# Supplemental Material
# Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data

Lisa Anne Hendricks[1] Subhashini Venugopalan[3] Marcus Rohrbach[1,2]
Raymond Mooney[3] Kate Saenko[4] Trevor Darrell[1,2]
[1] UC Berkeley, [2] ICSI, Berkeley, [3] UT Austin, [4] UMass Lowell

We present further empirical and qualitative results for both image and video description. For the image description task, we explore averaging weight vectors before transfer, illustrate errors made by the model when no unpaired text data is used during training and provide descriptions generated by DCC for a large variety of novel object categories in ImageNet. For the video description task, we include results when training the language model with an external text corpora and more DCC descriptions of novel objects in video.

## 1. Image Description

### 1.1. Transferring from Multiple Words

In the main paper, we transfer the most similar word in the paired image-sentence data to new objects in the multimodal unit. However, we could also average weights across multiple similar words before transfer. For direct transfer, averaging weights before transfer hurts performance substantially. In contrast, averaging weights before delta transfer does not significantly impact results (Table 1).

### 1.2. Transfer with No Language model

As mentioned in the paper, if unpaired text data is not used to train the language model, descriptions are poor because the caption model never learns good language features for new object categories. This is illustrated in Figure 1. Though a model which is trained on paired image-sentence data and unpaired image data can insert new words into a sentence, the generated sentences are not cohesive because the underlying language model has never seen the new object categories. For example, without training on unpaired text data, the model produces repetitive sentences like "A zebra with several zebra and zebra of zebra." or ungrammatical phrases like "a pizza bowl of food".

| | $\Delta$T (N=1) | $\Delta$T (N=5) | $\Delta$T (N=15) |
|---|---|---|---|
| F1 | 34.89 | 34.60 | 34.98 |
| BLEU-1 | 64.00 | 63.96 | 63.95 |
| METEOR | 20.86 | 20.88 | 20.88 |

Table 1: Image Description: Comparison of delta transfer method when averaging $N$ closest weight vectors before transfer. Averaging weight vectors before transfer has little impact on performance.

### 1.3. Qualitative Analysis of ImageNet Descriptions

**Comparison to sentences generated with no transfer** Figure 2 compares generated sentences for new object categories before and after transfer and also indicates which word in the paired image-sentence data is transferred to each new word. DCC does not simply substitute words seen in the image-sentence data with new object categories. Rather, subsequent words in the sentence are impacted by the use of a new vocabulary word. Consider the image of the candelabra in the top row of Figure 2. Without transfer, the model describes "A table with a vase of flowers in it.". However, after transfer, the model describes "A candelabra is sitting on a table in a room." Though the word vase is transferred to the word candelabra, candelabra is described differently. This is possibly because the language model learns to describe the words "vase" and "candelabra" in slightly different ways. For example, "flowers" are not frequently in candelabras. Furthermore, sentences generated before transfer do not necessarily include the known word which is transferred to the new category. Consider the image of the centrifuge in the second row of Figure 2. Though the word "refrigerator" is transferred to the word "centrifuge", the sentence generated before transfer does not include the word "refrigerator". However, the word "centrifuge" is accurately described after transfer.

**Successful Descriptions** We highlight sentences generated by DCC in Figure 3, 4, and 5. By placing different images of the same object side-by-side, we can compare

Without LM pretrain: A **zebra** with several **zebra** and **zebra** of **zebra.**
With LM pretrain: A **zebra** standing in a dirt lot with other animals.

Without LM pretrain: A close up of a **pizza** bowl of food.
With LM pretrain: A close up of a **pizza** with meat and cheese.

Figure 1: Image Description: Example captions generated by DCC direct transfer with and without pretraining the language model with unpaired text data. Without pretraining on text data, generated sentence fluency is poor. For example, the model will repeat the word "zebra" or insert a ungrammatical phrase like "pizza bowl of food".

how the same object is described in different contexts. For example, in Figure 3, a gecko is described as "A person holding a gecko in their hand" and "A gecko is standing on the branch of a tree" demonstrating that DCC is able to describe a single object in a variety of ways to reflect different visual contexts.

