blood pressure some of these drugs that we take actually will rot block this a a 2 or this angiotensin ii and because of that it will cause the kidneys to lower the amount of water tension and then lower the blood pressure because again more water less water in the blood the blood pressure goes down and so again that gets it warranted dehydration state and so those that are on hypertension a lot of times will then complain that they get thirsty a lot because of that and same with diabetes because you're getting more sugar in the blood and that stuff and that's a whole nother another situation |
|  | in detail your how this all works now with the hydrostatic skeleton this is fluid under the pressure of a close body compartment allows the muscles to contract and extend and so an essentially this works kind of like how her esophagus does with peristalsis and so again the use these rhythmic contractions using the fluid that allows them to extend and contract the muscle and allows it to move through the soil and so this is the main type of skeleton that most nigerians flatworms nematodes and annelids use to move through either water or soil in these situations and again under the hydrostatic skeleton | glands in these situations now most of these systems are based on the feedback loops in that and so you'll see that again typically what you see is either get a positive feedback or negative feedback so you either turn something on or shut something down depending on whether it is a response is needed and so the example that i have here is the negative feedback with calcium and so again typically in homeostasis you have normal calcium in the blood but if the calcium levels get too high your thyroid will release the sinkhole calcitonin and that will cause an increase of calcium into the bone which causes again a decrease in the amount of calcium in the intestines |

Figure 14: Figure-to-Text: Failure Case for Bio-1 (top-1 retrieval result shown on right)

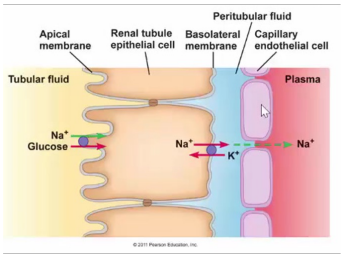
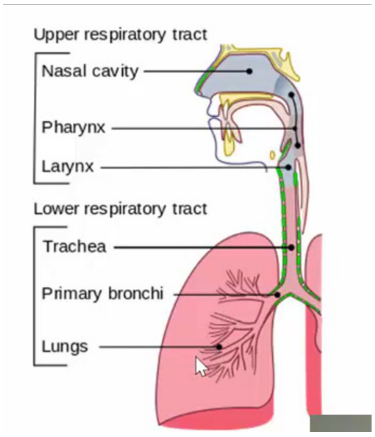
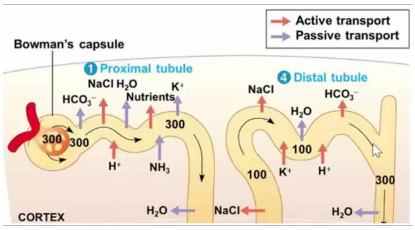
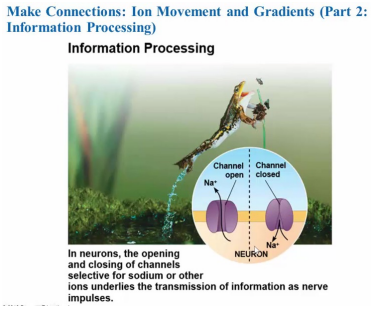
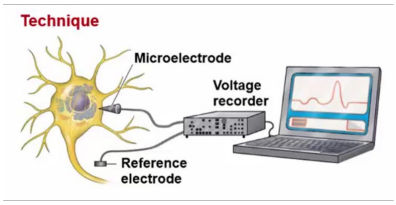
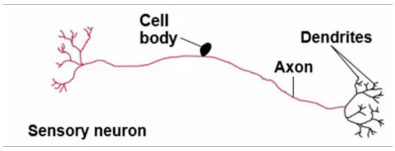
| Source Text | Ground Truth Figure | Retrieved Figure |
|---|---|---|
| now in most animals osmoregulation in metabolic waste disposal rely on transport epithelia and so again these are cells that help move things across membranes and so that's going to be the key step so when we talk about blood and capillaries again you're going to have these solutes going across the movement along with water and that's where you're going to see this and again these cells are specialized removing solutes across in controlled amounts across specific things and they have carrier protein… |  |  |
| process now we get up to the distal tube which is again next to the proximal tube and now you see more salt being pulled out more water being pulled out and then you get potassium hydrogen and proton ions in this case and again concentrations regulates those concentrations and then the ions contributes to the ph regulation in this case so that it stays neutral |  |  |
| place where you see koreans playing a big role so how are action potentials generated so we can measure the action potential by again using this the system and so you have this microelectrode attached to the neuron and we can measure with the reference electrode what is going on inside the inside the neuron now changes in the membrane potential occur because the neurons contain gated ion channels either open and closed due to stimuli and the voltage-gated ion channel opens and closes in response to the shift of the voltage across the plasma membrane… |  |  |

Figure 15: Text-to-Figure: Failure Cases for Bio-1 (top-1 retrieval result shown on right)

