# A Appendix

In this appendix, we provide additional results and ablations studies. The appendix comprises the following subsections:

1. Sec. A.1 studies the way that frame order is manifested in our captions
2. Sec. A.2, details on our approach for CLIP-based frame sampling
3. Sec. A.3, more results for image captioning
4. Sec. A.4, a stress-test analysis with non-homogeneous image sets
5. Sec. A.5, study of CLIP encoded text inversion
6. Sec. A.6, ablation study of different batch size
7. Sec. A.7, examples of evolution of sentences

Contestant singing and then the judges give him a score. **CLIP-S**: 0.653, **PP**: 40.82
Judges giving a score then the contestant sing. **CLIP-S**: 0.616, **PP**: 122.61

The ball goes in the hoop after the man throws it. **CLIP-S**: 0.732, **PP**: 20.87
The ball is thrown from the hoop to the man. **CLIP-S**: 0.694, **PP**: 22.07

Figure 6: Counterfactual examples, which reorder the events in generated sentences. The analysis shows that temporal knowledge is embedded in the language and in visual cues.

## A.1 Studying frame order

Despite showing video captioning as the main application in most of our experiments, our method is invariant with the order of frames in the sequences. This is a result of using a zero-shot strategy, in which the underlying models are trained per frame. Examining the results, the generated captions display a logical order. This is not surprising since ordering frames is typically not an extremely challenging visual understanding task and is also used as a self-supervised task [50]. Evidently, the information that exists in both the Language Model and the CLIP network is sufficient for generating sentences that adhere to the natural order of events.

To further examine this, we created counterfactual sentences in which the order of events is altered and measured the PP and CLIP scores. As can be seen in the examples of Fig. 6, changing the order leads to a drop in both scores. For example, judges giving a score before a contestant sings has a higher perplexity, or a player's pose can be used to identify them as someone who throws a ball and not someone who catches it.

## A.2 CLIP-based Frame Sampling for Video Captioning

This section describes our frame sampling methodology in more detail. We begin with the first frame as an anchor. The next frame (a third of a second later, due to the initial sampling) is included in the set if its distance from the anchor, in the space defined by the CLIP image encoder, exceeds a certain threshold $\lambda_{\text{frame}}$. The newly selected frame becomes the new anchor, and the selection process continues in the same manner. In all our experiments, we use $\lambda_{\text{frame}}$=0.9 as the matching threshold to pick a new frame.

In Fig. 7, we illustrate the CLIP-based mechanism we use to pick a novel and diverse frames. As a result of using CLIP image similarity, the method can find frames with very different content, e.g., where the environment or objects change. We highlight the selected frames with a red border. For example, the first row contains a frame depicting a pitcher, followed by a frame showing the catcher. Frames following this one are ignored until the ball is hit. In the following video, only four frames are selected, filtering out many repetitions. The strategy also works with animations, as shown in the third video.

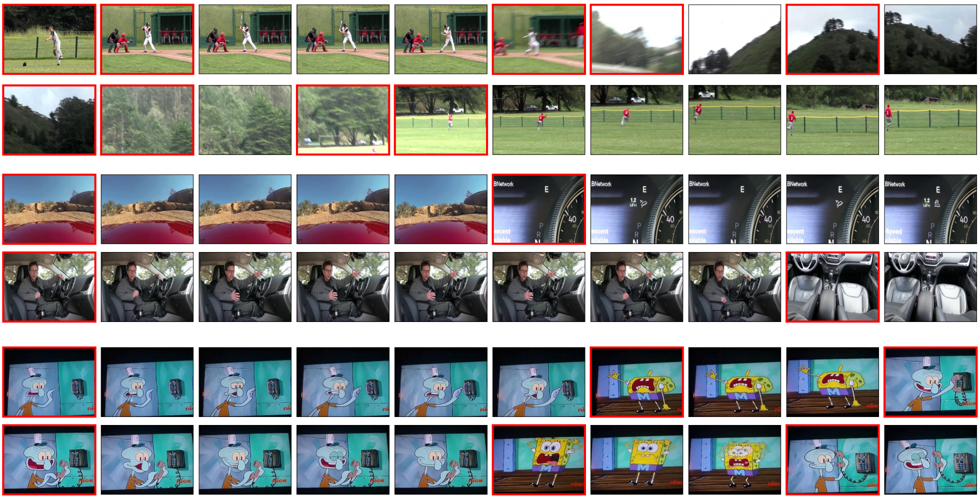


Figure 7: Illustration of our CLIP-based sampling strategy. The picked frames are outlined in red.

## A.3 More Results for Image Captioning

We used COCO's validation set (Karpathy splits) for qualitative and quantitative evaluations. Tab. 5 extends Tab. 2 and compares our method to state-of-the-art image captioning approaches, ZeroCap, and MAGIC. These approaches optimize at the token level resulting in a drop in the perplexity metric. Our sentence-level optimizations generate more fluent captions compared to token-level optimizations. We also assess supervised metrics aimed at language correspondence against human references. MAGIC excels on the supervised metrics. As the method fine-tunes the PLM on the human references, it allows the language to relate to their style. When evaluating captions using CLIP-based scores, performance drops.

Tab. 6 shows results for single-image captioning for different prompts, evaluated on the MS-COCO test-set [22]. We compare our method with another zero-shot method, ZeroCap. Unlike the baseline, our method is trained with a perturbed prefix instead of 'Image of a'. As a result, our method is more robust to prefix changes. Notably, the perplexity score is significantly higher when the prefix is removed from the baseline sentence (109.959 vs. 25.737). Furthermore, compared to the baseline's 0.870 CLIP score, our method has a higher CLIP score of 0.885. CLIPScoreRef is also improved (0.778 vs. 0.798), which means our caption matches human references better. In particular, we optimize complete sentences, resulting in a significant improvement in language fluency (19.049 vs. 25.737).

In Fig. 8, we demonstrate our method's zero-shot image captioning capabilities. We compare with ZeroCap. We find that ZeroCap's captions are more direct, whereas the narrative of our captions is more natural. For example, in the first row, on the left, the captions describe the girls and indicate that it is their summer vacation, whereas ZeroCap mentions what appears in the image. These results might come from the way we construct sentences. By letting the PLM construct sentences, we improve language fluency. ZeroCap, on the other hand, alternates each token to correspond to the image, which might hinder the language.

| | **Supervised Metrics** | | | | | **Unsupervised Metrics** | |
|---|---|---|---|---|---|---|---|
| Method | B@4 | M | C | S | CLIP-S$^{\text{Ref}}$ | CLIP-S | PP |
| VinVL [57] | 0.41 | 0.311 | 1.409 | 0.252 | 0.83 | 0.780 | 24.16 |
| BLIP [18] | 0.40 | 0.311 | 1.367 | 0.243 | 0.82 | 0.759 | 27.738 |
| | *Zero-Shot Methods* | | | | | | |
| ZeroCap [45] | 0.029 | 0.12 | 0.131 | 0.055 | 0.778 | 0.870 | 25.737 |
| MAGIC [42] | **0.129** | **0.174** | **0.493** | **0.113** | 0.763 | 0.737 | 37.126 |
| Ours | 0.022 | 0.127 | 0.172 | 0.073 | **0.798** | **0.885** | **19.049** |

Table 5: Quantitative results for image captioning methods. We evaluate supervised metrics that measure text correspondence to human references and unsupervised metrics that are computed without referring to human annotation.

| | | Supervised Metrics | | | | | | Unsupervised Metrics | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | Prefix | B@4 | M | C | R | S | CLIP-S$^{\text{Ref}}$ | CLIP-S | BLIP-S | PP |
| ZeroCap [45] | None | 0.021 | 0.1 | 0.139 | 0.207 | 0.051 | 0.760 | 0.821 | 0.604 | 109.959 |
| ZeroCap [45] | 'A' | 0.026 | 0.116 | 0.145 | **0.276** | 0.054 | 0.771 | 0.845 | 0.611 | 33.661 |
| ZeroCap [45] | 'Image of a' | **0.029** | 0.12 | 0.131 | 0.268 | 0.055 | 0.778 | 0.870 | 0.605 | 25.737 |
| Ours | None | 0.024 | **0.127** | **0.200** | 0.239 | 0.071 | 0.791 | 0.852 | **0.652** | 20.412 |
| Ours | Random | 0.022 | **0.127** | 0.172 | 0.228 | **0.073** | **0.798** | **0.885** | 0.651 | **19.049** |

Table 6: Quantitative results for image captioning on the MS-COCO test set using different prefixes.

**ZeroCap**: Group of hikers viewing the countryside.
**Ours**: The girls' view from a hilltop in Hampshire, teens enjoying their summer holidays.

**ZeroCap**: Remote car in the sunset.
**Ours**: The sun setting on television remote control of a vehicle.

**ZeroCap**: Group of Vietnamese male aging in the city of Lima.
**Ours**: Tibetan men in a traditional hat and sunglasses.

**ZeroCap**: Pizza served with wine in New York.
**Ours**: The pizza dinner in a restaurant, with food and wine served to guests.

**ZeroCap**: Dinosaur sand elephant in Beijing on May 1.
**Ours**: Fukushima's giant elephant sand sculpture being dug up by the city.

**ZeroCap**: 2013 ski racer.
**Ours**: Ski racer wearing a red top and shorts.

Figure 8: Examples of image captions.

**ZeroCap:** A genius CEO is not a genius in the world of of Silicon Valley billionaires
**MAGIC:** A man smiling while holding glasses of wine.
**Ours:** Microsoft billionaire and philanthropist Bill Gates, who is chairman of the foundation that has been criticized for supporting..

**ZeroCap**: A wall in the Chinese city of Gansu is a great hit hit.
**MAGIC**: A view of a big tower with a clock on it.
**Ours**: The world's largest wall in China, complete with a stunning view from above.

**ZeroCap**: A city in the Chinese blockchain network Zha dong (not a city in
**MAGIC**: A view of a city street from a tower.
**Ours**: Beijing's futuristic office building, which is expected to be one of the most expensive buildings in history.

**ZeroCap:** A city in Cairo taken from Shutterstock The Egyptian city of Cairo has been given a..
**MAGIC:** A view of a city street with a big, beautiful clock tower.
**Ours:** Cairo's ancient city center and its many wonders, including the pyramids.

**ZeroCap**: A historic Taj Mahal in India.
**MAGIC**: A view of a very big, luxurious,
**Ours**: Taj Mahal, which is a tourist destination in India's westernmost state.

**ZeroCap**: A map that shows the state is in the hands.
**MAGIC**: A red and green tour bus stands idly in the middle of a
**Ours**: Italian state logo on a map showing the country's borders, with its name and symbols of national identity.

Figure 9: More examples of our image captions on examples that require real-world knowledge.

## A.4 Stress-test with Non-homogeneous Image Set Captioning

In order to stress-test our method, we consider the task of captioning random image sets. The goal is to describe a set of images with one coherent sentence. We use the MS-COCO test-set [22] for images. The number of images varies between one and four, one being the conventional image captioning task.

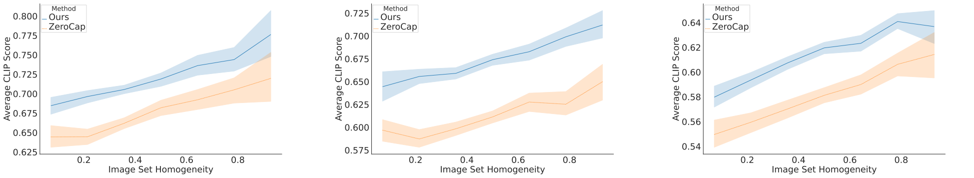
It is expected that the more heterogeneous the image set, the harder it is to generate a coherent caption. To quantify this, we measure the homogeneity of a set by the average CLIP score between the human captions of each image and the rest of the images in the set. Intuitively, this tells us whether the images depict similar concepts or not. In addition to comparing sets by their size, we also differentiate them based on their level of homogeneity. The results are shown in Fig. 10. The graphs reveal that the CLIP score increases with homogeneity, which is expected since a single sentence can describe homogeneous images better. Our approach has a higher CLIP score at all levels of homogeneity and for all set sizes.

