

## Appendix

Model	Mode	B-4	M	C	S
CLOSE w/Images	-	34.4	27.8	113.2	20.4
CLOSE w/Tuned Noise	S	28.6	25.2	95.4	18.1
CLOSE w/Tuned Noise	M	29.5	25.6	98.4	18.3
ESPER Style [78]	-	21.9	21.9	78.2	-
CLOSE w/o Noise	S	4.2	12.2	16.4	6.5
CLOSE w/o Noise	M	21.9	20.6	68.7	13.5
CLOSE	S	22.1	23.7	81.2	17.7
CLOSE	M	29.5	25.7	97.8	18.3

Table 1: Results on the caption test set in single-caption setting and multiple captioning setting, M indicates the multiple caption setting and S indicates the single caption setting.

Model	Yes/No	Num.	Other	All
CLOSE w/Images	83.2	44.8	54.9	65.4
CLOSE w/Tuned Noise	79.4	43.4	51.1	61.9
TAP-C <sub>ViT-B/16</sub> [57]	71.4	20.9	18.6	38.7
CLOSE	77.1	42.1	48.6	59.6
CLOSE w/o Noise	78.6	40.6	49.0	60.2

Table 2: Results on the VQA 2.0 test-dev set.

Model	Yes/No	Num.	Other	All
CLOSE w/Images	80.4	48.4	64.1	67.9
CLOSE w/Tuned Noise	78.2	46.0	59.5	64.3
CLOSE	74.9	45.2	59.2	62.9
CLOSE w/o Noise	76.8	36.8	53.9	59.8

Table 3: Results on the VQA-E validation set.

### 1. Hyperparameters

For all tasks, we fine-tune our model with the Adam optimizer [30] with a linear decaying learning rate starting at  $3e-4$ ,  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ , batch size of 128, and train for 8 epochs. We use beam search with a beam size of 5 for evaluations. When tuning the noise level, we select 0.04 for VQA, 0.08 for visual entailment and visual news, 0.14 for captioning in the single caption setting, and 0.04 for captioning in the multiple captioning setting.

### 2. Detailed Results

To facilitate more detailed comparisons with other works, we present results across more metrics of our evaluated datasets. In all tables, upper bounds that use images are shown above the dashed line.

**Captioning.** We present results in Table 1 for BLEU-4 [49],

Model	Val	Test
CLOSE w/Images	77.0	77.7
CLOSE w/Tuned Noise	75.9	75.9
CLIP Classifier [57]	67.2	66.6
CLOSE	75.9	75.9
CLOSE w/o Noise	68.7	68.2

Table 4: Results on the visual entailment test and validation set.

Model	B-4	M	R	C
VNC w/Images [40]	5.3	8.2	17.9	50.5
CLOSE w/Images	9.3	10.9	25	105.7
CLOSE	5.4	8.2	19.7	80.8
CLOSE w/o Noise	2.1	4.9	12.7	32.1

Table 5: Results on the visual news test set.

Model	Individual	Any
OpenAI Curie	58.8	85.0
GPT-J	42.7	81.9

Table 6: How often generated captions contain the target keywords when generating synthetic captions using different language models. The second column shows the success rate for individual generations, and the third column shows how often any caption in the 5 captions generated per a prompt contain both keywords.

METEOR [10], CIDEr [65] and SPICE [2].

**VQA.** We present results by question-type for VQA 2.0 in Table 2 and VQA-E in Table 3.

**Visual Entailment.** We present visual entailment results on the test and dev set in Table 4.

**Visual News.** We present results with BLEU-4 [49], METEOR [10], ROUGE [15] and CIDEr [65] following [40] in Table 5. To the best of our knowledge the previous best reported results is from Liu *et al.* [40] which does not make use of a pre-trained language model like CLOSE does. Qualitative results are show Section 5.

