Lecture Presentations Multimodal Dataset: Towards Understanding Multimodality in Educational Videos

Dong Won Lee[1] Chaitanya Ahuja[2] Paul Pu Liang[2] Sanika Natu[2] Louis-Philippe Morency[2]
MIT[1], Carnegie Mellon University[2]
https://github.com/dondongwon/LPMDataset

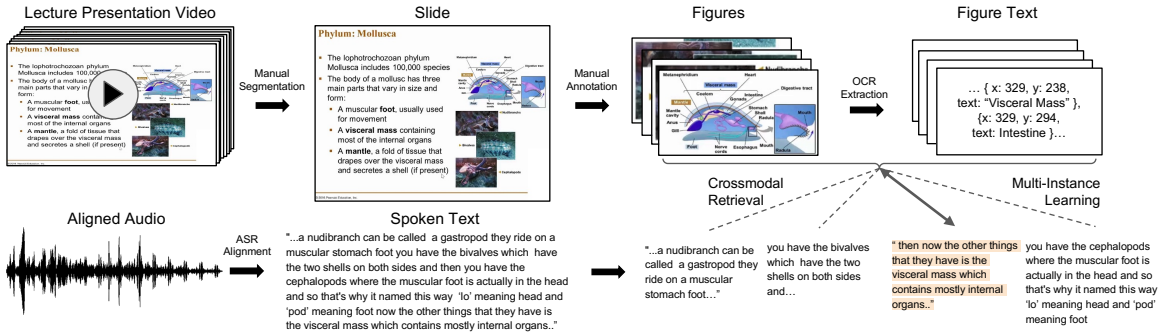


Figure 1: We present the Lecture Presentations Multimodal Dataset as a benchmark for developing AI technologies that can understand multimodal knowledge in educational content. Our diversely sourced and richly annotated dataset contributes three challenging vision-and-language research tasks: automatic retrieval of (1) spoken explanations given figures, (2) illustrative figures given spoken explanations and (3) generation of slide explanations. Through benchmarking existing and newly proposed models, we outline future research directions to bring us closer to intelligent and accessible tutoring aids.

Abstract

Many educational videos use slide presentations, a sequence of visual pages that contain text and figures accompanied by spoken language, which are constructed and presented carefully in order to optimally transfer knowledge to students. Previous studies in multimedia and psychology attribute the effectiveness of lecture presentations to their multimodal nature. As a step toward developing vision-language models to aid in student learning as intelligent teacher assistants, we introduce the Lecture Presentations Multimodal (LPM) Dataset as a large-scale benchmark testing the capabilities of vision-and-language models in multimodal understanding of educational videos. Our dataset contains aligned slides and spoken language, for 180+ hours of video and 9000+ slides, with 10 lecturers from various subjects (e.g., computer science, dentistry, biology). We introduce three research tasks, (1) figure-to-text retrieval, (2) text-to-figure retrieval, and (3) generation of slide explanations, which are grounded in multimedia learning and psychology principles to test a vision-language model's understanding of multimodal content. We provide manual annotations to help implement these tasks and establish baselines on them. Comparing baselines and human student performances, we find that state-of-the-art vision-language models (zero-shot and fine-tuned) struggle in (1) weak crossmodal alignment between slides and spoken text, (2) learning novel visual mediums, (3) technical language, and (4) long-range sequences. We introduce PolyViLT, a novel multimodal transformer trained with a multi-instance learning loss that is more effective than current approaches for retrieval. We conclude by shedding light on the challenges and opportunities in multimodal understanding of educational presentation videos.

1. Introduction

Students today commonly learn through multimedia, including online lecture presentation recordings, educational mobile applications, and other digital resources [28]. In particular, slide-assisted instruction through lectures has become predominant in educational settings [38, 42, 43] and is widely considered by teachers and students as the preferred instructional tool [42, 47]. The effectiveness of lec-

For all authors, work was done at CMU

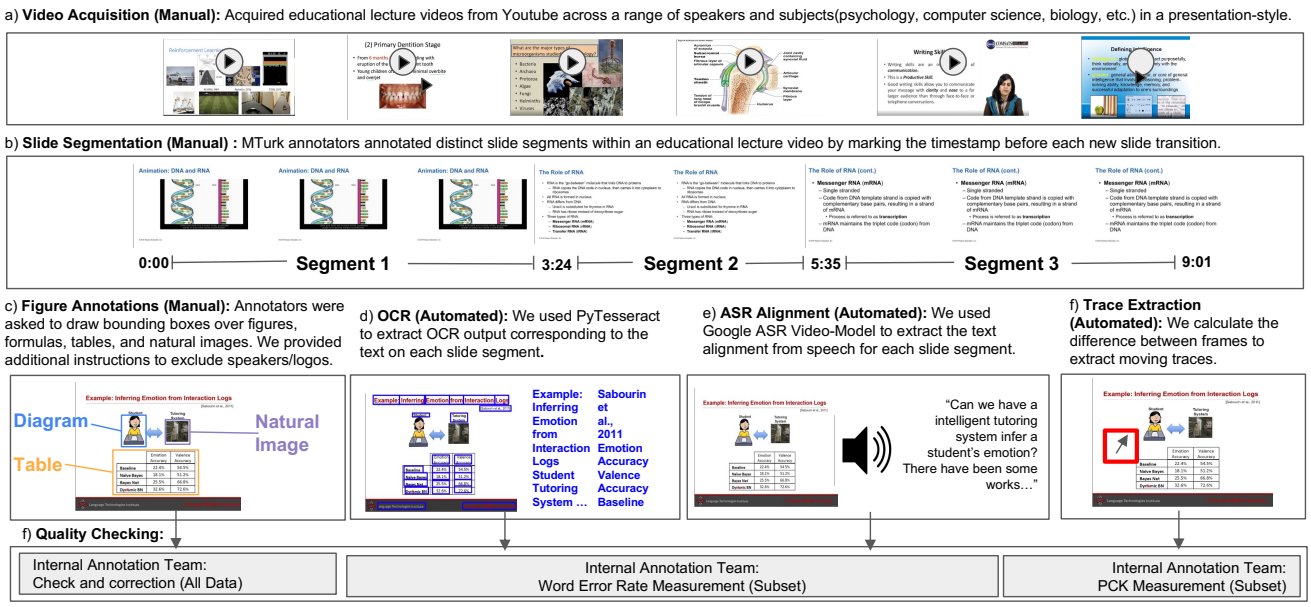


Figure 2: Overview of data collection and preprocessing with a summary of each step. Best viewed zoomed in and in color.

ture slides is supported by research in multimedia principles, which show that individuals learn more effectively from spoken (or written) language when accompanied by graphics rather than language in isolation [3, 29, 31, 33, 36]. The prevalence and effectiveness of lecture slides as an educational medium call for vision-and-language models that are also able to understand and communicate multimodal knowledge, in order to move closer towards intelligent teaching assistants [17].