In Fig. 11, an experiment similar to the one above, based on BERT-based perplexity, is conducted to measure language quality. We find that our method produces much better sentences. A particularly interesting case is that of homogeneous pairs (i.e., homogeneity level of 0.8). Only in this case, which involves two very similar images, does ZeroCap perform as well as we do (logPP of 3.0 vs. 3.5). This highlights the ability of our approach to generating coherent sentences that describe a set of images across various challenges.

In Fig. 12 we illustrate captions generated for sets of various sizes. We are able to identify and describe the content of two images even if there is no significant correlation between them. A stop sign and surfing images are translated to “Surfing stops...", while pictures of a toilet and ladies in formal attire are captioned with “The toilets at the wedding reception.”. Also, pictures of a sheep and a birthday cake are captioned with “Sheep’s birthday...”.

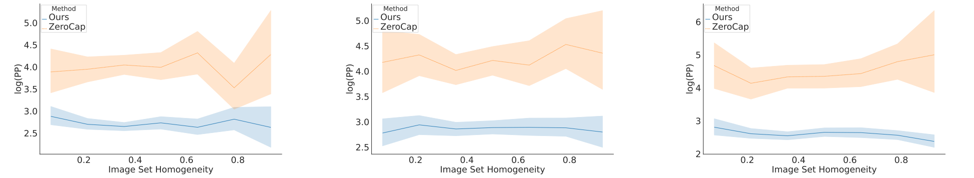
When three images can be described with a coherent story, the model can do so. As an example, for a set of images of a bus, a hotel bed, and a beach, our method generates the caption: “Photo of bus driver sleeping on the beach from the hotel.”. This caption grounds all the images while still creating a plausible narrative. In addition, even when real-world knowledge is necessary, e.g., a picture of Obama, the caption relates to it.

Our method was able to produce a coherent narrative even when dealing with a complex case of four images. It describes a narrative of an image of birds taken while cycling in Melbourne. We note that ZeroCap’s sentences tend to create an irrational context, e.g., “Captive Obama...,” which is perhaps the result of token-based optimization rather than sentence-based optimization.




(a) Two images (b) Three images (c) Four images

Figure 10: CLIP-score for different sets of images varying by size and set homogeneity.




(a) Two images (b) Three images (c) Four images

Figure 11: BERT-based perplexity score for different sets of images varying by size and set homogeneity.

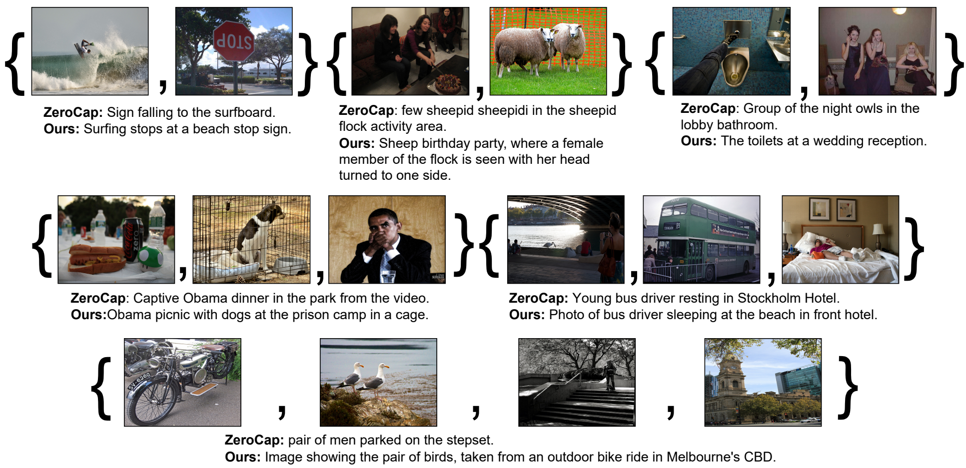



Figure 12: Examples of our image set captioning, for different set sizes. We compare our method with ZeroCap, another zero-shot method.

## A.5 Inversion of CLIP-encoded Text

We can treat our method as a general inversion technique from CLIP embedding to text. As such, a CLIP-encoded text can be directly inversed to measure our method's abilities. The benefit of this experiment is that instead of evaluating the captioning of CLIP-encoded visual cues using subjective captions provided by the annotators, here, the encoded caption text objectively reflects the CLIP-encoded representation. We use a set of 5k samples from Google's Conceptual Captions validation set [38]. We choose this dataset since in contrast with the curated style of other image caption annotations, Conceptual Caption images and their raw descriptions are harvested from the web and, therefore, represent a wider variety of styles.

In Tab. 4 in the main paper, we show that on all datasets, the perplexity of the generated captions is much lower in our case, showing that our method generates fluent captions. Moreover, the CLIP score is highest with our captioning. Thus, the inverted text is the closest to the original caption in CLIP space. We find that the perplexity performance of ZeroCap is low. Upon studying this, we find that ZeroCap has difficulty generating long sentences from encoded text. Also, we find that MAGIC performance is low compared to ours. Our method is significantly better on all metrics, especially on B@1, showing our approach's ability to find specific terms of the original caption.

Further, we show the inversed text of other captions and randomly selected captions below. Our captions pick up on the original concept much better than ZeroCap captions, which are often short. We also find that MAGIC vocabulary is limited, and the object it describes often does not exist in the original caption. For example, MAGIC switches cows to cats and shoes to skateboards. This might be because the PLM MAGIC employs has been fine-tuned using MS-COCO captions.

**Text Inversion of Captions from Conceptual Captions:**

**GT:** cow on a pasture in the mountains
**Ours:** Cow in mountains near the mountain range.
**MAGIC:** A cat looking at a camera in a field.
**ZeroCap:** Cow wandering mountains.

---

**GT:** wedding shoes from the 19th century .
**Ours:** Shoes worn in the nineteenth century wedding.
**MAGIC:** A skateboarder does a leg trick on a skateboard ramped
**ZeroCap:** Shoes made in 18000.

---

**GT:** the architects have designed thegreen heart of a contemporary city which responds to the diverse demands of 21st century public spaces
**Ours:** City design architecture combining the green and sustainable elements in neighbourhoods with architectural principles that
**MAGIC:** A living building with a green park at the heart of it.
**ZeroCap:** City in the 2030.

---

**GT:** music video performer performs at festival .
**Ours:** Music performance artist has been featured on stage and performed at festivals across the world
**MAGIC:** A picture of a person performing a very cool trick.

**ZeroCap:** Song artist.

---

**GT:** beautiful dog of breed labrador retriever sitting inside an open baggage
**Ours:** Labrador retriies are a beautiful puppy is great and loving mix.
**MAGIC:** A dog sits on luggage of a baggage claim.
**ZeroCap:** Cat labrador sitting in baggage.

---

**GT:** photo of western christian holiday on a dark background with greenery around it
**Ours:** Dark background image of holiday photo taken from the holidays picture backdrop, christmas scene
**MAGIC:** A picture of a black and white image of a black and white background
**ZeroCap:** Holiday in the dark background.

---

**GT:** wedding - bride and person on there wedding day
**Ours:** Bride photo wedding couple in the same time.
**MAGIC:** A couple of women dressed in wedding wedding wedding attire.
**ZeroCap:** Old wedding.

---

**GT:** spirit of an astronaut , boys looking thru a telescope
**Ours:** Telescope observing crew looking at the astronaut's eyes
**MAGIC:** A view of a plane and a couple of guys on the ground.
**ZeroCap:** Telescope.

---

**GT:** actor poses on red carpet of new movie
**Ours:** Movie premiere premieres actor's debut film has a huge hit and is the
**MAGIC:** A picture of a guy on a
**ZeroCap:** Lifetime star actor from.

---

**GT:** student and person plays symphony on the violin in this file photo .
**Ours:** Violin student playing in the music teacher, with a soloist is distinguished by
**MAGIC:** A picture of a couple playing music together.
**ZeroCap:** Concertmaster performing in the style of of a.

---

**GT:** preparation is the key to keeping a healthy , balanced diet .
**Ours:** Healthy preparation and eating right includes proper nutrition is always important.
**MAGIC:** A picture of a living room has some very very tasty looking pieces of
**ZeroCap:** Healthy eating day is based off.

---

**GT:** the benefits of outside free play - play is the most important work of early childhood .
**Ours:** Play outside education impacts preschool wellbeing is a lifelong approach.
**MAGIC:** A little kids play with a birthday cake.
**ZeroCap:** Fun playground with with.

---

## A.6 Experimental Setup and Ablation Study

The following settings are used: We set $\lambda$ to 0.8. During sentence generation, we pick one of the top-3 tokens at random. To avoid long repetitive sentences, the number of generated tokens per sentence was limited to 20. To avoid generating irrelevant entities, such as names, we reduce by 1 the logits of tokens with uppercase letters. Using a single Titan X GPU, all twenty sentence-generating iterations take approximately a minute.

To assess different hyperparameters, we use the MSVD [49] validation set, which consists of 100 videos. We examine two properties: (i) Video correspondence, which we examine with the Retrieval score, and (ii) language fluency, which we analyze with the BERT perplexity score. Additionally, we report CLIP Score and BLIP score, which measure image correspondence with the selected frames.

In Fig. 13, we study different values for $\lambda$, which controls the trade-off between CLIP loss ($\mathcal{L}_{\text{CLIP}}$) and language fluency loss ($\mathcal{L}_{\text{PLM}}$). Increasing the value of $\lambda$ decreases the Retrieval score. Our results show that $\lambda = 0.8$ provides a good trade-off between image correspondence and language fluency (i.e., low perplexity).

In Fig. 14, we ablate the learning rate (i.e., $\alpha$). Since optimization occurs during inference, the number of iterations is fixed, so a higher learning rate ensures convergence. In our experiments, we use $\alpha = 0.1$, which has the lowest perplexity, and the highest Retrieval score. Note that the graphs might be misleading due to the wide range of values. The method is relatively stable for this parameter.

In Fig. 15, we study different prompts. In our method, we perturbed a prompt for each generated sentence to increase robustness to different scenarios (e.g., image set captioning and videos). Note that while the option of no prefix at all results in a good performance, we find it less focused on the task of visual captioning.

In Fig. 16, we demonstrate how the CLIP score progresses during the generation process. We report the following statistics: (i) Mean is the average CLIP score at the given iteration across the set (ii) Max is the maximum CLIP score at the given iteration across the set (iii) Best Mean is the best mean score up to this iteration. The challenge of fitting to multiple visual cues can cause performance instability during optimization. Thus, we suggest selecting the sentence with the highest CLIP score from all the generated sentences.

In Fig. 17, we assess our CLIP-based sampling method. Our method employs CLIP's visual encoder to compute image similarity. The Retrieval score increases, as can be expected, with the CLIP image similarity. The perplexity score is relatively stable, but there is a trade-off between the two.