### 3. Generating Synthetic Captions using Language Models

In this section, we give more details about how we generate captions using language models and the results from Section 3.3. When generating captions, we use nucleus sampling [23] at  $p = 0.95$  and a temperate of 1, which we find generally improves results. It is not uncommon for the






Word	Image	Curie Model	COCO Model
pictured (100x)		a sandwich is pictured on a white background. CIDEr: 0.76	a sandwich is sitting on a white plate. CIDEr: 1.29
lays (100x)		a cat lays on a computer keyboard. CIDEr: 0.43	a cat is laying on a laptop computer. CIDEr: 1.94
cityscape (54x)		a clock with a cityscape in the background. CIDEr: 0.44	a clock on the side of a tall building. CIDEr: 1.95
person's (13x)		a tennis racquet is seen in a person's hand. CIDEr: 0.62	a close up of a person with a tennis racket CIDEr: 1.12
sunny (3.5x)		a sunny day with people flying kites. CIDEr: 0.09	a number of people on a beach with a kite CIDEr: 0.98

Figure 1: Examples of words that are over-produced by the captioning model trained on the OpenAI Curie synthetic captions relative to the model trained on the COCO captions. The first column shows the word and how much more common it is across captions generated for images in the COCO validation set. The remaining columns provide an example image and a caption from both models with the CIDEr score computed using human-annotated captions.

caption to fail to contain both input keywords, so we sample 5 captions for each prompt and then select a caption containing the keywords if one exists, and select one randomly otherwise. The in-context example captions are prefixed by randomly chosen words that exist within that caption (excluding stop words), and we use randomly selected captions from COCO training captions as the examples. During sampling, we randomly shuffle both the order of the in-context examples and what keywords are used as prefixes for those examples to improve the diversity of the outputs. If doing unigram sampling, we keep track of the distribution of words found in the captions generated so far, and sample new keywords in proportion to how under-represented they are, while never sampling over-represented words.

Statistics for how often the input keywords are correctly included in the caption are shown in Table 6. The success rate is less than 60%, although selecting from 5 generations brings the success rate up considerably. GPT-J is worse than OpenAI Curie, but sampling extra captions helps make up for this deficiency. Future work could integrate a constrained beam search method to address this difficulty [43].

We find that about 10% of GPT-J captions are not coherent or do not describe a visual scene, while these kinds of

captions almost never occur with OpenAI Curie. Overall, for GPT-J, producing 100k captions took about 50 GPU hours using a NVIDIA RTX A6000. For OpenAI Curie, each generation requires approximately 500 tokens per a query, so the total cost was about 100\$<sup>1</sup>. Both methods are far cheaper than annotating data.

As discussed, we observe stylistic differences occur between models trained on synthetic captions and models trained on COCO captions. A particular issue is that, while unigram sampling prevents words becoming under-represented, it still allows some words to become over-represented if the language model has a natural tendency to generate them. Figure 1 contains some examples where the model trained on OpenAI Curie captions uses words like “pictured”, “lays” or “cityscape” that almost never occur in COCO captions and thus lead to low quantitative scores even when used correctly. Interestingly, we find GPT-J is not as affected by this issue, which likely stems from differences in what data the language model was trained on. Nevertheless, the captions do still correspond well to the image content, as shown by reasonably good captioning scores despite these stylistic issues, showing it is possible to learn captioning using only synthetic data.

#### 4. The Relationship Between Image and Text Vectors

We perform a small case study by selecting four image/caption pairs that represent two different semantic changes in terms of animal species and positions (the result is shown in Figure 2) and examine how the image or text vectors shift according to these changes. We observe that text vectors move more consistently when either the species or positions of the animals change. This disparity is likely due to random shifts in image semantics that correlate with conceptual changes in the text, such as subtle alterations in the animals’ appearance, textures, or background.

We further analyze how image and text vectors typically differ by computing the differences between image/text pairs in an auxiliary corpus of COCO. We center these differences and apply PCA. The first two plots in Figure 3 show that the first few PCA dimensions explain a large portion of the variance in these differences, showing that differences often occur in similar directions. We also plot the Pearson correlation coefficient for the most related features in the third plot, showing that a number of these features are highly correlated. Indeed, image/text pairs tend to move in a structured manner that follows a particular “shape”. We capture this subtle relationship by studying the covariance matrix of the differences between text-image vectors. We then modify our Gaussian noise that is added to the text during training to better simulate this co-movement.

