Fill-in-the-Blank: A Challenging Video Understanding Evaluation Framework

Anonymous ACL submission

Two children throw _____ at each other as a video is captured in slow motion.
Correct answers: balloons, balloons filled with water, balloons of water, pink balloon, pink water balloon, things, water, water balloons, water-filled balloons

_____ sits at a drum set and practices playing the drums.
Correct answers: child, drummer, future drummer, girl, kid, little girl, little kid, musician, small child, young girl

A boy is trying to comb his hair while _____ dries it.
Correct answers: another person, friend, girl, his sister, his sister with hair dryer, person, young woman

Figure 1: Three examples from our dataset, each including three video frames, the caption, the blanked answers from the original caption together with the collected answers (all answers normalized, see Section 3.2).

Abstract

We propose fill-in-the-blanks as a video understanding evaluation framework. The task tests a model's understanding of a video by requiring the model to predict a masked noun phrase in the caption of the video, given the video and the surrounding text. To this end, we introduce a novel dataset consisting of 28,000 videos and fill-in-the-blank tests with multiple correct answers. The task and the dataset are challenging for the current state-of-the-art systems to solve. This task also does not share the weaknesses of the current state of the art language-informed video understanding tasks, namely: (1) video question answering using multiple-choice questions, where models perform relatively well because they exploit linguistic biases in the task formulation; and (2) video captioning, which relies on an open-ended evaluation framework that is often inaccurate because system answers may be perceived as incorrect if they differ in form from the ground truth.

1 Introduction

Despite current progress on multimodal (textual and visual) representations, *language-informed video understanding* is still a very challenging task for machine learning systems (Zhang et al., 2021; Li et al., 2021). This is due in large part to how the task is set up and how the corresponding datasets are built. Current video understanding datasets often have limited application value (e.g., multiple-choice questions (Lei et al., 2018; Tapaswi et al., 2016; Jang et al., 2017; Castro et al., 2020) do not reflect real-world tasks) or are based on subjective evaluation metrics (e.g., video captioning (Tran et al., 2016; Krishna et al., 2017; Zhou et al., 2018; Wang et al., 2019)) and are therefore hard to evaluate automatically, as the same caption can be expressed in different ways. In this paper, we address these limitations by introducing a new dataset that collects multiple perspectives on the same video, focusing on noun phrases as a proxy for different entities and their interactions in the video. Our data focuses on recall and tests the ability of models to capture a wide range of possible interpretations for a particular aspect of a video.

We construct a large fill-in-the-blanks dataset by systematically blanking captions from an existing video captioning dataset (Wang et al., 2019) and by providing additional correct answers for the blanks. VaTeX is a video captioning dataset that contains 40,000 10-second YouTube videos with 10 English captions per video. We build our video fill-in-the-blank dataset by blanking random noun phrases from one of the English captions for each video, from a subset of VaTeX with 28,000 videos. In an extensive analysis, we show that the blanked noun phrases are essential for understanding important visual aspects from the video. We

use crowdsourcing (Amazon Mechanical Turk) to gather additional possible correct answers for the blanks in the validation and test sets.

To address the fill-in-the-blanks task, we propose a Transformer-based (Vaswani et al., 2017) multimodal model. Our experiments show that our best multimodal model achieves a token-level F1 score of 71.4 while the F1 score of crowd workers is 82.5, indicating that this task is challenging for video and text understanding.

The contribution of this work is threefold: (1) We introduce a novel fill-in-the-blank task as a framework that addresses the drawbacks associated with previous approaches for video understanding. To facilitate the task, we offer a novel dataset of 28,000 videos and fill-in-the-blank captions with multiple correct answers. (2) We propose several unimodal baselines and two multimodal models for solving this task. (3) We provide a detailed analysis of the data to measure the diversity and complexity of the answers, and also compute an error analysis of the models' performance, to gain insights on what blanked captions and videos are hard for the models to solve.

2 Related Work

Language-informed video understanding is a complex task that has been extensively addressed in the multimodal (natural language and computer vision) machine learning research, by proposing diverse tasks and benchmarks.

Multiple-Choice Video Understanding. Multiple-choice benchmarks consist of distinguishing the only correct answer from a set of distractors, where the set of possible answers varies depending on the input. Video Question Answering (Video QA), a popular format, consists of answering questions based on the video content. Numerous multiple-choice Video Understanding benchmarks have been proposed such as TVQA (Lei et al., 2018), MovieQA (Tapaswi et al., 2016), TGIF-QA (Jang et al., 2017) (Repetition Action and State Transition tasks), LifeQA (Castro et al., 2020), PororoQA (Kim et al., 2017), MarioQA (Mun et al., 2017), VCQA (Zhu et al., 2017), VideoMCC (Tran et al., 2016), and ActivityNet QA (Yu et al., 2019). However, they provide choices and are thus easier to solve than writing free text. On top of this, the performance without the visual input is generally already high as models are able to exploit biases in the dataset (Agrawal et al., 2018) or they count on other modalities that overlap in functionality with the visual one.

Video Captioning. Video Captioning consists of generating a piece of text that describes a given video. There are multiple datasets for this task such as ActivityNet Captions (Krishna et al., 2017) (also features Dense-Captioning), YFCC100M (Thomee et al., 2016), (Alayrac et al., 2016), DiDeMo (Anne Hendricks et al., 2017), MSR-VTT (Xu et al., 2016), YouCook2 (Zhou et al., 2018), How2 (Sanabria et al., 2018), HowTo100M (Miech et al., 2019), VaTeX (Wang et al., 2019), TGIF (Li et al., 2016), MovieNet (Huang et al., 2020), LSMDC (Rohrbach et al., 2017), TGIF-QA (Li et al., 2016) (Frame QA task). Due to the diversity of captions provided, Video Captioning datasets do not achieve a high human performance and are thus hard to evaluate automatically with certainty (Aafaq et al., 2019).

