TEXT-IMAGE DE-CONTEXTUALIZATION DETECTION USING VISION-LANGUAGE MODELS

Mingzhen Huang[⋆] *Shan Jia*[⋆] *Ming-Ching Chang*[†] *Siwei Lyu*[⋆]

[⋆]University at Buffalo, State University of New York, NY, USA
[†]University at Albany, State University of New York, NY, USA

ABSTRACT

Text-image de-contextualization, which uses inconsistent image-text pairs, is an emerging form of misinformation and drawing increasing attention due to the great threat to information authenticity. With real content but semantic mismatch in multiple modalities, the detection of de-contextualization is a challenging problem in media forensics. Inspired by the recent advances in vision-language models with powerful relationship learning between images and texts, 0

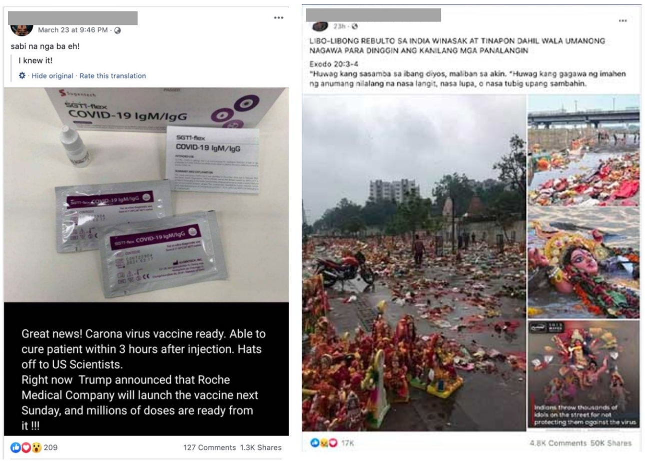
Index Terms— de-contextualization, online misinformation, out-of-text detection, text-image inconsistency

1. INTRODUCTION

The wide spread of online misinformation is a vexing threat to the information ecosystem and can greatly erode the public's trust of online information. As a vehicle to spread rumors and false information, misinformation can lead to many negative social impacts, such as misleading information, hate speech, racism, and psychological distress [1]. One widely used form of online misinformation is known as *de-contextualization*, where texts and images from different and/or unrelated contexts are composited together to generate false impressions. Fig. 1 shows two examples of real-life misinformation based on text-image de-contextualization.

There are many existing methods in media forensics that can be used to detect if images or texts are AI-generated or manipulated, *e.g.*, [2, 3, 4, 5, 6]. These methods work by exposing the signal-level inconsistency in the falsified texts or images, and can be used to detect de-contextualization when the texts or images are manipulated. However, when both the text and image are authentic but grafted to fabricate a new context, existing media forensics methods are not applicable, because the text or image itself in question does not have signal-level inconsistency, but there exists *semantic inconsistency* between them.

Currently, fact-checking by human operators offers the most effective solution to expose text-image de-contextualization. However, the process is expensive, exhaustive, time-consuming, and prone to errors [7]. On the other hand, it is desirable to develop machine learning based algorithms as an alternative solution for automatic detection of text-image de-contextualization.



[redacted]
March 23 at 9:46 PM ·
sabi na nga ba eh!
I knew it!

Great news! Carona virus vaccine ready. Able to cure patient within 3 hours after injection. Hats off to US Scientists.
Right now Trump announced that Roche Medical Company will launch the vaccine next Sunday, and millions of doses are ready from it !!!

[redacted]
23h ·
LIBO-LIBONG REBULTO SA INDIA WINASAK AT TINAPON DAHIL WALA UMANONG NAGAWA PARA DINGGIN ANG KANILANG MGA PANALANGIN
Exodo 20:3-4
"Huwag kang sasamba sa ibang diyos, maliban sa akin. "Huwag kang gagawa ng imahen ng anumang nilalang na nasa langit, nasa lupa, o nasa tubig upang sambahin.

