NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media

Grace Luo **Trevor Darrell** **Anna Rohrbach**

University of California, Berkeley

{graceluo, trevordarrell, anna.rohrbach}@berkeley.edu

Abstract

Online misinformation is a prevalent societal issue, with adversaries relying on tools ranging from cheap fakes to sophisticated deep fakes. We are motivated by the threat scenario where an image is used out of context to support a certain narrative. While some prior datasets for detecting image-text inconsistency generate samples via text manipulation, we propose a dataset where both image and text are unmanipulated but *mismatched*. We introduce several strategies for automatically retrieving convincing images for a given caption, capturing cases with inconsistent entities or semantic context. Our large-scale automatically generated NewsCLIPpings Dataset: (1) demonstrates that machine-driven image repurposing is now a realistic threat, and (2) provides samples that represent challenging instances of mismatch between text and image in news that are able to mislead humans. We benchmark several state-of-the-art multimodal models on our dataset and analyze their performance across different pretraining domains and visual backbones.

1 Introduction

Misinformation has reached new heights as sophisticated AI-based tools have come into the spotlight. For instance, it has become easy to generate images of people who "do not exist"[1] and create realistic deepfakes of existing people (Victor, 2021). Recent language models have become better at fooling people into believing that generated texts are from real people (Hao, 2020). However, simple and cheap image repurposing remains one of the most widespread and effective forms of misinformation (Fazio, 2020). Specifically, real images of people and events get reappropriated and used out of context to illustrate false events and misleading narratives by misrepresenting *who* is in the image, what is the *context* in which they appear, or *where* the event takes place. This method is effective since augmenting a story with an image has been shown to increase user engagement and make false stories seem true (Fenn et al., 2019). Here, we explore whether such a threat can be *automated*. We show that real world images can be automatically matched to captions to generate false but compelling news stories, a threat scenario that may lead to larger-scale image repurposing.

(a) Taylor Swift performs at Manchester Arena on June 24 2015 in Manchester England.

(b) A large group of people take part in a march to mark the 150th anniversary of Pickett s Charge at Gettysburg National Military Park.

(c) Prince William and Duchess Kate introduce their son to the world.

(d) Taxi drivers driving vacant taxis in queues cause traffic jams on a road in Wuhan city central China as part of a protest against Uber.

Figure 1: Consider the following examples and guess whether these are pristine news or automatically matched image-caption pairs. The solution and more discussion are given in text.

While synthetic media conceptually could be detected by doing unimodal analysis (e.g. a detector for GAN-generated images), in our case both the text and the image are real. Thus, determining whether an image-caption pair is pristine or falsified requires joint multimodal analysis of the image and text (consider Figure 1 and make your guess).

Prior work has proposed several datasets related to our problem statement. One line of work obtains out-of-context image-text pairs by manipulating the named entities within the text (Müller-Budack et al., 2020; Sabir et al., 2018). We find that in practice this may lead to linguistic inconsistencies,

[1] https://thispersondoesnotexist.com

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6801–6817
November 7–11, 2021.

providing sufficient signal for a text-only model to distinguish between pristine and falsified descriptions without looking at the images. One recent work on detecting out-of-context images (Aneja et al., 2021) focuses on a scenario where an image is accompanied by two captions (from two distinct news sources), and one has to establish whether the two captions are consistent. Here, we do not manipulate textual descriptions as we aim to minimize unimodal bias in our task. We do not assume that two captions are available per image, rather we focus on classifying each image-caption pair as pristine or falsified.

Specifically, we propose a large-scale automatically constructed dataset with real and out-of-context news based on the VisualNews (Liu et al., 2020) corpus. We consider several threat scenarios, designing matches based on: (a) *caption-image similarity*, (b) *caption-caption similarity*, where we retrieve an image with similar semantics to a given caption while the named entities between the source and the target are disjoint, (c) *person match*, where we retrieve an image that depicts a person mentioned in the source caption but pictured in a different context, and (d) *scene match*, where we retrieve an image that has the same scene type as the source image but depicts a different event [2]. We use the recent powerful multimodal model CLIP (Radford et al., 2021) and other image and text models to construct the **NewsCLIPpings Dataset**. To make our dataset more challenging, we introduce an adversarial filtering technique based on CLIP.

We benchmark several state-of-the-art multimodal models and analyze their performance on the NewsCLIPpings Dataset. We investigate the impact of the pretraining domains and various visual backbones. We conduct a human evaluation that shows humans find it challenging to distinguish between pristine and falsified samples from our dataset. We also perform a qualitative analysis with the help of visual salience to shed light onto the useful cues discovered by the models trained on our dataset. Our dataset is publicly available here: https://github.com/g-luo/news_clippings [3].

[2]This answers the question in Figure 1, i.e., examples a), b), c), d) correspond to these four threat scenarios in our dataset, so all four are falsified.

[3]Specifically, we provide pristine and falsified matches for captions/images, i.e. their identifiers within the VisualNews dataset. The copyright and usage rights of the data are subject to that of (Liu et al., 2020).

2 Related Work

We review several most relevant datasets in detail.

Some earlier proposed datasets for detecting multimodal misinformation are MultimodAl Information Manipulation dataset (MAIM) (Jaiswal et al., 2017) and Multimodal Entity Image Repurposing (MEIR) (Sabir et al., 2018). MAIM naively matches images to captions from other random images to create their falsified versions. MEIR introduces swaps over named entities for people, organizations and locations. One of their assumptions is that for each image-caption "package" there is an unmanipulated related package (geographically near and semantically similar) in the reference set. This allows verifying the integrity of the query package by first retrieving a related package and then comparing the two. This problem statement is different from ours, as we do not assume availability of a perfect reference set. A more recent work has proposed TamperedNews (Müller-Budack et al., 2020), a dataset where named entities specific to people, locations and events are swapped to other random[4] named entities within the article body. We show that such text manipulations lead to significant linguistic biases and the corresponding tasks can be solved without looking at the images (See Section 4 for more details).

Another recent work (Aneja et al., 2021) aims to detect when images are used out of context, somewhat similar to MEIR above. They collect a dataset where each image appears in two distinct news sources and thus is associated with two captions. Most of the collected data is not labeled, but a small subset has been manually annotated as in- or out-of-context. Their problem statement (analyzing image and two captions) is again different from ours.

One other work (Tan et al., 2020) tackles Neural News generation by replacing real articles with Grover (Zellers et al., 2019) generated text and real captions with synthetic ones. They do not mismatch the images, which remain relevant to the article's content. The impact of image analysis on this task is rather limited, while analyzing the captions and the article body is key to the best detection performance.

Finally, some work focuses on human-made fake news detection, such as FakenewsNet (Shu et al., 2020) and Fakeddit (Nakamura et al., 2019), etc. While these datasets contain important real world

[4]With some constraints, such as individuals of the same country and gender or locations within the same region.

examples of fake news, our focus is on exploring an *automated* threat scenario, where an image is automatically retrieved to match a given caption.