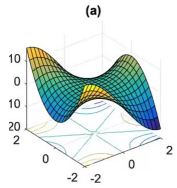
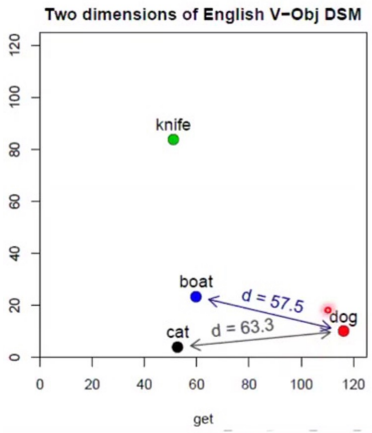
| Source Figure | Ground Truth Text | Retrieved Text |
|---|---|---|
|   | these saddle point i'm not giving as much in this kind of on purpose because now these days there are a lot of these that have been in a lot of districts have been included in these different libraries so you can | and beyond vqa there's a bunch of other data sets most of them similar in vain here's a reference you can put more of them if you're interested is there are image-based qa and |
|  | so let's summarize we've got two mvps we've seen how an agent with taking actions and going into different states and environment and observing rewards is a very general framework that i capsule has many of the real world reasoning tasks we know that we've seen that you know the goal of the agent is to maximize its cumulative reward x a discount factor and you want to learn the best policy organized over all possible policies that best maximizes your primitive reward each of these policies will define a particular distribution over actions that the agent should take from different states so here are the here's the main goal of reinforcement learning | so given all these real-world tasks we've tried to group our best we tried our best to group these applications into seven groups these are by no means exhaustive and this is by no means a perfect categorization but we like to share some of this categorization with you so that helps you localize what area you want to work on for your research research project and also the datasets that are within this area the first one is that on effective recognition affective computing building these computers are able to understand these human-centric behaviors like emotion and sentiment media description for image and video captioning multimodal qa which for the localizers need a description by… |
|   | distance between two values to these position is not as informative as the direction of it what it means the fact that dog is here here instead of here here here here are here the only reason it is at this position is that it happened more often in or coppers it happened hundred to eighteen times here and maybe 10 times here what is important is more the ratio between them is that the most respond the fact that it happened really often is not as much of a like a good descriptor of meaning of a word but the ratio between use and get is more useful so the angle and so like the distant itself just mean that dog was used more often the word in that corpus what | that and if you were to do on supervised i didn't put a slide for his but they are some very well-established algorithms that you can do to also study how to learn and put some kind of superb vision on the z so that eventually you keep the similarity as close to each other and there's quite a few of these note tuvok being one of them |

Figure 16: Figure-to-Text: Failure Case for ml-1 (top-1 retrieval result shown on right)

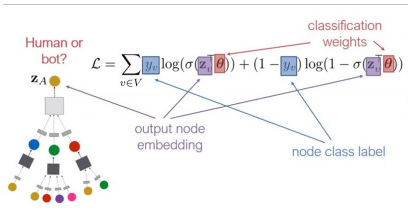
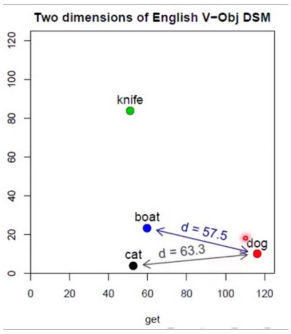
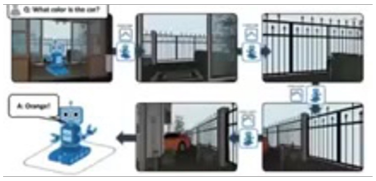
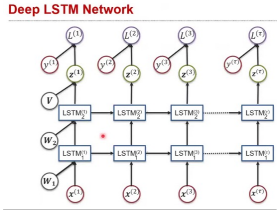
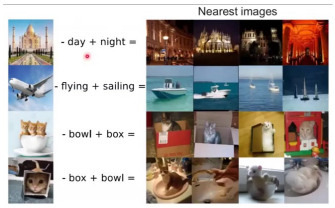
| Source Text | Ground Truth Figure | Retrieved Figure |
| --- | --- | --- |
| that and if you were to do on supervised i didn't put a slide for his but they are some very well-established algorithms that you can do to also study how to learn and put some kind of superb vision on the z so that eventually you keep the similarity as close to each other and there's quite a few of these note tuvok being one of them |  |  |
| so given all these real-world tasks we've tried to group our best we tried our best to group these applications into seven groups these are by no means exhaustive and this is by no means a perfect categorization but we like to share some of this categorization with you so that helps you localize what area you want to work on for your research research project and also the datasets that are within this area the first one is that on effective recognition affective computing building these computers are able to understand these human-centric behaviors like emotion and sentiment media description for image and video captioning multimodal qa which for the localizers need a description by only answering a question by only providing an answer about a specific question target to one specific area of the image multimodal navigation which really combines these aspects of reinforcement learning and robotics with understanding language and vision |  |  |
| bi-directional you can do bi-directional and multiple layers of bi-directional elmo had two of them |  |  |