Indians throw thousands of idols on the street for not protecting them against the virus

(a) (b)

Fig. 1. *Examples of misinformation with de-contextualization on social media. (a) A photo of rapid test kits from a South Korean company Sugentech was posted in Facebook to falsely claim that a new vaccine that can cure COVID-19 patients in 3 hours is now ready in March, 2021. (b) Photos taken during celebrations of Hindu festivals in India were posted to falsely claim that Indians were throwing away the figures for failing to protect them against COVID-19 in May, 2021.*

Several studies for image-text inconsistency detection, *e.g.*, MAIM [8], COSMOS [9], NewsCLIPpings [10], have been developed in the recent years and shown promising performance on benchmark datasets. Despite this progress, existing methods have the following limitations: 1) lack of relevance to real-life scenarios as most methods are evaluated on datasets with random mismatches; 2) lack of generalization as the evaluations are based on a single dataset; 3) lack of interpretability in showing how the detection model works. Motivated by the promising performance of vision-language models developed for other computer vision tasks such as VQA, annotation, and captioning, we conduct a comprehensive evaluation of *two* state-of-the-art models in identifying different kinds of image-text mismatches in de-contextualization, including the CLIP model [11] and the VinVL model [12] over three datasets. Our contributions include the following:

- We conduct a comprehensive evaluation of CLIP and VinVL vision-language models on three well-designed datasets for multimodal de-contextualization detection. Both intra-dataset and cross-dataset testing scenarios are considered.
- We visualize the detection results and summarize observations and insights in applying vision-language models for semantic inconsistency detection.

2. RELATED WORKS

Jaiswal et al. [8] present the first study on verifying the semantic integrity of multimedia. They propose the concept of a multimedia package, which contains an untampered image and related metadata. Based on the image-caption pairs in the packages, they employ deep multimodal representation learning models for jointly encoding images and captions from the untampered multi-media and perform anomaly detection on representations of query packages. A dataset named MAIM is created to include over 239K image-caption pairs with randomly mismatched falsified media. In [13], a deep multitask learning model is designed for image-repurposing detection, in which the image is authentic but the accompanying metadata has been manipulated. The overall method involves retrieval of one related multimedia package from a reference dataset first, followed by the comparison of the query package (with image, text, and GPS) to the retrieved one to determine the likelihood of manipulation. This method achieves an AUC of 0.88 on their MEIR dataset with swaps over named entities for people, organizations and locations.

Recently, McCrae et al. [14] focus on detecting semantic inconsistency between videos and captions in social media posts. They create a video-based dataset containing $4,000$ real-world Facebook news posts, and randomly swap in news captions from other posts to generate mismatched samples. Based on a multimodal fusion framework, they achieved 60.05% classification accuracy. Aneja et al. [9] create a large-scale dataset named COSMOS with 200K images and 450K captions, where each image is associated with two captions from two sources. They also use random-chosen text to generate mismatched image-caption pairs. Their self-supervised learning based scheme achieves a 85% out-of-context detection accuracy on the COSMOS dataset. Different from previous works using random mismatch for the inconsistency samples, Luo et al. [10] extend VisualNews [15] dataset to a large-scale automatically generated dataset named NewsCLIPpings, which contains 988K image-caption pairs for news media mismatch detection. Several strategies are considered for automatic retrieval of the suitable images for the given captions, including the caption-image similarity, caption-caption similarity, person match, and scene match, to capture cases with inconsistent entities or inconsistent semantic context. They evaluate two vision-language models, namely CLIP [11] and VisualBERT [16], on the proposed dataset, and achieve the classification accuracy of 60.23% and 54.81%, respectively.

3. VISION-LANGUAGE MODELS

This section presents how we apply the two vision-language models, namely CLIP and VinVL, to the multimedia inconsistency detection task. The motivations of using these two models are two-fold: (1) CLIP [11] has been pre-trained in a super scale corpus for image-text matching task. (2) VinVL [12] considers visual semantic information via an object detector and learns generic image-text representations.