We design the Lecture Presentations Multimodal Dataset (LPM Dataset) as a benchmark evaluating vision-and-language models' multimodal understanding of educational content. LPM Dataset contains over 9000 slides with natural images, diagrams, equations, tables and written text, aligned with the speaker's spoken language. These lecture slides are sourced from over 180 hours worth of educational videos in various disciplines such as anatomy, biology, psychology, speaking, dentistry, and machine learning. To benchmark the understanding of multimodal information in lecture slides, we introduce three research tasks of automatic retrieval of (1) spoken explanations for an educational figure (Figure-to-Text) and (2) illustrations to accompany a spoken explanation (Text-to-Figure) (3) generation of slide explanations. The tasks are strongly inspired by previous literature in multimedia learning [28, 51, 30] which state that meaningful learning takes place when one is able to organize verbal explanations (spoken words) combined with non-verbal knowledge representations (pictures) into a coherent mental model [32].

LPM Dataset and its tasks bring new vision and language research opportunities through the following technical challenges: (1) addressing weak crossmodal alignment between figures and spoken language (a figure on the slide is often related to only a portion of spoken language), (2) representing novel visual mediums of man-made figures (e.g., diagrams, tables, and equations), (3) understanding technical language, and (4) capturing interactions in long-range sequences. Through human and quantitative studies, we find that current multimodal models struggle with the aforementioned challenges. We work towards addressing weak alignment and novel visual mediums by introducing PolyViLT, a multimodal transformer trained with a multi-instance learning loss. Although PolyViLT presents some improvement, LPM Dataset still offers novel challenges that will spark future vision-and-language research in educational content modeling, multimodal reasoning, and question answering, thereby opening up pathways to exciting applications, such as an intelligent tutoring system that can utilize multimodal content to answer a student's question[19], a recommender system that automatically generates a slide on-the-fly as the speaker is speaking [46], or a evaluation system that provides feedback on the quality of the presentation [37].

2. Related Work

The effectiveness of lecture slides as a medium of transferring information can be attributed to five multimedia learning principles [15], which highlight the importance of multimodality. Firstly, the multiple representation principle states that individuals learn more effectively from graphics accompanied by spoken or written verbal information than solely spoken language. This principle is supported by

	Features				Size			Avail.
	Slide Segments	Slide Figures	Slide Text	Spoken Language	# Videos	# Hours	# Slides	
VLEngagment [4]					11568			✓
LectureBank [24]	✓(M)		✓(A)		1352		51,939	✓
ALV [14]	✓(A)			✓(A)			1498	✓
LectureVideoDB [12]	✓(M)		✓(M)		24		5000	✓
GoogleI/O [6]			✓(A)	✓(A)	209	174		✓
LaRochelle [48]	✓(A)		✓(A)	✓(A)	47	65	2350	
LPM Dataset (Ours)	✓(M)	✓(M)	✓(A)	✓(A)	334	187	9031	✓

Table 1: Comparison with existing lecture-based datasets, (A) represents automatic processing, (M) represents manual processing. LPM Dataset is the first of its kind to offer slide segmentation, aligned spoken language, slide text, and visual figures, while being publicly available.

dual-route processing mechanisms of working memory and comprehension processes, where integration of verbal and nonverbal information benefits formation of representations in working memory. [31, 29, 36, 3, 33]. Secondly, the contiguity principle expounds on reducing the spatio-temporal separation between different forms of information, which decreases the amount of effort required to build a coherent mental representation [5, 35]. Third, redundancy: the exposure to complementary but identical information in different modalities, improves learner's working memory (auditory, visual). Fourth, coherence: restricting the information presented to only essential information allows the learner to integrate key concepts and relationships [18]. Finally, structured signalling provides learner information regarding the overall hierarchical structure of the presentation [31].