**Errors** We highlight common errors generated by DCC. Figure 6 shows examples in which the description does not mention the new object category, but is still highly relevant. Sometimes, DCC produces accurate descriptions without mentioning the new object category. Other times, DCC correctly describes other elements of the scene, but misclassifies the new object category. For example, in Figure 6 right, the model classifies the "alpaca" as a "sheep".

Figure 7 provides examples of descriptions which do include a new object category, but describe it in the incorrect context. DCC commonly hallucinates objects which are likely given the object context. For example, when describing an amphitheater, DCC produces the sentence "A group of people standing around a amphitheater" even though no people are present in the image. Such errors can be caused because the language model learns that people and amphitheaters occur together or because the lexical classifier mistakenly classifies "people" in the image. When describing new types of buildings (e.g. chapel, fortress, or watchtower), DCC will frequently include the word "clock" in the description. This is likely because MSCOCO vocabulary words like "tower" and "building" which are transferred to new words like "chapel" are frequently pictured with clocks. Sometimes the model includes objects in the description which are not contextually likely. For example, when describing a little girl in a tutu, the model also describes an umbrella. This is likely because the visual features for "umbrella" overpower the language model. Describing objects out of context is common for images which include a single object and a monochromatic background, as is commonly seen in ImageNet images. Because only one object strongly activates the lexical layer, the model will frequently hallucinate objects which are not present in the image (e.g., "A man is sitting on a bench with a chainsaw" when there is no man or bench).

| Model (Video, WebCorpus LM) | METEOR | F1 |
|---|---|---|
| Baseline (No Transfer) | 28.4 | 0.0 |
| + DT | 28.4 | 13.3 |
| + ILSVRC Videos (No Transfer) | 28.8 | 0.0 |
| + ILSVRC Videos + DT | 28.8 | 13.7 |

Table 2: Video Description: METEOR scores across the test dataset and average F1 scores for the four held-out categories (All values in %). The DCC models were trained on videos with 4 objects removed and the language model were trained on WebCorpus sentences.

Figure 8 provides examples of grammatical errors. As mentioned in the main paper, grammatical errors arise when a poor word is chosen for transfer. Examples of poor grammatical sentences when the word "dog" is transferred to "foxhunting" and the word "frisbee" is transferred to "trampoline" are shown in Figure 8. Another common grammatical mistake is for the model to list two objects in a row such as a "vole bear" or for the object to repeat a phrase such as "a unicycle on a unicycle". One possible explanation for such errors is that the pretrained language model does not learn good language features for these words. Consequently, after a new word is generated, the model generates a poor subsequent word.

Finally, Figure 9 shows images with irrelevant captions.

## 2. Video Description

### 2.1. Empirical Results

As was presented for image description in the main paper, we also explore the effects of training the language model used for video description with out-of-domain unpaired text data in Table 2. Comparing to Table 4 in the main paper, METEOR drops when training the language model with out-of-domain unpaired text (29.1 to 28.8 when including ImageNet videos during training and using transfer). The F1-score also drops when trained with ImageNet videos, but without training on ImageNet videos the F1-score actually increases from 6.0 to 13.3.

### 2.2. Qualitative Results

We present qualitative results on ImageNet videos. The lexical classifier is trained with ImageNet and MSVD videos and the language model is trained with in-domain text data. In addition to the objects held-out in the main paper ("zebra", "hamster", "broccoli", and "turtle") we also describe videos which include the objects "fox" and "whale". The caption model is never provided any paired video-sentence data which include these objects during training. In Figure 10, we show the top five captions pre-

*Vase → Candelabra*
No Transfer: A table with a vase of flowers in it.
DCC: A **candelabra** is sitting on a table in a room.

*Refrigerator → Centrifuge*
No Transfer: A white and white cat sitting on top of a white toilet.
DCC: A white **centrifuge** is sitting on a table.

*Bird → Albatross*
No Transfer: A black and white bird sitting on a grass covered field.
DCC: A black and white **albatross** is sitting on a grassy field.

*Bicycle→ Unicycle*
No Transfer: A group of people riding on a street.
DCC: A group of people riding on a **unicycle**.