CLIP. The CLIP model [11] aligns features from different modalities (text and image) and aims to minimize the semantic gap for a pair of image and the text (caption). Pre-trained on a very large dataset with 400 million image-caption pairs collected from the internet, the CLIP model enjoys a strong robustness and generalization ability to a variety of multimodal downstream tasks, demonstrating competitive performance with a fully supervised baseline without the need for any dataset specific training. Given a pair of image and text, the CLIP model will output their cosine similarity, which can be used directly for the inconsistency measurement in the de-contextualization detection task. Specifically, we fine-tune the CLIP model on the image-text inconsistency datasets with a cross-entropy loss, which leads to the best performance for the binary classification task.

VinVL. The VinVL model [12] employs a detector to extract discrete information from images and pair them to the object names. Taken the given captions, the visual features of detected objects, and predicted object names as a triplet of input, the VinVL model is optimized by the contrastive loss and masked token loss for the vision language (VL) tasks. Pre-trained on a large corpus including COCO image-text retrieval [17], Conceptual Captions [18], SBU captions [19] and Flicker30k [20], VinVL can generate rich representations of visual objects, attributes, concepts, and therefore achieving a good generalization ability cross different tasks, *e.g.* VQA, GQA, image captioning, and image-text retrieval.

For the de-contextualization detection task, we finetune the VinVL model on the image-text inconsistency dataset with the same binary classification loss as we do for the CLIP model. Due to the rich visual objects that can be detected by the VinVL, we further build an additional image-text attention layer in VinVL. This not only guides the model to better focus on the the more important features for inconsistency detection, it also contributes to the performance interpretability, which shows different degrees of feature importance.

4. EXPERIMENTS

4.1. Experimental Setup

For a comprehensive performance evaluation, we compare the CLIP and VinVL models on three well-designed datasets with image-text semantic inconsistency.

NewsCLIPpings [10] dataset is a large-scale automatic retrieved image-text pairs from four news agencies: The

