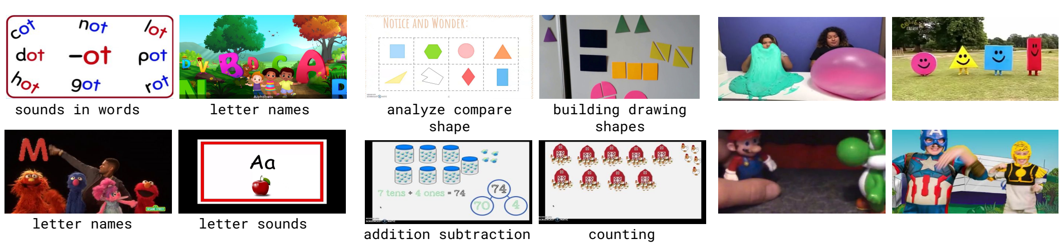
Detecting Educational Content in Online Videos by Combining Multimodal Cues

Anirban Roy Sujeong Kim Claire Christensen Madeline Cincebeaux
SRI International
{anirban.roy, sujeong.kim, claire.christensen, maddie.cincebeaux}@sri.com

Abstract

The increasing trend of young children consuming online media underscores the need for data-driven tools that empower educators to identify suitable educational content for early learners. This paper introduces a method for identifying educational content within online videos. We focus on two widely used educational content classes: literacy and math. We consider two levels: Prekindergarten and Kindergarten. For each class and level, we choose prominent codes (sub-classes) based on the Common Core Standards. For example, literacy codes include 'letter names', and 'letter sounds', and math codes include 'counting', and 'sorting'. We pose this as a fine-grained multilabel classification problem as videos can contain multiple types of educational content and the content classes can get visually similar (e.g., 'letter names' vs. 'letter sounds'). As the alignment between visual and audio cues is crucial for effective comprehension, we consider a multimodal video analysis framework to capture both visual and audio cues in videos while detecting the educational content. We leverage the recent success of the generative models to analyze audio and visual content. Specifically, we apply automatic speech recognition (ASR) to extract the speech from the audio and capture visual cues with descriptive captions. Finally, we fuse both cues to detect desired educational content. Our experiments show multimodal analysis of cues is crucial for detecting educational content in videos.

1 Introduction



(a) Frames from literacy videos (b) Frames from math videos (c) Frames from background videos

Figure 1: Sample video frames from the dataset Gupta et al. [2023]. We present the videos belonging to the **(a) literacy classes**, **(b) math classes**, and **(c) background**. Background videos do not contain educational content but share visual similarities with educational videos. The videos are labeled with sub-classes, e.g., `letter names` vs `letter sounds`.

As internet access continues to spread, and smart devices become ever-present, children are devoting more of their time to viewing online videos. A recent survey conducted on a national scale revealed that 89% of parents with children aged 11 or younger confirmed that their kids watch videos on

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

YouTube Auxier et al. [2020b]. Moreover, it is estimated that young children in the age range of two to four years consume 2.5 hours and five to eight years consume 3.0 hours per day on average Rideout and Robb [2019a,b]. Childhood is typically a key period for education, especially for learning basic skills such as literacy and math Hemphill and Tivnan [2008], Jordan et al. [2009]. Unlike generic online videos, watching appropriate educational videos supports healthy child development and learning Hurwitz [2019], Hurwitz and Schmitt [2020], Burkhardt and Lenhard [2022]. Hence, examining the content within these videos could offer valuable insights for parents, educators, and media creators aiming to enhance young children's access to high-quality educational videos, a factor that has been demonstrated to yield significant learning benefits Hurwitz [2019]. With the exponential growth of online content creation, automated methods for comprehending content become increasingly indispensable in achieving this objective.

