The Potential of Vision-Language Models for Content Moderation of Children's Videos

Syed Hammad Ahmed
Department of Computer Science
University of Central Florida
Orlando, USA
syed.hammad.ahmed@ucf.edu

Shengnan Hu
Department of Computer Science
University of Central Florida
Orlando, USA
shengnan.hu@ucf.edu

Gita Sukthankar
Department of Computer Science
University of Central Florida
Orlando, USA
gitars@eecs.ucf.edu

2023 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-4534-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMLA58977.2023.00186

***Abstract*—Natural language supervision has been shown to be effective for zero-shot learning in many computer vision tasks, such as object detection and activity recognition. However, generating informative prompts can be challenging for more subtle tasks, such as video content moderation. This can be difficult, as there are many reasons why a video might be inappropriate, beyond violence and obscenity. For example, scammers may attempt to create junk content that is similar to popular educational videos but with no meaningful information. This paper evaluates the performance of several CLIP variations for content moderation of children's cartoons in both the supervised and zero-shot setting. We show that our proposed model (Vanilla CLIP with Projection Layer) outperforms previous work conducted on the Malicious or Benign (MOB) benchmark for video content moderation. This paper presents an in depth analysis of how context-specific language prompts affect content moderation performance. Our results indicate that it is important to include more context in content moderation prompts, particularly for cartoon videos as they are not well represented in the CLIP training data.**

***Index Terms*—video content moderation, vision-language models, prompt engineering**

I. Introduction

Automated video content moderation is a necessity for protecting children from inappropriate videos uploaded to social media platforms. Every minute YouTube content creators upload around 300 hours of video [1]. In the United States, more than 80% children aged 11 years or under watch YouTube videos, out of which almost 50% have come across inappropriate content as reported by their parents [2]. Cartoon videos are especially problematic because they are attractive to young viewers. Scammers often create junk content that includes familiar cartoon characters in order to monetize their videos [3]. Children younger than 6 years of age are more prone to viewing harmful content as they are unable to distinguish between appropriate and inappropriate. Continuous exposure to malicious video content may have adverse effects on overall cognitive development [4].

YouTube and other video hosting platforms have tried to mitigate this serious problem of content moderation using automated techniques yet have not been successful in providing robust solutions. YouTube Kids was introduced as a means to publish child-appropriate videos, but YouTube acknowledges in [5] that they are unable to screen all malicious videos.

Supervised approaches to video content moderation rely on the existence of hand labeled datasets [6], [7]. These datasets are typically small and go out of date quickly as scammers mimic newly trending content. However, impressive zero shot performance on a variety of computer vision tasks such as object detection [8], action recognition [9], and depth estimation [10] has been achieved by joint vision-language models, such as CLIP (Contrastive Language-Image Pre training) [8]. This paper presents an evaluation of the performance of several CLIP variations on the Malicious or Benign benchmark for children's content moderation of cartoon videos [7]. The Malicious or Benign dataset includes challenging cases of malicious junk content videos that were designed to mimic popular educational cartoons. These videos lack explicit obscenity or violence but contain loud sounds, fast motions, and characters with scary appearances. Thus it is challenging to devise prompts that are sufficiently general to describe the disturbing activities. Reynolds and McDonell [11] present an overview of prompt programming strategies for large language models that illustrates how different prompting strategies can yield very different outcomes on downstream tasks. This paper makes the following contributions:

- Introduces a model (Vanilla CLIP with Projection Layer) that outperforms previous content moderation techniques on the MOB benchmark [7];
- Performs an in-depth analysis of how context-specific language prompts affect the content moderation performance of different CLIP variants;
- Proposes new benchmark prompt templates for the MOB Dataset.

II. Related Work

Previous work on the application of machine learning for video content moderation for children can be categorized into approaches that focus on a single modality including meta-data [12], video [13], [14], or text (user comments) [15] vs. approaches that leverage multimodal data [6], [16]. The Samba system [6] uses subtitles and meta-data to classify inappropriate videos, and Chuttur et al. [16] combine both user comments and images to detect inappropriate videos.