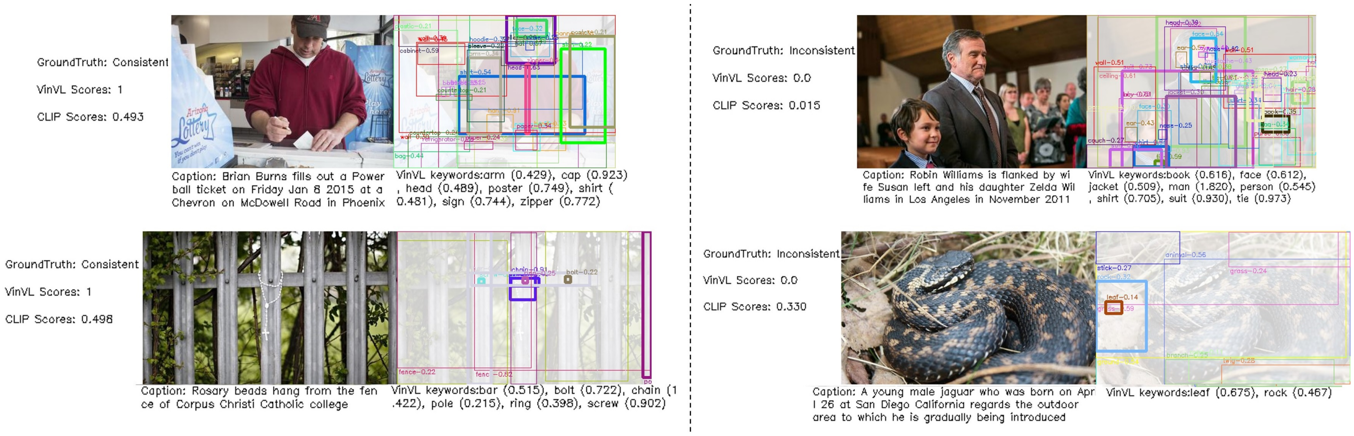
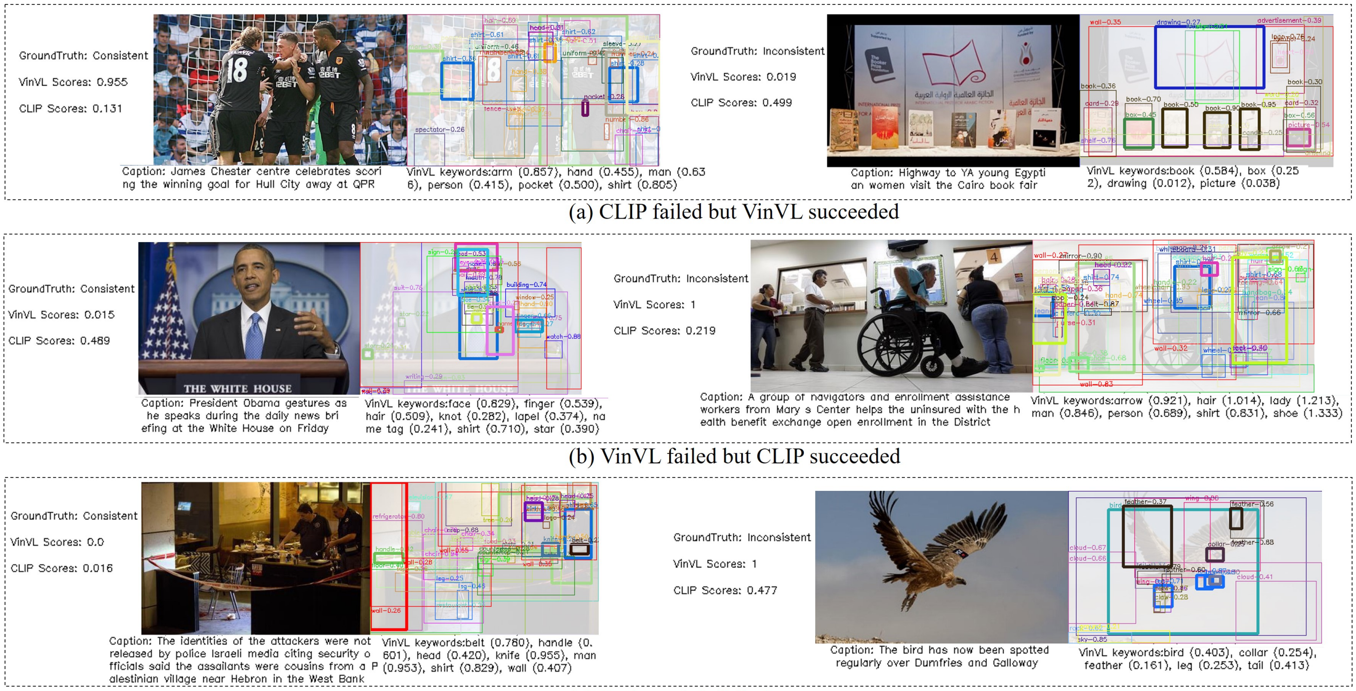

GroundTruth: Consistent
VinVL Scores: 1
CLIP Scores: 0.493
Caption: Brian Burns fills out a Power ball ticket on Friday Jan 8 2015 at a Chevron on McDowell Road in Phoenix
VinVL keywords:arm (0.429), cap (0.923), head (0.489), poster (0.749), shirt (0.481), sign (0.744), zipper (0.772)
GroundTruth: Inconsistent
VinVL Scores: 0.0
CLIP Scores: 0.015
Caption: Robin Williams is flanked by wife Susan left and his daughter Zelda Williams in Los Angeles in November 2011
VinVL keywords:book (0.616), face (0.612), jacket (0.509), man (1.820), person (0.545), shirt (0.705), suit (0.930), tie (0.973)
GroundTruth: Consistent
VinVL Scores: 1
CLIP Scores: 0.498
Caption: Rosary beads hang from the fence of Corpus Christi Catholic college
VinVL keywords:bar (0.515), bolt (0.722), chain (1.422), pole (0.215), ring (0.398), screw (0.902)
GroundTruth: Inconsistent
VinVL Scores: 0.0
CLIP Scores: 0.330
Caption: A young male jaguar who was born on April 26 at San Diego California regards the outdoor area to which he is gradually being introduced
VinVL keywords:leaf (0.675), rock (0.467)