3 The NewsCLIPpings Dataset

The objective of this work is to explore techniques for creating challenging, non-random image-caption matches that require fine-grained semantic and entity knowledge. As seen in Figure 2[5], misinformation in the wild is often extremely subtle and much more difficult than the random matches provided in prior synthetic datasets. In fact, general models that were not specifically trained or finetuned on the news domain can "solve" random news matches. We found that CLIP was able to achieve 97.39% Top-1 accuracy on a caption-image retrieval task with news images[6]. For comparison, a recent method TRIP (Thomas and Kovashka, 2020) reports a Top-1 accuracy of 73.78% on a similar task. As a result, we construct several splits that model specific threat scenarios seen in the real world, and we use CLIP ViT-B/32 *off-the-shelf* to filter out the less challenging samples.

In the following, we assume we have a pristine query pair (img_1, cap_1) and retrieve another pair (img_2, cap_2) to form a falsified pair (img_2, cap_1).

Preprocessing Our dataset is derived from VisualNews (Liu et al., 2020), a large-scale corpus which contains image-caption pairs from four news agencies (The Guardian, BBC, USA Today, and The Washington Post). We use spaCy NER (Honnibal and Montani, 2017) to label named entities in captions and the Radboud Entity Linker (REL) (van Hulst et al., 2020) to link them to their Wikipedia 2019 entries. We compute text embeddings using SBERT-WK (Wang and Kuo, 2020) and CLIP (Radford et al., 2021). We compute image embeddings with Faster R-CNN (Ren et al., 2015) and CLIP. We use a ResNet50 classifier trained on the Places365 dataset (Zhou et al., 2017) to get scene embeddings from images. We ensure that all matched samples are at least 30 days apart and that (cap_1, cap_2) have no overlapping named entities identified by spaCy and REL to prevent true matches, with the exception of the Person split, where we expect at least one "PERSON" entity to match.

[5]Examples found on https://www.snopes.com and https://www.politifact.com.

[6]We ran this on a random 40k subset of VisualNews and counted how often CLIP selected the true image vs. four random negative images.

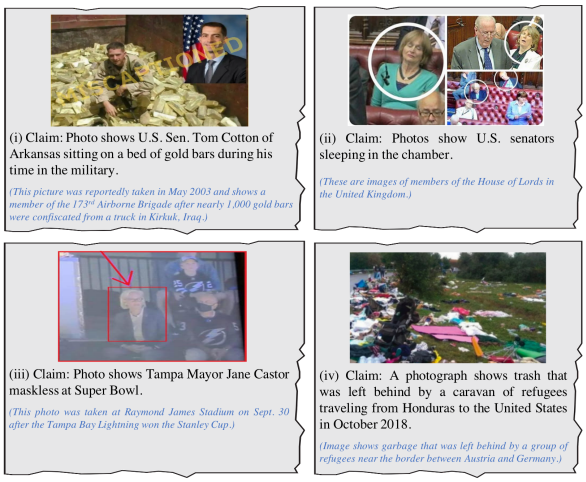


Figure 2: We are motivated by the real-world examples of images used out-of-context. Here we include *real* misinformation examples found online[5] which closely resemble the four threat scenarios in our dataset.

Query for Semantics Our first split models a threat scenario that queries for specific semantic content, with the intent to portray the subjects of the image as certain other named entities, see Figure 2 (i, ii). We consider two ways of getting the matches. **(a) CLIP Text-Image**: We rely on the state-of-the-art CLIP representation to retrieve samples with the highest CLIP text-image similarity between (img_2, cap_1). **(b) CLIP Text-Text**: We match samples with the highest CLIP text-text similarity between (cap_1, cap_2) and retrieve the corresponding img_2. See examples (a) and (b) in Figure 1.

Query for Person This split models a threat scenario that queries for a specific person, with the intent to portray them in a false context, as in Figure 2 (iii). We ensure that the person of interest is pictured: all considered samples must have "PERSON" entities in their captions and a person related Faster-RCNN bounding box detected in the image. To avoid cases where the query person is mentioned but unlikely to be pictured, we filter captions where the person is in the possessive form, the object of the sentence, or modify a noun as determined by spaCy's dependency parser. We ensure that the context is distinct: the Places365 ResNet similarity must be less than 0.9. Finally, we found that there were a number of unsolvable falsified samples where the caption could be plausibly matched with any image of the person of interest. We minimize the number of such "generic" captions: we finetune a BERT (Devlin et al., 2019) model on a small labelled subset of our training data to filter these captions from our matching process. **(c)**

SBERT-WK Text-Text: We match samples that mention the query person based on the lowest semantic similarity measured by their SBERT-WK score, a text-only sentence embedding. See example (c) in Figure 1.

Query for Scene This split models a threat scenario that queries for a specific scene, with the intent to mislabel the event, see Figure 2 (iv). All samples must have no "PERSON" named entities in the captions. This aims to filter headshots and other images with little scene information. **(d) ResNet Place:** We match samples with the highest Places365 image similarity, as determined by the dot product of their ResNet embeddings. See example (d) in Figure 1.

Merged Split This split mixes samples from all the splits to model a more realistic case where a variety of methods are used to generate out-of-context images, i.e. all types of mismatch may be encountered at test time. We merge the splits such that there is an equal number of samples from every split, and the captions and images across splits are disjoint.

Adversarial CLIP Filtering In the initial version of our dataset, we observed a distributional shift of CLIP Text-Image scores between the pristine (img_1, cap_1) and falsified (img_2, cap_1) samples. This makes sense, since it is not always possible to find a falsified image that is more convincing than the original. To reduce the difference between the two distributions, we use CLIP Text-Image similarity to adversarially filter our splits. For each pristine sample (img_1, cap_1) with CLIP Text-Image similarity CTI_p we have two options: (1) There may exist a set of falsified candidates (img_2, cap_1), where their score $CTI_f \geq CTI_p$, ordered for each of our splits: using (a) CLIP Text-Image, (b) CLIP Text-Text, (c) SBERT-WK Text-Text, (d) ResNet Place, respectively. (2) There exists a set of candidates where their score $CTI_f < CTI_p$, ordered in the same way. We select the top scoring sample from set (1), else we select the top sample from (2) if set (1) is empty. Finally, we remove the sample with $max(CTI_p - CTI_f)$ until we get a 50-50 ratio of samples from sets (1) and (2) since the larger the delta $CTI_p - CTI_f$ the more likely the falsified sample is of low quality. As a result, on a ranking task where CLIP off-the-shelf is given a caption and two images, it correctly chooses the pristine image 50% of the time by design.

NewsCLIPpings Dataset Statistics The detailed statistics for the proposed NewsCLIPpings Dataset are reported in Table 1. Each caption appears twice, once in a pristine sample then again in a falsified sample. Thus exactly half of the samples are pristine and half are falsified, and there is no unimodal text bias in the dataset. We report the total number of samples across splits *including any duplicates* as Total/Sum, and the number of *unique* text-image pairings as Total/Unique in Table 1.

Table 1: NewsCLIPpings Dataset Statistics.