*Boat → Shipwreck*
No Transfer: A group of people sitting on top of a beach.
DCC: A **shipwreck** is sitting on the beach.

*Fortress → Tower, Wall → Citadel*
No Transfer: A large stone building with a large clock on it.
DCC: A large stone **fortress** with a large stone **citadel**.

*Clock → Chime*
No Transfer: A large wooden bench with a large metal and a wooden table.
DCC: A large wooden bench with a large **chime**.

*Snow → Glacier*
No Transfer: A snow covered mountain in the snow.
DCC: A **glacier** with a mountain range of **glacier** and mountains.

*Hat → Cloak*
No Transfer: A woman in a red dress is standing in a red and white dress.
DCC: A woman in a red **cloak** standing in front of a red **cloak**.

*Flower → Chrysanthemum*
No Transfer: A large red and white vase with a pink flower.
DCC: A large red and white vase with a pink **chrysanthemum**.

*Vase → Mantlepiece*
No Transfer: A large room with a clock on a table.
DCC: A **mantlepiece** is shown in a room with a window.

*Computer → Supercomputer*
No Transfer: A man is standing in front of a large building.
DCC: A man is standing in front of a **supercomputer**.

*Knife → Chainsaw*
No Transfer: A man is standing in a field with a fence.
DCC: A man is standing in a field with a **chainsaw**.

*Clock → Doorbell*
No Transfer: A white and black cat sitting on a wooden bench.
DCC: A **doorbell** is sitting on a white wall.

Figure 2: Image Description: Comparison of descriptions generated by DCC for a variety of ImageNet objects with and without transfer. X → Y indicates that the known word X is transferred to the new word Y. DCC does not simply substitute new words in place of words present in image-text data. Further, the sentences generated by a model without transfer do not need to contain the word X for the sentences generated by a model after transfer to include the word Y.

dicted by DCC for videos which include "whale", "fox", and "hamster". For "whale" and "hamster", DCC predicts the correct object in the most probable caption. However, for "fox" the model predicts that "dog" is more probable than "fox", though "fox" is predicted in the second most probably caption.

Figure 11 compares descriptions generated by the model without transfer and after transfer. The model is correctly able to identify and generate sentences to describe "hamster", "lion", "turtle", "whale", and "zebra" after transfer. In comparison to the sentences produced by the model before transfer, the sentences generated after transfer describe the object and the context more accurately e.g. in the video containing a "whale", the sentence before transfer says "A woman is riding a jet ski" whereas after transfer DCC says "A whale is swimming" which appropriately describes the object (whale) and the context (swimming) correctly.

Figure 10 also includes an example where the model has difficulty choosing the correct object. The five highest probability results each include a different animal in the description which indicates that the model is unsure which object is present.