Fig. 2. *Visualization examples of successful de-contextualization detection cases of CLIP and VinVL. Note that the larger the scores of both models, the more consistent of image-text pairs will be. For the CLIP model, the score threshold is set as 0.390, while for VinVL, the threshold is 0.035. A larger number in the bounding boxes means a higher confidence score of object tag detection, and a larger weight in the detected keywords means more important to get the detection results.*


GroundTruth: Consistent
VinVL Scores: 0.955
CLIP Scores: 0.131
Caption: James Chester centre celebrates scoring the winning goal for Hull City away at QPR
VinVL keywords:arm (0.857), hand (0.455), man (0.636), person (0.415), pocket (0.500), shirt (0.805)
GroundTruth: Inconsistent
VinVL Scores: 0.019
CLIP Scores: 0.499
Caption: Highway to YA young Egyptian women visit the Cairo book fair
VinVL keywords:book (0.584), box (0.252), drawing (0.012), picture (0.038)
GroundTruth: Consistent
VinVL Scores: 0.015
CLIP Scores: 0.489
Caption: President Obama gestures as he speaks during the daily news briefing at the White House on Friday
VinVL keywords:face (0.829), finger (0.539), hair (0.509), knot (0.282), lapel (0.374), name tag (0.241), shirt (0.710), star (0.390)
GroundTruth: Inconsistent
VinVL Scores: 1
CLIP Scores: 0.219
Caption: A group of navigators and enrollment assistance workers from Mary s Center helps the uninsured with the health benefit exchange open enrollment in the District
VinVL keywords:arrow (0.921), hair (1.014), lady (1.213), man (0.846), person (0.689), shirt (0.831), shoe (1.333)
GroundTruth: Consistent
VinVL Scores: 0.0
CLIP Scores: 0.016
Caption: The identities of the attackers were not released by police Israeli media citing security officials said the assailants were cousins from a Palestinian village near Hebron in the West Bank
VinVL keywords:belt (0.780), handle (0.601), head (0.420), knife (0.955), man (0.953), shirt (0.829), wall (0.407)
GroundTruth: Inconsistent
VinVL Scores: 1
CLIP Scores: 0.477
Caption: The bird has now been spotted regularly over Dumfries and Galloway
VinVL keywords:bird (0.403), collar (0.254), feather (0.161), leg (0.253), tail (0.413)

(a) CLIP failed but VinVL succeeded

(b) VinVL failed but CLIP succeeded

(c) Both models failed

Fig. 3. *Visualization examples of failed de-contextualization detection cases of CLIP and VinVL.*

Guardian, BBC, USA Today, and The Washington Post. We report results on the Merged/Balanced set, which contains 85K balanced positive and negative samples. This dataset is divided into training, validation, and test subsets in the ratio of 10:1:1. The VinVL model was finetuned for 30 epochs and the CLIP for 10 epochs, both with a learning rate of 0.00002.

DARPA SemaFor Eval#1 benchmark (shortened to Eval#1) is a private dataset collected by Defense Advanced Research Projects Agency (DARPA) for the research of multimodal mismatch detection. It contains 1K images-article pairs only for testing (700 positive pairs and 300 negative pairs).