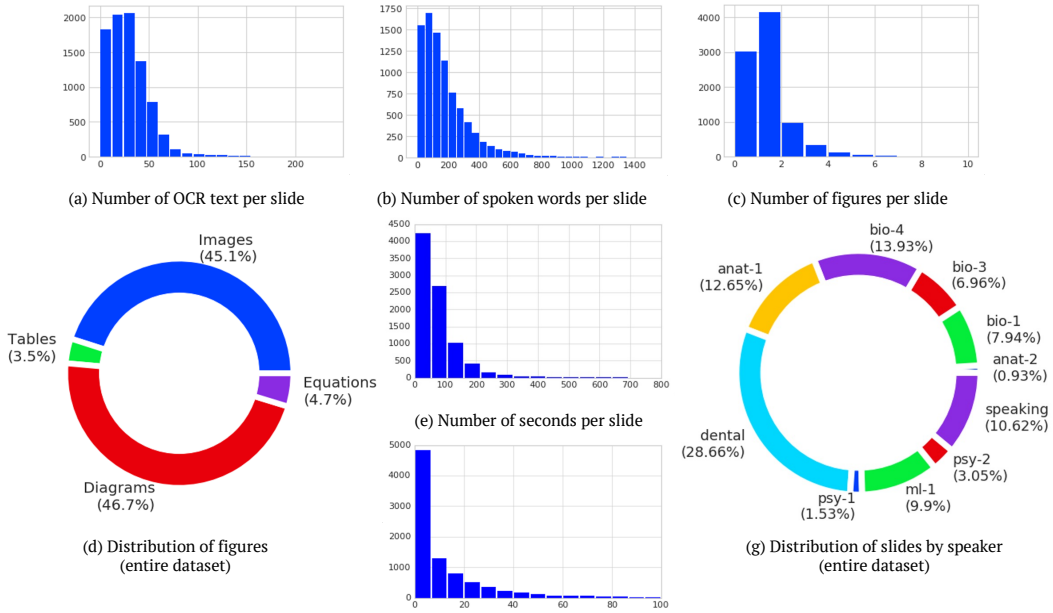
Given the effectiveness of lecture slides as a medium of presenting information, future vision-language models should be able to learn and extract from the rich information in lecture slides. LPM Dataset is the first to offer a large-scale dataset with aligned and complete modalities. We summarize and compare previous lecture slide datasets in Table 1. **LectureBank** [24] is a manually-collected dataset of lecture slides, consisting of 1352 online lecture PDF files from 60 courses in Computer Science. The dataset is annotated for each lecture's topic. It does not contain aligned transcripts and was primarily used to predict prerequisite relations for a given lecture slide. **ALV** [14] is a lecture video dataset of artificially-generated lectures, where transcripts from lectures are randomly split in fragments then assembled by stitching 20 randomly selected fragments from various videos. The resulting dataset only consists of transcripts. This work was developed for the purpose of evaluating lecture video fragmentation techniques. **VLEngagement** [4], is a dataset which was designed to study engagement in video lectures, where content-based (stop-word counts) and video-specific features (silence, video duration) are extracted from publicly available scientific video lectures. It only offers processed features, and does not contain raw language or pixel data. **LectureVideoDB** [12] is a dataset consisting of 5K frames of lecture videos, with annotated text characters developed for the purposed of text detection and recognition in Lecture Videos. No spoken language is provided. **GoogleI/O** [6] is a dataset consisting of 209 presentation videos from the Google I/O conferences between 2010-2012. It only offers textual information from the speech and the slides. The retrieval task is done at the video level, where entire transcripts are matched with all the text in a presentation. **LaRochelle** [48] contains 47 French lecture recordings at the university level and has been used to study video-level retrieval. In addition, the authors experiment with cross-modal retrieval where a bag of words approach is used for the text and visual tokens (image figures are not used). However, at this time, the data is not publicly available. **ScienceQA** [25] consists of multiple choice questions with aligned information from 261 lectures. However, the lectures are not sourced from video or educational slides, and are sourced from lectures purely based on written language, which do not offer any visual figures. **ChartQA** [27] presents 4.8K human-authored charts paired with 9.6K question-answer pairs to benchmark a model's question answering ability on man-made charts. Both ScienceQA and ChartQA offer different modalities and are incomparable to our dataset (hence, excluded from Table 1). To the best of our knowledge, LPM Dataset is the first of its kind to offer educational presentation videos with slide segmentation, aligned spoken language, slide text, and visual figures, while being publicly available for the research community.

3. Lecture Presentations Multimodal Dataset

The Lecture Presentations Multimodal Dataset is designed as a benchmark to develop vision-language models capable of understanding multimodal information present in lecture slides. Our dataset offers segmented slides, their aligned spoken language and visual elements (figures, diagrams, natural images, tables), and slide text.

3.1. Dataset Statistics

LPM Dataset consists of 9031 slides, 8598 figures, 28000 unique words, 1.6 million total words from 334 educational presentation videos with a total duration of 187 hours. As shown in Table 3(b), per slide, there are 186



(a) Number of OCR text per slide

(b) Number of spoken words per slide

(c) Number of figures per slide

(d) Distribution of figures (entire dataset)

(e) Number of seconds per slide

(f) Number of seconds of mouse trace per slide

(g) Distribution of slides by speaker (entire dataset)

Figure 3: Dataset statistics for Lecture Presentations Multimodal Dataset. 2(d) shows that our a large majority of our dataset consists of figures that include text (diagrams, equations, tables). Figure 2(g) shows that the variety of educational content covered by our dataset.

words spoken on average. Each slide's duration is an average of 72.6 seconds, as shown in Figure Table 3(e). Among the 8598 figures, there are 3877 (45.1%) natural images, 4018 (46.7%) diagrams, 301 (3.5%) tables, 402 (4.6%) equations, shown in Table 3(d). Each slide has an average of 0.94 (median of 1) figure, shown in Table 3(c), an average of 28.95 written words, displayed in Table 3(a). When available, we provide the mouse traces which hover over the region the speaker is describing. Globally, in 86.4% of the slides, speakers use the mouse traces at least once. There are 12.52 seconds of mouse trace data per slide on average, as shown in Table 3(f). Our dataset consists of 35 courses on biology, anatomy, psychology, dentistry, speaking, machine learning taught by 10 speakers. The distribution of the number of slides per speaker is shown in Table 3(g). Our dataset is designed to include some imbalances amongst topics and visual mediums such that it could also be used for continual learning, low resource domain adaptation and transfer learning. We release the full preprocessing pipeline for easy expansion into other topics, as we describe below.

3.2. Data Collection and Preprocessing

The LPM Dataset is developed from a curated list of lecture presentation videos, which are downloaded from YouTube [1]. Spoken language is extracted from speech via automatic speech recognition. We manually annotate for the slide segments as well as figure bounding boxes and corresponding labels in order to perform retrieval tasks between slide-level segments spoken text and individual visual figures. In addition, in order to utilize the language information in figures, the texts in the slide are automatically extracted via OCR. We extract the mouse trace location to enable researchers to utilize them as an additional grounding signal between visual objects and language. We share the full data preprocessing pipeline in our repository. A visual outline the data collection and processing steps taken to create LPM Dataset is shown in Figure 2. Detailed data collection and preprocessing steps can be found in Appendix A.

3.3. Problem Definition

To benchmark a vision-and-language model's understanding of multimodal educational content, we measure its ability to associate a visual figure with a spoken explanation. We carefully designed three tasks supported by multimedia learning literature: where (1) visual figures are retrieved given spoken language (Text-to-Figure), (2) spoken explanations are retrieved from visual figures (Figure-to-Text) and (3) generation of spoken explanations.

Meaningful multimedia learning takes place when students are able to organize verbal explanations (spoken words) combined with non-verbal knowledge representa-

[1]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 the creators were individually contacted about the inclusion of their content in our dataset. Implicit consent allows creators to be removed from the dataset at any time.