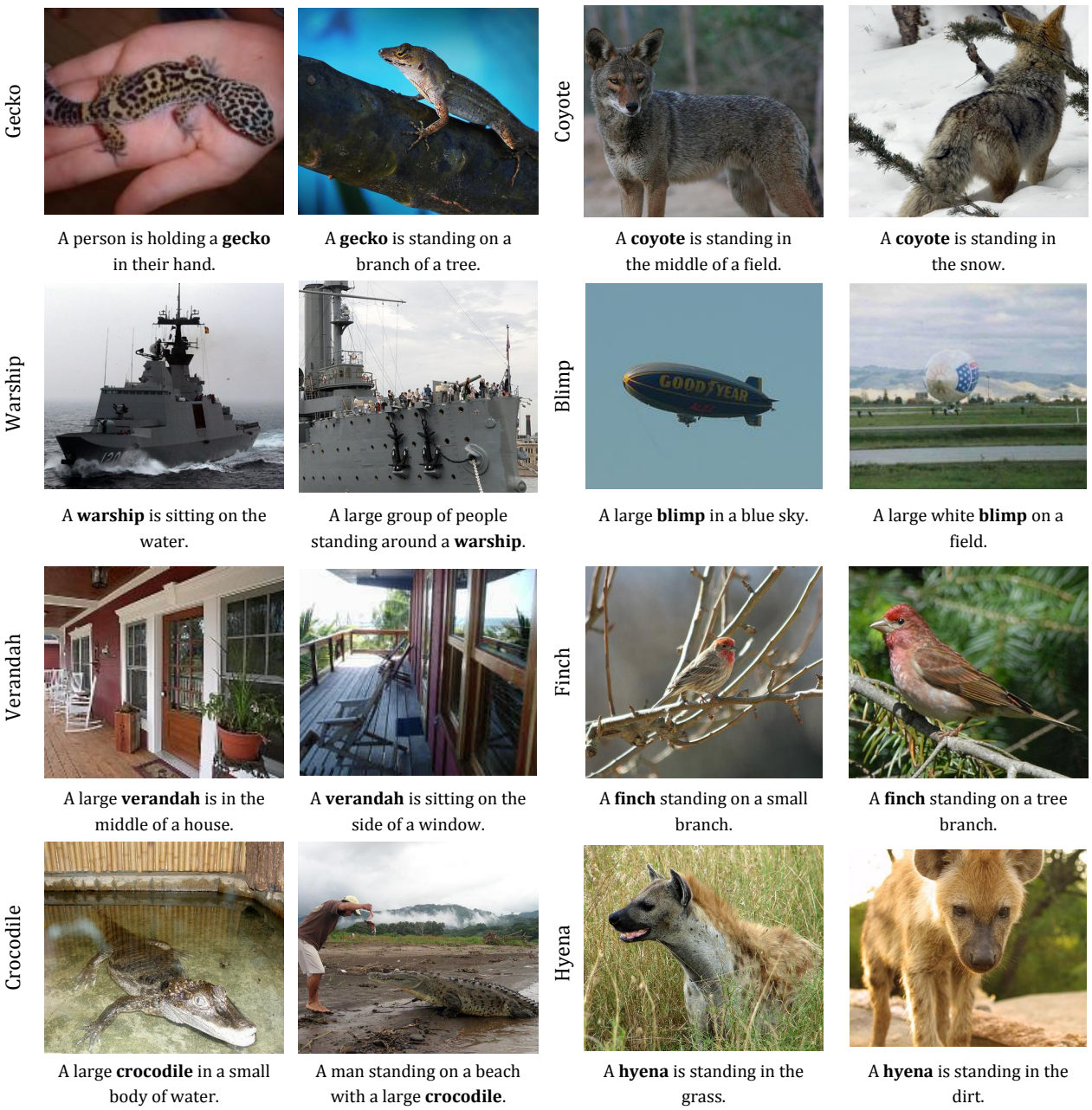


Figure 3: Image Description: Example DCC descriptions for eight different objects. DCC is able to describe objects in different contexts.