DARPA SemaFor HK2-CP5 benchmark (shortened to CP5) is also a private dataset collected by DARPA. Collected from twitter and some other social media, this dataset contains 2.6K images-caption pairs (2.2K pairs for training and 400 for testing). It consists mostly of sightings of airplanes but also contains other content such as ships, war zone images, etc. This dataset provides well designed image-caption mismatches, such as different aircraft type description from the image, inconsistent country in the text with image, etc. We finetune the VinVL model for 20 epochs and CLIP for 5 epochs, both with a learning rate of 0.00002 on this dataset.

Table 1. *Detection performance (%) on NewsCLIPpings*

Method	EER	ACC	AUC	FAR	FRR
VisualBERT	-	54.8	-	54.9	35.4
CLIP	37.2	62.6	67.2	37.3	37.3
Ours	**34.2**	**65.4**	**71.9**	**34.2**	**34.2**

4.2. Quantitative Results

We first compare the intra-dataset detection performance of the finetuned CLIP and VinVL models on the large-scale NewsCLIPpings dataset. Table 1 shows the performance, as well as compared with the VisualBERT model [10]. VinVL achieves better detection performance than the other two models. This can be attributed to that the NewsCLIPpings dataset contains highly-diverse data, where the VinVL model benefits from its rich representation of multiple visual objects and attributes in the image-text pairs.

To show the robustness of the VinVL and CLIP models in detecting different image-text inconsistencies, we compared the performance of these two models under the cross-dataset testing scenarios. Both finetuned on the NewsCLIPpings dataset, the two models demonstrate different generalizability in detecting de-contextualization on Eval#1 and CP5 data in Table 2. The CLIP model achieves higher accuracy and lower error rates on both datasets over VinVL, demonstrating stronger generalization ability in detecting unknown image-text inconsistencies.

We further compare the two models under different training schemes to demonstrate the influence of transfer learning on the detection performance. Table 3 first shows that the CLIP model outperforms the VinVL model on the CP5 dataset. We conjecture this is due to that the multimodal data in CP5 are mostly airplanes and similar others with a single object and simple background. Therefore, the CLIP model trained on similar but rich data demonstrated its superiority. Furthermore, the training scheme has a greater influence on the VinVL model, which is originally proposed for text-image retrieval tasks. Therefore finetuning on the de-contextualization data helps transfer the model to the new task. Pre-training on the CP5 data further improves the detection performance for both models.

4.3. Qualitative Results

To have a better understanding of how the vision-language models detect image-text inconsistencies, we also show visual results for successful cases in Fig. 2 and for failure cases in Fig. 3 on the NewsCLIPpings dataset. We show the detected objects in the VinVL model and their corresponding contributions to the final classification. The CLIP model does not afford such visualization. In general, VinVL tends to perform better on images with rich content, while CLIP works better for simple content samples. However, the ranges of scores of the two models differ significantly. This indicates that a simple combination of the two model outputs is unlikely to significantly improve the overall performance. Furthermore, we notice that the object labels in VinVL tend to have more generic semantic meanings (*e.g.*, person, man, animal) but the texts usually include more specific terms. The model seems to be able to capture some general correlation between the specific terms and generic categories. Furthermore, these examples also exhibit some intrinsic complexities in the detection of text-image de-contextualization. The definition of consistency itself is subjective and can often be a moving target, *e.g.*, in the failure case with inconsistent groundtruth in Fig. 3 (c), both the image and text showing a bird but they are inconsistent. This suggests that special care is required when using these models and interpreting the results.

Table 2. *Cross-dataset detection performance (%)*

Method	Train	Test	EER	AUC	FAR	FRR
CLIP	NewsCLIPpings	Eval#1	**38.5**	**66.4**	**38.4**	**38.7**
	NewsCLIPpings	CP5	**38.4**	**67.7**	**38.4**	**38.4**
VinVL	NewsCLIPpings	Eval#1	41.6	62.2	41.5	41.7
	NewsCLIPpings	CP5	45.5	56.3	45.5	46.0