Figure 17: Text-to-Figure: Failure Cases for ml-1 (top-1 retrieval result shown on right)

| Source Figure | Ground Truth Text | Retrieved Text |
|---|---|---|
|  | for most people once they get married one of the next things that they look for to is having babies or starting their own families we know that in today's western societies couples are having fewer and fewer children in part that's because of economics and you would think that oh it's because it's more expensive actually it's the opposite it's the fact that they are working and so less children is easier to take care of but also the fact that basically our children are living when we had high death rates of children you would have multiple children in hopes of getting a few to adulthood but today we pretty much figure our children are going to make it to adulthood we don't sit there and think as some people used to have to worry about is that 30 to 40 to 50% of their children weren't going to make it to adulthood so 50% of my children are going to make it to adulthood | when we know one of the biggest things that happening in adolescent is this change in physical maturation basically our body's going to go through this process of becoming from the child body to the adult body which basically means that we're going to start creating sexually mature bodies now what is it you're saying is that when we look at the bodies of young people the first thing so grows are like their hands their heads and their feet that's why we sometimes like a boys and they got these big old kwame feet the next bones are grow are going to be your tubular bones and then finally your trunk |
|  | by the time they're to the child's growth is really slow down and quite often they become much more finicky about what they want to eat the good thing is is that they still get all their nutrition generally mostly what people will say to you if you're talking to some of the experts we just tried to have the widest variety of diet that you can when the child is under 2 because the more they're introduced to new foods and other items the more likely they'll be eating it when they're too so if they peas one there's probably a good chance they're going to eat peas a to the problem is that we tend to maybe not feed the best food when they're younger | our satisfaction with a job is absolutely directly related to our age meaning is what we're looking for from the job will change as we grow older this chapter is mostly talking about middle-aged worker so let's kind of look at that as a specific and one thing is that in middle age as we said you're looking for a job for union with others and to help you within sort of your life as you're having it at the moment so this intrinsic reward becomes much more important than the extrinsic rewards and intrinsic means inside of yourself are you satisfied with what you're doing does you find it interesting are you challenged enough or challenge too much you might want to say so as we get older we do want certain things to help us with having a satisfactory life and that includes our job so as you begin to have children you might want to have a more flexible lifestyle |

Figure 18: Figure-to-Text: Failure Case for Psy-2 (top-1 retrieval result shown on right)

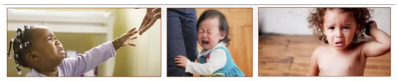
| Source Text | Ground Truth Figure | Retrieved Figure |
| --- | --- | --- |
| theory of attachment suggests that children come the world biologically programmed to form attachment with others this is because it helps them survive those who seem to be attached to somebody we're more likely to receive comfort and protection and they absolutely have shown that they're more likely to survive into adulthood this is not just seen in humans but this is also seen in the great apes scecina most of the mammals so when we have an attachment it helps our survival so infants produce this sort of innate social release of yer such as crying and smiling that sort of stimulates is an 8 caregiver response from us as adults now by the way the reason i'm saying caregiver is that it isn't always the mother i don't want this thought to be in your head that had has to be the mother this can be the father this can be the grandparent this can be an adopted person it doesn't have to be a biologically the mother so we use the word caregiver to be a more general type of scenario |  |  |
| for most people once they get married one of the next things that they look for to is having babies or starting their own families we know that in today's western societies couples are having fewer and fewer children in part that's because of economics and you would think that oh it's because it's more expensive actually it's the opposite it's the fact that they are working and so less children is easier to take care of but also the fact that basically our children are living when we had high death rates of children you would have multiple children in hopes of getting a few to adulthood… |  |  |

Figure 19: Text-to-Figure: Failure Cases for Psy-2 (top-1 retrieval result shown on right)