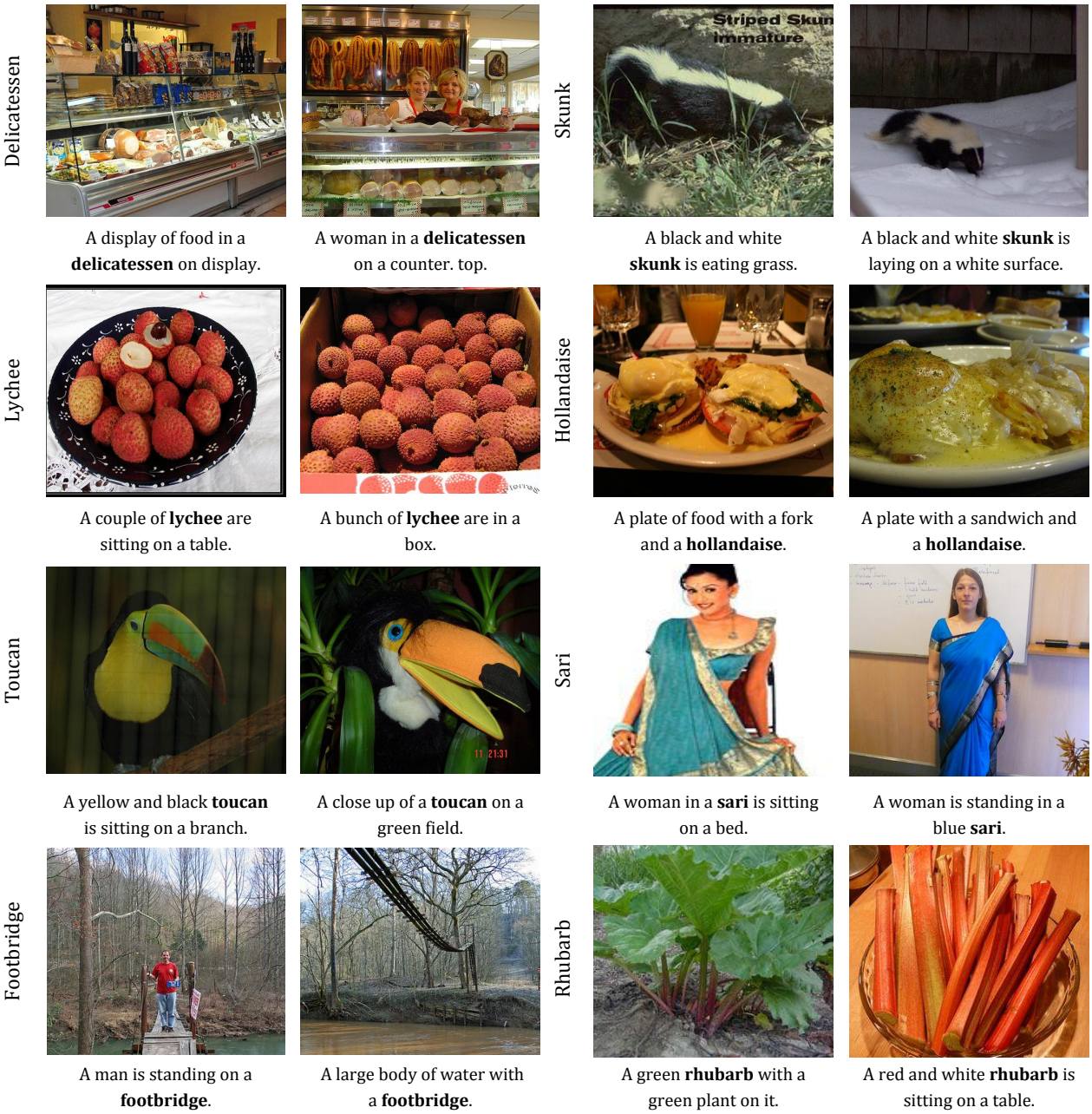


Figure 4: Image Description: Example DCC descriptions for eight different objects. DCC is able to describe objects in different contexts.