<sup>1</sup>At the current rate of 0.002\$ per 1k tokens on 11/16/2022

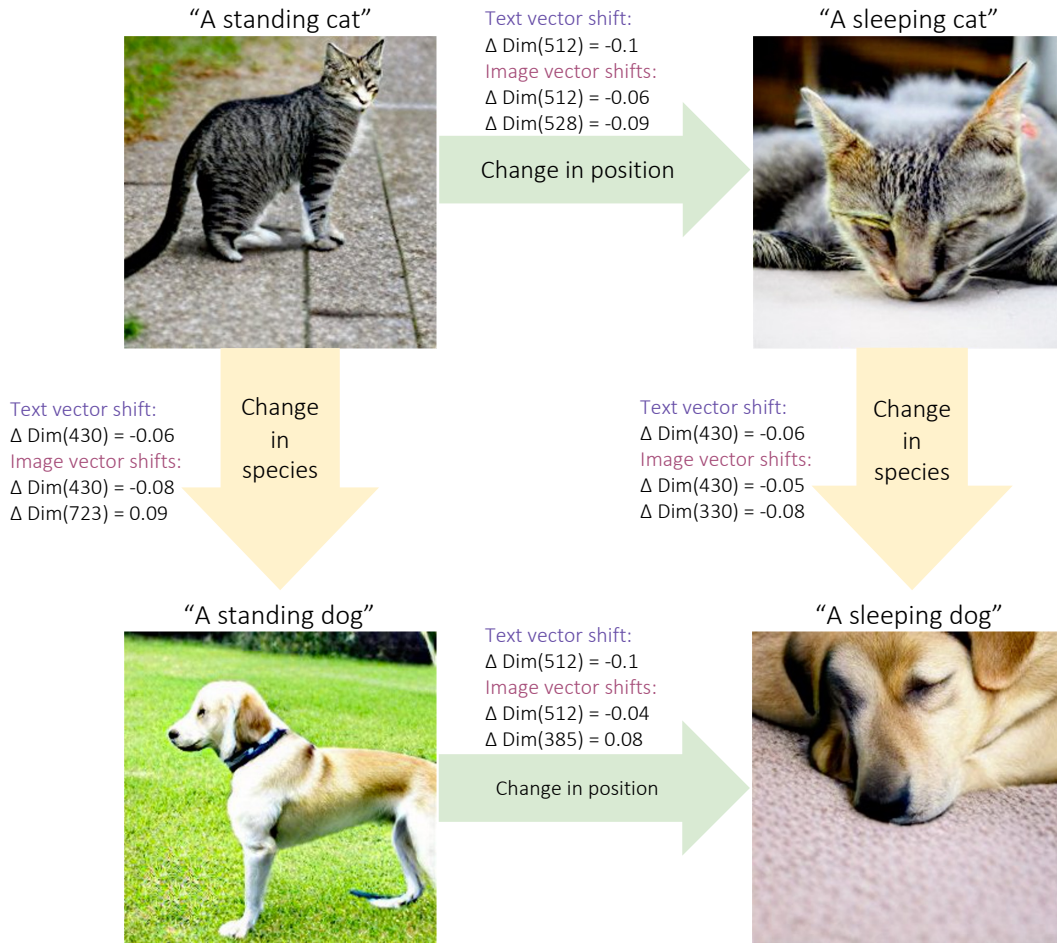


Figure 2: An example of how image/text feature vectors shift with a specific change in species (vertically) or position (horizontally). Text adjacent to each arrow shows any significant changes in the text (purple) or image (red) vector that occurred because of the shift.

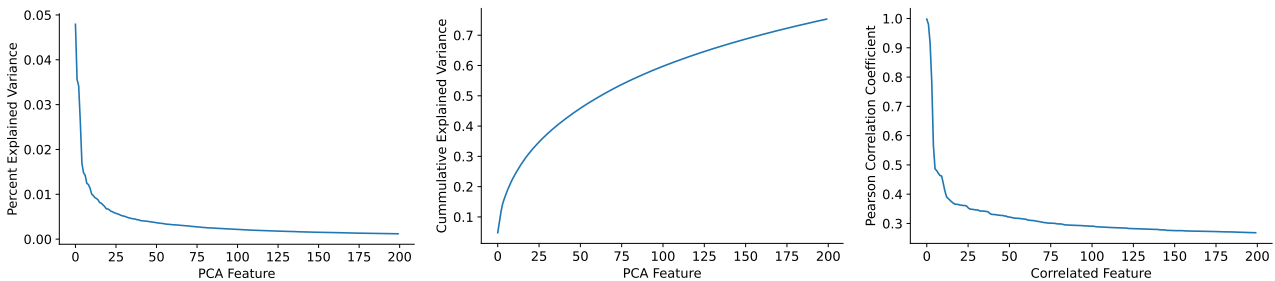


Figure 3: Plots analyzing the differences between image and text vectors for image/caption pairs in COCO captions. Only the first 200 features are shown.