Natural language supervision models such as CLIP [8] and ALIGN [17] combine video and text in a different way than

1946-0759/23/$31.00 ©2023 IEEE
DOI 10.1109/ICMLA58977.2023.00186

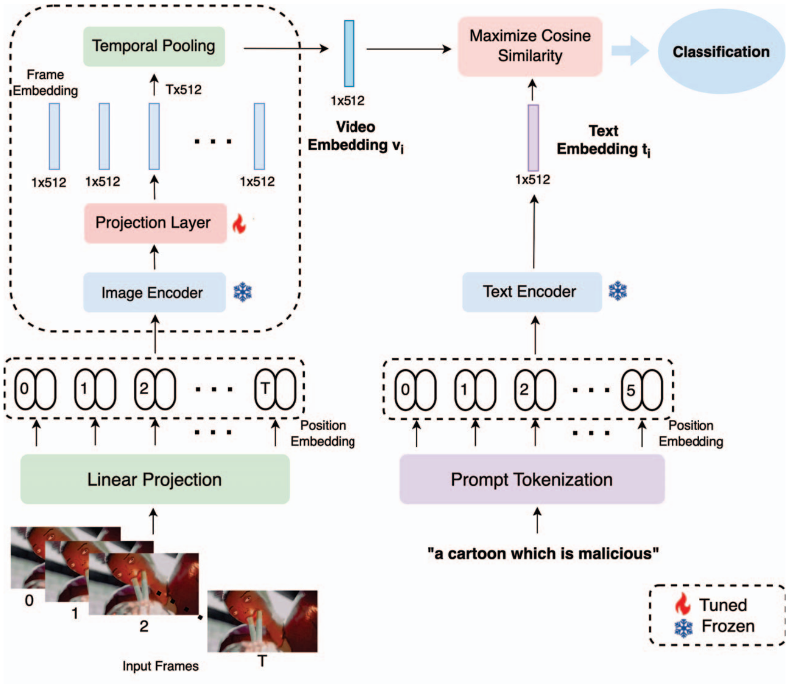

Fig. 1. Proposed architecture for language-assisted supervision of video content moderation. The image encoder produces frame-level embeddings with dimensions 1 x 512 after which Temporal Pooling finds the joint representation of all frames of the input video v_i,. Similarly, the encoder in the language branch of our model produces text embeddings, t_i after performing tokenization and positional encoding of the input prompt. Finally, cosine similarity is calculated between each (v_i, t_i) pair which maximizes the similarity for the video for the class for which the related prompt was learnt.

multimodal models. Rather than using aligned videos and subtitles as the classification input, a natural language prompt is used to describe the instance. The supervised learning process aims to maximize the similarity between the video embedding and prompt text embedding (Figure 1), rather than seeking to minimize classification loss. Zero-shot video classification approaches extract visual features in the different frames of a video using a backbone such as Convolutional 3D (C3D) [18], ResNet [19] or ViT [20], before learning a temporal behavior that maps visual representations to semantic embeddings.

Ahmed et al. [7] published a hand labeled dataset and a set of video benchmarks for classifying children's cartoon videos as malicious or benign. We demonstrate that our proposed method outperforms these benchmarks on the supervised learning content moderation task. Unlike other datasets, the MOB dataset includes malicious content that parodies children's educational videos. The dataset includes a list of audio and video features (summarized in Table I) that were used to characterize junk videos. In this paper we utilize these feature labels for our prompt generation.