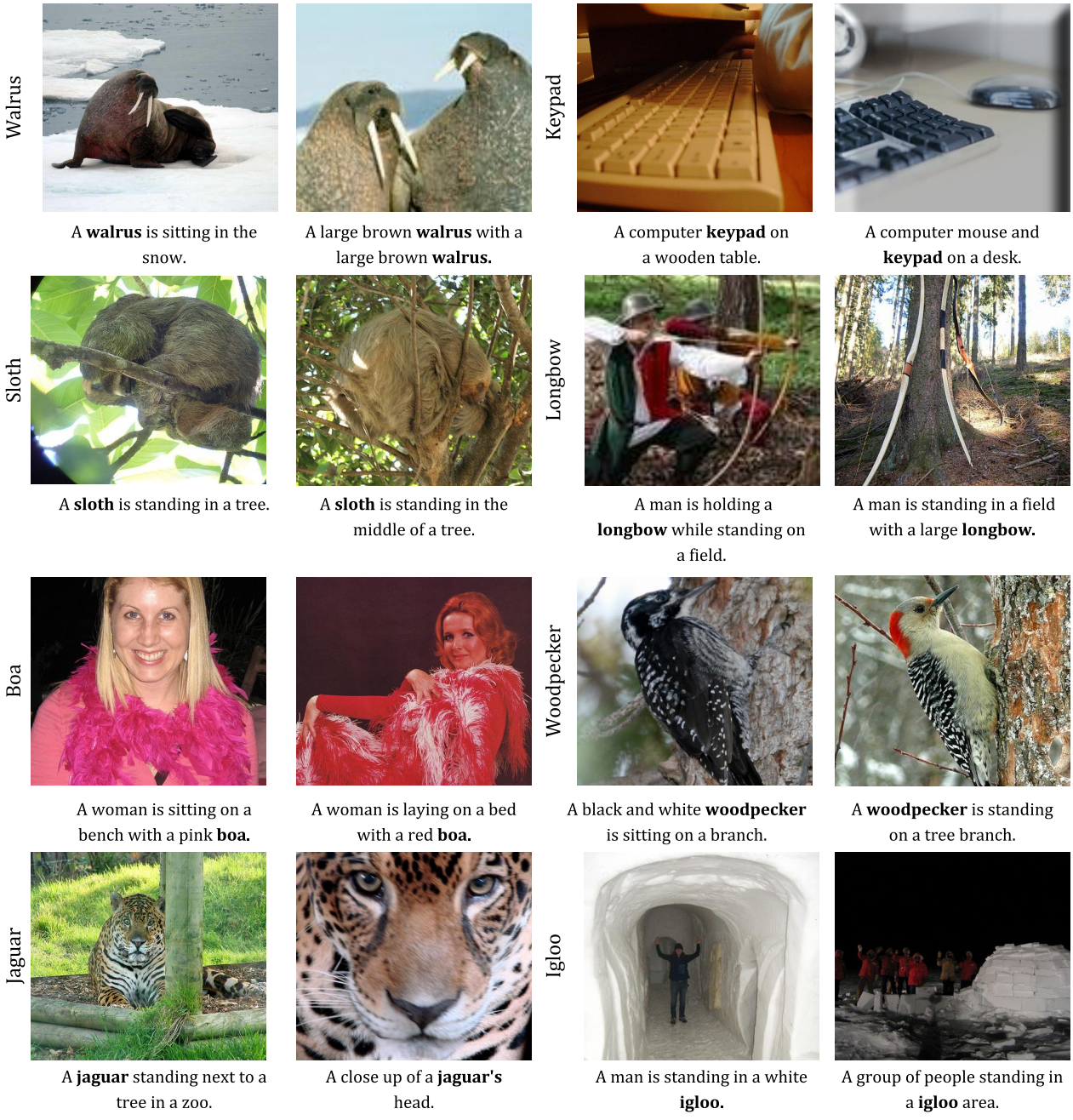


Figure 5: Image Description: Example DCC descriptions for eight different objects. DCC is able to describe objects in different contexts.


Figure 6: Image Description: Example descriptions generated by DCC which are relevant, but do not describe a new object category. Sometimes new objects are misclassified (e.g., "sheep" instead of "alpaca") but still describe the correct context. Other times, the new object category is not needed to accurately describe the image.


Figure 7: Image Description: Example descriptions generated by DCC in which a new object is described, but the context is described incorrectly. This is especially common in images, like the image of the chainsaw on the far right, which only include a single object on a monochromatic background.

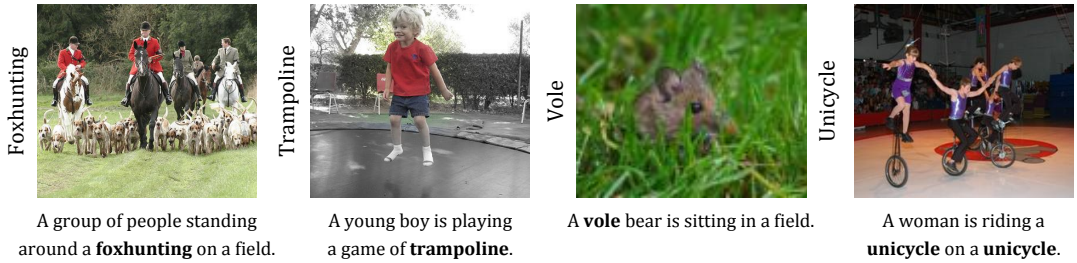


Figure 8: Image Description: Example descriptions generated by DCC with poor grammar. Poor grammar can be caused by a poor transfer word (the transfer words for "foxhunting" and "trampoline" are "dog" and "frisbee" respectively) or because the language model learns poor language features for the new object category.