In this study, our objective is to assess whether a given video includes educational material and to describe the nature of this content. We follow the Common Core Standards Association et al. [2010], Porter et al. [2011] to characterize age-appropriate educational content for young kids. Detecting educational content requires identifying multiple distinct types of content in a video while distinguishing between similar content types. The task is challenging as the education codes by Common Core Standards Association et al. [2010], Porter et al. [2011] can be similar such as 'letter names' and 'letter sounds', where the former focuses on the name of the letter and the latter is based on the phonetic sound of the letter. Also, understanding education content requires analyzing both visual and audio cues simultaneously as both signals are to be present to ensure effective learning Association et al. [2010], Porter et al. [2011]. This is in contrast to standard video classification benchmarks such as the sports or generic YouTube videos in UCF101 Soomro et al. [2012] Kinetics400 Smaira et al. [2020], YouTube-8M Abu-El-Haija et al. [2016], where visual cues are often sufficient to detect the different classes. Finally, unlike standard well-known action videos, education codes are more structured and not accessible to common users. Thus, it requires a carefully curated set of videos and expert annotations to create a dataset to enable a data-driven approach. In this work, we focus on two widely used educational content classes: literacy and math. For each class, we choose prominent codes (sub-classes) based on the Common Core Standards that outline age-appropriate learning standards Association et al. [2010], Porter et al. [2011]. For example, literacy codes include 'letter names', 'letter sounds', 'rhyming', and math codes include 'counting', 'addition subtraction', 'sorting', 'analyze shapes'. We present sample video frames corresponding to these codes in figure 1.

We formulate the problem as a multilabel video classification task as a video may contain multiple types of content that can be similar. We consider a multimodal content understanding framework to combine visual and audio cues from a video. Combining multimodal cues is crucial for detecting educational content because aligned visuals and audio are essential to effective literacy instruction. Furthermore, using multimodal cues improves the robustness of the model as individual modalities can be noisy or not sufficiently informative. The multimodal approach is shown to be effective for image-text matching and content understanding Datta et al. [2019]. We first extract visual and audio information from a given video, then develop separate machine learning models to classify videos based on each modality, and finally combine the modality-specific predictions to detect the educational codes in the video.

2 Related Works

Educational videos for early development. Providing young children (ages 0–8) with access to high-quality screen media represents a convenient means of promoting early math and literacy skills, particularly given that these children typically spend around 2.5 hours per day engaged with screen media Rideout and Robb [2019a], Auxier et al. [2020a]. Exposure to well-crafted educational media has been shown to yield positive outcomes in early learning. Controlled laboratory studies have demonstrated that young children can transfer the knowledge acquired from educational math videos to non-screen-based learning scenarios Aladé et al. [2016], Schroeder and Kirkorian [2016], Hurwitz [2019], Burkhardt and Lenhard [2022]. Moreover, in more naturalistic trials conducted in home settings over several weeks, these positive effects have proven to be enduring, even in less controlled environments Silander et al. [2016]. Furthermore, the benefits appear to persist over time, as evidenced by two longitudinal studies indicating that children who viewed educational programs during their preschool years exhibited stronger math performance, including self-reported grade point averages, course completion rates, and standardized test scores, lasting into adolescence Anderson et al. [2001], Wright et al. [2001].

Multimodal Learning. Supervised multimodal Learning typically relies on learning a common embedding based on the crowd-captioned datasets such as Flickr30k Young et al. [2014] and MS-COCO Captions Chen et al. [2015]. Some prior works such as OSCAR Li et al. [2020] and VinVL Zhang et al. [2021] have utilized pre-trained object detectors and multi-modal transformers to learn image captioning using supervised aligned datasets. BLIP Li et al. [2022] takes a hybrid approach where it bootstraps an image captioner using a labeled dataset and uses it to generate captions for web images. This generated corpus is then filtered and used for learning an aligned representation. ALign BEfore Fuse Li et al. [2021] highlights the importance of aligning text and image tokens before fusing them using a multi-modal transformer.