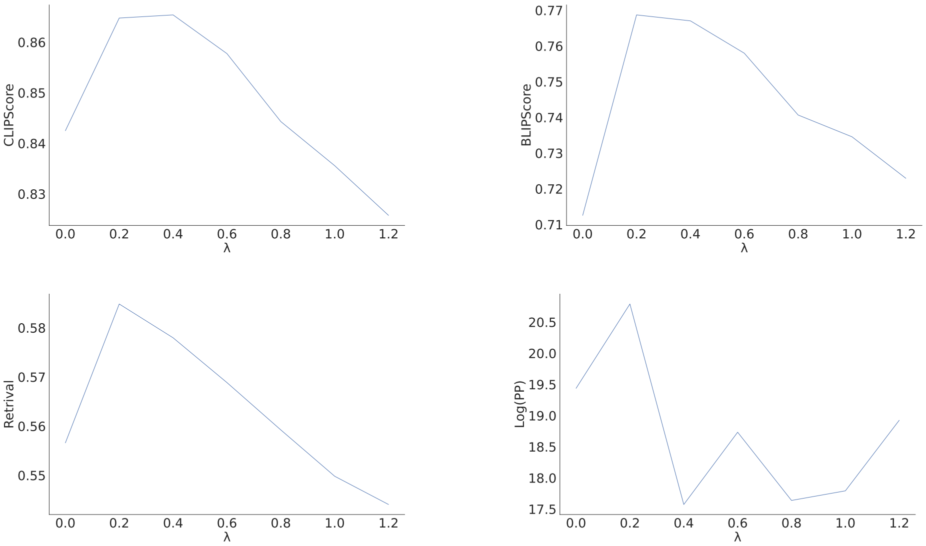



Figure 13: Ablation study for the hyper-parameter $\lambda$.

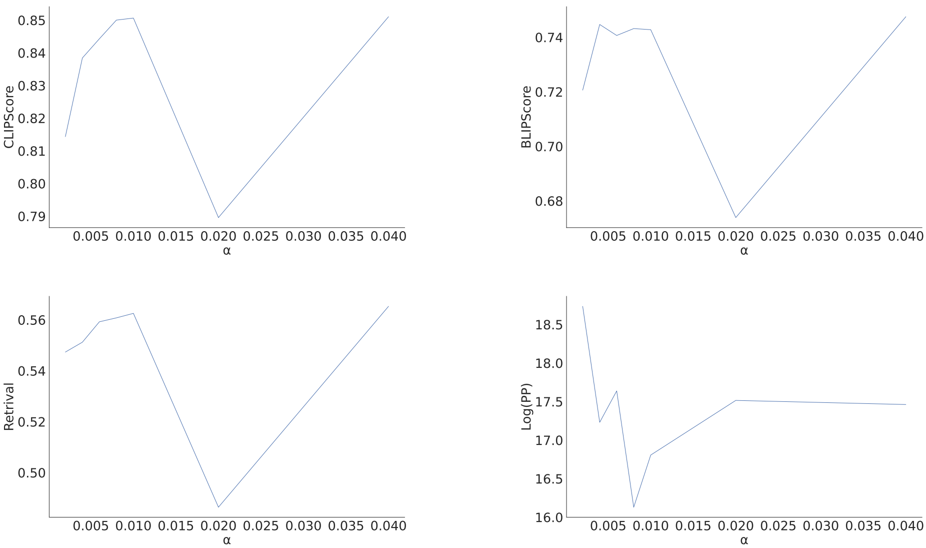



Figure 14: Ablation study for the learning rate, i.e., $\alpha$.

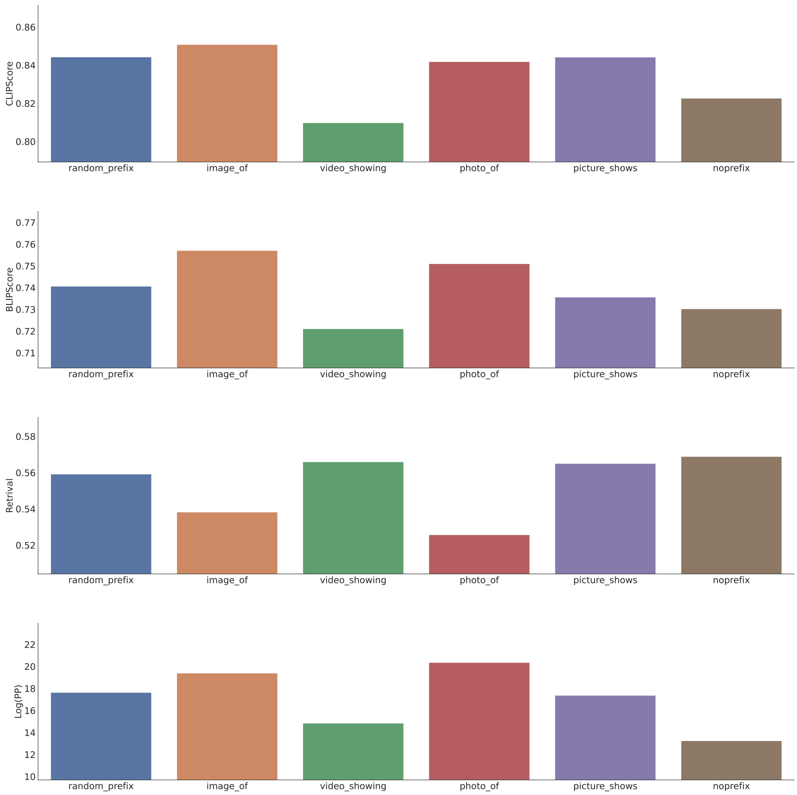
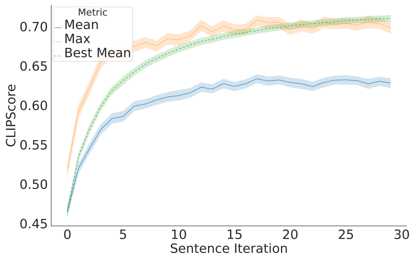
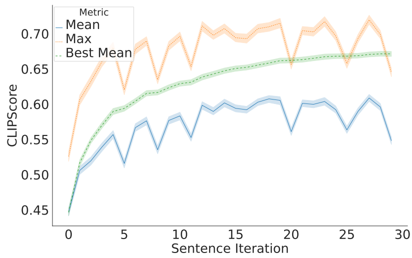


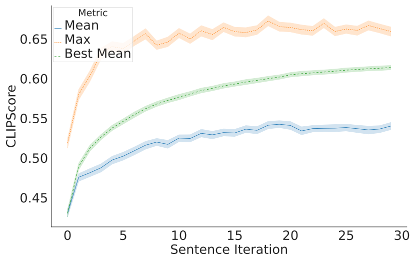

Figure 15: Ablation study for different prompts.




(a) Two images




(b) Three images




(c) Four images

Figure 16: CLIP Score progress over the generation process.

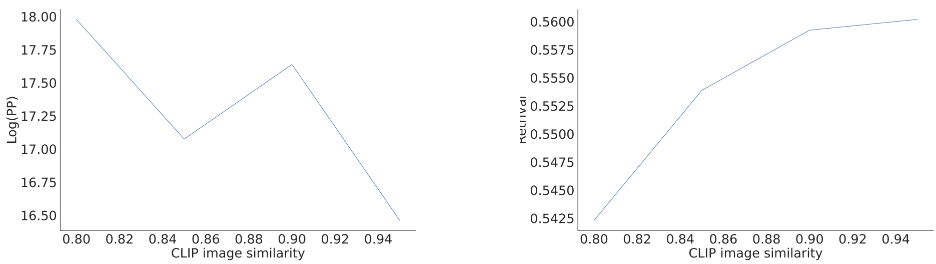



Figure 17: Ablation study for the CLIP-based frame selection method. We ablate different threshold values used to pick significant frames.

## A.7 Evolution of Sentences through Pseudo-token Optimization

In Fig. 18 we evaluate how sentences evolve during the inference process. We first note that, quantitatively, the CLIP score increases between generation iterations, e.g., on the left, at the fourth sentence iteration, the clip score is 0.58, while at the 16th iteration, the score is 0.89. This improvement can also be seen qualitatively. In the left video, we see that the sentence grounds the truck, which is visible in all frames, after four iterations. The model grounded both the truck and trailer in its eighth iteration. Only after the 16th iteration does the model recognize it as a Lego truck. We note an interesting failure case in which, after twenty iterations, the model incorrectly identifies the type of video as a trailer. The video on the right shows similar behavior. In the 16th sentence iteration, the CLIP score increased from 0.70 to 0.88. In the fourth iteration, the video was grounded to more abstract objects (e.g., soldier, battle, alien), while the eighth iteration identified the characters from Halo. As a final step, the model figures out that the video is an animated cartoon in the 16th iteration.

In Fig. 19 we illustrate progression of captions in a stress test of two different images that are not taken from the same video. The left image set shows two very different pictures of a toy bear and a baseball game. Earlier captions discuss the crowd and the dinner separately. The eighth iteration improves grounding, and the method recognizes the baseball game. A coherent narrative is built in the 16th iteration. It is described as a table of a pitcher at a dinner party. There are baseball cards on the table, and bears serve as a metaphor for phrasing a quote. For the right image set, after four iterations our method generates a caption that includes the word Pyongyang as the location and the word 'wildlife'. In the eighth iteration, the caption identifies the animal as a bird. As a result of detecting Pyongyang as the location, the bird is described as being from the DPRK. A reference is also made to the flowers.

In Fig. 20, we show more examples for the full generation process for videos. We present the frames selected by our CLIP-based sampling method for each video. Additionally, we report BERT-based perplexity score and CLIP score. The low perplexity score indicates that early sentences have good language, but subsequent sentences improve the CLIP score significantly. Our method can ground objects and generate coherent sentences in various contexts.

In Fig. 21, we illustrate the evolution of sentences, using two images. Interestingly, the method uses stories to weave the photos into a coherent story. For instance, the image of prison and a bedroom photo results in a caption about a prisoner's bedroom.

In Fig. 22, three images are employed, and the generation process is displayed. Often, creating a coherent sentence from three images is too challenging. Therefore, in those cases, it is better to choose the sentence based on the perplexity indicator rather than using the CLIP score. Thus, the language will be fluent as it describes a storyline without describing everything in every image. Fig 23 shows the same phenomenon when there are four images.

4th **iter** (0.58): Image shows a **truck driving** on the right side of this photo taken in late April.
8th **iter** (0.66): Image showing a **truck driver pulling the trailer** out from under **cars**.
16th **Iter** (0.89): Photo of **Lego truck driving around** in a **trailer** video.

4th **iter** (0.70): Image shows a **soldier** in **battle** with the **soldiers** from his past, including one scene of an **alien invasion** where he
8th **iter** (0.73): Image shows **Halo franchise** characters **fighting** the **aliens** in a video game trailer.
16th **Iter (**0.88): Image showing the **alien soldiers** from **Halo's animated cartoon**.

Figure 18: Evolution of video captions. We show the sentence with the highest CLIP score at different generation iterations. Grounded words are highlighted.

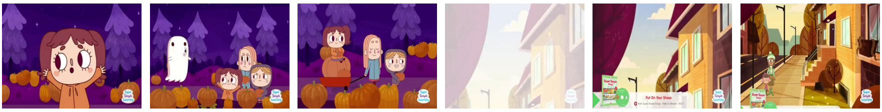
**4$^{th}$ iter** (0.57): Image of 'a few hundred people gathered in the crowd at dinner table, with some sitting around watching' says
**8$^{th}$ iter** (0.61): Image shows baseball player and former mayor of the city, who is now a member club dinner.
**16$^{th}$ Iter (**0.66): Picture shows pitcher's table at ballpark dinner party, baseball card says 'it was a hit for the bear,

**4$^{th}$ iter** (0.62): how the animal is being used as part of a conservation project by wildlife experts in Pyongyang.
**8$^{th}$ iter** (0.66): Photo shows Pyongyang's new bird sanctuary in the sky.
**16$^{th}$ iter** (0.71): Image shows DPRK's birds, flowers and wildlife statues at the memorial park in Pyongyang.