Fill-in-the-Blank Video Understanding. VideoBERT (Sun et al., 2019b), CBT (Sun et al., 2019a), UniVL (Luo et al., 2020), ActBERT (Zhu and Yang, 2020), and HERO (Li et al., 2020) methods propose to mask random parts of the input from text and video pairs for training. However, they do this only for the purpose of system training and do not use the framework to test and evaluate video understanding. The only exception is MovieFIB (Maharaj et al., 2017) who proposes a video fill-in-the-blank, based on LSMDC (Rohrbach et al., 2017) for both training and evaluation. However, they blank a single word, which makes it easier to guess; they evaluate correctness with a single ground truth answer per caption; and they focus on the movies domain (we focus on YouTube videos).

Concurrent Work. The most similar work to ours is VidQAP (Sadhu et al., 2021) that presents an evaluation framework to fill in blanks with phrases based on semantic roles based on ActivityNet Captions (Krishna et al., 2017) and Charades (Sigurdsson et al., 2016); unlike them, we design our benchmark to have a high human accuracy (avoid ActivityNet Captions as it is contextualized, collect multiple correct answers, and show a high human performance). Our work is also close to (Yang et al., 2021) on evaluating using free-form QA, however they present a small vocabulary and no human accuracy that serves as an upper bound for

the task.

Our work is novel because it presents a hard task (a considerable gap between human and best model performance) that measures a form of video understanding while at the same time it has high human performance due to the large number of possible correct answers we collected (∼13 per caption) from multiple annotators (∼9 per caption).

3 Video Fill-in-the-Blank Dataset

We construct a large video understanding dataset that can evaluate a model's ability to interpret and use multimodal context by requiring the models to "fill in" (generate) a "blank" (a missing constituent) in this context. We construct this "fill-in-the-blank" video dataset in two steps: (1) data generation, where we compile a large set of video-caption pairs with selectively blanked words; and (2) data annotation, where crowd workers provide additional valid answers for these blanks.

Note that we could also develop a "fill-in-the-blanks" dataset by completing only the first step: the data generation. However, this will result in only one valid answer (the original blanked word or phrase), which can lead to unfair evaluations that are too strict because of dismissing alternative correct answers (e.g., "child" provided as an answer where the blanked word was "kid"). Other than manual annotations, we found no high-quality method to automatically obtain additional correct answers (e.g., "building" and "t-shirt" from Table 7 are too dissimilar but both are correct, "pink" and "yellow" from Fig. 1 are semantically close but only one is correct).

3.1 Data Generation

The dataset is constructed starting with the VaTeX (Wang et al., 2019) dataset. VaTeX is a multilingual video captioning dataset, consisting of over 41,250 video clips, each of which is taken from a unique public YouTube video, and lasts around 10 seconds. For each video clip, there are 10 English and 10 Chinese captions associated with it.

We produce blanked captions by blanking noun phrases in the English captions in VaTeX. We chose to mask only noun phrases for three main reasons. First, noun phrases often require visual information to identify or understand. They cover a large variety of information in visual content, as their head noun can describe people, objects, scenes, events, and more. A model often needs to identify the related objects in the videos, as well as the objects' properties (e.g., color, number, or size) to fill the blank correctly.

Second, nouns are usually essential to the understanding of *visual* content and represent reliable predictors for a system's ability to understand a video. Other phrases, such as verbs or adjectives, can more easily be guessed only from the text while ignoring the visual information. To illustrate, consider the example "A woman _____ in the pool," where a model can easily predict that the blank should be "swims" only from the textual content (which would not be the case for "A woman swims in _____"), where the blank could be completed by sea, pool, lake, water, and so on.

Third, in preliminary experiments, we found that nouns lead to more robust annotations, as compared to e.g., adjectives, which can have low inter-annotator agreement due to their subjectivity. As an example, consider the phrase "A _____ hill stands behind the house." where the blank could be filled with a color property, a size property, and so on.

For each video, we choose the first English caption that contains at least one noun phrase as detected by spaCy[1] (Honnibal et al., 2020), and randomly blank one of these noun phrases to generate an instance. Accordingly, we generate our training, validation, and test data starting with the VaTeX v1.1 training set, a random subset of size 1,000 from the validation set, and a random subset of size 1,000 from the test set, respectively.

3.2 Data Annotation

We performed a crowdsourced annotation procedure to collect additional correct answers for each blank in the validation and test sets. As highlighted earlier, the main reason for collecting such additional annotations is to account for the natural diversity of language, and have multiple alternative answers for each blank.

We use Amazon Mechanical Turk (AMT) for the annotation. Figure 2 shows the annotation interface and a highlight of the data collection instructions (additional guidelines were provided, not shown here for space reasons). For each blanked caption, workers were presented with a video clip along with the corresponding masked caption. They were then asked to fill in the blank with a noun phrase.[2]

[1] We used the model `en_core_web_trf` from spaCy v3. An error analysis identified only three tagging errors in a sample of 247 sentences.

[2] We blanked multi-word spans for the task, rather than

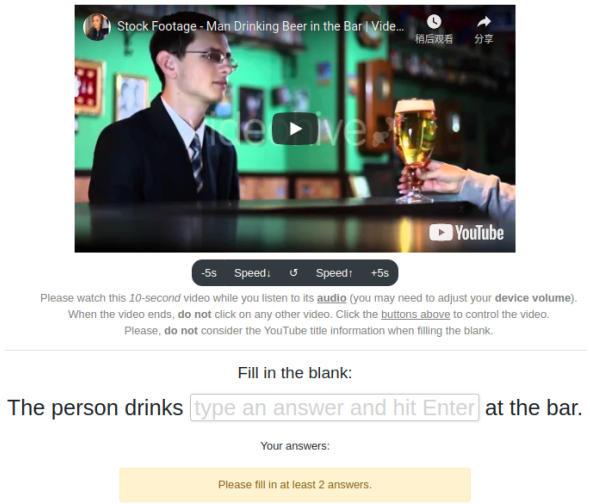


Figure 2: Annotation interface.

We also request annotators to provide answers in a confidence-descending order (the first answer should be the most natural one to the annotator).

We presented five videos in each Human Intelligence Task (HIT). Nine workers annotated each of them with at least two answers for each blank. We paid a bonus for each extra answer to each blanked caption, from the second one to the fifth one, to encourage them to provide more answers. We calculated a twelve-dollar hourly rate for a worker that provides at least five answers. We estimated the time to annotate one video to be 30 seconds. Consequently, the HIT pay rate was $0.2, which can result in a total of $0.5 with the added bonus. Additionally, we offered another type of bonus of $0.2 to the worker with the largest number of correct answers for every HIT, to encourage them to provide more than five answers.