III. Methodology

Figure 1 shows our general architecture for video content moderation. Videos are classified as malicious or benign based on a short time window of input frames. We compare the performance of different video embedding and prompt generation strategies for both the supervised and zero shot setting. This paper evaluates the performance of four variants of vision-language pretraining:

TABLE I
Malicious Video and Audio features identified in [7]

Video	**fast repetitive motions, scary/disgusting appearance, hurting/destruction/killing activity, obscene/indecent activity**
Audio	**loud music/noise, screaming or shouting, explosion or gunshot sounds, offensive language**

1) **Vanilla CLIP [8]** The simplest way to adapt the CLIP model for video classification is to use temporal pooling to create a combined representation of the T frames input to the model.
2) **ViFi CLIP [21]** Since CLIP was designed and trained for static images, the temporal representations and object interactions in videos are not explicitly modelled. To address this, ViFi simply performs video fine-tuning on the pre-trained CLIP model. Temporal pooling is used to aggregate representation across frames.
3) **AIM-CLIP [22]** AIM leverages the CLIP model's video encoders through fully trainable lightweight blocks called *adapters*. Thus, AIM trains on a substantially lower number of parameters. Normally AIM only uses the vision branch of the CLIP model to fine-tune its adapters on video datasets. In this paper, we incorporated the language transformer in the AIM model and then use it

to classify malicious or benign cartoon videos.

4) **ActionCLIP [9]** ActionCLIP was specifically designed for action recognition. It utilizes CLIP for video-text input and makes use of a multi-level prompt strategy.

A. *Supervised Learning*

Since full fine-tuning on videos is computationally expensive, we simply add a projection layer on top of the CLIP model while freezing the latter. Projection layers have been used by various models both in fine tuning scenarios and also for dimensionality reduction [23], [24], [25]. This helps the model adapt to the downstream task better, as only task-specific parameters are learnt by this layer. Our projection layer has 768 input nodes and 512 output nodes and connects on top of the image encoder which has 768 output nodes. We use ViT [20] which uses image patches of size 16 x 16; hence for 3 channels a patch can be fully represented by a vector of size $16 \times 16 \times 3 = 768$. We freeze the image and text encoders and leave the projection layer open for tuning as illustrated in Figure 1.

B. *Zero-shot Classification*

We also performed a zero-shot evaluation on different CLIP-based models. With the exception of Vanilla CLIP, we use respective models pre-trained on the Kinetics-400 dataset [26].

C. *Prompt Engineering and Ensembling*

Modifying the input text prompts can have a significant effect on CLIP's performance [8], [11]. After performing rigorous prompt engineering we identified seven prompt templates that empirically performed better than other text prompts during the trial-and-error process.

1) Adding cartoon context to default prompt: The standard prompt used in CLIP is *"a photo of a { }."*. As the reference dataset is cartoon only, based on this intuition we added the token *"cartoon"* and devised two new prompts: 1) *"a { } cartoon."* and 2) *"a photo of a { } cartoon."*.

2) Prompt candidate generation with cartoon context tokens: A typical prompt used in CLIP is "a photo of a { }". As discussed in Section III-C1, we add the token "cartoon" so that the tokens become context-specific as the reference dataset is of cartoons only. A general prompt format after adding context becomes: "a *clip-token* of a *context-token*". From the prompts published for other datasets we enumerated an initial list of clip-tokens and another list of context-tokens from the cartoon synonyms. Based on each prompt generated, we perform zero-shot performance analysis and selected the tokens contributing to the top performing prompts, to create a candidate list of tokens. The initial and final lists are shown in Table II.

The following formats were used to generate the prompt templates:

prompt formats:
'a *clip-tokens*$_i$ of a {} *context-tokens*$_j$.',
'a *clip-tokens*$_i$ of a *context-tokens*$_j$ which is {}.'
'a *context-tokens*$_j$ which is {}.'

TABLE II
INITIAL AND FINAL CANDIDATE PROMPT TEMPLATE TOKEN LIST USED TO GENERATE CONTEXT-SPECIFIC PROMPTS

clip-tokens (initial)	'photo', 'video', 'example', 'demonstration', 'image'
context-tokens (initial)	'cartoon', 'animation', 'caricature', 'comic', 'character'
clip-tokens (candidate)	'image', 'example'
context-tokens (candidate)	'cartoon', 'caricature', 'comic'

The prompt templates are generated by placing all pairs (*clip-tokens*$_i$, *context-tokens*$_j$) in the respective placeholders of each of the three prompt formats.