## L. Analysis on K Instance Representations

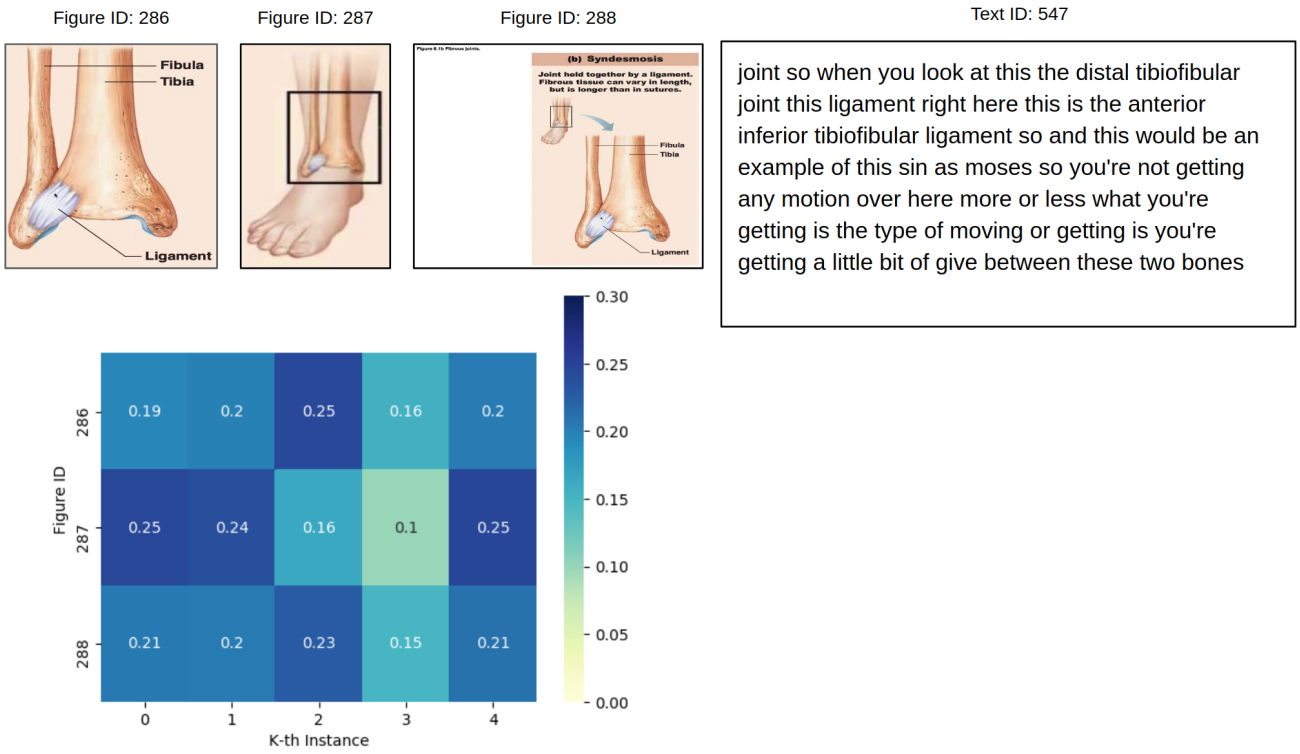



Figure 20: Successful Retrieval Case for Anat-1: The aligned figures and spoken language, with their corresponding IDs, are shown on top. A heatmap displaying the similarity scores of all K-instances of spoken language for each figure is shown on the bottom. In successful cases, the distribution of the similarity scores of K instances differ for each figure, potentially hinting that the representation has captured the many-to-one mapping between figures and spoken language.