We required workers to be in Canada or the United States,[3] and to have completed at least 1,000 HITs on AMT with at least 92% approval rate. The interface also checked that for a given worker and caption the answers were different. For this, we first normalized the answers by lower-casing, stripping punctuation and extra spaces, and removing the determiners "the", "a", and "an."

single-word noun phrases, because blanking a single noun at a time led us to less annotator agreement in preliminary experiments, likely due to less potential for overlap. E.g. annotator 1 might write "young boy" and annotator 2 might write "young child", which would have at least some overlap as compared to "boy" and "child" (no overlap).

[3]We restricted the task to these countries because it is a good proxy for proficient English speakers and because our task received lower-quality responses when we did not restrict the location.

Statistic	True labels	Answers
Unique labels per caption	∼	13.0 ± 4.14
Unique labels per caption per annotator	∼	2.63 ± 0.49
Characters per token	5.09 ± 1.89	5.27 ± 2.00
Tokens	1.47 ± 0.68	1.36 ± 0.68
Noun phrases	100%	95%
Visual word use (color, number, or size)	8.21%	3.31%

Table 1: Summary statistics for true labels and annotations; token counts computed after text normalization.

During the annotation, we manually reviewed a sample to identify cases of incorrectly tagged noun phrases (e.g., "inside" marked as a noun when it should be a preposition) and factually incorrect noun phrases (e.g., referring to bags as "eggs" without any evidence about the contents of the bags); we disqualified workers who consistently provided incorrect annotations. After collecting annotations, we filtered for noun phrases using the same method as before, based on whether the text is parsed as a noun phrase (including bare nouns, e.g. "man is walking"), a wh-phrases ("who is speaking"), a simple gerund ("eating is a good way to stay healthy"), or infinitive ("to eat is wonderful").

We compute summary statistics on the annotated data, to determine the degree of similarity with the label data. The statistics are shown in Table 1. We find that in general, annotators tend to provide ∼3 unique labels for the provided data. Compared to the true labels, annotators tend to use about the same number of tokens. Annotators also use visual words at a much lower rate than the true labels, possibly because the task encouraged the annotators to generate as many distinct nouns as possible without regard to descriptive information.

3.3 Annotation Analysis

To further validate the utility of the annotations collected in this study, we provide an extensive analysis of the annotations and ground truth labels.

We compute the most-frequent answers provided by the annotators and most-frequent ground-truth labels and find as expected that noun phrases related to "person" are the most frequent: the word "man" appears in 5.7% of total ground truth labels and 1.2% of total annotations (see Figure 5 in the Appendix). Note that our annotations have a long tail distribution, as the most-frequent noun phrase appears in only 1.2% of total annotations. In addition, we find that answers related to "person",

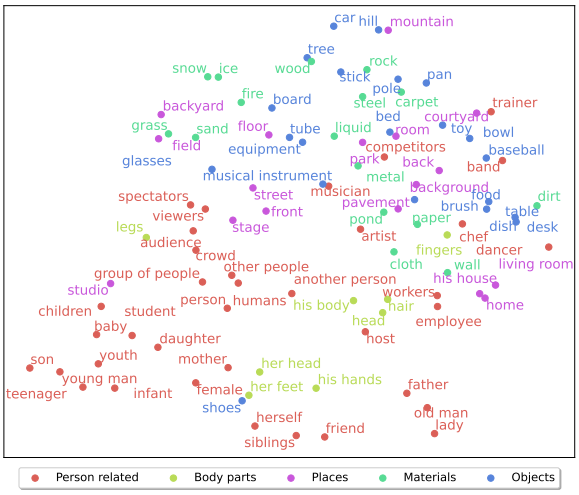


Figure 3: The t-SNE representation of the clustering of most frequent fill-in-blanks annotations. Each color represents a different cluster.

such as “another person" are not trivial, but on the contrary: e.g., in the third example from Fig. 1, a model has to reason about the actions of both persons and distinguish between them. The other two examples from Fig. 1 are also representative of how a model needs to understand both the video and the text in order to complete the blanks.

In Fig. 3 we show what kind of answers are depicted in the videos: we collect the top 100 most frequent annotations, represent them using the pre-trained model Sentence-BERT (Reimers and Gurevych, 2019) and then we cluster and plot them using t-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008). This analysis shows the diversity and complexity of answers that a model needs to fill in, demonstrating a strong video understanding. As expected, the cluster *Person-related* has the most answers, followed by the clusters: *Objects* (e.g., shoes, glasses), *Places* (e.g., mountain, street), *Materials* (e.g., metal, wood), and *Body parts* (e.g., fingers, head). Note also that the *Person-related* cluster, among more typical answers such as “male" and “female”, also contains complex and diverse answers such as “dancer”, “workers”, “musician” or “audience”.

3.4 Human Agreement

To establish a reference for machine models, we compute the agreement among annotators using the evaluation metrics described in Section 5.1, which we also use for model evaluation (Section 5.2).

Specifically, we apply a leave-one-out strategy to construct the “test set” and the “ground truth set.” We compare the first answer provided by each crowd worker (which is their most natural/confident answer) against the complete set of answers provided by the other crowd workers, using maximum F1 score (token overlap) and maximum exact match (EM) as agreement metrics, as described in Section 5.1.

Statistic	%
F1 first answers (per caption)	82.6 (± 15.7)
Exact Match first answers (per caption)	75.3 (± 19.7)
F1 first answers (per answer)	70.0 (± 11.9)
Exact Match first answers (per answer)	58.1 (± 16.3)

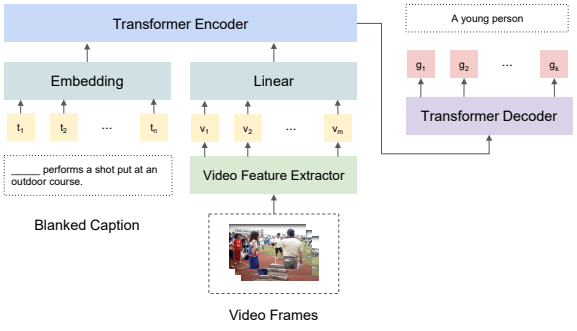
Table 2: Agreement statistics for answers (leave-one-worker-out-comparison; std. dev. in parentheses).