## 5. Visual News Qualitative Examples

We show some qualitative examples for visual news in Figure 4. We observe that close to 50% of time, the predicted captions can be more descriptive (i.e., they can include more details), indicating there is room for this visual

news captioner to grow. There are also some cases in which the predicted captions are better than the ones provided by human (the target captions). But overall, the general sense of both the news images and articles are present in the captions produced by CLOSE.





target caption: the trump family cuts the ribbon

predicted caption: the trump international hotel opened monday inside the old post office pavilion



target caption: lance armstrong waves after receiving the bronze medal in the men's individual time trials at the 2000 summer olympics in sydney

predicted caption: lance armstrong was stripped of a bronze medal won at the tour de france in 2000



target caption: watch a police officer save a man's life during his lunch break

predicted caption: video shows two officers helping a man who reportedly had life-threatening trouble with his foot



target caption: national grid has sought to play down the significance of the energy warning

predicted caption: national grid has used last resort emergency powers to tell companies to reduce their electricity usage

LocalStaying at the new Trump Hotel in D.C.? You'll pay a price beyond \$700 a night.By Petula DvorakA reservation at the new Trump International Hotel in the nation's capital will carry lots of baggage this fall — and not just the kind you would need to haul around the \$700 a night it is going to cost to stay at the swish new place.Emotionally and politically, the hotel that bears Donald Trump's name and opened Monday inside the Old Post Office Pavilion is already sparking fireworks. And protests right in front of the place.Stay at a Marriott. Book a Hyatt. So what? But consider a reservation at a Trump place — Hotels.com has the D.C. property just down the street from the White House on Pennsylvania Avenue at \$761 for Saturday night — and it gets all kinds of complicated. Endorse the Republican presidential nominee's hotel? "Never. Nope. Not a chance," said Becky Acton, who raised her middle finger at the place as she biked by Sunday night, on the eve of its soft opening. "I would never stay there. No matter what it costs."What about its bar, where wine is sold by the spoon? Or the daily Champagne sabering, where bottles are opened by sword?"No interest," she said.Acton is visiting the District from Columbia, Mo. And she stopped to gape at the Trump hotel as construction workers — many of them Latinos who have been on the receiving end of Trump's slurs against immigrants — rushed around in the dark, tile saws screaming when they cut marble outside the front doors in the final, frantic preparations.She shook her head as she pedaled away.The Klyder family had a different take."It's beautiful, like a castle," said Emily Klyder, 11, as she photobombed her mom's numerous pictures of the hotel Sunday night.

LONDON — On the day he went public with an admission of doping after years of denials, Olympic officials disclosed one more embarrassment for Lance Armstrong: He was stripped of a bronze medal won at the 2000 Sydney Games.The International Olympic Committee sent a letter to Armstrong on Wednesday night asking him to return the medal, just as it said it planned to do last month. The decision was first reported Thursday by The Associated Press.LANCE: Armstrong's admission part of long-term comeback planOn Monday, Armstrong taped an interview with Oprah Winfrey for broadcast Thursday and Friday on her network. A person familiar with the situation told the AP that the winner of seven straight Tour de France titles confessed to Winfrey to using performance-enhancing drugs.The timing of the IOC move, however, was not related to the TV interview.The IOC executive board discussed revoking the medal in December, but delayed a decision until cycling's governing body notified Armstrong he had been stripped of his seven Tour de France titles and all results since 1998. He then had 21 days to appeal.Now that the deadline has expired, the IOC decided to take the medal away. The letter to Armstrong was also sent to the U.S. Olympic Committee, which would collect the medal."Having had confirmation from UCI that Armstrong has not appealed the decision to disqualify him from Sydney, we have written to him to ask for the return of the bronze medal," IOC spokesman Mark Adams told the AP. "We have also written to USOC to inform them of the decision."Two months after winning his second Tour de France title in 2000, Armstrong took the bronze in Sydney in the road time trial behind winner and U.S. Postal Service teammate Vyacheslav Ekimov of Russia and Jan Ullrich of Germany.

