Lecture Presentations Multimodal Dataset: Towards Understanding Multimodality in Educational Videos

Dong Won Lee¹  Chaitanya Ahuja²  Paul Pu Liang²  Sanika Natu²  Louis-Philippe Morency²
MIT¹, Carnegie Mellon University²
https://github.com/dondongwon/LPMDataset

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-

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.
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
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 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 purpose 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 hours. As shown in Table 3(b), per slide, there are 186
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 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.

3.2. Data Collection and Preprocessing

The LPM Dataset is developed from a curated list of lecture presentation videos, which are downloaded from YouTube. 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 representations. Our approach includes automatically identifying 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.
Linear Projection of Flattened Patches

temporal lobe cerebrum ... anterior [MASK]

BERT

0 0 * 0 1 0 2 0 3 0 4 0 5 1 0 1 0 1 2 1 3 1 4 1 5

"this is a lateral view of the of the brain and the skull so again you can see the different parts of the brain where they're sitting the frontal lobe is over here the temporal lobe comes in the middle and then the the posterior lobe of the brain into condos all the way in the occipital well it's pouring the occipital your occipital bone that makes up your skull so let's move forward"

Bi-GRU

Local Feature Transformer Encoder

Repeat

\( K \)

Multiple Instance Learning Loss

Minimum Distance

Minimum Distance

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.

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
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.

<table>
<thead>
<tr>
<th>Models</th>
<th>Figure-to-Text</th>
<th>Text-to-Figure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall@1</td>
<td>Recall@5</td>
</tr>
<tr>
<td></td>
<td>Recall@1</td>
<td>Recall@5</td>
</tr>
<tr>
<td>Random</td>
<td>1.36 ± 0.22</td>
<td>7.63 ± 0.88</td>
</tr>
<tr>
<td>Greedy Text Matching</td>
<td>4.20 ± 0.01</td>
<td>10.9 ± 0.015</td>
</tr>
<tr>
<td>CLIP [40]</td>
<td>2.05 ± 0.7</td>
<td>7.4 ± 0.15</td>
</tr>
<tr>
<td>PVSE [45]</td>
<td>3.17 ± 0.68</td>
<td>12.44 ± 1.28</td>
</tr>
<tr>
<td>PVSE (BERT) [45, 9]</td>
<td>2.96 ± 0.76</td>
<td>10.96 ± 0.52</td>
</tr>
<tr>
<td>PCME [8]</td>
<td>2.31 ± 0.41</td>
<td>8.83 ± 0.34</td>
</tr>
<tr>
<td>PCME (BERT) [8, 9]</td>
<td>1.93 ± 0.26</td>
<td>8.27 ± 0.95</td>
</tr>
<tr>
<td>PolyViLT</td>
<td>4.94 ± 0.55</td>
<td>19.16 ± 0.69</td>
</tr>
<tr>
<td>PolyViLT w/ All Speakers</td>
<td>3.22 ± 0.64</td>
<td>10.65 ± 0.26</td>
</tr>
<tr>
<td>PolyViLT w/ Trace</td>
<td>3.85 ± 0.91</td>
<td>17.77 ± 1.88</td>
</tr>
<tr>
<td>PolyViLT w/o Fig. Lang.</td>
<td>4.29 ± 0.72</td>
<td>17.43 ± 0.72</td>
</tr>
<tr>
<td>PolyViLT w/o Fig. Image</td>
<td>3.79 ± 0.31</td>
<td>14.25 ± 0.71</td>
</tr>
</tbody>
</table>

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 figure. 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
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 $\lambda_m$, Diversity $\lambda_{div}$, Domain Discrepan$\gamma y$ $\lambda_{dom}$) 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 occurring 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 initially 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-
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.

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 Performance 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-
6.7. Slide Explanation Generation

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 traces 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].

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