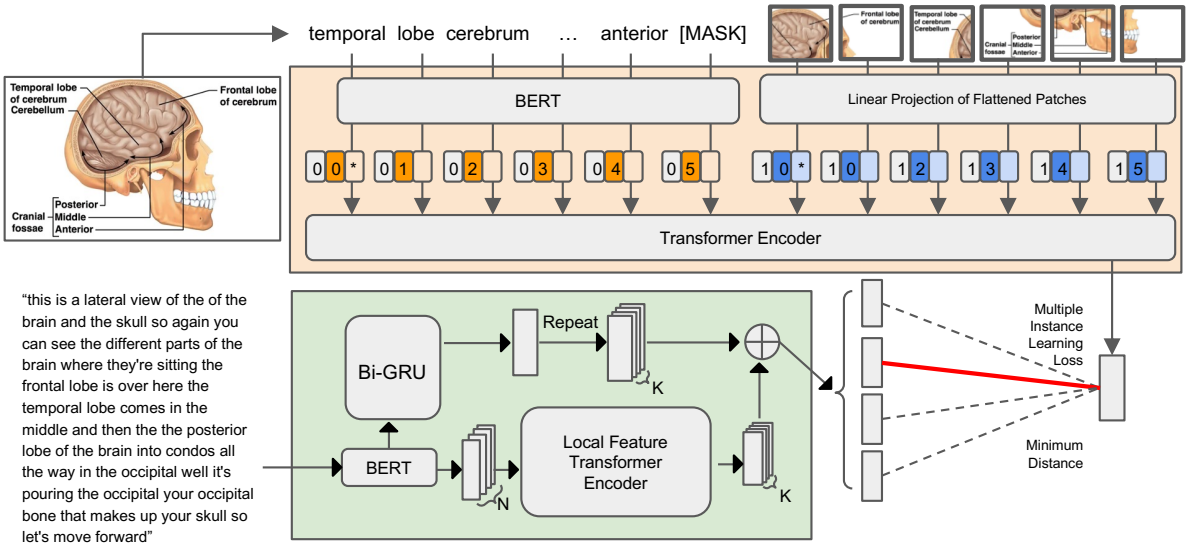


Figure 4: Overview of the key components of our proposed PolyViLT model. The diagram's text and image patches are input into a ViLT-based transformer encoder, and the spoken language BERT embedding is transformed into K representations. A MIL Loss [10] is used to address weak crossmodal alignment and find partially aligned instances.

tions (pictures) into a coherent mental model [32]. Furthermore, three cognitive processes that are essential for active learning involves selecting, organizing, and integrating the learnt material with existing knowledge [28, 51]. Therefore, crossmodal retrieval is a natural fit to evaluate a vision-language model's understanding of educational videos, as meaningful learning with video presentations, requires organizing and integrating text and pictures into a mental model. In the process of retrieving a text from a figure, the learner must select relevant verbal information from the given text, organize the information, and activate prior knowledge (from learnt representation) to retrieve the relevant image, which is aligned to the cognitive theory of multimedia learning [30]. In this work, we primarily focus on the two tasks of crossmodal retrieval, as the alignment between figures and text is a crucial pre-requisite problem that need be addressed to tackle the challenging problem of slide explanation generation. Furthermore, we provide baseline results and the codebase for slide explanation generation as described in Section 6.7.

In contrast to many prior crossmodal retrieval setups which assume one-to-one mappings between modalities [49], lecture presentations are unique in the presence of weak crossmodal alignment between spoken language and figures. There could exist $n > 1$ visual figures for a single spoken speech segment s and a figure could be aligned partially to the spoken segment (i.e., a part of the spoken segment is used to explain the figure). Thus, a core challenge lies in addressing weak crossmodal alignment. Formally, let $D = (S, V)$ be our dataset consisting of spoken language S and figures V. The goal is to learn an embedding space that can quantify the similarity between the figure and spoken language. As a result, given a segment of spoken language $s \in S$, one could retrieve the set of aligned visual figures $\{v_k, v_{k+1}, ..., v_{k+n}\} \subseteq V$. For Figure-to-Text, given a figure, we want to retrieve all of the transcriptions on each slide. For Text-to-Figure, given the spoken language of the entire slide, we want to retrieve all of the figures.

4. Retrieval Experimental Setup

We evaluate multiple state-of-art model's performance on text-to-figure and figure-to-text retrieval in comparison with human student performance. We are interested in understanding how current state-of-the-art models perform on different figure types (diagrams, images, equations, tables), long range sequences, and technical language. We also introduce PolyViLT, a multi-instance learning multimodal transformer that utilizes both vision and language information in slide figures.

4.1. Baselines

In addition to random selection and greedy text matching, we select previous baselines PVSE [45] and PCME [8] that are specifically designed for cross-modal retrieval in scenarios with weak alignment by using Multiple Instance Learning (MIL). These baselines are trained from scratch. We also measure the pre-trained CLIP [40] model's performance, as its zero-shot image-text matching performance is well recognized in the community.

CLIP [40] is an established baseline for image-text matching. We use pre-trained CLIP to embed pairs of figures and