Figure 19: Illustration of how sentences evolve when presented with two non-homogeneous images.

Figure 20: The evolution of captions for videos. (below and for multiple pages)

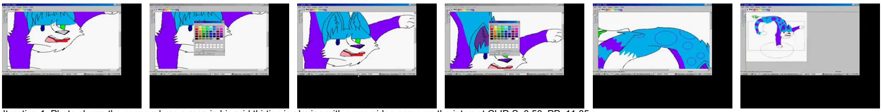
Iteration 1: Photo shows the moment an officer shot a man in front of his wife. CLIP-S: 0.43, PP: 17.33
Iteration 2: Photo shows the original poster's picture on his website, and it says he is a member of the film crew CLIP-S: 0.55, PP: 8.64
Iteration 3: Image of a man holding up two hands and the words, 'I am not an actor', in a film CLIP-S: 0.61, PP: 6.79
Iteration 4: Image shows a scene of the film being shown to an audience in movie theatre. CLIP-S: 0.68, PP: 15.42
Iteration 5: Image showing video camera footage of the kitchen appliance, which was discovered in a trailer at an apartment building on television CLIP-S: 0.70, PP: 10.98
Iteration 6: Video shows the scene of a kitchen fire that was filmed in front oven. CLIP-S: 0.56, PP: 16.72
Iteration 7: Photo shows the scene of an explosion at movie studio in a kitchen television set, circa late'sixties. CLIP-S: 0.69, PP: 10.77
Iteration 8: Image shows a television episode aired in the kitchen of an abandoned movie theatre. CLIP-S: 0.62, PP: 18.63
Iteration 9: Video shows an episode of television series, 'The cooking program,' in which a man is shown eating an episode CLIP-S: 0.71, PP: 7.41
Iteration 10: Picture shows the film in which a radio broadcast is played. CLIP-S: 0.71, PP: 26.28
Iteration 11: Picture showing kitchen door opening, with a radio dish on top. CLIP-S: 0.70, PP: 42.36
Iteration 12: Video of kitchen scene in movie, which was filmed during filming on the radio show. CLIP-S: 0.72, PP: 10.24
Iteration 13: Image showing the kitchen scene from a movie starring director of 'Star Wars episode. CLIP-S: 0.69, PP: 14.32
Iteration 14: Photo of a kitchen in the movie 'Dr. CLIP-S: 0.70, PP: 39.86
Iteration 15: Image shows a kitchen in the movie's opening scene. CLIP-S: 0.76, PP: 27.30
Iteration 16: Photo of the kitchen scene from film adaptation 'Starring a movie with me, starring an alien spy', which CLIP-S: 0.72, PP: 5.91
Iteration 17: Picture of the day, movie director's wife and cook in kitchen film from episode. CLIP-S: 0.72, PP: 9.35
Iteration 18: Picture showing a kitchen mirror film in the fridge of movie director. CLIP-S: 0.72, PP: 21.01
Iteration 19: Image shows the film's opening scene in which a kitchen stove is seen to have been turned into the microwave. CLIP-S: 0.81, PP: 7.57
Iteration 20: Image showing kitchen scene from the movie, where it was filmed on location in a diner. CLIP-S: 0.71, PP: 17.34



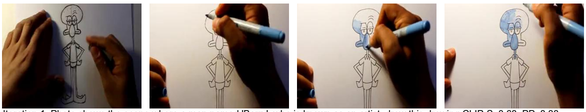
Iteration 1: Photo shows the moment that a man's face is seen in video of him being shot by police officer who was CLIP-S: 0.41, PP: 6.37
Iteration 2: Photo shows how much of the original image was removed from video. CLIP-S: 0.57, PP: 27.58
Iteration 3: Image of the week is a cartoon version that was released in episode two. CLIP-S: 0.66, PP: 12.93
Iteration 4: Image shows the animated trailer for episode 'Kaiyo's Revenge,' which was released on YouTube. CLIP-S: 0.63, PP: 7.68
Iteration 5: Image showing the episode from season finale of 'Lost Halloween', which aired on October, November and December episodes. CLIP-S: 0.62, PP: 6.82
Iteration 6: Video shows cartoon animation from a movie that was animated by the animators of 'Ghosts. CLIP-S: 0.64, PP: 8.47
Iteration 7: Photo shows a cartoon character being attacked by the animator, but it's not actually from that movie. CLIP-S: 0.59, PP: 8.96
Iteration 8: Image shows the cartoon character in his costume from 'Ghostbusters', but it was not animated by him, as CLIP-S: 0.57, PP: 7.12
Iteration 9: Video shows a clip of an animated version of the episode. CLIP-S: 0.67, PP: 27.15
Iteration 10: Picture shows the ghost of a man who appeared in video posted online. CLIP-S: 0.54, PP: 19.08
Iteration 11: Picture showing Halloween ghost story, but not a real one. CLIP-S: 0.59, PP: 30.27
Iteration 12: Video of Halloween episode on the anime series, 'Ghosts in a cartoon', which was released by animator CLIP-S: 0.66, PP: 7.71
Iteration 13: Image showing the animation on how to animate a character in this episode of. CLIP-S: 0.68, PP: 17.39
Iteration 14: Photo of Halloween episode by animator, via YouTube user jaylindie. CLIP-S: 0.69, PP: 12.58
Iteration 15: Image shows the animation of how a ghostly character would appear on screen. CLIP-S: 0.69, PP: 31.37
Iteration 16: Photo of the episode by animator and cofounder, Chris H. CLIP-S: 0.63, PP: 21.35
Iteration 17: Picture of the week, episode title and theme music by josh. CLIP-S: 0.60, PP: 16.65
Iteration 18: Picture showing animation of the ghostly characters in episode. CLIP-S: 0.72, PP: 47.03
Iteration 19: Image shows the animated animation that will appear in episode of 'Ghostly Halloween', a short film. CLIP-S: 0.70, PP: 7.30
Iteration 20: Image showing Halloween episode from the animated series, 'Scooby and Friends', on YouTube. CLIP-S: 0.73, PP: 9.63


Iteration 1: Photo shows the scene after a recent meal at an upscale restaurant in downtown. CLIP-S: 0.49, PP: 15.39
Iteration 2: Photo shows the contents of food items sold on shelves in an online supermarket. CLIP-S: 0.58, PP: 27.82
Iteration 3: Image of food ingredients in recipe above is from the menu at a restaurant called 'Chocolate, chocolate and cheese CLIP-S: 0.67, PP: 8.44
Iteration 4: Image shows a chicken eating the fried egg on an Argentinian breakfast dish. CLIP-S: 0.57, PP: 18.19
Iteration 5: Image showing the recipe in Spanish and English, which was published on a website called 'La de las dor CLIP-S: 0.77, PP: 7.64
Iteration 6: Video shows a recipe for homemade chocolate chip cookie dough. CLIP-S: 0.50, PP: 31.96
Iteration 7: Photo shows a food cart in front of the restaurant. CLIP-S: 0.49, PP: 39.87
Iteration 8: Image shows a breakfast tacos recipe that includes fried chicken fries and salsa beans enchiladas. CLIP-S: 0.55, PP: 14.37
Iteration 9: Video shows friesuitsupea, french and eggs being fried. CLIP-S: 0.55, PP: 27.86
Iteration 10: Picture shows food being prepared for breakfast at a cafe in Argentina, which has been banned by authorities after it was CLIP-S: 0.70, PP: 10.23
Iteration 11: Picture showing the food being sold in a restaurant advertising fries as they were fried and chips on the side. CLIP-S: 0.66, PP: 12.24
Iteration 12: Video of Spanish fries being served in breakfast advert. CLIP-S: 0.65, PP: 26.91
Iteration 13: Image showing a recipe for the breakfast chocolate chips. CLIP-S: 0.66, PP: 61.49
Iteration 14: Photo of the breakfast recipe on a menu at an Argentine restaurant, which is now being sold in the United states CLIP-S: 0.68, PP: 7.33
Iteration 15: Image shows a photo from the menu of fries, which were made with caramel sauce and fried eggs. CLIP-S: 0.70, PP: 12.52
Iteration 16: Photo of fries courtesy www recipes from the website. CLIP-S: 0.69, PP: 149.85
Iteration 17: Picture of the breakfast recipe, courtesy www recipes. CLIP-S: 0.67, PP: 52.34
Iteration 18: Picture showing the breakfast recipes in a Spanish restaurant, which was served with fries. CLIP-S: 0.69, PP: 13.73
Iteration 19: Image shows a breakfast menu from an advertisement for the fries in Buenos Aires, Argentina. CLIP-S: 0.72, PP: 13.72
Iteration 20: Image showing a recipe fries from the restaurant website of chef's son, who has since gone on a tirade CLIP-S: 0.72, PP: 8.88

Iteration 1: Photo shows the aftermath of a shooting at an event. CLIP-S: 0.42, PP: 45.56
Iteration 2: Photo shows off some of the new footage from video streaming site Twitch, showing a stream that is currently live on CLIP-S: 0.75, PP: 13.86
Iteration 3: Image of how it works, but not the actual streamer. CLIP-S: 0.70, PP: 28.95
Iteration 4: Image shows stream of video from the game being uploaded to twitch chat by player. CLIP-S: 0.74, PP: 22.09
Iteration 5: Image showing video of the streamer's server going down after a player called out for being banned. CLIP-S: 0.71, PP: 12.67
Iteration 6: Video shows the livestream of a man playing Minecraft. CLIP-S: 0.74, PP: 37.46
Iteration 7: Photo shows footage of player livestreamed in game. CLIP-S: 0.75, PP: 40.77
Iteration 8: Image shows the game streamer who uploaded a screenshot of himself playing. CLIP-S: 0.78, PP: 16.37
Iteration 9: Video shows livestream of the stream that was uploaded to twitch. CLIP-S: 0.74, PP: 22.26
Iteration 10: Picture shows livestream of the scene in question, uploaded to twitch. CLIP-S: 0.68, PP: 20.33
Iteration 11: Picture showing a stream player in the game Minecraft. CLIP-S: 0.80, PP: 41.00
Iteration 12: Video of Minecraft's gameplay livestream is now up on twitch, and it was just as hilarious. CLIP-S: 0.72, PP: 11.35
Iteration 13: Image showing the stream of players in a match. CLIP-S: 0.68, PP: 53.65
Iteration 14: Photo of gameplay footage courtesy twitchtv screenshot youtube. CLIP-S: 0.80, PP: 21.87
Iteration 15: Image shows the gameplay in action at a livestream of an early alpha build. CLIP-S: 0.84, PP: 21.73
Iteration 16: Photo of the livestream from a game in which we played. CLIP-S: 0.72, PP: 35.26
Iteration 17: Picture of the stream from a livestreamer's view on twitch. CLIP-S: 0.73, PP: 29.40
Iteration 18: Picture showing the livestream of an event hosted in game by a player called '. CLIP-S: 0.80, PP: 20.89
Iteration 19: Image shows gameplay footage from a livestream of the upcoming raid on an area server. CLIP-S: 0.79, PP: 17.06
Iteration 20: Image showing gameplay of the game, which was livestreamed on twitch. CLIP-S: 0.82, PP: 45.51