Weakly aligned text-image/video datasets scraped from the web such as Conceptual Captions Sharma et al. [2018] and WebVid-10M Bain et al. [2021] enable learning of multi-modal representations. CLIP Radford et al. [2021] applies a cross-modal contrastive loss to train individual text and image encoders. Everything at Once Shvetsova et al. [2022] is able to additionally utilize the audio modality and incorporates a pairwise fusion encoder which encodes pairs of modalities, as a result, 6 forward passes of the fusion model are required for 3 modalities. Frozen in Time Bain et al. [2021] is able to utilize both image-text and video-text datasets through the use of a Space-Time Transformer Visual Encoder. Visual Conditioned GPT Luo et al. [2022] uses a single cross-attention fusion layer to combine pre-trained CLIP text and visual features. Flamingo Alayrac et al. [2022] adds cross-attention layers interleaved with language decoder layers to fuse visual information into text generation. MERLOT Zellers et al. [2021, 2022] and Triple Contrastive Learning Yang et al. [2022] combine contrastive learning and generative language modeling to learn aligned text-image representations. Gupta et al. [2023] consider a class-prototypes based contrastive learning for classifying videos with multiple educational labels. Zhao et al. [2017] also consider multimodal cues to analyse online tutorial videos.

3 Proposed Approach

We pose the problem of detecting educational content in videos a multilabel video classification task. To address this, a multimodal content understanding framework is employed, which integrates both visual and audio information from the video. This combination of multimodal cues is particularly important for identifying educational content, as synchronized visuals and audio are crucial for effective literacy instruction. We present the framework in figure 2. The framework consists of two components: one for processing visual cues and another for processing audio cues. These components are described below.

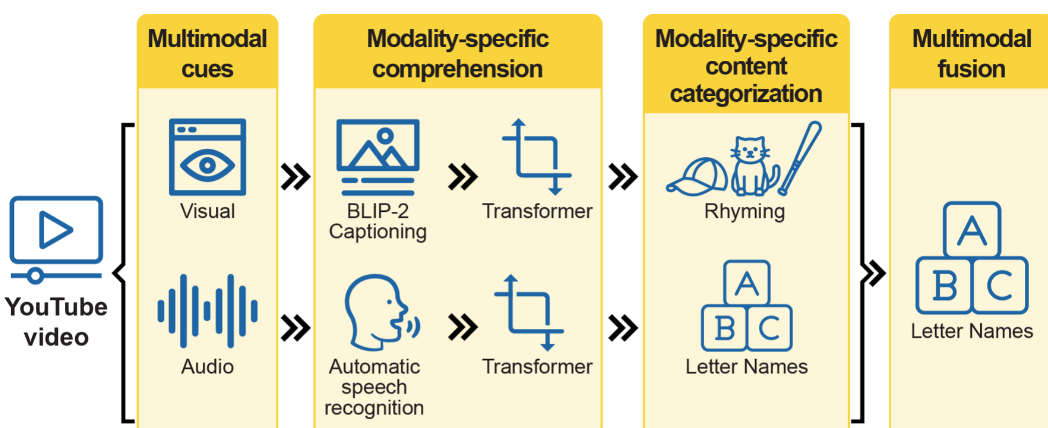
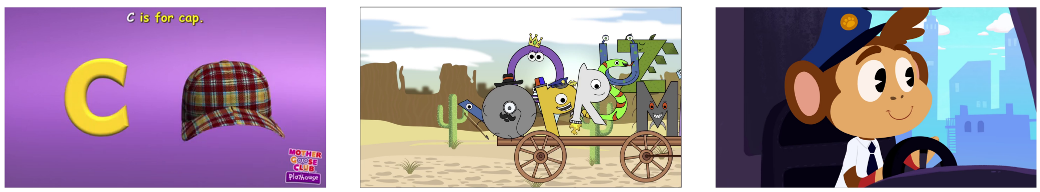


Figure 2: Proposed multimodal framework consists of two components. Top: processing visual cues, bottom: processing audio cues. Both the cues are fused to make final predictions.