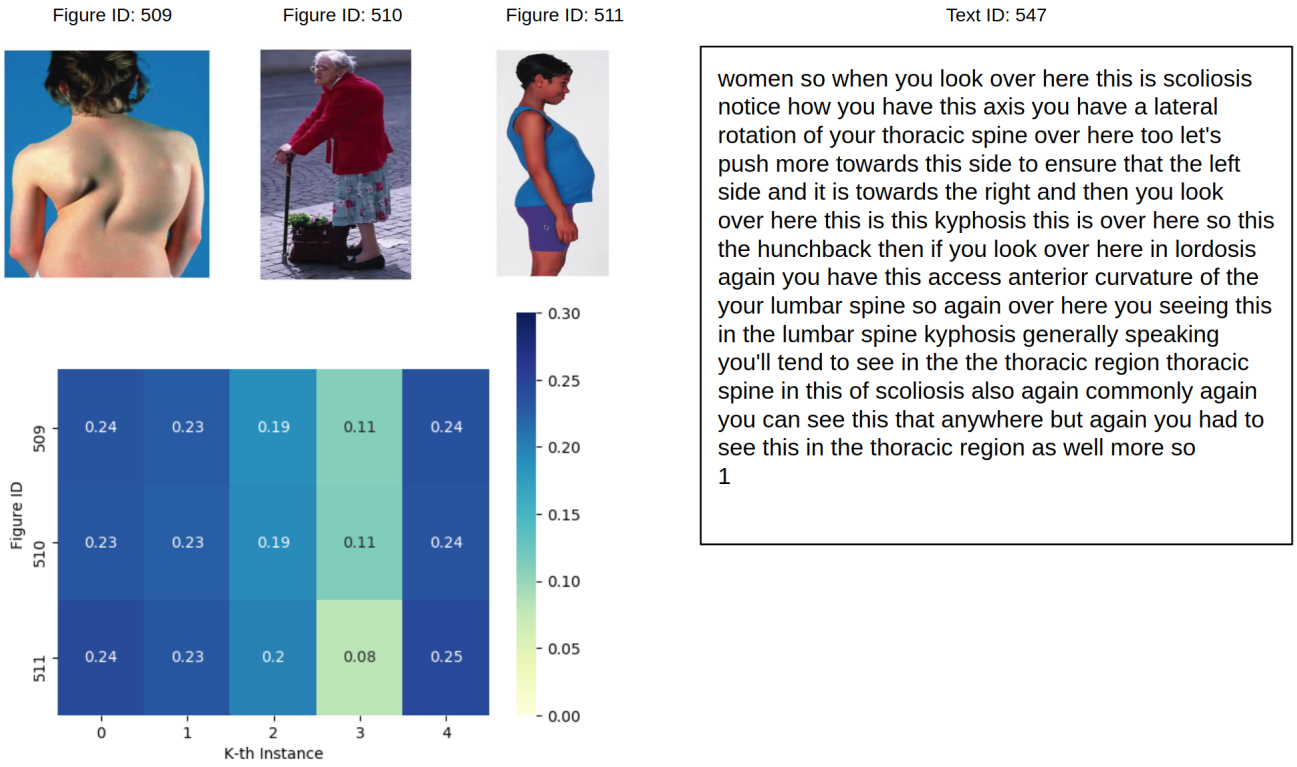



Figure 21: Successful Retrieval Case for Anat-1: he aligned figures and spoken language, with their corresponding IDs, are shown on top. A heatmap displaying the similarity scores of all K-instances of spoken language for each figure is shown on the bottom. In successful cases, the distribution of the similarity scores of K instances differ for each figure, potentially hinting that the representation has captured the many-to-one mapping between figures and spoken language.

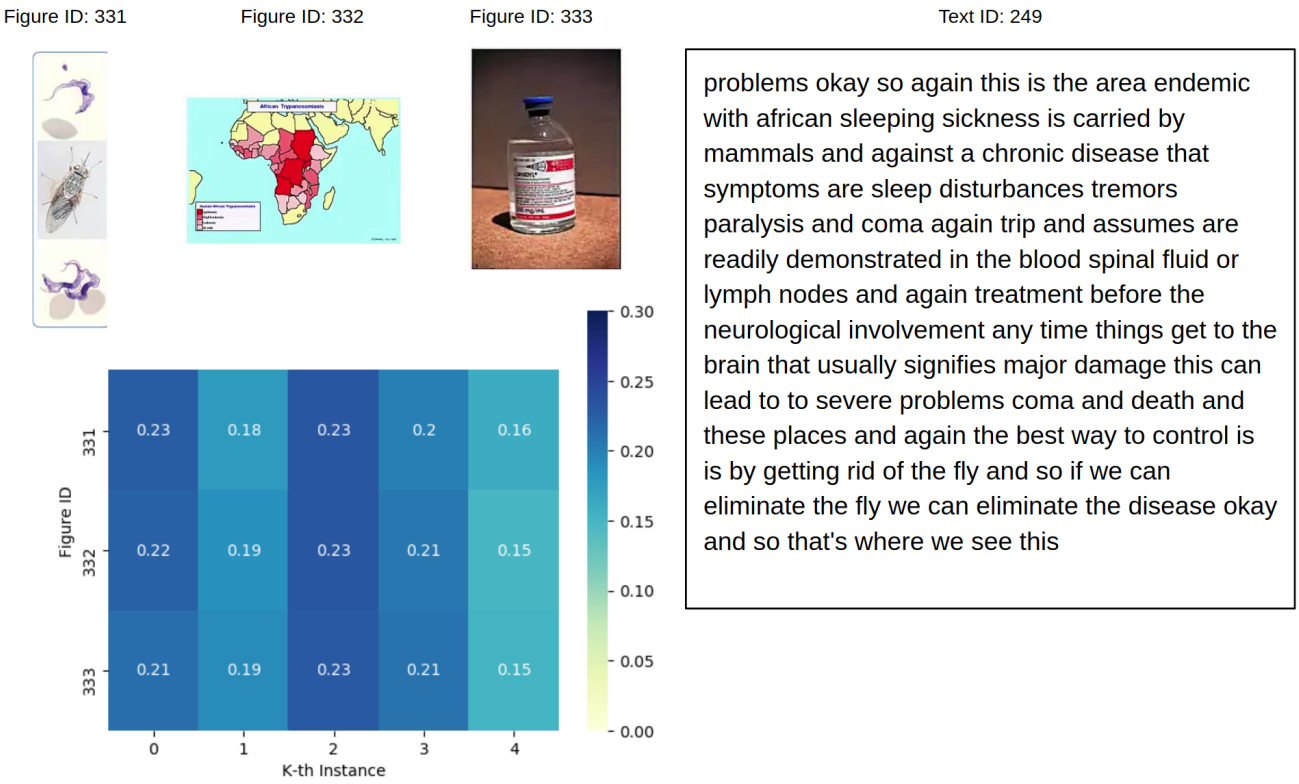



Figure 22: Successful Retrieval Case for Bio-1: he aligned figures and spoken language, with their corresponding IDs, are shown on top. A heatmap displaying the similarity scores of all K-instances of spoken language for each figure is shown on the bottom. In successful cases, the distribution of the similarity scores of K instances differ for each figure, potentially hinting that the representation has captured the many-to-one mapping between figures and spoken language.