Iteration 1: Photo shows the scene where a man in his mid thirties is playing with some video game on the internet CLIP-S: 0.50, PP: 11.05
Iteration 2: Photo shows a screenshot of the image that was posted on imgur. CLIP-S: 0.58, PP: 13.14
Iteration 3: Image of a video showing the animation on screen. CLIP-S: 0.81, PP: 56.22
Iteration 4: Image shows a video showing the animation that appears on screen in an animated GIF. CLIP-S: 0.77, PP: 18.24
Iteration 5: Image showing animation of the graphic design process in action. CLIP-S: 0.70, PP: 40.93
Iteration 6: Video shows footage from the camera that was used to create a fake image of an object being drawn by computer, CLIP-S: 0.66, PP: 8.90
Iteration 7: Photo shows video showing a graphic drawing of the image being created. CLIP-S: 0.77, PP: 23.14
Iteration 8: Image shows a drawing of an illustration from the game's website showing how to create animation using text. CLIP-S: 0.73, PP: 9.06
Iteration 9: Video shows a graphic animation that depicts the creation of an image. CLIP-S: 0.69, PP: 29.20
Iteration 10: Picture shows video of a man being hacked to death in the game. CLIP-S: 0.64, PP: 15.01
Iteration 11: Picture showing a screenshot animation drawing of the game character animations. CLIP-S: 0.76, PP: 20.64
Iteration 12: Video of animation drawing software created by animator, artist. CLIP-S: 0.77, PP: 23.74
Iteration 13: Image showing animation drawing of a character from the game. CLIP-S: 0.71, PP: 40.41
Iteration 14: Photo of a drawing from an animation program on youtube. CLIP-S: 0.75, PP: 22.22
Iteration 15: Image shows a browser window on an animated web page. CLIP-S: 0.72, PP: 49.42
Iteration 16: Photo of a drawing animation showing how to create an animated video using the browser. CLIP-S: 0.80, PP: 14.44
Iteration 17: Picture of the day, animated animation software that allows users to create a virtual world with text and images. CLIP-S: 0.71, PP: 9.64
Iteration 18: Picture showing animation of how the browser window would look like in a web page with mouse and text. CLIP-S: 0.71, PP: 9.92
Iteration 19: Image shows animation of the user painting with paint brush and drawing a cartoon character. CLIP-S: 0.78, PP: 23.24
Iteration 20: Image showing animation of a mouse's body moving through paint. CLIP-S: 0.68, PP: 26.33



Iteration 1: Photo shows the scene where a man named 'Barry', who is known as an artist, drew this drawing CLIP-S: 0.68, PP: 8.33
Iteration 2: Photo shows the character of a cartoon drawn by artist drawing from an animated animation series. CLIP-S: 0.80, PP: 13.97
Iteration 3: Image of a cartoon frog with the word 'animation drawing artist', drawn by Johnnie. CLIP-S: 0.73, PP: 11.58
Iteration 4: Image shows the cartoon's creator, artist and writer David J. CLIP-S: 0.72, PP: 25.28
Iteration 5: Image showing the drawing of an animation that shows a cartoon character with his head drawn on paper. CLIP-S: 0.85, PP: 14.15
Iteration 6: Video shows the animation artist, drawing a cartoon character with his hand drawn drawings of what he says is the shape CLIP-S: 0.88, PP: 7.42
Iteration 7: Photo shows a drawing of what looks like an animated cartoon version. CLIP-S: 0.77, PP: 23.74
Iteration 8: Image shows the animation being drawn in pencil, using an animated model and a computer drawing. CLIP-S: 0.79, PP: 12.60
Iteration 9: Video shows how the cartoonist drew a man with an erection on paper and then digitally inserted his penis into the CLIP-S: 0.74, PP: 11.40
Iteration 10: Picture shows cartoon drawing of a man with the head and body covered by black paint, who is then turned into CLIP-S: 0.70, PP: 11.01
Iteration 11: Picture showing a cartoonist drawing an outline of the character's head. CLIP-S: 0.82, PP: 34.65
Iteration 12: Video of cartoon character drawing a man's penis in the shape and size he wanted, then putting it on his CLIP-S: 0.74, PP: 9.45
Iteration 13: Image showing animation of the shape drawing on paper, with a pencil drawn by an unknown artist. CLIP-S: 0.79, PP: 15.25
Iteration 14: Photo of cartoon drawing by artist, showing the character in a sketch. CLIP-S: 0.81, PP: 23.83
Iteration 15: Image shows the drawing of an animated character, which was made by artist and writer. CLIP-S: 0.81, PP: 11.19
Iteration 16: Photo of the cartoonist, who drew a drawing on paper that is still being used in animation. CLIP-S: 0.77, PP: 8.70
Iteration 17: Picture of a drawing by the cartoonist who created this video, showing how he made his character's head animation CLIP-S: 0.81, PP: 10.70
Iteration 18: Picture showing a drawing animation of the animated character, who is known as 'Rickman's son', in a CLIP-S: 0.82, PP: 6.81
Iteration 19: Image shows a cartoon character drawing the outline for an animation of what it would be like to have a penis, CLIP-S: 0.80, PP: 12.16
Iteration 20: Image showing a cartoon drawing of the character, which has been drawn by artist and writer. CLIP-S: 0.78, PP: 10.48

Iteration 1: Photo shows the aftermath of a car accident in which one person was killed. CLIP-S: 0.48, PP: 15.24
Iteration 2: Photo shows the image of what you can see on your character screen. CLIP-S: 0.68, PP: 21.51
Iteration 3: Image of a player's character animation, with some gameplay footage from the game. CLIP-S: 0.79, PP: 13.41
Iteration 4: Image shows the gameplay in progress, with some animations and sound effects added by modding tool. CLIP-S: 0.78, PP: 12.80
Iteration 5: Image showing the gameplay of a modded version in progress. CLIP-S: 0.80, PP: 35.07
Iteration 6: Video shows the player character in action, with a sword drawn at him. CLIP-S: 0.81, PP: 15.42
Iteration 7: Photo shows the game being shown to a spectator. CLIP-S: 0.68, PP: 59.88
Iteration 8: Image shows the game in beta gameplay screenshot taken by developer Blizzard Entertainment on a server hosted at www. CLIP-S: 0.84, PP: 7.14
Iteration 9: Video shows gameplay in action on a mobile phone. CLIP-S: 0.77, PP: 45.60
Iteration 10: Picture shows the game's gameplay, with player using mouse. CLIP-S: 0.81, PP: 33.11
Iteration 11: Picture showing the player's avatar walking around in a game of Warcraft. CLIP-S: 0.90, PP: 12.52
Iteration 12: Video of the gameplay, with a mouse cursor and arrow keys. CLIP-S: 0.76, PP: 50.29
Iteration 13: Image showing gameplay of the game in action, with graphics and animation from Blizzard. CLIP-S: 0.82, PP: 16.50
Iteration 14: Photo of gameplay from the game, with animations added by me. CLIP-S: 0.75, PP: 27.21
Iteration 15: Image shows a player character running around the world, with his avatar's name and level of gameplay. CLIP-S: 0.81, PP: 11.93
Iteration 16: Photo of the game in progress gameplay video, courtesy MMO developer. CLIP-S: 0.75, PP: 18.91
Iteration 17: Picture of MMO combat avatar, with a sword and shield equipped in front facing view. CLIP-S: 0.91, PP: 13.75
Iteration 18: Picture showing the player avatar in combat mode and a quest log. CLIP-S: 0.87, PP: 16.77
Iteration 19: Image shows the player character in a combat avatar. CLIP-S: 0.85, PP: 45.01
Iteration 20: Image showing the player character standing up and casting his spells in a quest window. CLIP-S: 0.86, PP: 15.63


Iteration 1: Photo shows the aftermath of a shooting that left one person dead. CLIP-S: 0.48, PP: 23.49
Iteration 2: Photo showshow much you're getting in return. CLIP-S: 0.42, PP: 45.27
Iteration 3: Image of the week is this pic, and it's a little too close for comfort. CLIP-S: 0.51, PP: 10.80
Iteration 4: Image shows a guy in the video playing with baby boy's dick while he is being fucked on camera. CLIP-S: 0.68, PP: 7.53
Iteration 5: Image showing how a boy who was bullied in elementary school would have been treated if he had eaten chicken BBQ sauce CLIP-S: 0.69, PP: 8.38
Iteration 6: Video shows BBQ bison on grill in pot roast recipe from a backyard. CLIP-S: 0.68, PP: 13.20
Iteration 7: Photo shows a chicken pot pie on the barbecue. CLIP-S: 0.61, PP: 33.93
Iteration 8: Image shows the grill at BBQboy's backyard in Texas that was used to fry a bunch of chicken and beef CLIP-S: 0.79, PP: 5.05
Iteration 9: Video shows a barbecue boy who is now in trouble. CLIP-S: 0.76, PP: 20.89
Iteration 10: Picture shows BBQ at the barbecue in front of my boys video. CLIP-S: 0.75, PP: 12.66
Iteration 11: Picture showing the BBQ at home on a youtube video posted by my brother. CLIP-S: 0.80, PP: 15.80
Iteration 12: Video of BBQ boy playing with a fire in the kitchen. CLIP-S: 0.82, PP: 19.24
Iteration 13: Image showing the barbecue boy in a video of him eating. CLIP-S: 0.81, PP: 19.73
Iteration 14: Photo of BBQ boy in video posted on YouTube. CLIP-S: 0.80, PP: 27.92
Iteration 15: Image shows BBQ video of the boys cooking a pot boy. CLIP-S: 0.92, PP: 18.62
Iteration 16: Photo of BBQ chicken in the video posted on Facebook by a friend. CLIP-S: 0.71, PP: 22.41
Iteration 17: Picture of BBQ video posted to facebook by a boy named 'B. CLIP-S: 0.83, PP: 11.46
Iteration 18: Picture showing barbecue chicken in a video posted by the boy's father on his website, which was later edited to CLIP-S: 0.79, PP: 6.70
Iteration 19: Image shows BBQ boy's brother cooking a chicken sandwich at his house on the episode. CLIP-S: 0.80, PP: 8.87
Iteration 20: Image showing BBQ chicken being cooked by a man who is also the owner of barbecue boys. CLIP-S: 0.79, PP: 7.71

Iteration 1: Photo shows the scene where a man in his late teens is talking about how he wants to be president. CLIP-S: 0.58, PP: 7.96
Iteration 2: Photo shows Hillary's campaign chairman says she is 'not a journalist', but her own spokesman said he was fired CLIP-S: 0.56, PP: 6.44
Iteration 3: Image of the man who shot down a reporter on live television, saying he would be willing to talk about the CLIP-S: 0.69, PP: 12.66
Iteration 4: Image shows the former prime minister, speaking to a conference in his office at an interview with journalist and broadcaster Alan CLIP-S: 0.65, PP: 13.51
Iteration 5: Image showing the reporter's interview on television in which he was interviewed by a correspondent from an unknown source, who CLIP-S: 0.76, PP: 7.92
Iteration 6: Video shows reporter saying he was shot at the scene of an interview with a man who claimed to have killed journalist CLIP-S: 0.64, PP: 11.56
Iteration 7: Photo shows a woman in her hotel room drinking alcohol and smoking cigarette. CLIP-S: 0.46, PP: 17.96
Iteration 8: Image shows the interview in question posted on a video website. CLIP-S: 0.69, PP: 28.14
Iteration 9: Video shows interviewer asking for interview with reporter in drunk driving incident, then drinking drink and smoking cocaine at the time CLIP-S: 0.65, PP: 6.13
Iteration 10: Picture shows a drunk driver who was drinking at the time of his interview with interviewer. CLIP-S: 0.75, PP: 11.91
Iteration 11: Picture showing interviewer drinking reporter's drink interview with the man who killed him in a drunken car. CLIP-S: 0.78, PP: 7.39
Iteration 12: Video of reporter interviewing a drunk driver in the backseat is shown on his website. CLIP-S: 0.63, PP: 12.06
Iteration 13: Image showing the reporter interviewing him at his home in a drunk driving incident. CLIP-S: 0.75, PP: 13.17
Iteration 14: Photo of a reporter drinking coffee in the background, drunk. CLIP-S: 0.67, PP: 30.11
Iteration 15: Image shows a reporter drinking from the interviewer, which was drunk at that point. CLIP-S: 0.77, PP: 12.93
Iteration 16: Photo of the interviewer drunk driving in a car interview with journalist, who was not drinking. CLIP-S: 0.76, PP: 8.77
Iteration 17: Picture of the interviewee drinking beer in front a journalist. CLIP-S: 0.70, PP: 31.89
Iteration 18: Picture showing a drinker at the bar in front of journalist. CLIP-S: 0.67, PP: 22.61
Iteration 19: Image shows reporter drinking beer from a glass of water. CLIP-S: 0.66, PP: 48.32
Iteration 20: Image showing interviewer, journalist and reporter interviewing a drunkard. CLIP-S: 0.70, PP: 21.72