Table 2 shows the inter-annotator agreement. We show the mean values of the agreement metrics per-caption and per-answer (recall there are multiple answers per caption, so in the former case we first average among the answers within the caption, then across the captions). The higher rates of agreement at the caption level, compared to the answer level, indicate a high amount of answer diversity among the workers.

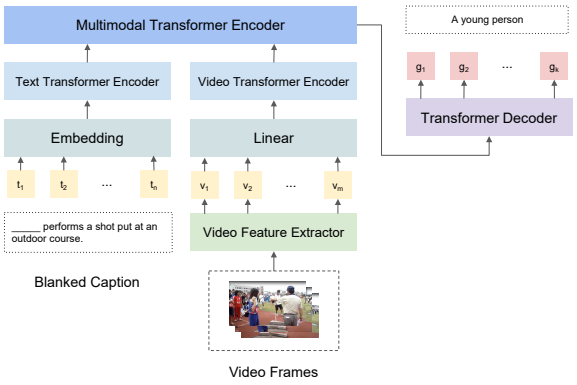
To validate the quality of the crowdsourced annotations, we also compare them against human annotations collected from two trusted annotators (both Computer Science researchers). We sample 200 captions from the validation set and ask these two annotators to perform the same labeling task that the MTurk workers performed, then compare their agreement with the crowdsourced data. The annotators obtain a per-caption average of 90.2% F1 score and 49.0% exact match accuracy, comparable with the agreement scores obtained by the workers.

4 Multimodal Method for Video Fill-in-the-Blanks

We propose an encoder-decoder multimodal method to perform the task of video fill-in-the-blanks. We first encode the text and visual modalities together to get a semantic representation of the blanked caption and video. The decoder uses the semantic representation to generate text corresponding only to the answer to the blank. To correctly generate an answer, a model needs to learn which parts of videos relate to the missing parts of the caption. To accomplish this, we use the original transformer architecture (Vaswani et al.,



(a)



(b)

Figure 4: (a) early-fusion multimodal model for video fill-in-the-blanks (b) late-fusion multimodal model for video fill-in-the-blanks

2017), whose self-attention mechanism is particularly effective for encoding relations within an input sequence and have been shown to perform well in many language understanding tasks.

We consider two types of encoders, namely the early-fusion encoder and the late-fusion (two-stream) encoder. The structure of our multimodal model with an early-fusion encoder is shown in Fig. 4a. The input to the model consists of the tokenized blanked caption text $t_1, \ldots, t_n$, as well as a representation of the video consisting of multiple video sequence features $v_1, \ldots, v_m$ from a video feature extractor. The blanked captions are embedded by an embedding layer. The video features are projected into the encoder by a linear layer. We use a special token to represent the masked phrase and another one to separate the input text and video sequences. We add positional embeddings to each input token or video feature to represent the sequence order, and another embedding to indicate whether it belongs to the text or video sequence similarly to BERT (Devlin et al., 2019).

The late-fusion model is shown in Fig. 4b. We use a late-fusion that first encodes the language and video separately and then jointly. The modalities may benefit from learning independently about their own context before using them together.

4.1 Implementation Details

For the video encoder, we use the existing I3D (Carreira and Zisserman, 2017) features (size 1024 every 8 consecutive frames) provided by the VaTeX dataset (Wang et al., 2019), in which videos were sampled at 25 fps. We initialize our multimodal model using T5 (Raffel et al., 2020), given its ability to fill in variable-length blanks. T5 is an encoder-decoder Transformer (Vaswani et al., 2017) model that is a good starting point as it provides state-of-the-art performance on text-only tasks and it was pretrained to fill arbitrary-length text spans that were previously masked. Building upon T5 allows our model to not only leverage the pre-trained large-scale language models that already have strong language abilities but also to fuse it with visual inputs. We initialize the early-fusion model with pretrained `T5-base` weights. For the late-fusion model, we use `T5-base` for the text encoder and for the decoder. We use two one-layer transformers to encode videos and fuse text and video features, and the weights of these two transformers are randomly initialized. Following T5 model implementation, the special token `<extra_id_0>` is used to represent the blanked phrase, and `<\s>` is used to separate the text and video sequences. The generated output follows T5 output format: the special token `<extra_id_0>` followed by the predicted text for the blanked phrase. See Appendix B.1 in the Appendix for more details.

4.2 Baselines

We compare our model to the following baselines.

Most Frequent Answer. The baseline makes use of the most frequent answer in the training set ("a man") as the answer to all the blanked captions during evaluation.

Text-based Transformer. Previous visual question-answering datasets found that a text-only model can nearly match the performance of the multimodal system (Antol et al., 2015). We analyze the degree to which language alone can contribute to our video understanding framework by conducting experiments based on text-only models. We use the off-the-shelf T5-base transformer model (Raffel et al., 2020) as our baseline

model. We use both a zero-shot model (not trained on our data) and a fine-tuned model. For the latter, we use the `base` model v1.1 later released by Google because it performed better in our experiments on the validation set. The decoding hyperparameters are the same as in the multimodal models, except the beam size is 8 for both the zero-shot one and 2 for the fine-tuned variant as we obtained the best validation results for each one using these beam sizes.

Single video feature. We consider using a single I3D feature per video to see how well can the model do with a small portion of the video. Based on a study of 50 randomly sampled videos, the blanked entity in the caption appeared 95% of the time at the third second of the video (see Fig. 11 in the Appendix). For this method, we pick the I3D feature which corresponds roughly to it and apply it to the proposed multimodal methods instead of using all the video features. Note I3D takes a window of 16 frames as input, which in our case corresponds to 640 milliseconds, centered at the mentioned moment within the video. This can be seen as a small generalization to the Image Understanding task, that considers a single image (frame).

5 Experiments and Results

We perform experiments and evaluations using the dataset described in Section 3.