MLocalWatch police on a lunch break save a man choking on a Subway sandwichBy Justin Wm. MoyerSometimes, a Subway sandwich does not go down smoothly. And when a timely Heimlich maneuver is required, police in Virginia are there to help.That's what Fairfax County police said after posting video of two officers on a lunch break helping a man who reportedly had life-threatening trouble with his foot-long during a June 30 visit to the fast-food chain."Officer Mulhern and Officer Weaver were taking a quick lunch break in a local Subway," the statement, posted to YouTube, said. "A man approached them who was in distress. Officers noticed that the man was choking and sprung into action."As the video's title put it, the officers were "trained and always ready." "Officer Mulhern utilized his training and administered back blows and the Heimlich maneuver," the statement said. "Due to the officers' quick actions, the man's airway was cleared and all were able to finish enjoying their \$5 footlongs."The man, who was not identified, survived — and got a friendly pat on the back from Officer Mulhern.

National Grid has for the first time used "last resort" emergency powers to tell companies to reduce their electricity usage in an effort to avoid the risk of blackouts. It asked firms to reduce their power demand immediately, issuing a so-called demand-side balancing reserve notice to companies that have signed a contract to say they will take part in the demand reduction scheme. A spokesman said this measure had never been used before, while the grid has previously said it would "only be used as a last resort, after all other actions available in the market have been exhausted". Earlier on Wednesday, National Grid issued an urgent request for energy companies to make more power available after multiple breakdowns at UK power stations. Power firms were asked to supply an extra 500 megawatts between 4.30pm and 6pm, a period when power demand surges, with some people still at work and others arriving home and turning the lights on. The owner of Severn power station, Calon Energy, sold electricity to the National Grid at £2,500 per megawatt hour during the afternoon, industry sources confirmed, compared with the typical price at that time of about £60. National Grid issued the original request by sending a "notification of inadequate system margin", a warning that there was not enough power in reserve to keep the lights on in the event of an unforeseen emergency. Shortly before 6pm, National Grid issued a further statement saying suppliers had responded to its urgent request and 40MW of extra power had been ordered, so the NISM had been withdrawn. "This is one of the routine tools that we use to indicate to the market that we would like more generation to come forward for the evening peak demand period," the company said. "The issuing of a NISM does not mean we were at risk of blackouts. It means that we needed the safety cushion of power in reserve to be higher."



**target caption:** this dec 16 photo shows president obama pausing during a speech at an interfaith vigil for the victims of the sandy hook elementary school shooting in newtown conn

**predicted caption:** president obama must decide where to go big on gun control



**target caption:** tourists follow the pathway

**predicted caption:** a view of the san cristobal bridge in the sierra nevada



**target caption:** very few developers have spent time making apps for google glass

**predicted caption:** google's glass app allows users to track their progress in real time



**target caption:** responding to ad blocking google

**predicted caption:** google's accelerated mobile pages project is at its early stages

WASHINGTON — It's hardly a secret that Barack Obama, like every president no doubt, muses about his ultimate legacy and spot in the presidential pantheon. He approaches his second term confronting tough and shifting challenges that will play big roles in shaping the rest of his presidency and his eventual place in history. In the coming months, Obama will have to decide where to be ambitious, where to be cautious, and where to buy time. He draws political strength from his surprisingly easy re-election in a bad economy. It's partly offset, however, by Republicans' continued control of the House, plus their filibuster powers in the Senate. Some of the big issues awaiting the president's decisions are familiar, long-simmering problems. They include immigration and the need for a tenable balance between taxes, spending and borrowing. Another issue, gun control, jumped to the national agenda's top tier this month following the massacre of first-graders and teachers in a Connecticut school. And the issue of climate change remains unresolved. Veteran politicians and presidential historians say it's almost impossible for Obama to "go big" on all these issues. Indeed, it might prove difficult to go big on even one. While some counsel caution, others urge the president to be as bold and ambitious as possible. "Americans are yearning for leadership," said Gil Troy, a presidential scholar at McGill University. As a president dealing with policy, he said, Obama has generally failed to give "that visionary, powerful address that we came to know and love and expect in the 2008 campaign." Rather than let Congress take the lead on big issues, as it did in drafting the 2009 health care overhaul, Obama should be more forceful in pushing new legislation or using his executive powers.