Figure ID: 555

Figure ID: 556

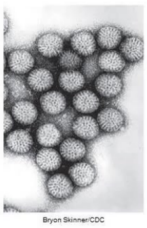
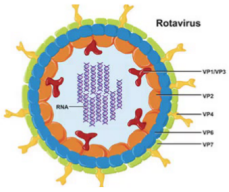
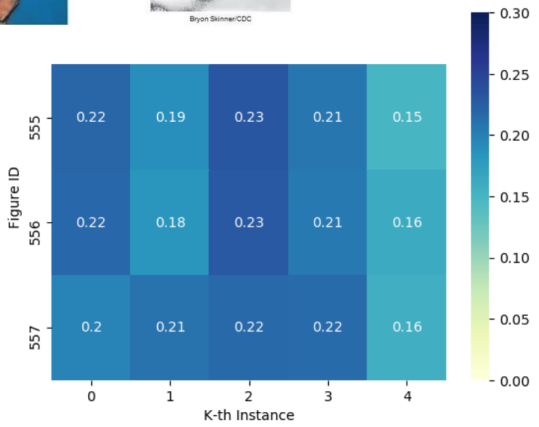


Figure ID: 557




Text ID: 385

little kids so again these are the non envelopes single are segmented double-stranded rna viruses the rio viruses are usually unusual double stranded rna and the two are the rio and roto virus has which lead to infantile diarrhea which leads to severe dehydration and so that's where we worry about these again both of these guys can cause either cold or oral fecal transmission in these cases where you get again mortality and morbidity and infants due to this severe dehydration that's associated with these viruses in these cases so again with the real virus caused an upper respiratory infection like a cold and then also diarrhea that goes along with that and so that's what you see with there and these are very common the ones same ones that you find on cruise ships that again people will then have these massive outbreaks of diarrhea and cruise ships that you hear about where then everyone has to sit and there are in their cabin for three days three to four days because they had a diarrhea outbreak in so you probably in those situations can't leave your cabin anyways because you're so sick and and that and have constant diarrhea so it's probably not a fun time to get any one of these viruses so again we'll leave it at there but typically more problems smaller kids due to the dehydration 0

Figure 23: Successful Retrieval Case for Bio-1: The aligned figures and spoken language, with their corresponding IDs, are shown on top. A heatmap displaying the similarity scores of all K-instances of spoken language for each figure is shown on the bottom. In successful cases, the distribution of the similarity scores of K instances differ for each figure, potentially hinting that the representation has captured the many-to-one mapping between figures and spoken language.