5.1 Evaluation Metrics

We use exact match accuracy and ROUGE-1 F1 score (token-level) (Lin, 2004) to evaluate the output of the generation models and to evaluate human agreement (Section 3.4). For the exact match, we count a generated text string as correct if it has at least one string-level match among the provided annotations. For the token-level F1, we compute the token overlap (true positives) between the generated text string and each annotation, normalized by the sum of the true positives and average of the false negatives/positives, then compute the maximum across all annotations. For all evaluations, we computed the metrics based on the normalized text (i.e., without articles).

5.2 Results

We evaluate our multimodal model's visual understanding ability by comparing its performance with the text-only baseline and human performance. The results from the fill-in-the-blank task are shown in Table 3. The text-only model's accuracy and F1 score is low, indicating that the language bias is controlled in our dataset. The multimodal model outperforms the text-only baselines in both exact match accuracy and F1 score, meaning that our multimodal model is able to learn video features relevant to caption language during training. We also note that the early-fusion multimodal model (T5 + 13D) slightly outperforms the late-fusion multimodal model, which suggests that the model learns more effectively without extra encoders (see Fig. 4b). Both the early-fusion and the late-fusion multimodal models perform worse with a single I3D feature. This suggests that the model benefits from the whole video to correctly answer the caption.

There is also a large performance gap between the multimodal model performance and human performance. Therefore, there is plenty of space for improvements to achieve human performance, and the video fill-in-the-blank task is worth investigating in future visual understanding research.

Method	val EM	val F1	test EM	test F1
BASELINES				
Most Frequent Answer	15.4	45.1	16.4	45.3
T5 zero-shot	39.3	52.0	37.4	49.2
T5 fine-tuned	58.0	73.8	54.5	70.9
OUR MULTIMODAL MODELS				
T5 + 1f I3D	59.2	74.7	54.3	70.5
T5 + I3D	**60.2**	**75.0**	**56.2**	**71.4**
Late-fusion T5 + 1f I3D	53.7	70.3	50.3	67.6
Late-fusion T5 + I3D	53.5	69.7	51.6	67.8
UPPER BOUND (HUMAN AGREEMENT)				
leave one worker out	75.3	82.6	75.0	82.5
new humans*	49.0	90.2	n/a	n/a

Table 3: Results on the validation set. EM stands for Exact Match, and F1 is the token-level F1 score (both in percentage). *1f* refers to the variant of the multimodal model with a single I3D feature. The new humans performance is measured from a random sample of size 200. See Section 3.4 for more details on the human baselines.

5.3 Error Analysis

Results per Semantic Label. To measure how well the model understands different patterns in the caption data, we compare the predictions generated for blanks corresponding to words of different semantic categories (the rest of the answers gen-

erally belong to the same category as the blanked words). Two of the authors annotated the ground-truth labels for common non-overlapping semantic categories, including people, passive entities, and locations, similar to Semantic Role Labeling (SRL) categories. *Person* corresponds to answers about people, *Passive entity* represents passive entities such as objects, *Pronoun* includes subject or object pronouns, *Location* corresponds to places in general, *Preposition* includes noun phrases inside multi-word prepositions (e.g., "order" in "in order to"), *Action* involves actions over time ("a handstand" in "perform a handstand"), *Audio* refers to noun phrases indicated through audio ("the procedure" in "the person describes the procedure"), *Abstract* corresponds to high-level concepts (e.g., "a great time"), *Event* are long-running processes ("a party"), and *Other* correspond to instances hard to label for the annotators (e.g., "a video").

We list the categories and their distribution/size in Table 4, and we also show the performance for the best text-only zero-shot method (T5 zero-shot), text-only fine-tuned method (T5 fine-tuned), and multimodal method (T5 + I3D). The results of T5 zero-shot show some categories can be easily predicted, without fine-tuning on the dataset, such as *Preposition*, *Pronoun*, and *Event*. However, fine-tuning T5 on our dataset accomplishes improvements for nearly all categories. The multimodal (T5 + I3D) model presents improvements for categories such as *Person* and *Abstract* nouns but performs worse for others such as *Audio* and *Action*. This finding follows from the fact that understanding higher-order audio and visual concepts requires complex reasoning, for which the video-aware model may need more training. In general, *Action* and *Passive entity* will likely require extra attention in future work, considering the comparatively low performance for these categories.

Best model vs. Human performance. We want to measure where our best model (T5 + I3D) fails and humans perform well, in order to gain insights on how to improve our models for future work. We find three main types of wrong predictions: The most common error is predicting "man" instead of "women", followed by predicting "person" instead of "child" or "baby". The majority of the remaining errors are predictions close to the ground truth answers such as "dance" instead of "exercise", "pillow" instead of "sheets", "rug" instead of "sand", "floor" instead of "court", "knife" instead of "spatula" or "basketball game" instead of "wrestling".

Category	Size (%)	T5 zs	T5 ft	T5 + I3D
Passive entity	40.4	52.9	**63.6**	**63.6**
Person	33.4	37.0	81.8	**83.2**
Pronoun	6.1	73.5	**85.6**	84.3
Location	5.5	55.1	74.5	**75.4**
Preposition	4.5	81.6	95.7	**97.5**
Action	3.9	47.8	**65.5**	59.9
Audio	2.5	56.4	**73.0**	63.6
Abstract	2.2	59.6	70.0	**77.9**
Other	1.5	56.9	75.0	**83.7**
Event	1.0	70.0	68.0	**84.0**

Table 4: F1 scores on the validation set for blanks with different semantic categories, in descending order based on their size. The results correspond to the best T5 zero-shot, T5 fine-tuned, and T5 + I3D models.

Based on these types of errors, in future work, the model would benefit from pre-training on unbiased data (both gender and age) and also from pre-training on a large-scale multimodal (language and video) dataset, to learn about more diverse situations and objects.

6 Conclusions

This paper introduced a fill-in-the-blanks evaluation framework for video understanding. The framework addresses drawbacks of alternative video understanding tasks, such as multiple-choice visual question answering or video captioning.

The paper made three important contributions. First, we generated a large dataset consisting of 28,000 videos and fill-in-the-blanks tests, building upon an existing video captioning dataset with a new set of manual annotations, using a modified annotation framework to encourage diverse responses among annotators. This process can be easily replicated to create new fill-in-the-blank data for other datasets and tasks. Second, we conducted extensive analyses of the dataset and the manual annotations, to measure the quality of the annotations, and to understand the patterns and limitations of the data. Finally, we introduced a multimodal model that fuses language and visual information and found that the video-aware models significantly outperform the text-only models. Notably, we found a consistent gap between model performance and human performance, which suggests room for improvement among future models in the task of fill-in-the-blank for video understanding.