Split	Train	Val	Test
(a) Semantics/CLIP Text-Image	453,128	47,248	47,288
(b) Semantics/CLIP Text-Text	516,072	53,876	54,164
(c) Person/SBERT-WK Text-Text	17,768	1,756	1,816
(d) Scene/ResNet Place	124,860	13,588	13,636
Total/Sum	1,111,828	116,468	116,904
Total/Unique	816,922	85,609	85,752
Merged/Balanced	71,072	7,024	7,264

Table 2 provides a comparison to the most related prior datasets, highlighting the key differences, such as the image-text mismatch procedure used in each dataset.

Table 2: Comparison to prior related datasets. Size is the total number of unique samples across all splits.

Dataset	Data	Source	Mismatch	Size
MAIM (Jaiswal et al., 2017)	Caption, Image	Flickr	Random	239k
MEIR (Sabir et al., 2018)	Caption, Image, GPS	Flickr	Text entity manipulation	57k
TamperedNews (Müller-Budack et al., 2020)	Article, Image	BreakingNews	Text entity manipulation	776k
COSMOS (Aneja et al., 2021)	Caption, Image	News Outlets	Two sources (3k labeled)	453k
NewsCLIPpings (Ours)	Caption, Image	VisualNews	Automatic retrieval	988k

Dataset Examples Here, we provide a few samples from the NewsCLIPpings Dataset. Figure 3 compares the matches from each split for the same query caption and pristine image. Our diverse methods of computing similarity result in different weightings for concepts, displaying the realm of plausible images for a given caption. In (1), CLIP Text-Image matches "parliament" to Tennessee's governor speaking to a General Assembly, CLIP Text-Text matches "Angela Merkel" to Ingeborg Berggreen-Merkel speaking, and SBERT-WK Text-Text finds a match of Angela Merkel at a summit. In (2), CLIP Text-Image matches "tsunami" to a flooding in New York, CLIP Text-Text matches "Japan" to the president of a Japanese company, and ResNet Place matches "earthquake" to a destroyed highway after an earthquake in Chile[7].

[7]The Person split and Scene splits have no shared pristine

Figure 3: Comparison of the retrieved matches for the same query caption obtained within our four splits.

4 Experiments

We start by describing our experimental setup and then present the results of our benchmarking study.

4.1 Experimental Setup

Model Architectures For our base models we rely on CLIP (Radford et al., 2021) and VisualBERT (Li et al., 2019). We include VisualBERT as it is a representative recent model and is an appropriate baseline for addressing the semantic mismatch tasks.

CLIP passes image and text through separate encoders that are trained to generate similar representations for related concepts. The model is pretrained on a web-based corpus of 400M image-text pairs using a contrastive loss, in which the cosine similarity of true image-text pairs is maximized.

VisualBERT passes image and text through a shared series of transformer layers to align them into one embedding space. For its bounding box features, we use a Faster-RCNN model (Ren et al., 2015) trained on Visual Genome with a ResNeXT-152 backbone. For pretraining, we only use the Masked Token Loss reported by Li et al. (2019), which masks each text token with probability 0.15. We pretrain VisualBERT either on the 3M image-caption pairs from Conceptual Captions (Sharma et al., 2018), based on alt-texts from web images stripped of all named entities, or on the 1M pairs from the VisualNews (Liu et al., 2020), based on captions from the news images.

Implementation Details Our task is to *classify* each image-caption pair as pristine or falsified. We fine-tune both models as we train the classifiers. When finetuning, we use a learning rate of 5e-5 for the classifier and 5e-7 for other layers. We train with a batch size of 32 for 88k steps for the Semantics splits and 44k steps for the Person and Scene splits. We report *classification accuracy* over all samples (All) and separately for the Pristine and Falsified samples. We also report model performance at varying false alarm rates via ROC curves.

4.2 Experimental Results

In this section, we benchmark several methods on our proposed dataset to assess its difficulty. First, we compare the performance of unimodal vs. multimodal models to ensure that methods cannot exploit unimodal biases. Next, since we leverage CLIP ViT/B-32 to make our dataset challenging, we explore whether our task could be solved by a different model specifically pretrained on the news domain, leveraging a different backbone, or with more model parameters. In our final experiment, we train a single model on the union of all splits (Total/Sum in Table 3), while all the other experiments report the performance of the *distinct models trained on each split individually*. All tables in this section evaluate on the same test set per split.

Unimodal Model Performance One motivation for this work is that several prior works rely on automatic text manipulation to generate mismatched media. We argue that entity manipulation can introduce linguistic biases. We trained a text-only BERT model (Devlin et al., 2019) on *just the named entities* of the TamperedNews dataset (rather than the full articles) and achieved comparable results to the original paper's image-and-text based system (Müller-Budack et al., 2020). For their "Document Verification" task, where the goal is to select one out of two articles given an image, we were able to achieve 90% versus their 93% on the Persons Country Gender (PsCG) split. For the Outdoor Places City Region split, GCD(25, 200), we were able to

samples since all matches either do or do not have 'PERSON" named entities depending on the split.

Table 3: Classification performance on the test set for the following models: (I) Image-only CLIP (w/ ViT-B/32), (II) Multimodal CLIP (w/ ViT-B/32), (III) VisualBERT-CC pretrained on the Conceptual Captions dataset, (IV) VisualBERT-VN pretrained on the Visual News.

	(I) CLIP Image-Only	(II) CLIP			(III) VisualBERT-CC	(IV) VisualBERT-VN		
Split	All	All	Pristine	Falsified	All	All	Pristine	Falsified
(a) Semantics/CLIP Text-Image	0.5471	0.6698	0.7543	0.5853	0.5413	0.5774	0.6770	0.4778
(b) Semantics/CLIP Text-Text	0.5247	0.6939	0.7409	0.6469	0.5714	0.5949	0.6591	0.5307
(c) Person/SBERT-WK Text-Text	0.5000	0.6101	0.6178	0.6024	0.5947	0.6333	0.7247	0.5419
(d) Scene/ResNet Place	0.5391	0.6821	0.7835	0.5807	0.5636	0.6112	0.6693	0.5532
Merged/Balanced	0.5288	0.6023	0.7007	0.5039	0.5482	0.5863	0.7841	0.3885

achieve 96% versus their 76%. This suggests that text manipulation can introduce biases that make the use of images unnecessary.

To avoid unimodal biases, our dataset is *balanced* with respect to its captions (every caption is used once in a pristine sample and again in a falsified sample). Since we do not have such constraint on our images, we ran an image-only CLIP model (i.e. zeroing out the text inputs to CLIP) to verify that there is minimal visual bias. Based on our findings and due to the smaller size of the Person split (c), we additionally balance this particular split with respect to images, which means any image-only model is expected to achieve exactly 50% accuracy on this split. As shown in Table 3 (I), overall the image-only CLIP model obtains slightly above chance performance, significantly lower than the full image-text model, Table 3 (II).