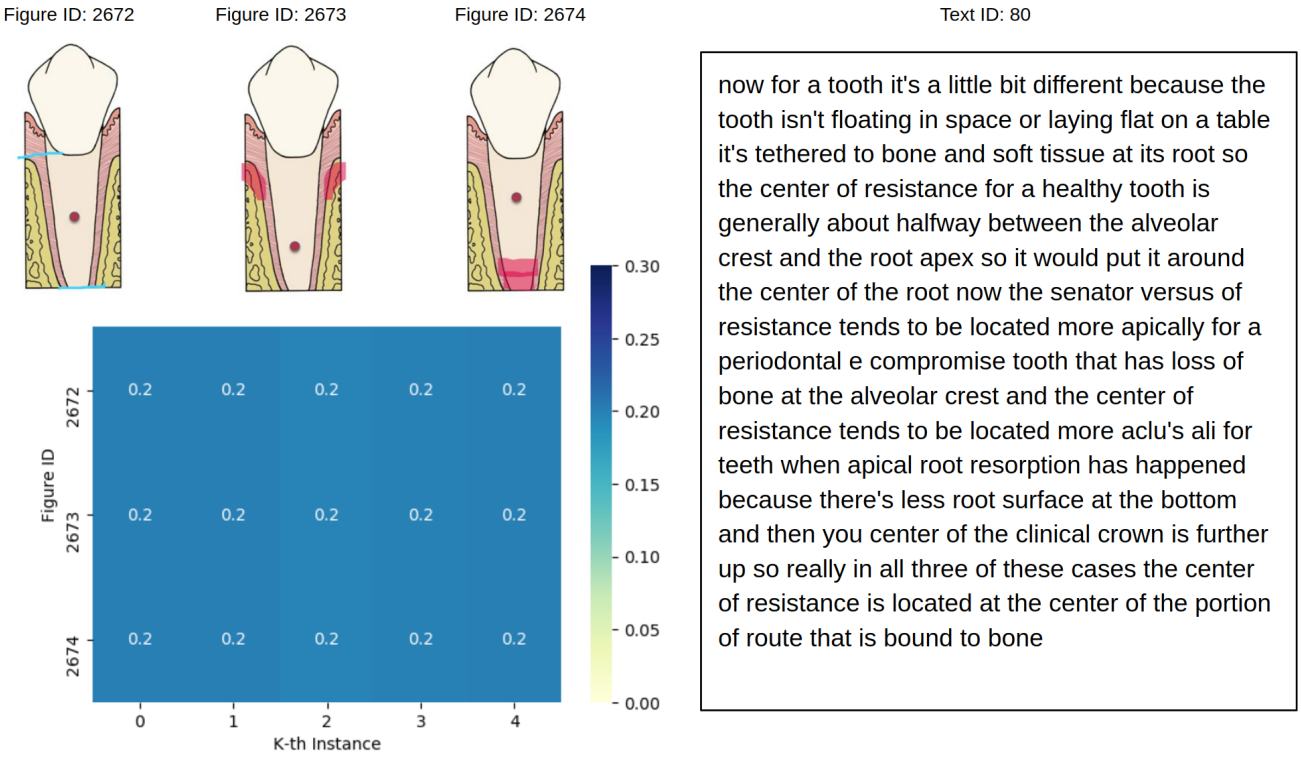



Figure 24: Failed Retrieval Case for Dental: The aligned figures and spoken language, with their corresponding IDs, are shown on top. A heatmap displaying the similarity scores of all K-instances of spoken language for each figure is shown on the bottom. In failed cases, the distribution of the similarity scores of K instances are identical for each figure, hinting that the separate K instances has not learned to disambiguate the different figures. In this specific case, we see that the images look very similar, which could be a reason for failure.

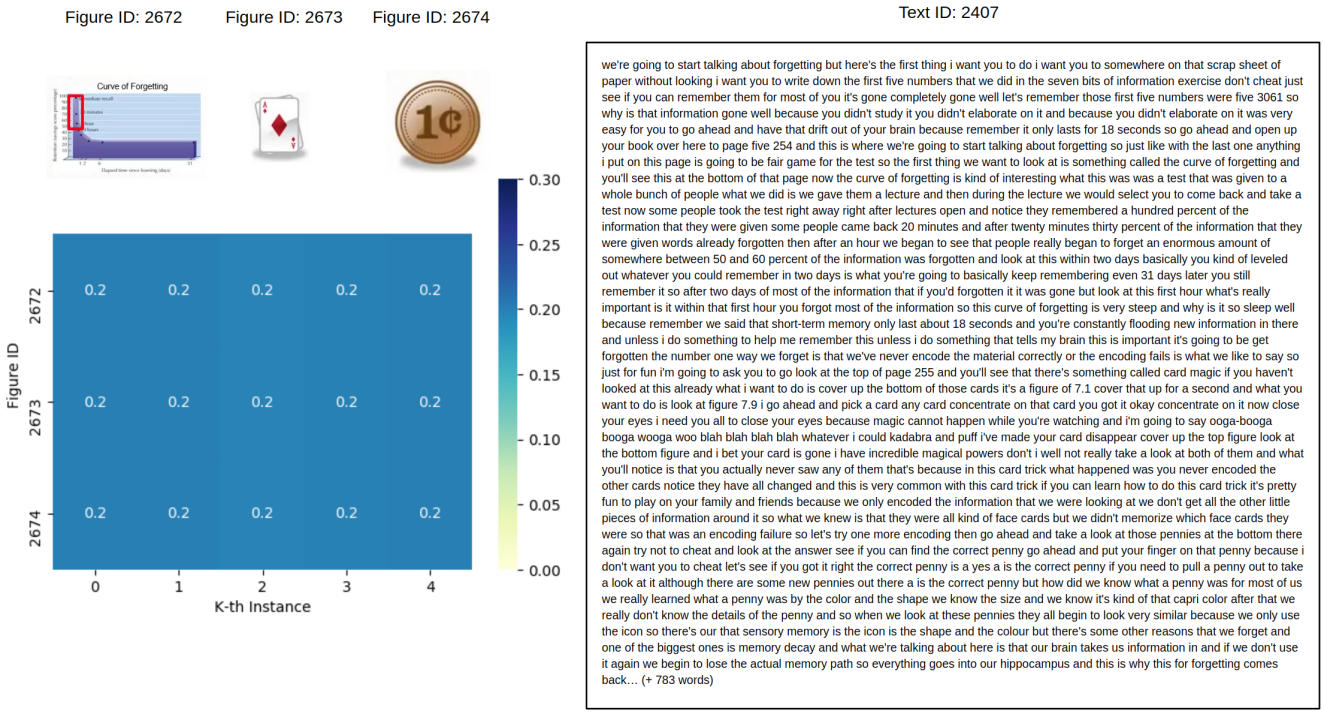



Figure 25: Failed Retrieval Case for Psy-1: The aligned figures and spoken language, with their corresponding IDs, are shown on top. A heatmap displaying the similarity scores of all K-instances of spoken language for each figure is shown on the bottom. In failed cases, the distribution of the similarity scores of K instances are identical for each figure, hinting that the separate K instances has not learned to disambiguate the different figures. In this specific case, the spoken language phrase is extremely long, which current baselines are unable to handle.