Our dataset and code are available at https://anonymous.4open.science/r/video-fill-in-the-blank-754D/.

References

Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Comput. Surv.*, 52(6).

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4971–4980.

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 64–73.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5803–5812.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308.

Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. 2020. LifeQA: A real-life dataset for video question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4352–4358, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. MovieNet: A holistic dataset for movie understanding. In *Computer Vision – ECCV 2020*, pages 709–727, Cham. Springer International Publishing.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2758–2766.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: Video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2016–2022. AAAI Press.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 706–715.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online. Association for Computational Linguistics.

Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. 2021. VALUE: A multi-task benchmark for video-and-language understanding evaluation. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A New Dataset and Benchmark on Animated GIF Description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4641–4650.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Huaishao Luo, Lei Ji, Botian Shi, H. Huang, N. Duan, Tianrui Li, X. Chen, and M. Zhou. 2020. UniViLM: A unified video and language pre-training model for multimodal understanding and generation. *ArXiv*, abs/2002.06353.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6884–6893.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640.

Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. 2017. MarioQA: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2867–2875.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123(1):94–120.

Arka Sadhu, Kan Chen, and Ram Nevatia. 2021. Video question answering with phrases via semantic roles. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2460–2478, Online. Association for Computational Linguistics.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision – ECCV 2016*, pages 510–526, Cham. Springer International Publishing.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. VideoBERT: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Du Tran, Maksim Bolonkin, Manohar Paluri, and Lorenzo Torresani. 2016. VideoMCC: a new benchmark for video comprehension. *arXiv preprint arXiv:1606.07373*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1686–1697.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. ActivityNet-QA: A dataset for understanding complex web videos via question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9127–9134.

Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. 2021. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9837–9846.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. 2017. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3).

Linchao Zhu and Yi Yang. 2020. ActBERT: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8746–8755.

A Dataset

A.1 Most-Frequent Noun Phrases

We report the most-frequent noun phrases in the original labels and in the annotations we collected, in Fig. 5. The most frequent nouns for both label sets tend to reference people, which makes sense considering the content of the videos. In the annotation data, we see a greater variety of synonyms for the same kind of person ("male", "man", "guy"), likely a result of the task definition, which encourages paraphrasing.

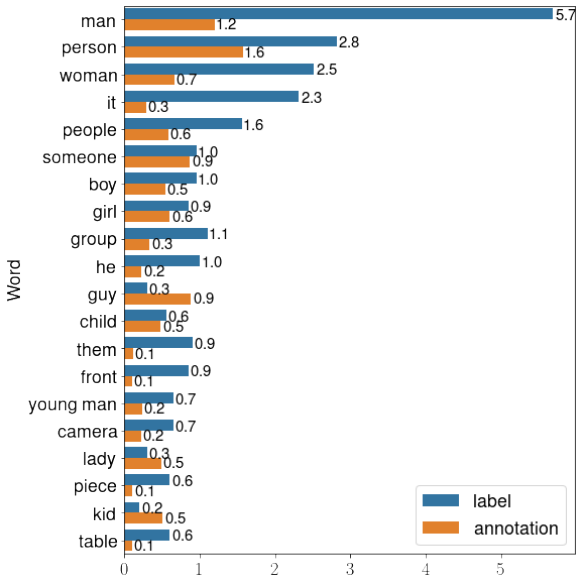


Figure 5: Top 20 nouns for ground-truth labels and annotations in validation and test data.

A.2 Part-of-speech Distribution

We compare the rate of use of words in different part-of-speech categories for the ground-truth labels and the annotations, using the same parser specified earlier to label part-of-speech tags in the noun phrases. The distributions are shown in Fig. 6, and we see that the annotations have roughly the same rate of part-of-speech tag use in all categories, except among adjectives and pronouns where the ground-truth labels have a higher rate of use. This is likely an artifact of the data collection strategy, which encouraged annotators to generate unique noun phrases rather than phrases with adjectives or pronoun references.

A.3 Part-of-speech Sequence Distribution

Although the candidate answers collected from crowd workers consist of noun phrases, they may include different part-of-speech (POS) sequences within the noun phrases. The distributions of POS sequences in Fig. 7 show that the annotators tended to write "bare" nouns without extra determiners and proper nouns, more than the original labels. This makes sense considering that the task asked annotators to provide many unique nouns without consideration for the nouns' structure.

A.4 Dependency categories

Due to the sampling process, some of the labels occur in different syntactic contexts, e.g. in a prepositional phrase in "A woman does push-ups on _____"

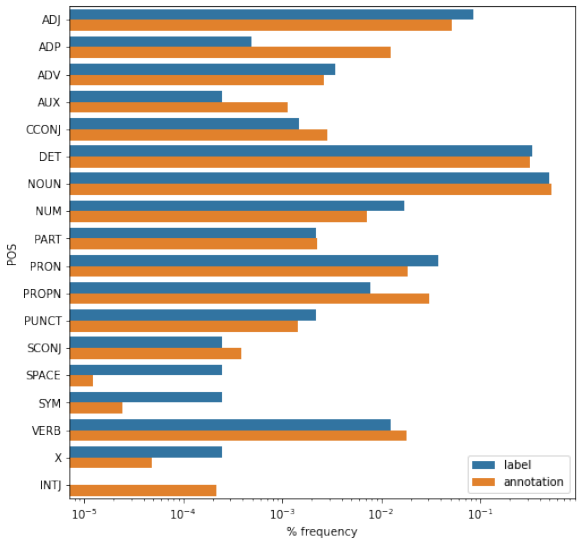


Figure 6: Relative frequency of part-of-speech tags in annotated and ground-truth labels.