3) Prompt generation using all combinations of cartoon features: The MOB dataset includes information about the presence or absence of the malicious features shown in Table I. Similarly, we devise a list of malicious and benign feature tokens as follows:

malicious feature tokens: [”fast-moving”, ”scary”, ”disgusting”, ”hurting”, ”destructive”, ”killing”, ”obscene”, ”indecent”]

benign feature tokens: [”good”, ”friendly”, ”happy”, ”joyful”, ”singing”, ”enjoying”, ”loving”, ”caring”, ”playing”, ”funny”]

These lists form a third list of tokens and a candidate list is generated similar to the algorithm described in Section III-C2. The prompt format used is:
'a *clip-tokens*$_i$ of a {} *context-tokens*$_j$ which is {} and *feature-tokens*$_k$.'

4) Prompt candidate generation based top-performing frequent token pairs: The zero shot evaluation of the exhaustive prompt templates generated from the initial lists described in Section III-C2 is performed on all pairs (*clip-tokens*$_i$, *context-tokens*$_j$). The prompt templates with higher than the median performance score (accuracy) are formulated as a list of items in a transaction and given as input to the Apriori rule association algorithm [27] which finds the frequent itemsets of size 2. All those frequent pairs become candidate pairs which are then used to generate all prompt templates using the prompt formats discussed in Secion III-C2.

5) Prompts of other relevant datasets: We also performed a performance evaluation for the MOB dataset [7] using the exact prompt templates as published for other datasets. We evaluated prompts of UCF101 [28], CIFAR10 and CIFAR100 [29], and FER2013 (FacialEmotionRecognition2013) dataset [30]. FER2013 gave the best performance for specific prompt templates for some of the benchmarks.

IV. EVALUATION

This section presents results of our evaluation vs. the MOB benchmarks [7]. The MOB dataset is annotated with domain-specific features that commonly occur in inappropriate children's cartoon videos. It includes 1875 clips, each 10 seconds in length and frame rate = 25 fps. The dataset has two classes: malicious and benign.

For fair comparison, we used the ViT-B/16 backbone for all CLIP variants discussed in this paper. ViT-B/16 refers to

TABLE III

Supervised setting: The columns show the different prompts for which we ran our experiments with the benchmarks on the MOB Dataset. The prompts for the frequent item-set combinations outperform other prompts.

Model	**Text Prompt Templates**						
	’a photo of a {}.’	**’a {} cartoon.’**	**’a photo of a {} cartoon.’**	**clip+context token pair combinations**	**feature-based token combinations**	**frequent item-set combinations**	**FER-2013**
Vanilla (PL)	78.9	78.1	77.8	76.3	75.2	**80.3**	78.5
ViFi-CLIP (PL)	71.0	71.3	69.9	69.2	71.3	**71.7**	**71.7**
AIM-CLIP (PL)	68.5	65.5	67.7	66.7	66.3	**69.9**	67.7

TABLE IV

Zero-shot setting: The prompts using multiple combinations of context tokens give the best zero-shot classification score.

Model	**Text Prompt Templates**						
	’a photo of a {}.’	**’a {} cartoon.’**	**’a photo of a {} cartoon.’**	**clip+context token pair combinations**	**feature-based token combinations**	**frequent item-set combinations**	**FER-2013**
Vanilla	62	65.6	67	**68.5**	67.7	67	66.3
ViFi-CLIP	53.4	57.3	**58.8**	58.4	58.1	53.4	56.3
AIM-CLIP	**64.2**	60.9	**64.2**	62.0	54.8	60.9	60.9
ActionCLIP	56.6	54.8	54.8	55.6	53.8	**60.9**	**60.9**

ViT’s Base model with an image input resolution of 224 x 224, patch sizes 16 x 16, 12 layers, 12 attention heads, and 12 layers and 8 heads for text input. In the supervised setting, training was performed for 20 epochs with a batch size of 16 and the number of frames set to 16. Adam optimizer with learning rate 1e-4 was used for all experiments. Experiments were run on a 64-bit system with NVIDIA GeForce RTX 3090 GPU and 12th Gen Intel Core i7 CPU with 64 GB of memory.