Iteration 1: Photo shows the scene where a car is driven by someone with an automatic transmission. CLIP-S: 0.57, PP: 13.51
Iteration 2: Photo shows a car in front of the building, and video showing its driver being shot dead. CLIP-S: 0.61, PP: 11.01
Iteration 3: Image of the video posted by police car on camera, showing Mazda cars speeding down highway at high speed. CLIP-S: 0.70, PP: 11.45
Iteration 4: Image shows the vehicle in question is a Mazda, but it's not. CLIP-S: 0.70, PP: 15.03
Iteration 5: Image showing Mazda cars in a photo posted on its website. CLIP-S: 0.68, PP: 25.65
Iteration 6: Video shows the car being driven at a high rate of speed in front. CLIP-S: 0.70, PP: 18.47
Iteration 7: Photo shows Mazda's new car concept video, showing the cars in motion. CLIP-S: 0.75, PP: 12.50
Iteration 8: Image shows Toyota's logo on video screen during commercial. CLIP-S: 0.73, PP: 32.23
Iteration 9: Video shows the moment a car crashes into pedestrians during rush hour traffic on an episode of 'Star Trek,' which CLIP-S: 0.59, PP: 6.50
Iteration 10: Picture shows the Mazda commercial advertisement for'moto x car video game', featuring footage from a movie. CLIP-S: 0.76, PP: 6.80
Iteration 11: Picture showing the car in which a man was shot by an unknown person. CLIP-S: 0.60, PP: 14.20
Iteration 12: Video of a car crash in the movie industry is now available online for all to watch. CLIP-S: 0.64, PP: 9.71
Iteration 13: Image showing a video of the Mazda commercial, released in Japan. CLIP-S: 0.86, PP: 19.47
Iteration 14: Photo of Mazda's new video teaser trailer for the upcoming movie, which will be shown in a few minutes on CLIP-S: 0.79, PP: 8.18
Iteration 15: Image shows the moment police shot Mazda commercial footage of a car crash in which two teens were killed. CLIP-S: 0.72, PP: 7.87
Iteration 16: Photo of Mazda's new concept car, which is based on the upcoming film trailer. CLIP-S: 0.68, PP: 11.76
Iteration 17: Picture of Mazda commercial driver in the video, which was released by a local station on their website. CLIP-S: 0.74, PP: 7.18
Iteration 18: Picture showing Mazda's teaser video of the upcoming movie, 'Mazdaspeed', in a commercial. CLIP-S: 0.82, PP: 7.71
Iteration 19: Image shows a car being filmed by police on the road in which it was later found. CLIP-S: 0.62, PP: 10.24
Iteration 20: Image showing the video posted by police of a man driving on red lights in Mazda commercial, according to the video CLIP-S: 0.76, PP: 7.30

Figure 21: The evolution of captions for two images in an image set. (below and for multiple pages)

Iteration 1: the aftermath of a massive fire in downtown Toronto. CLIP-S: 0.32, PP: 22.61
Iteration 2: how to get started in the industry, with a simple and easy way of meeting people. CLIP-S: 0.51, PP: 10.13
Iteration 3: the night from a helicopter flying over sea ice on board ship at sunset in harbour. CLIP-S: 0.34, PP: 7.87
Iteration 4: the dinner table dining group of four people at restaurant, which was closed to public view. CLIP-S: 0.55, PP: 6.82
Iteration 5: a dinner at the waterfront restaurant, Pierpiers. CLIP-S: 0.64, PP: 21.19
Iteration 6: the sea pier in downtown Portland's harbour on Sunday evening, when a group of young people gather for CLIP-S: 0.65, PP: 10.51
Iteration 7: the dining group of sea bass tasting dinner at downtown. CLIP-S: 0.59, PP: 15.78
Iteration 8: the dining room of a restaurant in downtown, where people were discussing what to do after the pier was CLIP-S: 0.61, PP: 8.82
Iteration 9: boat piergoers enjoying lunch at sea restaurant, harbour views. CLIP-S: 0.58, PP: 10.66
Iteration 10: a dinner boat at sea with its pier. CLIP-S: 0.56, PP: 31.34
Iteration 11: the pier tasting dinner at a restaurant in San Francisco, where they were told by their guests to leave CLIP-S: 0.62, PP: 10.90
Iteration 12: dinner at the waterfront restaurant where a man was stabbed by his own dogs has been posted on social media CLIP-S: 0.53, PP: 7.76
Iteration 13: a tasting room at the pier of 'dinner club', where people were eating in a boat, CLIP-S: 0.62, PP: 8.54
Iteration 14: a pier in San Diego with dinner birds. CLIP-S: 0.64, PP: 35.63
Iteration 15: the tasting of pier food at an outdoor restaurant in downtown. CLIP-S: 0.62, PP: 16.12
Iteration 16: the dinner tasting at Pierpont, a waterfront pub on Long Island's harborside. CLIP-S: 0.69, PP: 7.13
Iteration 17: a dinner party at the pier, tasting wine from nearby restaurants. CLIP-S: 0.71, PP: 25.30
Iteration 18: tasting pierogi at a waterfront restaurant, which has been named the'most beautiful in town', with CLIP-S: 0.65, PP: 9.25
Iteration 19: a dinner club, where guests are served by the harbour pier. CLIP-S: 0.68, PP: 14.07
Iteration 20: the harbourside dining pier in a restaurant. CLIP-S: 0.62, PP: 30.55

Iteration 1: the scene where a man named 'Johnnie,' who was convicted of murdering his wife. CLIP-S: 0.62, PP: 6.70
Iteration 2: the room in which a former prison inmate was held for years. CLIP-S: 0.75, PP: 14.25
Iteration 3: a prison cell in the basement room, where prisoners are allowed to shower naked or sleep on beds with CLIP-S: 0.70, PP: 17.90
Iteration 4: a bedroom window that is believed to be used as prison cell number one. CLIP-S: 0.79, PP: 16.59
Iteration 5: bedroom in prison, from the book 'Prisoners and their families. CLIP-S: 0.81, PP: 15.08
Iteration 6: inmates at prison on a bed in solitary confinement. CLIP-S: 0.66, PP: 43.24
Iteration 7: bedroom of a prisoner in prison for killing and raping women. CLIP-S: 0.77, PP: 10.09
Iteration 8: prison bed in a bedroom where prisoners is sleeping. CLIP-S: 0.74, PP: 31.17
Iteration 9: prisoner sleeping with bed in prison yard, where it was discovered she had been locked for a month and CLIP-S: 0.64, PP: 20.49
Iteration 10: room where prisoners was held for a month before trial. CLIP-S: 0.73, PP: 40.63
Iteration 11: the bedroom of a prisoner who was held in solitary for nearly three years, and is now being read CLIP-S: 0.78, PP: 6.89
Iteration 12: the prison where they are being held in. CLIP-S: 0.62, PP: 31.29
Iteration 13: the bedroom of a jailed prisoner in prison. CLIP-S: 0.79, PP: 33.98
Iteration 14: bedroom prison in the early morning hours. CLIP-S: 0.66, PP: 76.22
Iteration 15: bedroom in which prisoners were imprisoned for writing books on the walls of their cells, according to a report CLIP-S: 0.82, PP: 8.03
Iteration 16: prison via Shutterstock, the author's home. CLIP-S: 0.70, PP: 45.50
Iteration 17: bedroom in jail cell, with books on the bedrooms door. CLIP-S: 0.81, PP: 18.90
Iteration 18: bedrooms bedroom of prison cell, bedside book. CLIP-S: 0.70, PP: 50.40
Iteration 19: room in which prisoners were sleeping, but the bedrooms of a bedroom. CLIP-S: 0.77, PP: 20.16
Iteration 20: the bedroom of imprisoned inmate in a prison cell. CLIP-S: 0.78, PP: 21.20

Iteration 1: the scene of a man's death in his home kitchen. CLIP-S: 0.46, PP: 10.67
Iteration 2: food and beverage items that have been sold in supermarkets. CLIP-S: 0.58, PP: 28.01
Iteration 3: food in the breakfast menu at an airport restaurant, which is a staple item for many people in the CLIP-S: 0.62, PP: 8.87
Iteration 4: fruit, fruits are shown on a supermarket menu. CLIP-S: 0.67, PP: 36.52
Iteration 5: food items being prepared in a supermarket fruit aisle. CLIP-S: 0.59, PP: 43.64
Iteration 6: breakfast fruits and fruit juice, which were served in the morning meal. CLIP-S: 0.58, PP: 18.89
Iteration 7: breakfast fruit and vegetable bar, with the 'healthy food', a large portion of which is served to CLIP-S: 0.62, PP: 20.78
Iteration 8: breakfast food shop in the supermarket chain. CLIP-S: 0.65, PP: 97.76
Iteration 9: fruit breakfast at the supermarket in a restaurant. CLIP-S: 0.69, PP: 28.49
Iteration 10: fruit at a cafe in the northern town, pictured on Tuesday. CLIP-S: 0.62, PP: 12.22
Iteration 11: fruit and vegetable market breakfast in front of a bakery, lunch buffet or diner serving pancakes. CLIP-S: 0.68, PP: 12.55
Iteration 12: Breakfast at the restaurant, which is a fruit salad. CLIP-S: 0.60, PP: 42.54
Iteration 13: fruits and veggies being prepared in a restaurant breakfast buffet at lunch. CLIP-S: 0.65, PP: 13.16
Iteration 14: fruits and vegetables breakfast in a diner. CLIP-S: 0.68, PP: 126.99
Iteration 15: breakfast fruits and vegetables in a shopping centre. CLIP-S: 0.64, PP: 70.03
Iteration 16: the breakfast menu at a grocery store, courtesy food. CLIP-S: 0.64, PP: 21.91
Iteration 17: breakfast fruits, fruit juices from a bakery in the morning. CLIP-S: 0.64, PP: 22.78
Iteration 18: breakfast fruit in a basket, with the price of produce on each side. CLIP-S: 0.64, PP: 42.42
Iteration 19: the fruit and vegetables in a supermarket at breakfast. CLIP-S: 0.68, PP: 38.12
Iteration 20: breakfast at a supermarket with fruits and veg, which is served in the same way. CLIP-S: 0.68, PP: 28.63