Figure ID: 1159

$$\pi_{[k+1]}(a|s) = \arg\max_a \sum_{s'} p(s'|s,a)\left[r(s,a,s') + \gamma V^{\pi}_{[k]}(s')\right]$$

Figure ID: 1160

$$\pi_{[k+1]}(a|s) = \arg\max \sum p(s'|s,a)\left[r(s,a,s') + \gamma V^{\pi}_{[k]}(s')\right]$$

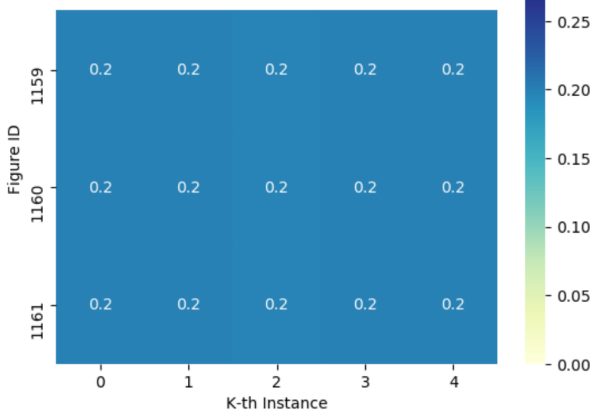



Figure ID: 1159

Text ID: 798

so that gives you a list these policies values for a given policy pie and then i'm also try to improve my policy because my policy initially is also like a random policy right so a random policy i'm going to iterate such a left and right hand sides are equal to get my optimal state values for random policy but my goal is to also improve my policy so i'm going to take my policy at time k i'm going to try to obtain a better policy at time k plus 1 and i'm going to obtain this better policy by exactly using the estimates of my state value functions at time k so we call this equation was how we went from given these your state values to actually your best policies using a one-step look ahead basically narc max over our actions and with each action evaluated based on how good it is under your value functions so this is a unit iterative algorithm your three variables here again your policies which are iterating okay of the k + 1 and for each set of policies pie you're also going to to compute your state value functions and these two steps are known as policy evaluation so evaluating how good a policy is so given your pies evaluate your vis and policy improvement so given a good estimate of your visas at time k try to obtain a better policy by k plus 1 and you're going to repeat these two steps until you converge and at least in this case where all of these are observable your transitions and all and summations are you can do them exactly this is actually guaranteed to converge to your optimal policy and this algorithm is known as policy duration

Figure 26: Failed Retrieval Case for Ml-1: The aligned figures and spoken language, with their corresponding IDs, are shown on top. A heatmap displaying the similarity scores of all K-instances of spoken language for each figure is shown on the bottom. In failed cases, the distribution of the similarity scores of K instances are identical for each figure, hinting that the separate K instances has not learned to disambiguate the different figures. In this specific case, the given figures are highly technical mathematical equations.

Figure ID: 81

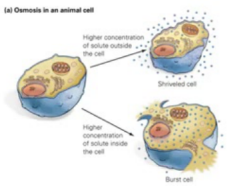



Figure ID: 82

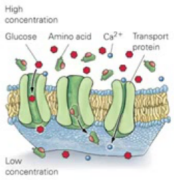
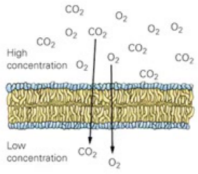



Figure ID: 83




Text ID: 71

so once again the three types of passive transport diffusion facilitated diffusion and osmosis remember none of these taking the extra energy in the form of atp and all three of them the substances move from higher to lower concentration

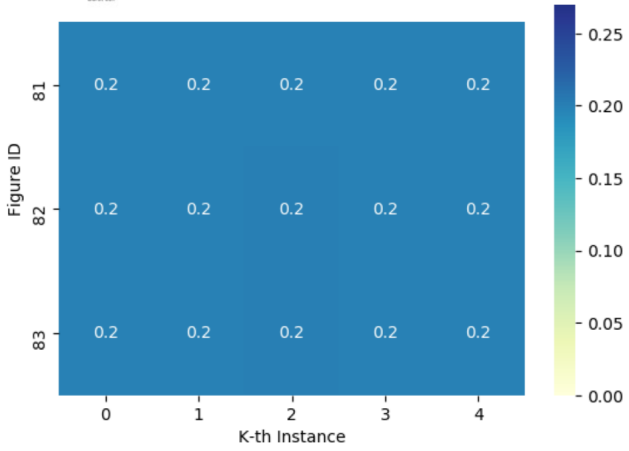



Figure 27: Failed Retrieval Case for bio-3: The aligned figures and spoken language, with their corresponding IDs, are shown on top. A heatmap displaying the similarity scores of all K-instances of spoken language for each figure is shown on the bottom. In failed cases, the distribution of the similarity scores of K instances are identical for each figure, hinting that the separate K instances has not learned to disambiguate the different figures. In this specific case, the given figures are complex diagrams.