A. Supervised Learning

In this section we present results of our proposed supervised language pretraining setting where we use a learnable projection layer to classify videos as malicious or benign. Table III presents the overview of results for supervised classification with different prompting strategies. We trained and tested three CLIP-based models, out of which the best performing benchmark was our adapted Vanilla CLIP which is labelled with the suffix *(PL)*. ViFi-CLIP (fine-tuned on the Kinetics-400 [26] video dataset) performed the second-best while AIM-CLIP, our adaptation of AIM for CLIP, stood last in terms of accuracy. The best overall supervised classification results for the MOB Dataset are achieved by our proposed Vanilla-CLIP PL with a testing accuracy of 80.3% which beat all previously published benchmarks from [7]. Table V provides a summary of the comparison vs. the other benchmarks.

B. Prompt Generation Strategies

The columns in our tables show the results of different prompt generation strategies. In all our tables, *clip+context token pair combinations* refers to the prompts discussed in Section III-C2. *Feature-based token combinations* refers to prompts explained in Section III-C3, and *frequent item-set combinations* refers to text prompt templates generated through the method described in Section III-C4. Interestingly for all these three benchmarks, the prompt templates generated using the frequent item-set combinations approach give the best results on supervised learning.

We evaluated the zero-shot performance of different prompt generation strategies in Table IV. Although there is no single dominant prompting strategy, the context-based prompts show better performance during zero-shot inference. The best prompt template is the set of prompts generated using both, clip and cartoon context tokens. For both settings, the top results for each benchmark include prompt templates generated by the techniques which use cartoon context-based tokens. This is important since the initial CLIP dataset primarily contains natural images rather than cartoons.

TABLE V

Supervised setting vs. other benchmarks

Benchmark	**Accuracy (%)**
VTN [31]	77.9
I3D [32]	72.1
ConvLSTM [33]	69.7
Vanilla (PL)	**80.3**

V. Conclusion and Future Work

In this paper we discussed how the problem of video content moderation for children of ages 1-5 years can be addressed using state-of-the-art language supervision techniques. We explored the usage of a foundation model for language pre-training, CLIP (Contrastive Language–Image Pre-training), and performed evaluations on several CLIP variants in supervised and zero-shot settings. In supervised settings, we employed projection layers to improve training on the cartoon video dataset. We also highlighted how language prompts which include context-specific tokens can affect performance on different video classification models. Lastly, we propose the benchmark prompt templates for MOB Dataset for training and evaluating joint vision-language models like CLIP.

We believe that less well defined video analysis problems such as content moderation pose a significant challenge to prompting strategies. It is challenging for humans to verbally define what makes a video inappropriate. There is a famous quote about pornography by a Supreme Court Justice: "I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description, and perhaps I could never succeed in intelligibly doing so. But I know it when I see it." [34] In the future we plan to apply prompt learning approaches where the prompts are learnable parameters, and also explore the performance on cartoon datasets when fine-tuning is done on other pre-trained video datasets.

REFERENCES

[1] D. Donchev, "40 Mind Blowing YouTube Facts, Figures and Statistics – 2023," accessed 2023-07-24. [Online]. Available: https://fortunelords.com/youtube-statistics/

[2] B. Auxier, "Parenting children in the age of screens," Jul 2020. [Online]. Available: https://www.pewresearch.org/internet/2020/07/28/parenting-children-in-the-age-of-screens/

[3] D. D. Placido, "YouTube's "Elsagate" illuminates the unintended horrors of the digital age," Forbes, 11 2017.