Models	Figure-to-Text			Text-to-Figure		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
Random	1.36 ± 0.22	7.63 ± 0.88	15.81 ± 0.7	2.15 ± 0.61	8.64 ± 1.10	16.38 ± 1.91
Greedy Text Matching	4.20 ± 0.01	10.9 ± 0.015	11.9 ± 0.01	4.20 ± 0.01	10.9 ± 0.015	11.9 ± 0.01
CLIP [40]	2.05 ± 0.7	7.4 ± 0.15	17.65 ± 1.02	1.58 ± 0.56	6.89 ± 1.18	13.78 ± 0.55
PVSE [45]	3.17 ± 0.68	12.44 ± 1.28	22.01 ± 0.61	2.81 ± 0.27	11.87 ± 1.24	21.2 ± 0.63
PVSE (BERT) [45, 9]	2.96 ± 0.76	10.96 ± 0.52	18.54 ± 0.99	2.43 ± 0.05	11.21 ± 1.11	18.51 ± 1.10
PCME [8]	2.31 ± 0.41	8.83 ± 0.34	16.43 ± 0.67	2.12 ± 0.36	8.68 ± 0.14	16.9 ± 1.10
PCME (BERT) [8, 9]	1.93 ± 0.26	8.27 ± 0.95	15.76 ± 1.64	1.93 ± 0.26	8.36 ± 1.08	15.85 ± 1.77
PolyViLT	**4.94 ± 0.55**	**19.16 ± 0.69**	**30.35 ± 0.55**	**6.14 ± 1.25**	**23.19 ± 0.68**	**33.22 ± 1.73**
PolyViLT w/ All Speakers	3.22 ± 0.64	10.65 ± 0.26	20.54 ± 0.32	2.00 ± 0.29	9.85 ± 1.45	19.19 ± 1.77
PolyViLT w/ Trace	3.85 ± 0.91	17.77 ± 1.88	28.26 ± 1.78	5.38 ± 0.78	19.66 ± 2.39	32.26 ± 0.59
PolyViLT w/o Fig. Lang.	4.29 ± 0.72	17.43 ± 0.72	27.83 ± 0.35	4.32 ± 1.11	19.89 ± 1.71	31.77 ± 0.48
PolyViLT w/o Fig. Image	3.79 ± 0.31	14.25 ± 0.71	24.70 ± 1.34	6.14 ± 0.89	19.30 ± 2.07	29.39 ± 2.80

Table 2: Comparison between PolyViLT vs previous state-of-the-art models for crossmodal retrieval with multiple instance learning across all dataset for 3 random seeds, standard deviation bars are reported. PolyViLT outperforms all previous state-of-the-art approaches by a large margin. We also find that training speaker-specific models outperforms training collectively across all speakers. We run ablation studies by masking the figure image (w/o Fig. Image) and language (w/o Fig. Image) to find a drop in performance.

text and rank their similarity scores for retrieval.

PVSE [45] is designed to model one-to-many alignment for crossmodal retrieval, by encoding visual and text features as K possible embeddings and training with a multiple instance loss that rewards weak cross-modal alignment (i.e., the best pair among K^2 pairs is rewarded).

PCME [8] handles pairwise semantic similarities and uncertainty in crossmodal retrieval. It models each modality as probabilistic distributions in a common embedding space using Hedged Instance Embeddings [34] and utilizes a soft version of the contrastive loss to handle weak alignment.

4.2. PolyViLT: Proposed Retrieval Model for Weak Image-Text Alignment

On top of these baselines, we introduce Polysemous-ViLT (or PolyViLT), which is designed to handle vision and language inputs (e.g., diagrams) and weak cross-modal alignment. Previous approaches were designed specifically for the task of crossmodal retrieval on datasets consisting of only natural images and text. However, to perform well on retrieval problems involving figures, models must utilize text information present in the figure, as they could provide valuable signals to the model. Our approach utilizes local feature transformers in PVSE [45], a multi-instance learning loss [10] and a ViLT encoder [21] to utilize both vision and language information in figures. We refer the readers to Figure 4 for details of the architecture.

ViLT Figure Encoder We utilize the ViLT model [21] as a backbone encoder to contextualize the text and vision information present in figure. Given an image of a figure, we utilize the text (from OCR output) within the bounding box occupied by the figure. The text input is tokenized with BERT [9], and patches of the diagram image is flattened and linearly projected, and inputted to a transformer encoder.

Multiple Instance Learning (MIL) To account for the weak crossmodal alignment between figures and spoken language, we represent the spoken language with K embeddings, capturing different words of the speech, inspired by local feature transformers in [45], The local K embedding are combined with global information via residual connections. Then, we utilize the MIL objective [10], which assume that there is a partial match between a figure and K local embeddings of the spoken language.

4.3. Human Student Performance

To measure human student performance and probe the difficulty of each topic with non-expert performance, human students are shown 10 figure-caption pairs (to prevent attention loss) for 10 speakers and 3 seeds. For Figure-to-Text, a student is shown one figure image and all of the aligned spoken language. Then, they are asked to select the most relevant spoken language. For Text-to-Figure, the annotator is shown one spoken language, all the figure and is asked to select the most relevant figure. There are exactly 300 figures, 197 spoken language segments and we measure Recall@1 performance. For fair comparison, all baseline models are evaluated with Recall@1 again with identical samples from the human study. We share human evaluation results as a part of our dataset.

5. Training Details

We use PyTorch as the auto-differentiation library to train all our models. For each speaker, with split the data such that a random 80% is used as training data and the

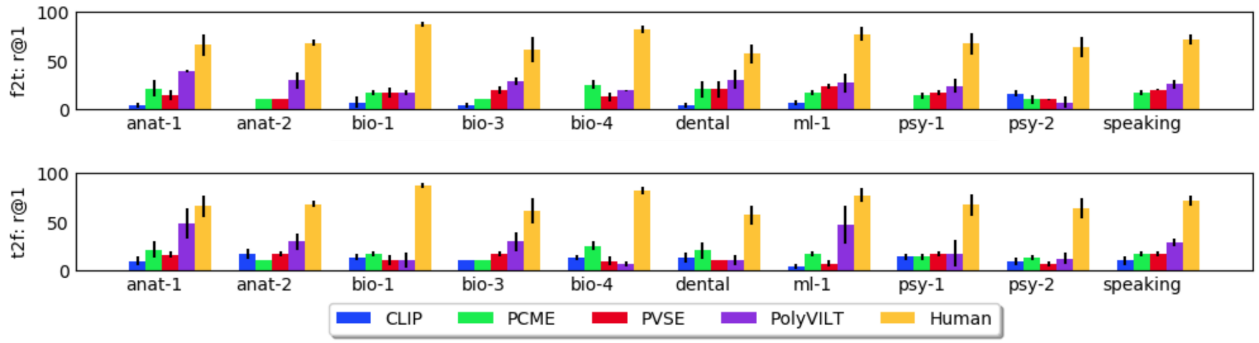


Figure 5: Comparison of baselines and PolyViLT against human student performance in Recall@10 for (Top) Figure-to-Text and (Bottom) Text-to-Figure retrieval.

remaining 20% is used for test (the data is split according to each random seed). In our experiments, we use the following hyperparameters. We train for 100 epochs, and our batch size is 8. We also utilize the 3 losses (MIL with a margin parameter λ_m, Diversity λ_{div}, Domain Discrepancy λ_{dom}) as motivated in [45], we refer the audience to the original paper for the formulation of these losses. We use the default parameters, $\lambda_m = 0.1$, $\lambda_{div} = 0.01$, $\lambda_{dom} = 0.01$. For the number of locally guided features K, as shown in Figure 4, we use $K = 5$. Further finetuning on these hyperparameters is a future direction of study to boost performance. As mentioned in Section 4.2, we use a pre-trained backbone ViLT encoder from HuggingFace, by the original authors, which has been trained on masked language modelling and image-text matching ('ViLT-b32-mlm-itm') [52, 21]. We will release the full code base with our default hyperparameters. The average model train runtime was around 8 hours on Titan X 1080 GPUs.