Processing visual cues. For capturing the visual cues, we select key frames from a video and generate a detailed caption describing the visual content for the frames. A caption corresponding to a frame describes the primary activity, objects, attributes of the objects (such as color, shape, count), and interactions between the objects (such as a cartoon character pointing to a letter). We combine the captions for the frames to generate a detailed description of the video. We use the BLIP-2 Li et al. [2022, 2023] model for generating captions. This is a generative multimodal model that combines images and texts. A frozen vision transformer Dosovitskiy et al. [2021] is employed to capture the visual content and a frozen large language model (LLM) is employed to capture text queries. A lightweight query transformer is trained to connect two modalities by learning to attend text-informed visual features in order to generate appropriate responses. The model is trained on a large-scale data set of 234 million image and caption pairs, achieving state-of-the-art performance in generating descriptive captions for novel images such as frames from YouTube videos. We present examples of generated captions in figure 3. Note that the captions capture the objects (such as cap and car), identify the object type (such as a character C), the style of the image (such as cartoon), and actions (such as driving). We concatenate the captions from the frames to generate a detailed textual description of the video. Similar consecutive captions are removed to avoid including redundant captions. Finally, we apply a multiclass text transformer model Vaswani et al. [2017] to detect the educational codes in the video.



Caption: a cartoon image of a hat with the letter C on it

Caption: a cartoon wagon with many different characters on it

Caption: a cartoon monkey in uniform driving a car

Figure 3: Examples of captions for video frames generated by BLIP-2Li et al. [2023].

Processing audio cues. To process the audio cues we first extract the audio from the video. As online videos often include a background track, we extract the voice from the audio to avoid including spurious audio signals Hennequin et al. [2020]. For the voice track, we apply automatic speech recognition (ASR) to generate text from the audio. We use Whisper Radford et al. [2023] for ASR. Whisper is an encoder-decoder based transformer model that is trained for multiple tasks including ASR and translation. Finally, we apply a multiclass transformer Vaswani et al. [2017] to detect the educational codes in the transcription text. The transformer model classifies videos based on the occurrence of words and the relationships between words, such as the order in which they appear. For example, in videos classified as containing letter names, the letters tend to appear in alphabetical order.

Fusion of visual and audio cues. The video-based model and audio-based model are trained separately, and the predictions are combined to generate the final predictions in a late-fusion manner. For each video, we have two scores for the content categories: one indicating the likelihood the video belongs to the categories based on the visual content, and the other based on the audio content. We use a weighted sum of the classifications from the video-based and audio-based models to determine the final class predictions as

$$p_c = \alpha p_c^V + (1 - \alpha) p_c^A \tag{1}$$

where p_c^V and p_c^A are the prediction scores for a video corresponding to a category class c from the video-based and audio-based models, respectively. $\alpha \in [0, 1]$ is used to control the contribution of the models to determine the final prediction score p_c. We experimentally determine $\alpha = 0.5$, i.e., equally weighting both modalities, results in the best performance.

4 Experiments

Datasets. We perform our experiments on a dataset of Youtube videos. The dataset consists of more than 200 hours of expert-annotated videos with literacy and math classes suitable for kindergarten(K)

Literacy classes	Audio only	Video only	Multimodal
Follow words	69	67	86
Sight words	74	72	75
Letter sounds	66	52	64
Sounds in words	72	65	74
Letter names	82	78	83
Letter in words	53	49	61
Rhyming	98	85	98
Average	74	67	77

Table 1: Accuracy(%) on literacy classes at the pre-Kindergarten level.

Literacy classes	Audio only	Video only	Multimodal
Follow words	73	79	76
Sight words	57	63	64
Letter sounds	76	67	79
Sounds in words	76	62	73
Letter names	83	76	81
Letter in words	55	55	59
Rhyming	98	96	100
Average	74	71	76

Table 2: Accuracy(%) on literacy classes at the Kindergarten level.

and pre-kindergarten(pre-K) levels. These videos are selected from Youtube and annotated by expert education researchers. To ensure reliability, we train the annotators before labeling the videos and check inter-annotator agreement (more than 95% agreement) for selecting the final set of videos.