[4] K. Habib and T. Soliman, "Cartoons' effect in changing children mental response and behavior," *Open Journal of Social Sciences*, vol. 03, no. 09, p. 248–264, Jan 2015. [Online]. Available: https://doi.org/10.4236/jss.2015.39033

[5] "Important info for parents about YouTube Kids," accessed 2023-07-25. [Online]. Available: https://support.google.com/youtubekids/answer/6130561

[6] L. Binh, R. Tandon, C. Oinar, J. Liu, U. Durairaj, J. Guo, S. Zahabizadeh, S. Ilango, J. Tang, F. Morstatter, S. Woo, and J. Mirkovic, "Samba: Identifying inappropriate videos for young children on YouTube," in *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2022, p. 88–97.

[7] S. H. Ahmed, M. J. Khan, H. M. U. Qaisar, and G. Sukthankar, "Malicious or Benign? Towards Effective Content Moderation for Children's Videos," in *Proceedings of the International FLAIRS Conference*, vol. 36, 2023.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[9] M. Wang, J. Xing, and Y. Liu, "ActionCLIP: A new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.

[10] R. Zhang, Z. Zeng, Z. Guo, and Y. Li, "Can language understand depth?" in *Proceedings of the ACM International Conference on Multimedia*, 2022, pp. 6868–6874.

[11] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," 2021.

[12] M. Gkolemi, P. Papadopoulos, E. Markatos, and N. Kourtellis, "YouTubers Not MadeForKids: Detecting Channels Sharing Inappropriate Videos Targeting Children," in *ACM Web Science Conference*, 2022, p. 370–381.

[13] K. Yousaf and T. Nawaz, "A deep learning-based approach for inappropriate content detection and classification of YouTube videos," *IEEE Access*, vol. 10, p. 16283–16298, 2022.

[14] S. Singh, R. Kaushal, A. B. Buduru, and P. Kumaraguru, "Kidsguard: fine grained approach for child unsafe video representation and detection," in *Proceedings of the ACM/SIGAPP Symposium on Applied Computing*, 2019, p. 2104–2111.

[15] S. Alshamrani, A. Abusnaina, M. Abuhamad, D. Nyang, and D. Mohaisen, "Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in YouTube," in *Companion Proceedings of the Web Conference*. Association for Computing Machinery, 2021, p. 508–515.

[16] M. Y. Chuttur and A. Nazurally, "A multi-modal approach to detect inappropriate cartoon video contents using deep learning networks," *Multimedia Tools and Applications*, vol. 81, no. 12, p. 16881–16900, May 2022.

[17] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," 2021.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan, "Fine-tuned CLIP models are efficient video learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6545–6554.

[22] T. Yang, Y. Zhu, Y. Xie, A. Zhang, C. Chen, and M. Li, "AIM: Adapting image models for efficient video action recognition," in *The International Conference on Learning Representations*, 2023.

[23] P. Kaliamoorthi, S. Ravi, and Z. Kozareva, "Prado: Projection attention networks for document classification on-device," in *Conference on Empirical Methods in Natural Language Processing*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202785536

[24] C. Sankar, S. Ravi, and Z. Kozareva, "Proformer: Towards on-device lsh projection based transformers," *arXiv preprint arXiv:2004.05801*, 2020.

[25] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Advances in neural information processing systems*, vol. 26, 2013.

[26] W. Kay, J. Carreira, K. Simonyan, and B. Zhang, "The Kinetics Human Action Video Dataset," 2017.

[27] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc of VLDB*, vol. 1215. Santiago, Chile, 1994, pp. 487–499.

[28] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," 2012.

[29] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[30] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*, M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, Eds., 2013, pp. 117–124.

[31] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3163–3172.

[32] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[33] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[34] P. Lattman, "The origins of Justice Stewart's I know it when I see it," 2007. [Online]. Available: https://www.wsj.com/articles/BL-LB-4558