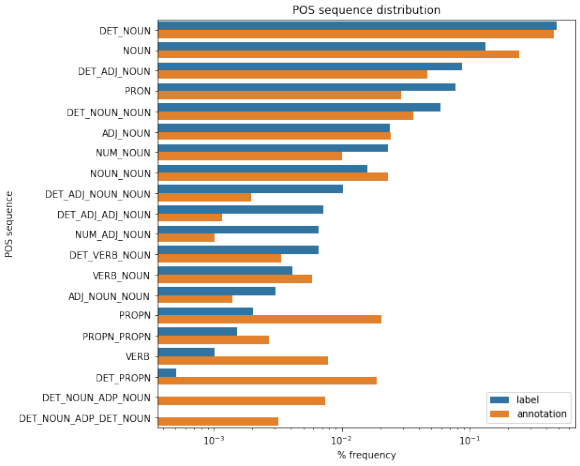


Figure 7: Relative frequency of POS tag sequences in annotated and ground-truth labels.

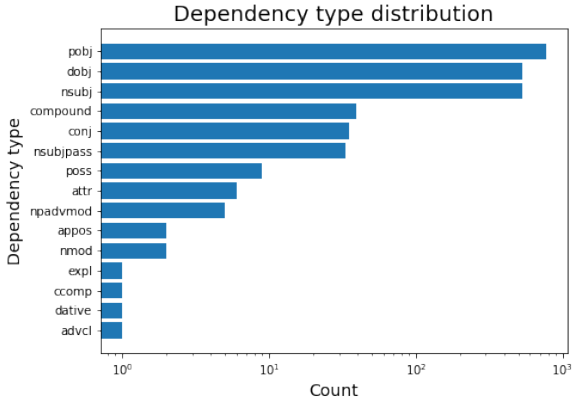


Figure 8: Dependency category counts (per caption).

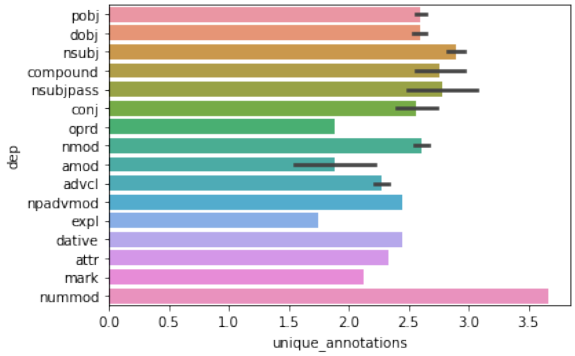


Figure 9: The unique number of labels provided for masked noun phrases grouped by dependency category (categories sorted by frequency).

or as a subject in "_____ at a driving range demonstrating..." (see Fig. 1). We plot the distribution of dependency categories in Fig. 8, which shows that nouns occur in a wide range of positions but mostly occur in a preposition, subject, and direct object positions.

Next, we test whether certain syntactic contexts tend to attract more labels from the annotators than others, by computing the mean unique number of labels per annotator within each syntactic context (based on the dependency parse connected to the masked NP). The distribution is shown in Fig. 9. Captions that mask noun phrases which occur in preposition (`pobj`) and direct object (`dobj`) positions tend to attract slightly fewer unique labels per annotator than the next most-frequent categories, subject (`nsubj`) and compounds (`compound`). This intuitively makes sense, since annotators would likely have fewer options for noun phrases when faced with a preposition or a direct object, as opposed to the less restrictive subject noun position.

A.5 Gender Representation

Often, language processing models can learn to encode social bias due to non-representative training data, such as image captions for photos of men and women taken in stereotypical environments (Zhao et al., 2017). We find a slight gender gap in our own data: by using a gender word list, we find that about 10.9% of the original masked labels are male-related words in contrast to 6.2% that are female-related, and 9.1% of the annotations are male-related while 5.9% are female-related. We note that the gender imbalance is less severe for the annotations than for the original labels, and the annotations do in fact use more gender-neutral hu-

man words than the labels (6.6% for annotations vs. 6.0% for original labels). While some of the annotators may undoubtedly have some bias in terms of their decisions, some of the bias may also result from the original video clips. We acknowledge this limitation as a direction for future work in collecting video caption data.

We used the following lists for gendered words, which were chosen to be in similar semantic categories (e.g. male "brother", female "sister", neutral "sibling"):

- Male-oriented words: "boy", "brother", "father", "guy", "he", "him", "himself", "his", "male", "man", "son"
- Female-oriented words: "daughter", "female", "girl", "her", "herself", "lady", "mother", "she", "sister", "woman"
- Gender-neutral words: "adult", "baby", "child", "human", "kid", "parent", "people", "person", "sibling"

A.6 Spatiotemporal Trends of the Blanked Entities

One of the authors of this paper randomly sampled 50 videos to analyze spatiotemporal information on the blanked entities. Figures 10 to 12 show trends on where, when, and for how long the blanked entities appear in the videos. As expected, the blanked entity generally appears at the center of frames, with a small tendency to be on the lower side. We observe that around 93% of the time the blanked entity appears between seconds 2 and 4 of the video but that there is still a high chance (75%) of seeing it at any given moment. 68% of the time the blanked entities appear for the entire duration of their corresponding video.

B Experiments and Results

B.1 More Implementation Details

We use the T5 model from the HuggingFace Transformers library (Wolf et al., 2020). We train the model with Adam (Kingma and Ba, 2015) on a V100-16Gb with a batch size of 64 for 10 epochs (4,000 steps) using a learning rate of 1e-4 with a warm-up of one epoch and a linear decay. The training time is short, less than an hour. We compute the loss as the cross-entropy between the model-generated output and the originally blanked phrase.

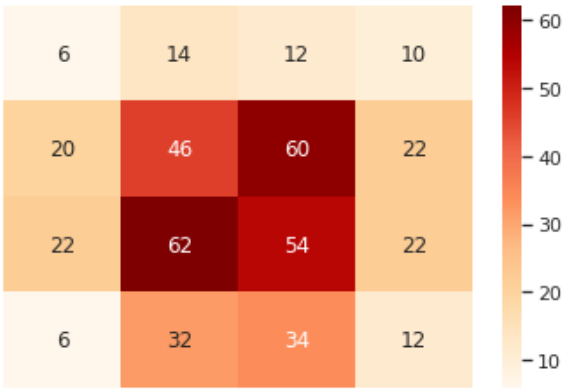


Figure 10: A heat map showing where the blanked entity appears within the video if we divide each frame into a 4 by 4 grid, for a sample of 50 videos. A blanked entity is counted for a given cell if any part of the entity is present in that cell at any moment of its corresponding video (so multiple cells can be counted because it is big enough or because there is movement).