In California, the drought is so much bigger than not being able to water your lawn. We've heard about California's historic drought for years, but today the game changed. While standing on a patch of dry grass in the Sierra Nevada that should have been a snowpack, California Gov. Jerry Brown announced the state's first-ever mandatory cuts in water usage. The state has been working to trim water use since Brown proclaimed a drought emergency last year, but it wasn't enough. More than 98% of the state remains in some level of drought. The water restrictions will affect everything from golf courses to public streets. Campuses, cemeteries and other large landscapes are going to have to make significant cuts in water use. Fifty million square feet of lawns throughout the state will have to be replaced with drought-tolerant landscaping. Families in homes where wells have run dry will have to be relocated. "It's a different world," Brown said. Welcome to California's new normal. What's in #TheShortList: • The "religious freedom" bill in Arkansas has divided the governor's family • How much top NCAA basketball coaches get paid • Controversy around video purported to be from inside doomed Germanwings flight • What it really means to be "smartphone-dependent" Short on time? Listen to the audio version of #TheShortList: Arkansas governor's son urged him not to sign 'religious freedom' bill The "religious freedom" bill in Arkansas is so divisive, it's even split Gov. Asa Hutchinson's own family: His son Seth joined the state's growing opposition to the bill and signed the petition urging him to veto it. Today, the governor said he won't sign the bill in its current form. "It has been my intention all along to have House Bill 1228 to mirror the federal act," Hutchinson said. "The bill that is on my desk at the present time does not ... mirror the federal law." He was referring to the federal Religious Freedom Restoration Act signed by then-president Bill Clinton in 1993.

More than two years after it first introduced Glass to the world, Google is bringing the futuristic device to the UK. But although software developers have had many months to cook up ideas for the eyewear, there are still just a few dozen apps available. Some analysts say the controversial spectacles lack a "killer app" - the one function that will make the average user rush out and buy a pair. Ben Wood, from CCS Insight, says he sees Glass as a "science project," and a "window into what's possible in the future", but by no means a commercial product. But some developers have been giving it a good go, using Glass to... Help the hard of hearing Students at Georgia Tech university have created an app that recognises speech and turns it into on-screen captions, in real time. Conversations appear in text on Glass, allowing the wearer to read what is being said, if they didn't quite catch it the first time. The same boffins are working on a similar app that will be able to translate languages in real time. Fantastique. Put out fires As Patrick Jackson, a firefighter from North Carolina in the US, exhibited in a video, emergency service personnel using Glass will be able to summon up critical data such as floor plans and aerial imagery before they enter a burning building. A company called Mutualink is also testing an app that will allow medics to view a patient's medical records as they arrive on the scene of an accident, and police could use Glass to view footage from security cameras, or record alterations. Make your run more fun Race Yourself, an augmented reality app, aims to make running more fun by letting you compete against a version of yourself. A 3D avatar running at the speed of your last run appears in Glass, allowing the wearer to gauge their progress in real time. You can even race against friends and celebrities, or, if you dare, against rampant zombies. Enhance your gallery visits Knowing very little about art need not be a barrier to enjoying a cultural trip.

Google is attempting to counter the threat from ad-blocking and rivals Facebook and Apple by radically improving the loading speed of web pages on smartphones and tablets. Accelerated Mobile Pages aims to simplify the structure of mobile web pages and place the data needed to deliver them closer to users both physically and virtually in a bid to achieve almost "instant" delivery of articles to anywhere in the world. The project is at its early stages and while the company says it hopes to launch AMP next year, no date has been set. Google said it was unveiling the project early because it wants to work with publishers, the advertising industry and other web platforms such as Facebook and Twitter. The code for AMP will be made public, allowing any company or organisation to tailor it to their own needs. The company said it wanted to collaborate with the wider industry to develop a framework for what works effectively on mobile devices. Early demos of AMP articles show pages with less clutter, such as related stories, but that could change as the project progresses. Google's head of news and social products Richard Gingras said: "This is about making the world wide web great again ... to make sure all users can access the vibrant ecosystem of the world wide web, and get it in a near instant fashion everywhere in the world." The initial plan for AMP has come from the Digital News Initiative, a collaboration between Google and eight European publishers including the Guardian, Les Echos in France, El Pais in Spain and the Daily Mail in the US. The shift away from desktops to mobile devices has left many publishers struggling to get their articles and other content on to people's devices quickly. Those slow speeds are blamed for the rise of ad-blocking software, which stops many publishers from earning money from advertising online.

Figure 4: Examples of visual news captions produced by CLOSE trained on text captions and news articles alone, and then applied zero-shot to news images and articles.