6. Results

6.1. Model and Human Performance

The in-domain performance of all models can be seen in Table 2. We first implement a greedy text matching baseline, where we assign a match between spoken language and figure by considering the greatest number of commonly occuring words. Our proposed model, PolyViLT, outperforms this approach by a significant margin. We note that 46% of our dataset contains figures with no text, which reflects the challenging multimodal nature of our task. We also refer the reviewer to Appendix I, where we test PolyViLT's performance for varying difficulties of text-matching measured by tf-idf. We find that PolyViLT performs better for cases with easier keyword identifiability than harder cases. PolyViLT outperforms previous state-of-the-art vision-language models in both figure-to-text retrieval and text-to-figure retrieval. The second best performing model is PVSE [45], which further justifies our reasoning behind utilizing local feature transformers and the MIL loss. Surprisingly, CLIP's zero-shot performance often is worse than Random, which indicates that large-scale pre-training on natural image-text pairs may not be sufficient for our task. We find that fine-tuning CLIP on our dataset yields a performance boost (6.21% increase for text-to-figure Recall@10 shown in Appendix H). We refer the readers to Appendix F, where we conduct out-of-domain experiments where we train on a source speaker and evaluate on a different target speaker. To test speaker and topic independence, we report the results for 3 speaker pairs (bio-1 $\rightarrow$ bio-3, anat-1 $\rightarrow$ anat-2, psy-1 $\rightarrow$ psy-2) on similar domains and a pair from different domains (bio-1 $\rightarrow$ psy-1) and find that the baseline models perform better than random, which demonstrate the baselines' generalizability from one domain/speaker to another. The detailed results for each speaker can be found in Appendix E We also provide human student retrieval performance in Figure 5. We see that all methods fall well below human students' performance, even PolyViLT, the closest method, is 47.68% worse for text-to-figure retrieval and 43.63% worse for figure-to-text retrieval, which demonstrates the challenging nature of our dataset. Below, we perform error analysis to uncover the concrete challenges presented in LPM Dataset.

6.2. Performance on Novel Visual Mediums

We first investigate the impact of novel visual mediums such as man-made figures (e.g., diagrams, tables, and equations) on model performance. We report Recall@10 scores conditioned on each type in Table 3, and find that PolyViLT outperforms other baselines for most figure types. Interestingly, we can see that for natural images, previous approaches perform worse than PolyViLT. Whereas we intially suspected that PolyViLT's main advantage is in its use of text information, it outperforms previous approaches even when no text information is used. This indicates that the usage of a ViT encoder [11] is superior over using local and global feature transformers as proposed in PVSE [45] and PCME [8] even for natural images. We also find models struggle, particularly on equations. As mentioned in Sec-

Models	Figure-to-Text: Recall@10				Text-to-Figure: Recall@10			
	Diagram	Image	Table	Equation	Diagram	Image	Table	Equation
CLIP [40]	6.2 ± 0.57	5.77 ± 0.73	6.2 ± 4.36	2.83 ± 1.11	6.5 ± 1.27	6.0 ± 0.22	6.9 ± 2.5	3.5 ± 0.96
PVSE [45]	8.2 ± 0.93	9.6 ± 0.57	7.27 ± 0.29	**12.27 ± 3.27**	7.6 ± 1.3	10.33 ± 1.76	6.97 ± 4.15	4.47 ± 4.66
PCME [8]	6.0 ± 0.37	6.9 ± 0.22	6.3 ± 3.28	2.93 ± 3.27	5.9 ± 0.49	6.87 ± 0.26	6.3 ± 3.28	2.93 ± 3.27
PolyViLT	**18.53 ± 1.65**	**15.2 ± 0.91**	**15.83 ± 2.67**	5.53 ± 5.37	**18.53 ± 1.89**	**20.13 ± 0.7**	**19.17 ± 6.34**	**9.97 ± 3.48**

Table 3: Comparison of recall@10 scores for baselines conditioned on types of figures, mean and standard deviations are reported for 3 seeds across all speakers. PolyViLT outperforms previous baselines in most cases, except for equation text-to-figure retrieval.

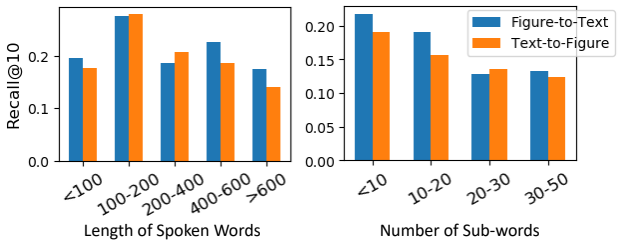


Figure 6: (Left) PolyViLT performance drops for very short or very long sequences, (Right) or with increasing number of subwords (technical terms).

tion 4.2, this could be attributed to the significant domain difference between the pretraining domain (natural images, non-educational language) of ViLT [21]. PVSE [45] is initialized with random weights, therefore is unaffected.