Multimodal Model Performance We report results for the multimodal CLIP-based classifiers in Table 3 (II). (Again, we repeat that here we train distinct classifiers for each split individually.) CLIP tends to "over-predict" pristine labels, indicating that many falsified samples are highly realistic and plausible. The Person split appears the most challenging, which could be partly explained by having the least number of samples. The Merged split, which contains an equal proportion of all four splits, is as difficult as its most difficult sub-split, seen by how CLIP classifies correctly 60% of the time compared with 61% for the Person split.

On the other hand, VisualBERT-CC (pretrained on Conceptual Captions) in Table 3 (III) performs the best on the Person split with a performance of 59% that approaches CLIP's. This indicates that the Person split primarily requires semantic understanding, and that a model with no knowledge of named entities can compete with a model that is strong at recognizing celebrities and other named entities. As expected, on all other splits that test entity understanding VisualBERT-CC performs on average 10% worse than CLIP.

Pretraining VisualBERT on News Domain We also compare the performance of VisualBERT pretrained on Conceptual Captions (VisualBERT-CC) vs. VisualNews (VisualBERT-VN). In Table 3 (III vs. IV) we observe that in-domain data, including named entities, provides a 3-5% boost uniformly across all splits. Even more, with a training corpus less than 1% the size of CLIP's training data, VisualBERT-VN is able to exceed CLIP performance on the Person split and approach CLIP performance on the Merged split. In fact, the largest gap between VisualBERT-VN and CLIP remains in the Semantics splits, where named entity understanding is crucial. Hence, through these results we can observe that VisualBERT-VN is strongest at semantic reasoning while CLIP is strongest at named entity recognition, which makes sense given their architectures (more deeply interactive VisualBERT-VN vs. more shallow CLIP).

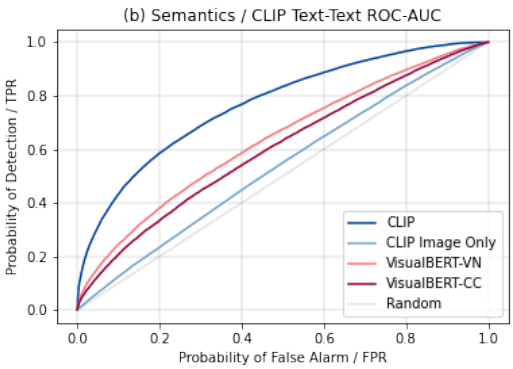


Figure 4: Semantics/CLIP Text-Text ROC Curve

ROC Curves We also include the ROC curves for the softmaxed logits produced by the models, see Figure 4, 5. We see that the trends for

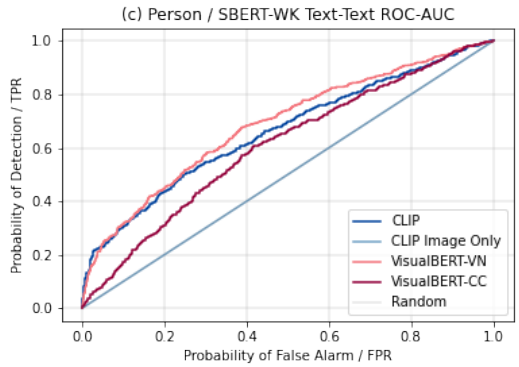


Figure 5: Person/SBERT-WK Text-Text ROC Curve

these curves are consistent with the model rankings recorded in Table 3, with CLIP outperforming the other models by a wide margin on Semantics/CLIP Text-Text in Figure 4 across all false alarm rates. For Person/SBERT-WK Text-Text in Figure 5, VisualBERT-VN has virtually identical performance as CLIP at low false alarm rates. However, for systems that can tolerate more false alarms VisualBERT-VN shows a small advantage.

Comparing CLIP Models Recall that we used CLIP ViT-B/32 to construct our dataset. Here, we investigate whether our dataset could be "solved" by an existing CLIP model with a different backbone (ViT-B/32 vs. RN50) or a bigger model (RN50 vs. RN101). In Table 4, we observe that RN50 performs slightly better than ViT-B/32 across the board, with at most 2% performance difference between the two. We also see that while RN101 has more parameters than RN50, it only provides a small 1-2% boost on most splits. The split where RN101 achieves the largest improvement (4%) is the Merged/Balanced split, which aligns with the need for a model to capture more complex patterns to classify samples from multiple generation methods. Although we used a specific CLIP model architecture during dataset generation, we see that our dataset is still challenging for models with different architectures and more parameters.

Evaluating A Single Unified Model Finally, we explore whether it is beneficial to combine various splits during training. Unlike Tables 3, 4 which evaluate *separate* models trained on each individual split, here we evaluate a *single* model trained on all the splits jointly, see Table 5. The Total/Sum set (introduced in Table 1) combines the samples from all the splits, so that it is balanced with respect to pristine and falsified labels but has different proportions of each type, e.g. around 87% of samples are from the Semantics splits (a,b).

Table 4: Comparing different CLIP backbones, classification performance (test set).

Split	Model	All	Pristine	Falsified
(a) Sem/CLIP T-I	ViT-B/32	0.6698	0.7543	0.5853
	RN50	**0.6824**	**0.7461**	**0.6188**
	RN101	0.6765	0.7444	0.6085
(b) Sem/CLIP T-T	ViT-B/32	0.6939	0.7409	0.6469
	RN50	0.7182	0.7486	0.6878
	RN101	**0.7244**	**0.7442**	**0.7046**
(c) Per/SB-WK T-T	ViT-B/32	0.6101	0.6178	0.6024
	RN50	0.6123	0.7357	0.4890
	RN101	**0.6393**	**0.7004**	**0.5782**
(d) Scene/RN Place	ViT-B/32	0.6821	0.7835	0.5807
	RN50	0.7004	0.7765	0.6244
	RN101	**0.7137**	**0.7712**	**0.6562**
Merged/Balanced	ViT-B/32	0.6023	0.7007	0.5039
	RN50	0.6162	0.6836	0.5487
	RN101	**0.6597**	**0.6768**	**0.6426**

Table 5: CLIP (ViT/B-32) test set classification performance when training a single model with all the available training samples, i.e. Total / Sum in Table 1.

Split	All	Pristine	Falsified
(a) Semantics/CLIP Text-Image	0.6651	0.7582	0.5720
(b) Semantics/CLIP Text-Text	0.6457	0.7563	0.5351
(c) Person/SBERT-WK Text-Text	0.6399	0.7434	0.5363
(d) Scene/ResNet Place	0.6824	0.7778	0.5870
Merged/Balanced	0.6611	0.7574	0.5647

Comparing Table 3 (II) with Table 5, we note that the Person split experiences a 2% boost in performance even though it represents only 1% of the training data. Clearly, it benefits from the other sample types. We also note the 5% degradation in performance for the Semantics/CLIP Text-Text, likely due to the challenges in learning to address several mismatch types at once[8]. Finally, we see a boost of almost 6% for the Merged/Balanced set, showing the benefit of training in a unified setting for this more realistic split. One other trend we notice is that the Pristine accuracy seems to overall benefit more than the Falsified accuracy.