For literacy, both pre-kindergarten and kindergarten levels have the same set of seven classes. The classes and the number of videos per class are as follows: Follow words(175 for pre-K and 204 for K), Sight words(441 for pre-K and 228 for K), Letter sounds(223 for pre-K and 297 for K), Sounds in words(259 for pre-K and 282 for K), Letter names(341 for pre-K and 341 for K), Letter in words(161 for pre-K and 161 for K), and Rhyming(89 for pre-K and 89 for K). The math pre-kindergarten classes and the number of videos per class are as follows: Counting(318), Written numerals(343), Addition and Subtraction(79), Building and drawing shapes(30), Shape identification(185), Subitizing(304), Patterns(615), Cardinality(168), Analyzing and comparing shapes(90), Comparing groups(79), Measurable attributes(203), Sorting(80), and Spatial language(346). The math kindergarten classes and the number of videos per class are as follows: Counting(318), Written numerals(347), Addition and Subtraction(79), Building and drawing shapes(81), Shape identification(190), Cardinality(168), Analyzing and comparing shapes(90), Comparing groups(79), Measurable attributes(203), Sorting(80), Spatial language(346).

Experimental setup. We consider 75% of the videos for each class for training and 25% for tests. We consider three random splits of data for experiments and results are presented as the average of these three setups. As our goal is to detect educational videos for children, we consider precision as the metric to focus on only reliable predictions.

Results. Results for both pre-K and K levels for literacy classes are shown in table 1 and table 2, respectively. Results for both pre-K and K levels for math classes are shown in table 3 and table 4, respectively. We compare our results with the baselines where only one modality is considered, i.e., either audio or video cues. Note that the multimodal variant achieves better performance overall. Due to the variations in the number of videos per code and inter-code similarities, we notice a variation in accuracy numbers among the codes. Furthermore, the effectiveness of multimodal cues varies across the codes due to the relative importance of visual and audio cues in expressing the code.

Math classes	Audio only	Video only	Multimodal
Counting	86	67	84
Written numerals	80	70	81
Addition and Subtraction	98	73	98
Building and drawing shapes	89	71	83
Shape identification	89	62	87
Subitizing	73	60	77
Patterns	93	89	93
Cardinality	66	56	73
Analyzing and comparing shapes	89	51	89
Comparing groups	95	51	89
Measurable attributes	88	61	90
Sorting	87	60	90
Spatial language	79	67	78
Average	85	64	86

Table 3: Accuracy(%) on math classes at the pre-Kindergarten levels.

Math classes	Audio only	Video only	Multimodal
Counting	83	69	83
Written numerals	79	72	81
Addition and Subtraction	98	84	98
Building and drawing shapes	77	56	87
Shape identification	88	71	89
Cardinality	66	54	70
Analyzing and comparing shapes	89	63	89
Comparing groups	97	72	97
Measurable attributes	91	57	91
Sorting	95	69	93
Spatial language	79	64	79
Average	86	66	87

Table 4: Accuracy(%) on math classes at the Kindergarten levels.

5 Conclusion

We have proposed an approach for detecting educational content in online videos. The problem is formulated as a multilabel video classification task and we have considered a multimodal video analysis framework to address this. Our framework consists of two components: one for processing visual cues and another for processing audio cues. We fuse the predictions from two components to generate final predictions. We evaluate our approach on a large-scale expert-annotated educational video dataset. Our experiments indicate that multimodal analysis is important to detect educational content in videos and this outperforms the baselines where only a single modality is considered.

References

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.

Fashina Aladé, Alexis R Lauricella, Leanne Beaudoin-Ryan, and Ellen Wartella. Measuring with murray: Touchscreen technology and preschoolers' stem learning. *Computers in human behavior*, 62:433–441, 2016.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Daniel R Anderson, Aletha C Huston, Kelly L Schmitt, Deborah L Linebarger, John C Wright, and Reed Larson. Early childhood television viewing and adolescent behavior: The recontact study. *Monographs of the society for Research in Child Development*, pages i–154, 2001.