6.3. Technical Language and Long Sequences

We investigate the effects of technical language beyond commonly spoken and written text on model performance. The right figure in Fig. 6 shows the number of subwords tokenized by HuggingFace's BERT Tokenizer [9, 52], which represents the number of Out-of-Vocabulary (OOV) tokens, a proxy measure for how much external knowledge is required to understand technical language. With an increasing number of subwords, there is a drop in performance, indicating that our models struggle to quickly acquire technical information or require external knowledge to perform well. Furthermore, our dataset poses challenges in capturing information in long range language sequences due to its educational nature. On the left of Fig. 6, we report Recall@10 scores conditioned on the number of spoken words. PolyViLT's performance peaks between 100 and 200 words, and decreases with increasingly longer spoken phrases, or very short spoken phrases (under 100, where we find phrases that often do not contain enough information to disambiguate from different figures). This calls for a need to develop models for extremely long-range and short-range sequences. We refer the readers to Appendix K and Appendix J where we display qualitative and quantitative analysis of how current baselines fail when technical knowledge or understanding of long range interactions are required.

6.4. Impact on Performace due to OCR/ASR errors

We find 100 samples each of figure-text pairs with correct and incorrect OCR/ASR, then human and PolyVilT's r@1 scores are calculated. For correct OCR/ASR, model: 0.343, humans: 0.765. For incorrect OCR/ASR, model: 0.270, human: 0.837. Humans are more robust to incorrect OCR/ASR, whereas the model suffers a performance drop. We investigate human's performance with incorrect OCR/ASR further. In the correct case, we find an average of 170.8 ASR tokens and 15.6 OCR tokens, whereas in the incorrect case, we find an average of 243.9 ASR tokens and 14.4 OCR tokens. We hypothesize that the greater number of ASR tokens in the incorrect case provides more informative context that human annotators could exploit even with incorrect ASR/OCR. When we control the number of ASR to be equivalent (at less than 75 tokens), we find that the human performance is similar at 61.9% for correct ASR/OCR vs. 62.5% for incorrect ASR/OCR cases. However, even with correct OCR/ASR, the model has comparable performance to that of incorrect OCR/ASR, which indicates that its limitation is not due to errors in OCR/ASR but more likely due to other challenges; such as the novel medium of man-made figures, highly technical language. We include these annotations in the dataset such that users can analyze how OCR/ASR errors might impact performance.

6.5. Importance of MIL objective

We investigate the effects of using a Multi-Instance Learning (MIL) objective to handle ambiguous alignment by comparing PolyViLT with and without the MIL objective in Fig. 7. "No MIL" is the case where we optimize using the standard triplet ranking objective [13, 22]. Consistently, across all 3 speakers, we see that MIL is useful and leads to performance boosts by handling weak crossmodal alignment. In Appendix L, we provide additional analysis in regard to the learnt alignments, by investigating the K representations. To specifically find out whether the MIL objective was successful at disambiguating figures amongst K representations, we display the aligned spoken language and figures for a given slide, and show the calculated similarity scores for the K instances of spoken language for each figure. We see that in successful cases, the distribu-

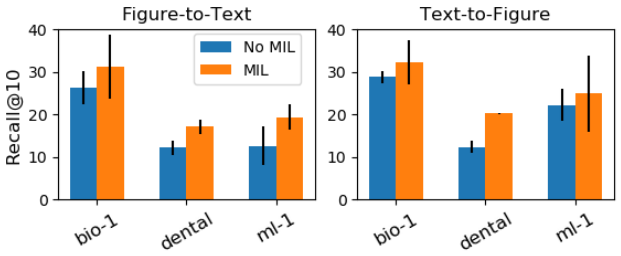


Figure 7: Using MIL to handle weak crossmodal alignment leads to performance boosts.

tion 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. However, in failed cases, the distribution of the similarity scores of K instances are identical for each figure, hinting that the separate K instances have not learned to disambiguate the different figures. A potential way to enforce the K instances to be different for each given figure by increasing the diversity hyperparameter λ_{div}, which penalizes the redundancy among K instances.

6.6. Using Mouse Trace as a Grounding Signal

Our aim is to provide the users with all available modalities and provide an opportunity to study mouse traces, which are known to be related to deictic gestures, eye gaze and grounding of language and vision. Thus, we experiment with utilizing mouse trace as an additional grounding signal to capture crossmodal alignment and represent mouse traces as a one-hot vector with length equivalent to the spoken language sequence. For the indices corresponding to words when the mouse hovered over the figure, we assign it the value 1, indicating that the spoken word is directly aligned to the given figure and is conceptually similar to hard attention. We re-parameterize this categorical distribution with a Gumbel-Softmax [20], and use a dot-product attention with skip connections to fuse spoken language and mouse traces. The result for this model is shown in Table 2, as 'PolyViLT + Trace'. For certain speakers, the inclusion of mouse-trace data offers better performance. We refer the readers to Appendix E for speaker-specific studies. Future work should aim at better utilizing the valuable information in mouse traces as a grounding signal [23, 39].

6.7. Slide Explanation Generation

LPM Dataset and text-to-figure retrieval enables generating slides from spoken language, where individual figures need to be retrieved one-by-one, which is a reflection of how speakers refer to figures in presentations (a segment within an explanation refer to a single figure). We also attempt the more ambitious task of slide explanations generation. We experiment with generating captions for each slide via fine-tuning a pre-trained ViT [11] encoder and GPT-2 [41] decoder with the next word prediction objective. We refer the reader to Table 6, where we see that the ROUGE-L scores for anat-1 is 12.3, bio-1: 9.2, psy-1: 6.6. Their low performance indicates that crossmodal retrieval and its challenges needs to be first addressed before tackling the more challenging task of generation. Nonetheless, our dataset is the first to enable the task of slide caption generation.

7. Conclusion and Discussion

We present the Lecture Presentations Multimodal Dataset as benchmark for developing vision-and-language models that can understand multimodal knowledge in educational videos. Our diversely sourced and richly annotated dataset contributes three challenging research tasks as a step towards educationally relevant goals: (1) automatic retrieval of spoken explanations given figures, (2) automatic retrieval of illustrative figures given spoken explanations (3) generation of spoken slide explanations. Through benchmarking existing and newly proposed models, we outline future research directions in tackling weak crossmodal alignment, novel visual mediums, technical language, and long-range sequences to step closer towards intelligent and accessible tutoring aids.