5 Additional Analysis

In this section, we gain further insights into the quality of our dataset via human evaluation and saliency map analysis. With the human evaluation,

[8]We hypothesize that this may be due to the joint training with the Person samples – if a model does not know who the pictured individual is, then the mismatches in Semantics/CLIP Text-Text may look similar to those in the Person split, as they both are matched using only textual information.

we assess whether our dataset could fool humans and pose a realistic threat. We also assess whether our dataset may have "unsolvable" true matches that in fact do not misrepresent anything. With our qualitative saliency map analysis, we investigate if the automatic models are learning to leverage high level semantic or entity cues after training on our dataset.

Table 6: Human Performance on 200-sample subset of Merged/Balanced. "Optimistic" accuracy is defined as at least 1 worker gave the correct answer.

	All	Pristine	Falsified
Average	0.656	0.962	0.350
Optimistic	0.845	1.000	0.690

Human Performance Here, we estimate the difficulty of the proposed task for *humans*, aiming to assess how convincing our automatically matched images and captions are. We randomly select a set of 200 samples from the Merged/Balanced split, with an equal number of samples from all types (50 from each split, where 25 are pristine and 25 are falsified). We conduct our evaluation on Amazon Mechanical Turk[9]. For each image-caption pair we ask 5 workers the following three questions: (a) "Could this image belong to the given caption?" (Yes/No), (b) "How confident are you in your answer?" (1: Very, 2: Somewhat, 3: Not at all), (c) "Would it help to use a search engine to be more confident?" (Yes/No). Note, that we specifically instruct the workers **not** to use search engines, to prevent them from discovering the original news articles on the Web. The key takeaways from the evaluation are as follows. (1) The average accuracy over all samples is 0.656, while the most "optimistic" accuracy (at least 1 worker gave the correct answer) is 0.845. This clearly shows that the task is not easy for humans. For reference, our CLIP model trained on the Total/Sum set (Table 5) achieves 0.6650 on these 200 samples, essentially *matching human performance*. (2) Humans are much better in recognizing pristine than falsified samples, with an average accuracy of 0.962 and 0.350 respectively. This shows that they are often misled by our falsified matches. (3) The "optimistic" accuracy for falsified samples is 0.690, meaning that majority are still solvable with just the prior knowledge of those workers. Among the 31 samples that all workers classified incorrectly, we estimate that 67% are answerable with additional knowledge of person identity and other context cues. (4) While the average confidence score is 1.755, the confidence on the correctly vs. incorrectly predicted samples is 1.658 and 1.940 respectively (lower is better), i.e. humans were more confident on samples they predicted correctly. (5) The average accuracy when there is a reported "need to use a search engine" is 0.589 vs. 0.760 otherwise. This shows that the humans do better when they encounter familiar concepts vs. less familiar ones. Hence, additional search is likely to boost the results, as we have observed in our own internal analysis. (6) Across the four types of mismatch, the easiest for humans is Scene, followed by Semantics/CLIP Text-Text, Semantics/CLIP Text-Image, and finally the Person split. Interestingly, this overall aligns with the trends observed for the automatic methods.

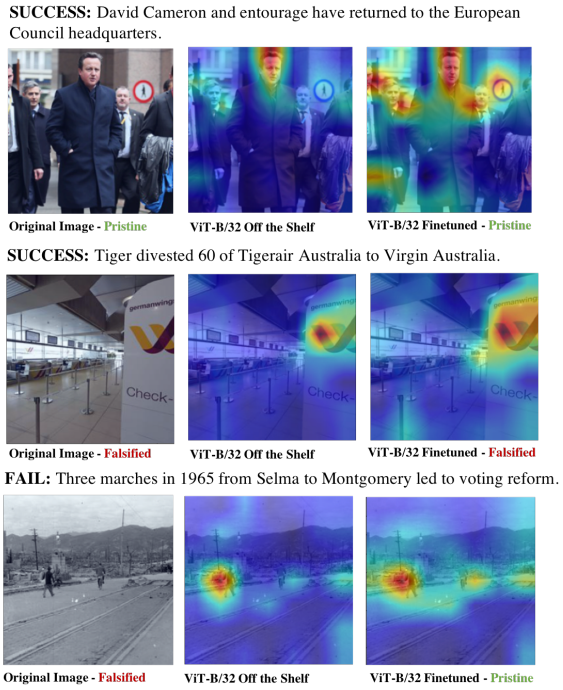


Figure 6: Qualitative examples of CLIP ViT-B/32 success and failure cases with salience visualization.

Qualitative Analysis Finally, we analyze CLIP ViT saliency maps and prediction using the method presented in Chefer et al. (2021). We select examples from the 200 samples used in our human evaluation. As seen in Figure 6, finetuning on our dataset often forces CLIP to focus on salient objects mentioned in the caption beyond the person of interest, for example expanding its attention from David Cameron

[9]www.mturk.com

to "entourage." CLIP also has a number of capabilities off-the-shelf that require minimal finetuning, for example sign reading and logo recognition in the case of distinguishing "Tigerair Australia" from "Germanwings." We also present a failure case where CLIP focuses on the two people in the foreground when the caption talks about "marches." Evidently the model does find one point of support – the photo looks like it could have been taken in 1965 – but it fails to identify the absence of a crowd which would have been a "red flag" for a human. Similar failure cases, such as a falsified caption that mentions Mitt Romney and Rand Paul at a rally but only pictures Romney, highlight how our dataset can be particularly challenging because pristine news is an ambiguous domain that often mentions entities but does not picture them.

6 Conclusion

We introduced **NewsCLIPpings**, a large-scale automatically constructed dataset for classifying news image-caption pairs as real or out-of-context. We found CLIP to be effective for the dataset construction and recognition of mismatches. By design, unimodal models cannot solve our task, while multimodal ones require named entity and semantic knowledge to do well on our diagnostic splits. Our Merged set aims to model the more realistic diversity of image-caption mismatches in the wild.

From our experimental results, we find that the ResNet backbone offers a modest performance boost compared to a ViT-B/32 model. We find that a CLIP ViT model is able to match human performance on a small subset of our Merged / Balanced split, and that our task is generally difficult with an average 66% human accuracy.

Our training data could be used to augment and increase the training data size of human-made falsified news, which often lack ground truth labels or sufficient scale. Overall, we show that it is possible to automatically match plausible images for given input captions, and we present a challenging benchmark to foster the development of defenses against large-scale image repurposing.

Acknowledgements. This work was supported in part by DoD including DARPA's XAI, LwLL, and/or SemaFor programs, as well as BAIR's industrial alliance programs.

7 Ethical Considerations

Here, we discuss ethical considerations regarding our dataset. Image repurposing is a prominent societal issue that lacks sufficient training data, as both human generation and annotation are costly. Even more, our work aims to be proactive in characterizing this new threat of machine generated misinformation and proposes a number of solutions that serve as baselines for detection. By presenting a number of techniques and key observations about falsified out-of-context news, we hope that our dataset serves as a net benefit for society.