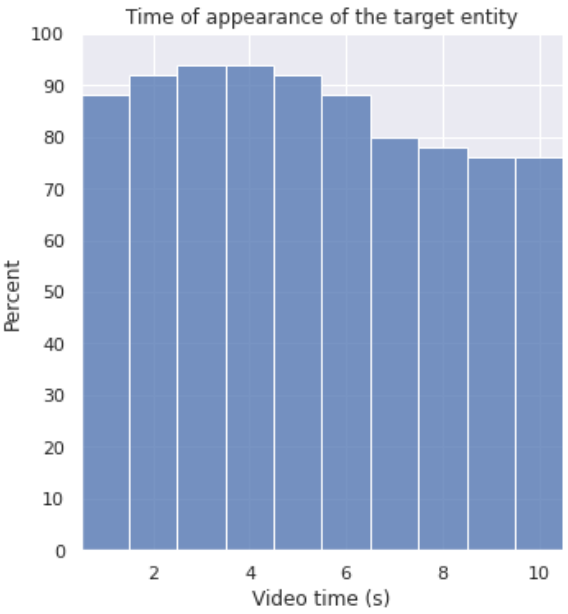


Figure 11: The Percentage that the blanked entity appears at a given time in its corresponding video for a sample of 50 videos. The time is divided into one-second buckets, and a bucket is counted if it appears in any of its frames.

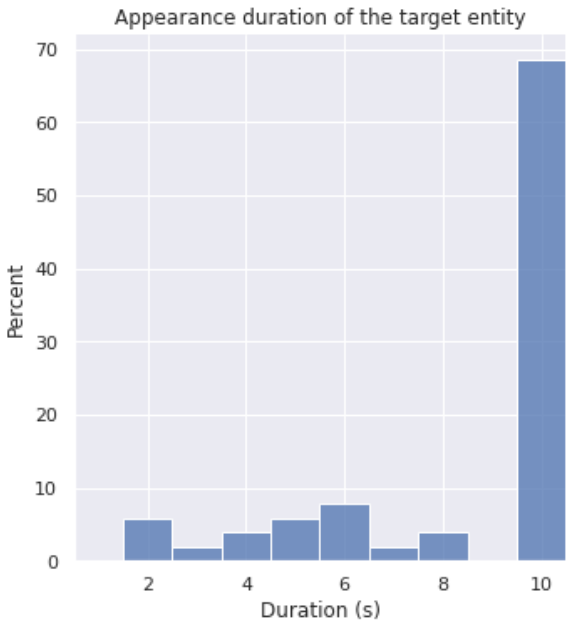


Figure 12: Distribution of the total time that each blanked entity is seen within its video, for a sample of 50 videos.

For test-time decoding, we use beam search with a beam size of 4 for the early-fusion model and 8 for the late-fusion one, with a maximum token length of 10. We stop the decoding early, if an example has seen as many complete hypotheses as the beam size (beam search early-stopping[4]). We penalize the repetitions of bigrams within a decoded text. For each example, we choose the first beam that is a noun phrase, as detected by spaCy (Honnibal et al., 2020), or the first one if none. We show the effect of varying the beam size in Appendix B.2. We find that modifying the beam search early-stopping property does not lead to major performance changes.

B.2 Beam Search

Table 5 shows the effect of varying the beam size during the beam search decoding. In all cases, using a beam search of at least size 2 is better than a greedy search. However, the results are marginally better or inconclusive when using beam size 4 or 8. This is probably related to the phenomenon described by Meister *et al.* (Meister et al., 2020) in which beam search does get us closer to the true maximum a posteriori solution but the answers actually start to get worse after a certain point.

[4] https://huggingface.co/transformers/internal/generation_utils.html#transformers.BeamSearchScorer

	1	2	4	8
T5 fine-tuned	72.9	**74.2**	73.8	73.8
T5 + I3D	73.0	74.0	**74.3**	74.2
Late-fusion T5 + I3D	69.0	69.6	**69.7**	**69.7**

Table 5: F1 scores on the validation set for the beam sizes 1 (greedy search), 2, 4, and 8.

	EM	F1
`t5-small`	20.2	37.1
`t5-base`	34.9	50.2
`t5-large`	43.5	59.5
`t5-3b`	**44.9**	**62.6**

Table 6: Results on the validation set for different model sizes of the T5 text-only zero-shot model.

B.3 Model size

In Table 6 we show the result of changing the T5 model size for the text-only zero-shot baseline. We note we could not fit the model variant `t5-11b` into GPU memory. As expected, we note an increase in the evaluation metrics as the model capacity increases.

B.4 Qualitative Analysis

We show in Table 7 several examples of answers correctly predicted by the best multimodal method but incorrectly answered by the best text-only method. Even though the answers provided by the text-only method are plausible by just looking at the text, they do not make sense with the given videos. In the second example, one can quickly tell the person is not at a gym but instead is in some kind of indoor room. For these examples, the multimodal method seems to have identified what is visually important.

	 A person at the top of _____ with ropes hanging down.	 A guy is by the stairs in _____ doing the moonwalk in socks.	 A man is showing and describing a rock sample to _____.
correct answers	adirondacks, cliff, climb, frozen waterfall, gully, hill, ice, icy cliff, ledge, **mountain**, ravine, slope, snow	building, doors, entryway, foyer, his home, his house, home, house, living room, **room**, shorts, t-shirt	audience, **camera**, consider where its hinge goes, describe how it looks, discuss its hinge, explain his viewers, his audience, his followers, his subscribers, his viewers, people, students, viewer, viewers
T5 fine-tuned	a tree (0)	a gym (0)	a woman (0)
T5 + I3D	a mountain (100)	a room (100)	a camera (100)

Table 7: Examples of instances correctly predicted by the best multimodal method but incorrectly predicted by the best text-only method. The F1 score obtained by each answer is shown in parentheses. The correct answers are shown normalized and separated by commas while the model predictions are shown verbatim. From each video, we show a single frame illustrating the key moment.