Iteration 1: the moment an officer fired his gun at a car carrying two men who had been arrested. CLIP-S: 0.39, PP: 10.88
Iteration 2: the world is getting a little bit more crowded. CLIP-S: 0.45, PP: 27.58
Iteration 3: the year is a little bit different than what's on your car. CLIP-S: 0.49, PP: 14.02
Iteration 4: the bus stop at an abandoned railway station, with its wooden sign advertising a 'free camping spot', CLIP-S: 0.50, PP: 7.63
Iteration 5: a man carrying the bus on his back, which was spotted in front of several homes along lake shore CLIP-S: 0.42, PP: 10.25
Iteration 6: the pond in lake trout ponds, which were closed for hiking and camping. CLIP-S: 0.47, PP: 10.12
Iteration 7: bus driver enjoying scenic lake in a restaurant picnic area. CLIP-S: 0.60, PP: 13.38
Iteration 8: bus advert promoting the trail hikers' campsite in scenic area near Lake. CLIP-S: 0.53, PP: 8.68
Iteration 9: a couple hiking together in the park. CLIP-S: 0.57, PP: 48.67
Iteration 10: the bus that hikers and drivers have to use for commercial advertising in a scenic lake resort. CLIP-S: 0.58, PP: 6.65
Iteration 11: a bus advert for the restaurant in downtown lakefront village, which has been seen hiking hikers and cyclists CLIP-S: 0.55, PP: 6.84
Iteration 12: hikers enjoying hiking in the woods, which are popularly known as buses and busparks. CLIP-S: 0.60, PP: 11.06
Iteration 13: bus buses parked at a pond in the woods. CLIP-S: 0.46, PP: 30.70
Iteration 14: bus hikers in a pond outside downtown restaurants, hiking buses. CLIP-S: 0.51, PP: 21.10
Iteration 15: bus advert hiking hikers in the park near a restaurant. CLIP-S: 0.54, PP: 35.13
Iteration 16: a bus advert for the hike in front pond by photographer, via Facebook. CLIP-S: 0.58, PP: 9.20
Iteration 17: hikers in bus on the riverfront, with a couple kissing. CLIP-S: 0.57, PP: 19.81
Iteration 18: hikers in a pond with their bus. CLIP-S: 0.59, PP: 78.44
Iteration 19: bus advert in the woods near a pond, hikers enjoying romantic picnic. CLIP-S: 0.64, PP: 9.83
Iteration 20: the hikersbus and its trailer park in rural France, with signs advertising hiking. CLIP-S: 0.50, PP: 8.37

Figure 22: The evolution of captions for three images in an image set. (below and for multiple pages)



Iteration 1: the aftermath of a plane crash in southern France, where one passenger was injured. CLIP-S: 0.45, PP: 9.29
Iteration 2: the front view of aircraft flying in a formation over runway at airport. CLIP-S: 0.44, PP: 25.07
Iteration 3: the day from a plane landing at sea in front yard. CLIP-S: 0.49, PP: 16.82
Iteration 4: plane landing pad in downtown area, which was owned and operated by a company called 'airport. CLIP-S: 0.48, PP: 10.78
Iteration 5: the bus station in downtown office building. CLIP-S: 0.43, PP: 37.39
Iteration 6: the plane landing at a restaurant in downtown hotel room, which is now being searched by police. CLIP-S: 0.49, PP: 7.14
Iteration 7: the runway landing of aircraft at sea port. CLIP-S: 0.50, PP: 31.44
Iteration 8: plane in the parking lot at their home apartment. CLIP-S: 0.41, PP: 16.26
Iteration 9: a van flying into the kitchen of home office cafe restaurant on runway in front room. CLIP-S: 0.63, PP: 7.86
Iteration 10: a runway of the house in question. CLIP-S: 0.53, PP: 131.14
Iteration 11: runway of a bus that crashed on the island. CLIP-S: 0.49, PP: 18.11
Iteration 12: a van flying over the runway at an apartment complex in central London. CLIP-S: 0.49, PP: 11.46
Iteration 13: the restaurant's interior, which is being investigated as part of a runway cafe development. CLIP-S: 0.56, PP: 9.22
Iteration 14: truck runway in front kitchen van pier. CLIP-S: 0.59, PP: 134.91
Iteration 15: the truck runway of a restaurant in front apartment complex, where passengers were seen boarding buses. CLIP-S: 0.53, PP: 11.91
Iteration 16: the apartment van on Facebook by airline flight crew via www. CLIP-S: 0.60, PP: 15.00
Iteration 17: Flight lounge in the kitchen floor vanishes. CLIP-S: 0.48, PP: 65.35
Iteration 18: aircraft at the airport in front of apartment building. CLIP-S: 0.47, PP: 23.46
Iteration 19: a truck flying over the restaurant's pier, which was destroyed by an explosion. CLIP-S: 0.52, PP: 11.61
Iteration 20: the runway at airport restaurant and catering truck rental company. CLIP-S: 0.60, PP: 17.48



Iteration 1: the aftermath of a shooting that left one person dead. CLIP-S: 0.44, PP: 19.40
Iteration 2: how you can get it from here, if i'm not wrong. CLIP-S: 0.49, PP: 16.35
Iteration 3: a girl with her face on the backside. CLIP-S: 0.52, PP: 37.02
Iteration 4: a young man wearing the uniform of youth football team playing against his parents' friends and relatives in a CLIP-S: 0.54, PP: 6.86
Iteration 5: a young man in the streets, wearing his shirt off at night. CLIP-S: 0.39, PP: 10.87
Iteration 6: the player's team of players trying to play soccer in front a wall with no helmet. CLIP-S: 0.52, PP: 8.25
Iteration 7: young boy in the hat of soccer player who is now playing. CLIP-S: 0.56, PP: 17.44
Iteration 8: a young man wearing hat and sunglasses. CLIP-S: 0.47, PP: 81.85
Iteration 9: the players celebrating after winning cup soccer hat trick against club's youth academy. CLIP-S: 0.50, PP: 10.04
Iteration 10: young boy in the game's final moments, with his head on top. CLIP-S: 0.61, PP: 12.10
Iteration 11: the football team eating a pizza and beer after winning cupcake. CLIP-S: 0.47, PP: 13.09
Iteration 12: the game's food and drinks menu, which was introduced in early's. CLIP-S: 0.47, PP: 7.44
Iteration 13: a soccer hat football club in the background. CLIP-S: 0.58, PP: 40.78
Iteration 14: food and football cake from the game. CLIP-S: 0.55, PP: 85.02
Iteration 15: a football game being played in the kitchen. CLIP-S: 0.61, PP: 31.95
Iteration 16: a soccer ball cake by the sports food restaurant. CLIP-S: 0.52, PP: 26.68
Iteration 17: cake from the soccer game between kids in school caps, with a hat. CLIP-S: 0.56, PP: 17.27
Iteration 18: the soccer ball from a hat trick in which players are given food cake. CLIP-S: 0.60, PP: 9.67
Iteration 19: the soccer ball hat trick cake in a restaurant. CLIP-S: 0.49, PP: 22.59
Iteration 20: a soccer ball hat cake, the top is decorated in red with chocolate and then covered by a soccer CLIP-S: 0.60, PP: 15.59

Iteration 1: the scene after a few minutes of eating. CLIP-S: 0.51, PP: 28.43
Iteration 2: a young girl with her hair cut and dressed up as food at the restaurant where they serve pizza. CLIP-S: 0.45, PP: 8.25
Iteration 3: pizza oven food recipe from the kitchen. CLIP-S: 0.52, PP: 101.58
Iteration 4: a bird feeders in front of an open oven. CLIP-S: 0.59, PP: 25.28
Iteration 5: the food truck kitchen birds in front of a restaurant. CLIP-S: 0.47, PP: 44.22
Iteration 6: birds flying over pizza oven in kitchen. CLIP-S: 0.62, PP: 99.52
Iteration 7: food birds in ovens at home, including a turkey sandwich. CLIP-S: 0.56, PP: 19.12
Iteration 8: a pizza bird perched in front of the fireplace. CLIP-S: 0.57, PP: 18.04
Iteration 9: birds eating pizza at a restaurant in southern Turkey on the way to their dinner table, but not everyone CLIP-S: 0.59, PP: 5.64
Iteration 10: birds eating pizza, fireplace decor and more in the kitchen. CLIP-S: 0.66, PP: 11.13
Iteration 11: fireplace ovens and birds in the garden. CLIP-S: 0.52, PP: 56.40
Iteration 12: the birds is from an oven in which they were served dinner, courtesy fireplace. CLIP-S: 0.63, PP: 11.48
Iteration 13: fireplace birds and pizza ovens in a restaurant, courtesy of the kitchener. CLIP-S: 0.59, PP: 12.48
Iteration 14: the day by chef at a restaurant in which they were serving. CLIP-S: 0.53, PP: 13.13
Iteration 15: a pizza oven in the fireplace of one restaurant. CLIP-S: 0.57, PP: 21.71
Iteration 16: fireplace birds in the kitchen at home by chef and author, food. CLIP-S: 0.55, PP: 10.77
Iteration 17: birds flock fireplace pizza oven in the kitchen. CLIP-S: 0.60, PP: 41.41
Iteration 18: the fireplace pizza oven in a flock of birds. CLIP-S: 0.62, PP: 30.76
Iteration 19: the fireplace in a restaurant at home. CLIP-S: 0.54, PP: 38.95
Iteration 20: the birds in a pizza oven, courtesy of chef. CLIP-S: 0.65, PP: 16.47

Iteration 1: the moment an officer shot dead a man in his sleep. CLIP-S: 0.44, PP: 15.21
Iteration 2: the baby in a pink diaper and toddler wearing an orange shirt, which is not his. CLIP-S: 0.49, PP: 14.14
Iteration 3: baby boy in baseball cap, batting gloves and helmet during his first game with team after birth of son CLIP-S: 0.49, PP: 8.31
Iteration 4: the birthday party for his daughter, a toddler who died of leukemia. CLIP-S: 0.56, PP: 18.22
Iteration 5: the pitcher pitching in a birthday celebration. CLIP-S: 0.64, PP: 42.36
Iteration 6: pitcher pitching birthday cake for baby who died at party in. CLIP-S: 0.61, PP: 9.86
Iteration 7: baby pitching birthday party for pitcher who died at age of three in the park on a cake. CLIP-S: 0.64, PP: 5.66
Iteration 8: pitcher birthday cake in the backyard of his son's house. CLIP-S: 0.52, PP: 35.76
Iteration 9: the birthday boy pitching for a child in baseball uniform, but it's just not quite right. CLIP-S: 0.57, PP: 6.62
Iteration 10: pitcher pitching for a cake in front of his house. CLIP-S: 0.62, PP: 39.20
Iteration 11: pitcher birthday cake being tossed in the stands at age of baby's toddler. CLIP-S: 0.56, PP: 17.90
Iteration 12: toddler pitching birthday cake baby's age in the hospital. CLIP-S: 0.60, PP: 18.86
Iteration 13: the pitcher's birthday cake in a baseball game. CLIP-S: 0.59, PP: 17.65
Iteration 14: pitcher baby batter's birthday cake, circa the time when he was born. CLIP-S: 0.61, PP: 10.39
Iteration 15: pitcher's birthday, with his arm thrown out to the left by toddler in diaper suit. CLIP-S: 0.70, PP: 15.03
Iteration 16: pitcher Birthday baby by the diaper boy's mom, circa 'throwing pitch', c. CLIP-S: 0.63, PP: 6.87
Iteration 17: pitcher's diaper from birth day photo. CLIP-S: 0.61, PP: 25.68
Iteration 18: toddler's diaper, with the pitcher and catcher on top pitching. CLIP-S: 0.55, PP: 9.12
Iteration 19: a birthday cake toddler in the background of this photo. CLIP-S: 0.57, PP: 20.04
Iteration 20: a pitcher pitching birthday cake to child, who throws the ball back in diaper. CLIP-S: 0.60, PP: 12.83

Figure 23: The evolution of captions for four images in an image set. (below and for multiple pages)