Table 3. *Detection performance (%) on CP5 dataset*

Method	Train	EER	AUC	FAR	FRR
CLIP	-	37.2	68.6	37.0	37.9
	NewsCLIPpings	38.4	67.7	38.4	38.4
	NewsCLIPpings+CP5	**33.7**	**73.0**	**34.1**	**33.7**
VinVL	-	53.6	46.2	53.6	54.0
	NewsCLIPpings	45.5	56.3	45.5	46.0
	NewsCLIPpings+CP5	37.7	66.5	37.0	39.8

5. CONCLUSIONS

The detection of text-image de-contextualization is a challenging problem not only lies in the multimodal semantic mismatches but also due to the immature problem definition and datasets for evaluation. This paper evaluates two powerful vision-language models for image-text inconsistency detection on three datasets. Experimental results under both intra- and cross-dataset evaluation show the effectiveness as well as performance differences of the fine-tuned vision-language models in the binary classification task. We conclude that the current vision-language models are reasonably effective in this task. However, there is much room for improvement, and our subsequent work will focus on how to combine different models to accommodate the semantic gaps between the text and image to achieve more robust and interpretable results.

6. ACKNOWLEDGEMENT

This work is supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0123. The authors would like to thank Matt Turek, Arslan Basharat, Kirill Trapeznikov, and Sam Blazek in the project for providing data for our evaluation.

7. REFERENCES

[1] Sana Ali, “Combatting against covid-19 & misinformation: A systematic review,” *Human Arenas*, pp. 1–16, 2020.

[2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, “Fake news detection on social media: A data mining perspective,” *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[3] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva, “Detection of gan-generated fake images over social networks,” in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2018, pp. 384–389.

[4] Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay, “The limitations of stylometry for detecting machine-generated fake news,” *Computational Linguistics*, vol. 46, no. 2, pp. 499–510, 2020.

[5] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush, “Gltr: Statistical detection and visualization of generated text,” *arXiv preprint arXiv:1906.04043*, 2019.

[6] Reuben Tan, Bryan A Plummer, and Kate Saenko, “Detecting cross-modal inconsistency to defend against neural fake news,” *arXiv preprint arXiv:2009.07698*, 2020.

[7] Maria D Molina, S Shyam Sundar, Thai Le, and Dongwon Lee, ““fake news” is not simply false information: a concept explication and taxonomy of online content,” *American behavioral scientist*, vol. 65, no. 2, pp. 180–212, 2021.

[8] Ayush Jaiswal, Ekraam Sabir, Wael AbdAlmageed, and Premkumar Natarajan, “Multimedia semantic integrity assessment using joint embedding of images and text,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1465–1471.

[9] Shivangi Aneja, Chris Bregler, and Matthias Nießner, “Cosmos: Catching out-of-context misinformation with self-supervised learning,” *arXiv preprint arXiv:2101.06278*, 2021.

[10] Grace Luo, Trevor Darrell, and Anna Rohrbach, “Newsclippings: Automatic generation of out-of-context multimodal media,” *arXiv preprint arXiv:2104.05893*, 2021.

[11] 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,” *arXiv preprint arXiv:2103.00020*, 2021.

[12] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao, “Vinvl: Revisiting visual representations in vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.

[13] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan, “Deep multimodal image-repurposing detection,” in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1337–1345.

[14] Scott McCrae, Kehan Wang, and Avideh Zakhor, “Multi-modal semantic inconsistency detection in social media news posts,” *arXiv preprint arXiv:2105.12855*, 2021.

[15] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez, “Visualnews : Benchmark and challenges in entity-aware image captioning,” 2020.

[16] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang, “Visualbert: A simple and performant baseline for vision and language,” *arXiv preprint arXiv:1908.03557*, 2019.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[18] 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)*, 2018, pp. 2556–2565.

[19] Vicente Ordonez, Girish Kulkarni, and Tamara Berg, “Im2text: Describing images using 1 million captioned photographs,” *Advances in neural information processing systems*, vol. 24, pp. 1143–1151, 2011.

[20] 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*, vol. 2, pp. 67–78, 2014.