Limitations and Ethics: There exists an imbalance in the distribution amongst topics (most of our data fall under science and math) and types of visual mediums (small proportion of quantitative figures, tables and equations). Our dataset does not contain other extraneous sources of information such as animations, websites, or virtual whiteboards. While we believe that LPM Dataset is a sufficient first step towards tackling AI understanding of multimodal educational content, a robust and diverse dataset will require a broader variety of topics, mediums and information types. Hence, we share the full data preprocessing pipeline to enable expansion. To increase the transparency of our dataset, we share a Datasheet for Datasets [16] in the supplementary. Our work is intended to support research in AI teaching aids that could help address the shortage of teachers and democratization of education. Along the benefits, there are risks such as changing traditional educational settings, misinformation spread and unforeseen economic impacts.

Acknowledgement

This work was partially supported by the National Science Foundation (Award #1750439, #1722822), National Institutes of Health and NTT Japan. We thank the creators of the educational videos: Edward Kerschen, Mental Dental, Virtual Comsats, R. J. Birmingham, Lynda Kiesler, Fuzail Majoo, Carla Sweet, Rita Marcon, LP Morency. We also thank Multicomp group members at CMU, who provided with insight and feedback.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher. A realistic dataset for performance evaluation of document layout analysis. In *2009 10th International Conference on Document Analysis and Recognition*, pages 296–300. IEEE, 2009.

[3] Alan Baddeley. Working memory: looking back and looking forward. *Nature reviews neuroscience*, 4(10):829–839, 2003.

[4] Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. Vlengagement: A dataset of scientific video lectures for evaluating population-based engagement. *arXiv preprint arXiv:2011.02273*, 2020.

[5] Paul Chandler and John Sweller. Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4):293–332, 1991.

[6] Huizhong Chen, Matthew Cooper, Dhiraj Joshi, and Bernd Girod. Multi-modal language models for lecture video retrieval. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1081–1084, 2014.

[7] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE, 2018.

[8] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Kartik Dutta, Minesh Mathew, Praveen Krishnan, and C. V. Jawahar. Localizing and recognizing text in lecture videos. In *ICFHR*, 2018.

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

[14] Damianos Galanopoulos and Vasileios Mezaris. Temporal lecture video fragmentation using word embeddings. In *International Conference on Multimedia Modeling*, pages 254–265. Springer, 2019.

[15] Joanna Garner and Michael Alley. How the design of presentation slides affects audience comprehension: A case for the assertion-evidence approach. *International Journal of Engineering Education*, 29(6):1564–1579, 2013.

[16] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[17] David Griol and Zoraida Callejas. An architecture to develop multimodal educative applications with chatbots. *International Journal of Advanced Robotic Systems*, 10(3):175, 2013.

[18] Shannon F Harp and Richard E Mayer. The role of interest in learning from scientific text and illustrations: On the distinction between emotional interest and cognitive interest. *Journal of educational psychology*, 89(1):92, 1997.

[19] Richang Hong et al. Multimedia question answering. *IEEE MultiMedia*, 2012.

[20] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[21] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[22] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[23] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-image generation grounded by fine-grained user attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 237–246, 2021.

[24] Irene Li, Alexander R Fabbri, Robert R Tung, and Dragomir R Radev. What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6674–6681, 2019.

[25] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *arXiv preprint arXiv:2209.09513*, 2022.

[26] Zelun Luo, Wanze Xie, Siddharth Kapoor, Yiyun Liang, Michael Cooper, Juan Carlos Niebles, Ehsan Adeli, and Fei-Fei Li. Moma: Multi-object multi-actor activity parsing. *Advances in Neural Information Processing Systems*, 34:17939–17955, 2021.

[27] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[28] Richard E Mayer. Multimedia learning. In *Psychology of learning and motivation*, volume 41, pages 85–139. Elsevier, 2002.

[29] Richard E Mayer and Richard B Anderson. Animations need narrations: An experimental test of a dual-coding hypothesis. *Journal of educational psychology*, 83(4):484, 1991.

[30] Richard E Mayer and Logan Fiorella. 12 principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In *The Cambridge handbook of multimedia learning*, volume 279. Cambridge University Press New York, NY, 2014.

[31] Richard E Mayer and Roxana Moreno. Aids to computer-based multimedia learning. *Learning and instruction*, 12(1):107–119, 2002.

[32] Roxana Moreno and Richard Mayer. Interactive multimodal learning environments. *Educational psychology review*, 19(3):309–326, 2007.

[33] Roxana Moreno and Richard E Mayer. Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of educational psychology*, 91(2):358, 1999.

[34] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. *arXiv preprint arXiv:1810.00319*, 2018.

[35] Fred Paas, Alexander Renkl, and John Sweller. Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional science*, 32(1/2):1–8, 2004.

[36] Allan Paivio. *Mental representations: A dual coding approach*. Oxford University Press, 1990.

[37] Yi-Hao Peng, et al. Say it all: Feedback for improving non-visual presentation accessibility. In *CHI*, 2021.

[38] Annie Piolat, Thierry Olive, and Ronald T Kellogg. Cognitive effort during note taking. *Applied cognitive psychology*, 19(3):291–312, 2005.

[39] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, pages 647–664. Springer, 2020.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[42] Gabriel B Reedy. Powerpoint, interactive whiteboards, and the visual culture of technology in schools. *Technology, Pedagogy and Education*, 17(2):143–162, 2008.

[43] April Savoy, Robert W Proctor, and Gavriel Salvendy. Information retention from powerpoint™ and traditional lectures. *Computers & Education*, 52(4):858–867, 2009.

[44] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007.

[45] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.

[46] Edward Sun, et al. D2S: Document-to-slide generation via query-based text summarization. In *NAACL*. ACL, 2021.

[47] Joshua E Susskind. Powerpoint's power in the classroom: Enhancing students' self-efficacy and attitudes. *Computers & education*, 45(2):203–215, 2005.

[48] Nhu Van Nguyen, Mickal Coustaty, and Jean-Marc Ogier. Multi-modal and cross-modal for lecture videos retrieval. In *2014 22nd International Conference on Pattern Recognition*, pages 2667–2672. IEEE, 2014.

[49] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[50] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.

[51] Merlin C Wittrock. Generative processes of comprehension. *Educational psychologist*, 24(4):345–376, 1989.

[52] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[53] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.