How was the data collected? Our dataset was automatically generated using features from CLIP (Radford et al., 2021), SBERT-WK (Wang and Kuo, 2020), a ResNet50 trained on Places365 (Zhou et al., 2017), and other additional metadata. Our dataset is composed of intelligent automatic matches for the VisualNews (Liu et al., 2020) images and captions that occur in the real world.

What are the intellectual property rights? The copyright and usage rights of our dataset are subject to that of VisualNews (Liu et al., 2020). Our key contribution is an automatic dataset generation *approach* that can be applied to any news dataset.

How did we address participant privacy rights? N/A

Were annnotators treated fairly? Did we require review from a review board? N/A

Which populations do we expect our dataset to work for? As our dataset is composed of news from The Guardian, BBC, USA Today, and The Washington Post, it largely focuses on events and people from Western countries like the US and the UK in addition to world news.

What is the generalizability of our claims? We expect our results regarding the dangers of automatic image repurposing and experimental results of model detection performance to primarily apply to Western news in the English language.

How did we ensure dataset quality? We made a number of design choices noted in Sections 3,4 to remove any unimodal biases and ensure that samples would be challenging for recent AI models. We also conducted a human evaluation and found that our matches were challenging for humans.

What is the climate impact? For dataset construction, excluding the fixed cost of embedding extraction, matching takes about 1-2 hours with 8 GPUs to process 400k pristine samples. This is an estimated 1.28 kg CO_2 eq emissions. Each finetuning experiment takes about 9 hours on one GPU, which is an estimated 0.97 kg CO_2 eq emissions. Estimations were conducted using the MachineLearning Impact calculator presented in Lacoste et al. (2019).

What are the potential dataset biases? Here, we discuss the models used to compute matching scores during dataset generation to understand the potential biases that may appear in our dataset.
CLIP: Radford et al. (2021) conducts a number of experiments on race and gender prediction to investigate potential biases learned from their large-scale web corpus. They report that CLIP had significant disparities when classifying individuals from different races into crime-related and non-human categories. They also report gender differences when attaching appearance-related terms and occupations to photos of Members of Congress.
SBERT-WK: Since SBERT-WK dissects the embedding of BERT (Devlin et al., 2019) to compute semantic similarity, any biases present in the model are from those learned by BERT. Prior works note that BERT does encode a number of social biases, including a strong association between gender and career/family or math/arts (Kurita et al., 2019).
Places365: From our qualitative analysis, we found that the Places365 dataset exhibited race and age associations with certain scene labels. As a result, we only used the ResNet embeddings when computing matches and completely ignored the labels to somewhat mitigate these biases.

How might this work contribute to the spread of disinformation? We acknowledge that our work can be misused to mass-generate repurposed images. Adversaries can now automate both synthesizing an inflammatory news piece using recent models like GPT-3 (Brown et al., 2020) or Grover (Zellers et al., 2019), and they can retrieve an appropriate image using models we present such as CLIP (Radford et al., 2021), SBERT-WK (Wang and Kuo, 2020), and ResNet (Zhou et al., 2017).

However, we argue that our work cannot be immediately used to generate targeted attacks. An adversary would either have to generate synthetic captions (since our models are trained on human-made captions, this would mean a domain shift) or manually write captions (which is time and money consuming) tailored for their narrative. Although we demonstrate that automatic image repurposing

can be convincing to humans, additional effort is required to produce *malicious* image-caption pairs.

References

Shivangi Aneja, Christoph Bregler, and Matthias Nießner. 2021. Catching out-of-context misinformation with self-supervised learning. *arXiv:2101.06278*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Lisa Fazio. 2020. Out-of-context photos are a powerful low-tech form of misinformation. https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959/.

Elise Fenn, Nicholas Ramsay, Justin Kantner, Kathy Pezdek, and Erica Abed. 2019. Nonprobative photos increase truth, like, and share judgments in a simulated social media environment. *JARMAC*, 8(2):131–138.

Karen Hao. 2020. A college kid's fake, ai-generated blog fooled tens of thousands. this is how he made it. https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Ayush Jaiswal, Ekraam Sabir, Wael AbdAlmageed, and Premkumar Natarajan. 2017. Multimedia semantic integrity assessment using joint embedding of images and text. In *Proceedings of the ACM international conference on Multimedia (MM)*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

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

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visualnews: A large multi-source news image dataset. *arXiv:2010.03743*.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *ACM ICMR*.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep multimodal image-repurposing detection. In *ACM MM*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188.

Reuben Tan, Kate Saenko, and Bryan A Plummer. 2020. Detecting cross-modal inconsistency to defend against neural fake news. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Christopher Thomas and Adriana Kovashka. 2020. Preserving semantic neighborhoods for robust cross-modal retrieval. In *ECCV*.

Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. Rel: An entity linker standing on the shoulders of giants. In *ACM SIGIR*.

Daniel Victor. 2021. Your loved ones, and eerie tom cruise videos, reanimate unease with deepfakes. https://www.nytimes.com/2021/03/10/technology/ancestor-deepfake-tom-cruise.html/.

Bin Wang and C-C Jay Kuo. 2020. Sbert-wk: A sentence embedding method by dissecting bert-based word models. *IEEE/ACM TASLP*, 28:2146–2157.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE TPAMI*.

A Appendix to "NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media"

In the following we include additional details regarding dataset construction (Section A.1), additional experimental results (Section A.2), and dataset examples (Section A.3).

A.1 Dataset Construction

Here, we go over additional implementation details for dataset construction. (1) To ensure the quality of our dataset, we filter pristine samples such that they have between 5 and 30 words, at least 2 named entities, and non-corrupted images. (2) We split our 509,730 pristine image-caption pairs into chunks of size ∼40k for train/val/test. We only computed matches across these disjoint chunks, which allowed us to parallelize the generation process. (3) We precomputed features such as spaCy NER, Radboud Entity Linking, SBERT-WK text embeddings, CLIP text embeddings, CLIP image embeddings, Faster R-CNN bounding boxes, ResNet50 place embeddings. (4) We ran Algorithm 1, the NewsCLIPpings matching algorithm. (5) We removed low quality samples by balancing the number of samples where the CLIP text-image score is higher for the pristine vs falsified pair as described in our Adversarial CLIP Filtering process in Section 3.

Algorithm 1 NewsCLIPpings Matching

Input: Dataset D, scoring function $sim \in$ {CLIP Text-Image, CLIP Text-Text, SBERT-WK Text-Text, ResNet Place}, split type $split$.
Output: Matches M

$M \leftarrow \{\}$
for each $p = (i_1, c_1) \in D$ **do**
  $M_H, M_L \leftarrow \{\}, \{\}$
  $D* \leftarrow$ sort(D, key=sim)
  for each $(i_2, c_2) \in D^*$ **do**
    $f \leftarrow (i_2, c_1)$
    if filter($p, f, split$) **then**
      continue
    else if $CTI(f) > CTI(p)$ **then**
      $M_H \leftarrow M_H \cup \{f\}$
    else
      $M_L \leftarrow M_L \cup \{f\}$
    end if
  end for
  $M \leftarrow M \cup \{(M_H \cup M_L)[0]\}$
  $M \leftarrow M \cup \{p\}$
end for

A.2 Additional Results

ROC Curves We include ROC curves for the splits not depicted in Section 4 in the main paper. Note that the rankings for model performance across all these splits are consistent with Table 3. We see that CLIP vastly outperforms the other models in the Semantics/CLIP Text-Image split whereas model performance is very similar in the Merged/Balanced split. This makes sense since CLIP was used as the scoring function for matches in Semantics/CLIP Text-Image, whereas a diverse set of scoring functions were used for Merged/Balanced.