National Governors Association et al. Common core state standards. *Washington, DC*, 2010.

Brooke Auxier, Monica Anderson, Andrew Perrin, and Erica Turner. Parenting children in the age of screens. 2020a.

Brooke Auxier, Monica Anderson Andrew Perrin, and Erica Turner. Parenting children in the age of screens. Technical report, Pew Research Center, 2020b.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

Johanna Burkhardt and Wolfgang Lenhard. A meta-analysis on the longitudinal, age-dependent effects of violent video games on aggression. *Media Psychology*, 25(3):499–512, 2022.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2601–2610, 2019.

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. *ICLR*, 2021.

Rohit Gupta, Anirban Roy, Claire Christensen, Sujeong Kim, Sarah Gerard, Madeline Cincebeaux, Ajay Divakaran, Todd Grindal, and Mubarak Shah. Class prototypes based contrastive learning for classifying multi-label and fine-grained educational videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19923–19933, 2023.

Lowry Hemphill and Terrence Tivnan. The importance of early vocabulary for literacy achievement in high-poverty schools. *Journal of Education for Students Placed at Risk*, 13(4):426–451, 2008.

Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. doi: 10.21105/joss.02154. URL `https://doi.org/10.21105/joss.02154`. Deezer Research.

Lisa B Hurwitz. Getting a read on ready to learn media: A meta-analytic review of effects on literacy. *Child Development*, 90(5):1754–1771, 2019.

Lisa B Hurwitz and Kelly L Schmitt. Raising readers with ready to learn: A six-year follow-up to an early educational computer game intervention. *Computers in Human Behavior*, 104:106176, 2020.

Nancy C Jordan, David Kaplan, Chaitanya Ramineni, and Maria N Locuniak. Early math matters: kindergarten number competence and later mathematics outcomes. *Developmental psychology*, 45 (3):850, 2009.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020.

Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. A frustratingly simple approach for end-to-end image captioning, 2022. URL `https://arxiv.org/abs/2201.12723`.

Andrew Porter, Jennifer McMaken, Jun Hwang, and Rui Yang. Common core standards: The new us intended curriculum. *Educational researcher*, 40(3):103–116, 2011.

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

V Rideout and M. B. Robb. The common sense census: Media use by tweens and teens. common sense media. Technical report, Common Sense Media, 2019a.

Victoria Rideout and Michael B. Robb. The common sense census: Media use by tweens and teens, Oct 2019b. URL `https://www.commonsensemedia.org/research/the-common-sense-census-media-use-by-tweens-and-teens-2019`.

Elizabeth L Schroeder and Heather L Kirkorian. When seeing is better than doing: Preschoolers' transfer of stem skills using touchscreen games. *Frontiers in Psychology*, 7:1377, 2016.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL `https://aclanthology.org/P18-1238`.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022.

Megan Silander, Savitha Moorthy, Ximena Dominguez, Naomi Hupert, Shelley Pasnik, and Carlin Llorente. Using digital media at home to promote young children's mathematics learning: Results of a randomized controlled trial. *Society for Research on Educational Effectiveness*, 2016.

Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset, 2020.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

John C Wright, Aletha C Huston, Kimberlee C Murphy, Michelle St. Peters, Marites PiÃ±on, Ronda Scantlin, and Jennifer Kotler. The relations of early television viewing to school readiness and vocabulary of children from low-income families: The early window project. *Child development*, 72(5):1347–1366, 2001.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL `https://aclanthology.org/Q14-1006`.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16375–16387, June 2022.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021.

Baoquan Zhao, Shujin Lin, Xiaonan Luo, Songhua Xu, and Ruomei Wang. A novel system for visual navigation of educational videos using multimodal cues. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1680–1688, 2017.