Iteration 1: the scene where a man in his early thirties was killed by police officers who were not aware CLIP-S: 0.46, PP: 5.72
Iteration 2: a very interesting and unique picture of the ancient temple at Kabbalah in Jerusalem on a hill CLIP-S: 0.36, PP: 6.50
Iteration 3: a man being attacked in front garden, police said on Wednesday night. CLIP-S: 0.45, PP: 15.39
Iteration 4: up a the top of this page is very important to me, and it's not just a matter CLIP-S: 0.48, PP: 6.18
Iteration 5: a restaurant, but it's the first thing that pops out of your mouth when you see this photo CLIP-S: 0.55, PP: 6.77
Iteration 6: pizza church in a restaurant on the street, where he was shot dead. CLIP-S: 0.51, PP: 17.52
Iteration 7: a man's face, and his wife in front of the restaurant with pizza on her head. CLIP-S: 0.47, PP: 6.49
Iteration 8: pizza delivery man being taken into church in the street, where he is shown feeding animals. CLIP-S: 0.51, PP: 12.68
Iteration 9: animal being fed elephant meat pizza and then eaten by church altar priest. CLIP-S: 0.61, PP: 15.90
Iteration 10: Pope Francis praying at a church chapel. CLIP-S: 0.40, PP: 83.52
Iteration 11: pizza delivery elephant's penis being removed from chapel altar. CLIP-S: 0.54, PP: 19.15
Iteration 12: pizza delivery elephant chapel being destroyed by the church's owner. CLIP-S: 0.55, PP: 21.35
Iteration 13: pizza delivery hall in the chapel of St. CLIP-S: 0.53, PP: 74.06
Iteration 14: the chapel elephants taken by photographer's son. CLIP-S: 0.51, PP: 80.58
Iteration 15: the chapel in its elephant enclosure, which was built by pizza chef and animal rights activist. CLIP-S: 0.60, PP: 6.94
Iteration 16: Pizza Hut's elephant mascot, which is being filmed by a team from the University of California. CLIP-S: 0.53, PP: 7.60
Iteration 17: pizza chapel elephant trainer in his yoga pose, which was posted to social media. CLIP-S: 0.60, PP: 13.00
Iteration 18: the elephant chapel, which has been used for pizza and yoga since it was built by a man. CLIP-S: 0.61, PP: 5.70
Iteration 19: Pizza Hut in the chapel of a church elephant sanctuary. CLIP-S: 0.58, PP: 28.28
Iteration 20: Pizza elephant chapel, which is a free yoga restaurant and church in the city of London. CLIP-S: 0.62, PP: 10.43

Iteration 1: the aftermath of a bomb explosion in central Baghdad on Wednesday, which killed more than half. CLIP-S: 0.37, PP: 13.90
Iteration 2: a small amount of light in the air, but it is not visible from ground bus. CLIP-S: 0.57, PP: 8.26
Iteration 3: bus on ground, which is not a part and the water level at sea. CLIP-S: 0.48, PP: 11.58
Iteration 4: the bus station tower, with a fountain on its side and an archway. CLIP-S: 0.49, PP: 16.89
Iteration 5: the fountain in front of a large tree. CLIP-S: 0.49, PP: 30.38
Iteration 6: the bus stop fountain at an outdoor water tower. CLIP-S: 0.55, PP: 20.72
Iteration 7: bus tower at the entrance of park in front parking garage on a busy street. CLIP-S: 0.47, PP: 11.34
Iteration 8: the water fountain in front of bus station. CLIP-S: 0.55, PP: 26.88
Iteration 9: busker fountain, which was created in the park. CLIP-S: 0.54, PP: 30.23
Iteration 10: a busker fountain in the middle of busy street with buses and tram lines. CLIP-S: 0.49, PP: 9.00
Iteration 11: bus driver and students marching in front of the fountain. CLIP-S: 0.50, PP: 30.50
Iteration 12: the fountain tower in downtown bus field. CLIP-S: 0.58, PP: 28.25
Iteration 13: bus tower, the fountain and field in a stadium. CLIP-S: 0.55, PP: 23.81
Iteration 14: the bus stop fountain in front tower at night. CLIP-S: 0.42, PP: 21.99
Iteration 15: the bus tower, a monument to buses and water. CLIP-S: 0.59, PP: 19.33
Iteration 16: Fountain by bus stop in the background. CLIP-S: 0.54, PP: 133.53
Iteration 17: the day fountain at Busk Field, where buses are parked. CLIP-S: 0.58, PP: 13.70
Iteration 18: bus driver tower, the fountain at left and a monument to victims of buses in front yard. CLIP-S: 0.53, PP: 13.03
Iteration 19: the bus tower in front of buses parked outside a school. CLIP-S: 0.48, PP: 14.17
Iteration 20: the bus tower fountain in front of a statue on campus, with buses flying over it. CLIP-S: 0.53, PP: 8.15

Iteration 1: the moment two men were shot dead by a group of people. CLIP-S: 0.39, PP: 16.97
Iteration 2: the animal's body being eaten by elephant at its enclosure. CLIP-S: 0.50, PP: 12.57
Iteration 3: a baby elephant in an aquarium at the zoo. CLIP-S: 0.35, PP: 23.34
Iteration 4: a baby being fed by elephants in northern Thailand, where the animals are considered sacred to some tribes. CLIP-S: 0.45, PP: 6.54
Iteration 5: a woman in the village of Chagang. CLIP-S: 0.41, PP: 20.75
Iteration 6: elephants eating a cake with an elephant in it. CLIP-S: 0.52, PP: 92.28
Iteration 7: elephant cake in the kitchen, butchee is still alive. CLIP-S: 0.57, PP: 13.45
Iteration 8: a cake in the middle of this year's elephants, but it is not yet eaten. CLIP-S: 0.56, PP: 7.15
Iteration 9: elephant bear being eaten by elephants in Thailand's'meat cake,' but the restaurant owner says it is CLIP-S: 0.48, PP: 7.34
Iteration 10: elephant cake being prepared by pizza delivery man in a restaurant near elephants' enclosure at an amusement park. CLIP-S: 0.51, PP: 9.33
Iteration 11: pizza cake being prepared by a bear, but the elephant is seen eating from its own mouth and not CLIP-S: 0.59, PP: 9.79
Iteration 12: elephants being fed cake pizza by a bear, which was filmed in the forest. CLIP-S: 0.57, PP: 14.54
Iteration 13: elephants eating cake at the zoo in southern Thailand. CLIP-S: 0.47, PP: 25.72
Iteration 14: a bear cake in the kitchen by my friend, which is not edible. CLIP-S: 0.51, PP: 9.98
Iteration 15: a cake bear in front elephants, which are not allowed to eat them but the elephant's owner says CLIP-S: 0.59, PP: 6.74
Iteration 16: elephants by the photographer's wife, who is wearing cake and pizza. CLIP-S: 0.56, PP: 11.71
Iteration 17: Pizza cake with elephants, elephant's head and cheese on top by the photographer. CLIP-S: 0.49, PP: 12.96
Iteration 18: pizza elephant in a bear suit and cake on the back of its head, taken by photographer 'p CLIP-S: 0.45, PP: 11.56
Iteration 19: a pizza elephant cake at an outdoor restaurant in Bangkok, Thailand. CLIP-S: 0.47, PP: 13.21
Iteration 20: elephants in Thailand, which is not a cake pizza. CLIP-S: 0.59, PP: 18.74

Iteration 1: the scene where a man was shot in front of his house, circa early 'nineties. CLIP-S: 0.41, PP: 7.15
Iteration 2: a painting by art artist and illustrator, the image of which is on display at a mural in CLIP-S: 0.55, PP: 13.79
Iteration 3: the day by photojournalist, author and artist. CLIP-S: 0.52, PP: 33.41
Iteration 4: the new artwork on horse wall mural at a local church. CLIP-S: 0.49, PP: 13.33
Iteration 5: the wall mural on display in a gallery, which is part of an exhibition called 'The art horse CLIP-S: 0.55, PP: 7.14
Iteration 6: a man painting horses on horseback. CLIP-S: 0.53, PP: 63.62
Iteration 7: the wall decor at a local hotel. CLIP-S: 0.43, PP: 31.94
Iteration 8: flowers and graffiti decorating a house in front of the fireplace. CLIP-S: 0.50, PP: 23.91
Iteration 9: horse paintings hanging on the wall in a hotel room window at home. CLIP-S: 0.44, PP: 13.44
Iteration 10: horses in a horse bed, fireplace and garden wall. CLIP-S: 0.54, PP: 28.01
Iteration 11: horse flowers on the wall in a fireplace. CLIP-S: 0.52, PP: 30.54
Iteration 12: horses flowers in the beach house fireplace. CLIP-S: 0.51, PP: 97.33
Iteration 13: flowers and fireplace walls in a beachfront hotel. CLIP-S: 0.50, PP: 44.60
Iteration 14: horse beach wall decor from the fireplace in an office building at a hotel. CLIP-S: 0.48, PP: 11.91
Iteration 15: a horse riding flowers in front of the fireplace. CLIP-S: 0.58, PP: 25.24
Iteration 16: flowers by the beachfront horse wall in downtown. CLIP-S: 0.49, PP: 51.60
Iteration 17: horse fireplace by the wall in a beach house. CLIP-S: 0.50, PP: 22.53
Iteration 18: fireplace at beach house on horseback. CLIP-S: 0.55, PP: 40.80
Iteration 19: graffiti fireplace in the beach room of a house on horseback horses. CLIP-S: 0.51, PP: 11.70
Iteration 20: flowers and horses at the beach house of former surf horse. CLIP-S: 0.54, PP: 37.22

## A.8 Web-Scale Models Limitations

**Before viewing the examples below readers are advised that they contain harsh language.**

A limitation of large-scale models is that they can sometimes generate sexist, or otherwise toxic language. Before viewing the examples, readers are advised that they contain harsh language (see Appendix Fig. 24). CLIP and GPT-2 may be at fault because they use web-based, uncurated data [10]. It is advisable to be aware of these weaknesses before deploying our method or any other method that uses these models.

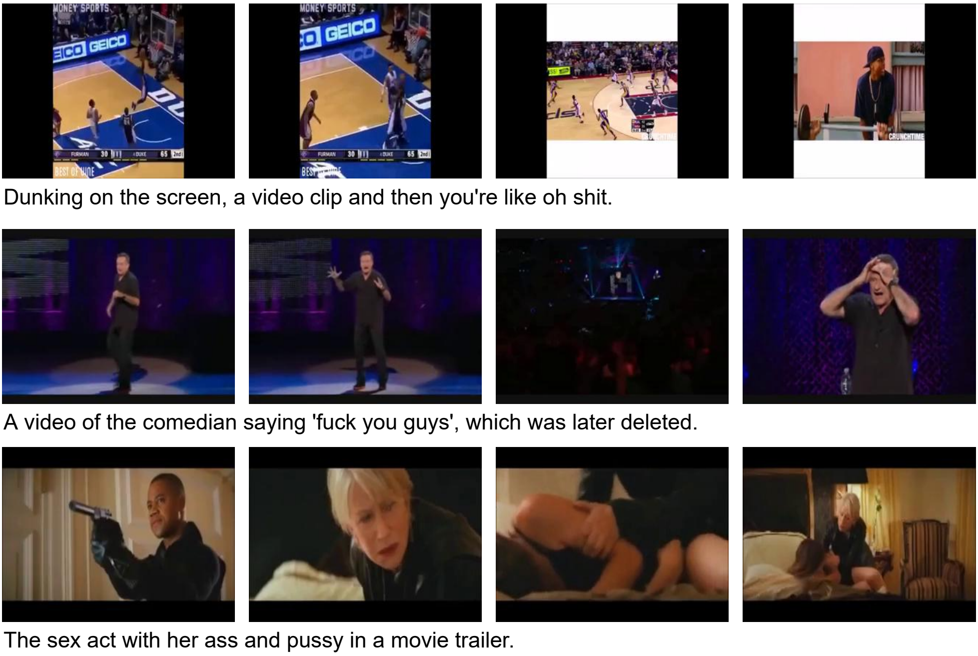



Figure 24: Example of harsh language being generated by our model. This illustrates a limitation of web-scale models.