Finetuning CLIP Representation We also experiment with freezing a varying number of CLIP layers to determine how useful CLIP's original knowledge and representation is for each split. We compare three models: RN50-all-frozen (RN50-af, no CLIP layers finetuned), RN50-lower-frozen (RN50-lf, final few layers finetuned) [10], and RN50 (all layers finetuned). In Tables 7 and 8, we observe that finetuning all layers (RN50) benefits the Semantics splits (a,b) while freezing some or all layers (RN50-af/lf) benefits the Person and Scene splits (c,d). We posit that this result could be related to the training data size – since the Semantics splits are significantly large we can meaningfully finetune all layers whereas in the others we do not have enough data to do so.

Table 7: Comparing different finetuning strategies for CLIP, classification performance (val set).

	Model	All
(a) Semantics/CLIP Text-Image	RN50-af	0.6354
	RN50-lf	0.6706
	RN50	**0.6808**
(b) Semantics/CLIP Text-Text	RN50-af	0.6826
	RN50-lf	0.6954
	RN50	**0.7138**
(c) Person/SBERT-WK Text-Text	**RN50-af**	**0.6606**
	RN50-lf	0.6560
	RN50	0.6167
(d) Scene/ResNet Place	RN50-af	0.6965
	RN50-lf	**0.7034**
	RN50	0.6999
(e) Merged/Balanced	RN50-af	0.6642
	RN50-lf	**0.6684**
	RN50	0.6468

[10] The exact layers we finetune are ["visual.layer4", "visual.attnpool", "transformer.resblocks.11", "ln_final", "text_projection", "logit_scale"].

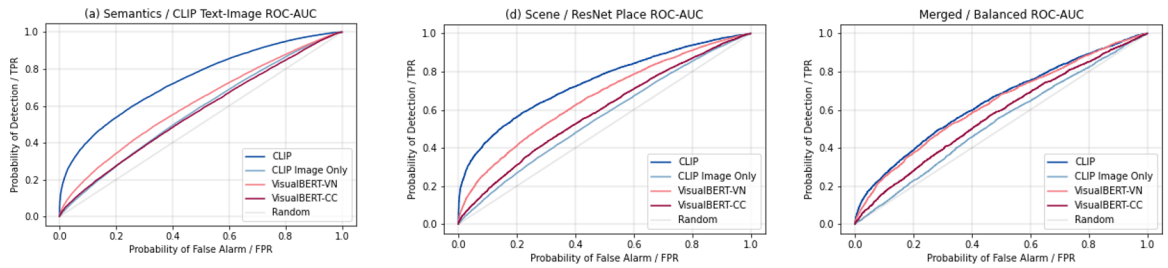


Figure 7: Semantics/CLIP Text-Image, Scene/ResNet Place, Merged/Balanced ROC Curve

Table 8: Comparing different finetuning strategies for CLIP, classification performance (test set).

Split	Model	All	Pristine	Falsified
(a) Sem/CLIP T-I	RN50-af	0.6372	0.6677	0.6068
	RN50-lf	0.6500	0.6840	0.6163
	RN50	**0.6824**	**0.7461**	**0.6188**
(b) Sem/CLIP T-T	RN50-af	0.6860	0.7167	0.6554
	RN50-lf	0.7005	0.7211	0.6800
	RN50	**0.7182**	**0.7486**	**0.6878**
(c) Per/SB-WK T-T	**RN50-af**	**0.6669**	**0.6641**	**0.6696**
	RN50-lf	0.6547	0.6575	0.6520
	RN50	0.6123	0.7357	0.4890
(d) Scene/RN Place	RN50-af	0.6945	0.7543	0.6346
	RN50-lf	**0.7028**	**0.7646**	**0.6411**
	RN50	0.7004	0.7765	0.6244
Merged/Balanced	**RN50-af**	**0.6732**	**0.6726**	**0.6737**
	RN50-lf	0.6657	0.6952	0.6363
	RN50	0.6162	0.6836	0.5487

Ensembling Performance Next, we compare the predictions made by our best VisualBERT and best CLIP-based model (Table 9). Specifically, we ask whether these very different models exhibit distinct behavior and have complementary skill sets. First, we assess how often both models are correct (Overlap) or at least one of them is correct (Union). As we see, the overlap tends to be near 40%, while the "optimistic" union boosts accuracy to over 80%. This clearly shows that the two models indeed have complementary strengths. We build a simple ensemble of VisualBERT and CLIP, in which we average the normalized logits of each model and classify accordingly (Avg). This scheme only gives a 1-2% boost, which possibly indicates that CLIP is more "opinionated" and yields more disparate logits per class, forcing the ensemble to perform similarly to CLIP. This shows that leveraging both models may be promising but is non-trivial and should be explored by future work.

Results on the Validation Set We include the results that correspond to Tables 3, 4, 5 of the main paper but evaluated on the validation set, see Tables 10, 11, 12. Note, that the validation and test sets of our dataset were generated using identical techniques and should be comparable in composition. As a result, these tables demonstrate the same overall trends as reported in our main paper.

Table 9: Exploring complementarity of VisualBERT-VN and CLIP-RN101, reported overall classification performance (test set).

Split	VB-VN	CLIP-RN101	Avg	Overlap	Union
(a) Sem/CLIP T-I	0.5774	**0.6765**	0.6747	0.4341	0.8197
(b) Sem/CLIP T-T	0.5949	**0.7244**	0.7234	0.4663	0.8530
(c) Per/SB-WK T-T	0.6333	0.6393	**0.6553**	0.4576	0.8150
(d) Scene/RN Place	0.6112	0.7137	**0.7230**	0.4723	0.8527
Merged/Balanced	0.5863	0.6597	**0.6662**	0.4145	0.8315

A.3 Additional Dataset Details

Next, we report the amount of overlap across splits of our dataset (see Table 13) and find that they are relatively distinct, with a maximum of 11% overlap in falsified matches for the two Semantics splits.

Figures 8 and 9 depict *randomly* selected samples from the train/val/test set of each respective split. Note how matches are highly plausible in our Semantics splits (a,b) but can be solved if the identity of a person is known or even with subtle semantic cues (e.g. an American flag when the caption describes a European person in the top left or that the image shows a letter when the caption describes a banner on the bottom left for Semantics/CLIP Text-Image). Also note, how in the Semantics/CLIP Text-Text, textual concepts from a query may impact the resulting retrieved image based on its own caption, e.g., "flood" in the top-right example. Note the challenging examples from our Person split (c), including some rather ambiguous ones such as top-right with Hillary Clinton, that require recognizing the context in which the person appears. Finally, in our Scene split (d) we show that the events in captions may be plausibly "illustrated" by purely leveraging visual scene similarity. Note again, that these are *randomly* selected samples from our dataset.