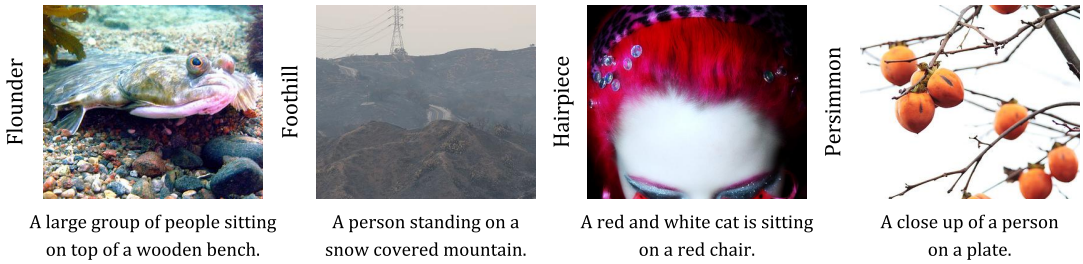


Figure 9: Image Description: Though most descriptions are relevant, some descriptions are incorrect.

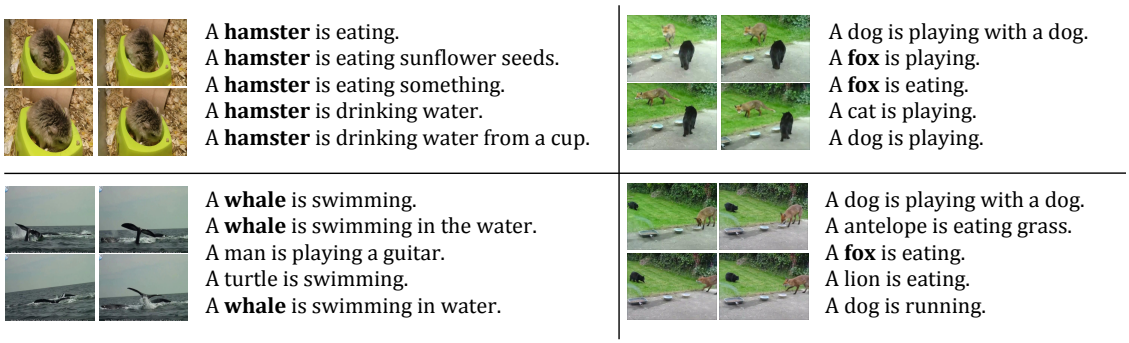


Figure 10: Video Description: Five most likely captions generated by DCC on novel objects unseen in paired training data. Captions are sorted by likelihood with the top caption corresponding to the most likely caption.

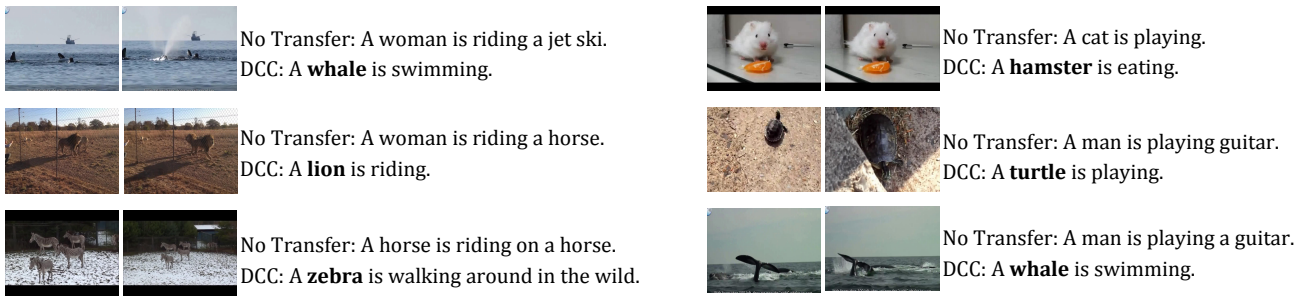


Figure 11: Video Description: Comparison of descriptions generated by DCC for some of the objects in the ImageNet video dataset with and without transfer. Note that our model does not simply substitute added words in the place of words present in paired image-sentence data.