Table 10: Classification performance on the val set for the following models: (I) Image-only CLIP (w/ ViT-B/32), (II) Multimodal CLIP (w/ ViT-B/32), (III) VisualBERT-CC pretrained on the Conceptual Captions dataset, (IV) VisualBERT-VN pretrained on the Visual News.

	(I) CLIP Image-Only	(II) CLIP	(III) VisualBERT-CC	(IV) VisualBERT-VN
Split	All	All	All	All
(a) Semantics/CLIP Text-Image	0.5472	0.6711	0.5451	0.5773
(b) Semantics/CLIP Text-Text	0.5266	0.6921	0.5738	0.5963
(c) Person/SBERT-WK Text-Text	0.5000	0.6259	0.5666	0.6139
(d) Scene/ResNet Place	0.5424	0.6803	0.5558	0.6128
Merged/Balanced	0.5185	0.6048	0.5504	0.5893

Table 11: Comparing different CLIP backbones, classification performance (val set).

	Model	All
(a) Semantics/CLIP Text-Image	ViT-B/32	0.6475
	RN50	**0.6808**
	RN101	0.6741
(b) Semantics/CLIP Text-Text	ViT-B/32	0.6921
	RN50	0.7138
	RN101	**0.7189**
(c) Person/SBERT-WK Text-Text	ViT-B/32	0.6099
	RN50	0.6167
	RN101	**0.6395**
(d) Scene/ResNet Place	ViT-B/32	0.6803
	RN50	0.6999
	RN101	**0.7153**
(e) Merged/Balanced	ViT-B/32	0.6048
	RN50	0.6468
	RN101	**0.6676**

Table 12: CLIP (ViT/B-32) val set classification performance when training a single model with all the available training samples, i.e. Total/Sum in Table 1.

Split	All
(a) Semantics/CLIP Text-Image	0.6632
(b) Semantics/CLIP Text-Text	0.6445
(c) Person/SBERT-WK Text-Text	0.6395
(d) Scene/ResNet Place	0.6858
Merged/Balanced	0.6640

Split	(a)	(b)	(c)	(d)
(a) Semantics/CLIP Text-Image	1.0000	0.1133	0.0000	0.0813
(b) Semantics/CLIP Text-Text	0.1133	1.0000	0.0000	0.0888
(c) Person /SBERT-WK Text-Text	0.0000	0.0000	1.0000	0.0000
(d) Scene/Place Label	0.0813	0.0888	0.0000	1.0000

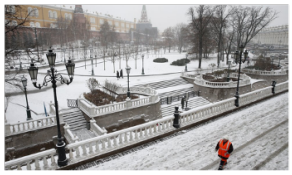
Table 13: Ratio of exact overlap across splits.

(a) Semantics / CLIP Text-Image

Query: Energy Commissioner Gunther Oettinger said cuts greater than 20 would deindustrialise Europe
Retrieved: Republican Scott Brown will focus on financial services and commercial real estate for the law firm Nixon Peabody


Query **Retrieved**

Query: Muscovites will now have to try alternative winter getaways such as Thailand or Vietnam
Retrieved: A man walks with his luggage over the Brooklyn Bridge in the snow on Thursday in New York




Query **Retrieved**

Query: A news banner in New York announces casualties in the Boston blast
Retrieved: A note left for the victims of Malaysian Airlines flight MH17 at Schiphol Airport

Query **Retrieved**

Query: Coronation Street s Hayley and Roy Cropper played by Julie Hesmondhalgh and David Neilson
Retrieved: Ken and Deirdre married for the first time in 1981


Query **Retrieved**

(b) Semantics / CLIP Text-Text

Query: Stay classy San Diego Kristen Wiig joins the Anchorman team
Retrieved: CBS Sports Radio host Dana Jacobson who once worked for ESPN


Query **Retrieved**

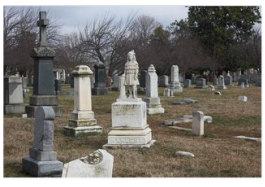
Query: A Pakistani army helicopter evacuates flood survivors from a field in Medain
Retrieved: Flood victims wade through a flooded field as they head toward a boat to be evacuated to dry land following heavy rain in Jhang Pakistan


Query **Retrieved**

Query: Librarian with the Lincoln Financial Foundation Collection Jane Gastineau is photographed with a print of the famous photograph taken by spirit photographer William H Mulmer
Retrieved: Irene Fogel Weiss holds a photo of her that was taken at Auschwitz by two Nazi guards


Query **Retrieved**

Query: The grave of Marion Kahlert in Washington s Congressional Cemetery
Retrieved: The monument of James Smithson in the Protestant or English cemetery in Italy

Query **Retrieved**

Figure 8: Randomly selected (a), (b) samples from the train/val/test samples of each respective split.

(c) Person / SBERT-WK Text-Text

Query: Palestinian president Mahmoud Abbas listens to Qatari foreign minister Sheikh Hamad bin Jassim alThani during a meeting of the Arab League yesterday
Retrieved: Mahmoud Abbas signed the ICC s founding treaty on Wednesday


Query **Retrieved**

Query: Amy Schumer accepts the Critics Choice MVP award at the 21st annual Critics Choice Awards in Santa Monica Calif on Jan 17 2016
Retrieved: Amy Schumer rethinks her life after falling for Bill Hader s sports doctor in Trainwreck


Query **Retrieved**

Query: Democratic presidential candidate Hillary Clinton speaks at a United Food and Commercial Workers International union Legislative and Political Affairs conference May 26 2016 in Las Vegas
Retrieved: In this Dec 3 2014 file photo former Secretary of State Hillary Rodham Clinton speaks at Georgetown University in Washington


Query **Retrieved**

Query: Dissident Chinese artist Ai Weiwei speaks during an interview at his hotel in Beijing on March 24
Retrieved: Ai Weiwei in Melbourne No question is hard to answer


Query **Retrieved**

(d) Scene / ResNet Place

Query: Santander The Spanish bank is rebranding its UK businesses including Abbey
Retrieved: A demonstrator holds a poster during a World March Against Monsanto event in Lisbon in May


Query **Retrieved**

Query: In this video released Thursday a Kurdish Iraqi pesh merga soldiers are shown in captivity before one of them was executed by Islamic State militants
Retrieved: Indian TV channels are freely available in Nepal


Query **Retrieved**

Query: Customers get a look at products at Microsoft s popup store in Manhattan in October
Retrieved: iPhone is ringing up profits for Apple improving quarterly earnings by 95 to 6bn The number of iPhones sold has doubled from a year ago but iPad sales have stuttered


Query **Retrieved**

Query: The 2013 Ford CMax Energi the plugin version
Retrieved: 2 Honda s Odyssey minivan


Query **Retrieved**

Figure 9: Randomly selected (c), (d) samples from the train/val/test samples of each respective split.