ICSVR: Investigating Compositional and Syntactic Understanding in Video Retrieval Models

Avinash Madasu Vasudev Lal
Cognitive AI, Intel Labs
{avinash.madasu, vasudev.lal}@intel.com

Abstract

Video retrieval (VR) involves retrieving the ground truth video from the video database given a text caption or vice-versa. The two important components of compositionality: objects & attributes and actions are joined using correct syntax to form a proper text query. These components (objects & attributes, actions and syntax) each play an important role to help distinguish among videos and retrieve the correct ground truth video. However, it is unclear what is the effect of these components on the video retrieval performance. We therefore, conduct a systematic study to evaluate the compositional and syntactic understanding of video retrieval models on standard benchmarks such as MSRVTT, MSVD and DIDEMO. The study is performed on two categories of video retrieval models: (i) which are pre-trained on video-text pairs and fine-tuned on downstream video retrieval datasets (Eg. Frozen-in-Time, Violet, MCQ etc.) (ii) which adapt pre-trained image-text representations like CLIP for video retrieval (Eg. CLIP4Clip, XCLIP, CLIP2Video etc.). Our experiments reveal that actions and syntax play a minor role compared to objects & attributes in video understanding. Moreover, video retrieval models that use pre-trained image-text representations (CLIP) have better syntactic and compositional understanding as compared to models pre-trained on video-text data. The code is available at https://github.com/IntelLabs/multimodal_cognitive_ai/tree/main/ICSVR.

1. Introduction

Video-retrieval (VR) is the task of retrieving videos for a given text caption or given a video, retrieve the corresponding text caption. This involves understanding important details such as **objects & attributes** (Eg: two women and a man, red shirt guy), **actions** (Eg. playing, standing, talking etc.) in the text caption and the video. In vision it is referred to as compositional reasoning [5, 18, 24, 29, 30], i.e. representing the image or video requires the understanding of primitive concepts that make them. In the recent years, new benchmarks [6, 23, 27, 38] have been proposed to measure the compositional capabilities of foundational image models. The compositionality in these models is measured by creating new text captions from the original text captions using word ordering [48], word substitutions [39], negative pairs [21], image-text mismatch [45].

When compared to images, measuring compositionality is a lot harder in videos. There are multiple reasons to this: First, videos are made-up of time-series image frames with multiple **objects & attributes** and **actions** unlike images. Therefore, methods like creating negative pairs, mismatching pairs etc. used for evaluating compositionality in image-language models have very limited scope. Second, even though tasks based on video question answering (VQA) [6, 23] have been proposed to measure the compositionality, recent studies [3, 9, 13, 22, 42] have shown that these datasets exhibit single frame bias. Most of the previous works [21, 39, 48] focus on understanding the compositionality of image-text models. It mainly involves experimenting with **objects & attributes** in the text captions and retrieving the images. However, **actions** play a crucial role in when retrieving videos using text captions. Another important aspect which is often overlooked in the previous studies is the **syntactics**. For example consider the query *"a guy wearing a red shirt drives a car while talking"*, the **objects & attributes** are **guy**, **red shirt** and **car**, the **actions** are **wearing**, **driving** and **talking** and rest of the words (**a**, **while**) form the **syntactics** of the text captions. The video retrieval models can comprehend such queries because of the accurate **syntactic** and compositionality (**objects & attributes** and **actions**).

Now consider the following scenarios of the text captions in which (i) objects & attributes are missing (*a wearing a drives a while talking*) (ii) actions are missing (*a guy a red shirt a car while*) and (iii) syntactics are missing (*guy wearing red shirt drives car talking*). This begs an important question: **What is the effect of each of these scenarios on the video retrieval performance?**

To address this question, we propose a detailed study to evaluate the syntax and compositional understanding of video retrieval models. For this study we create a comprehensive test bed to evaluate the state-of-the-art video retrieval models for compositonality and syntactics. We base this investigation along three axes: Objects & attributes, actions and syntactics. We propose a set of 10 tasks for these categories: four tasks to evaluate the knowledge of VR models for objects & attributes (§3.1.1), three tasks for testing action understanding (§3.1.2) and finally, three tasks for syntactic capabilities (§3.2). Table 1 describes these tasks with an example. We perform a comprehensive evaluation on 12 state-of-the-art video retrieval models belonging to two categories (§4.1): The first category of models such as Frozen-in-Time (FiT) [2], MCQ [15] etc. are pre-trained on large scale video datasets and fine-tuned for video retrieval. The second category uses pretrained image features like CLIP for video retrieval namely CLIP4Clip [33], CLIP2Video [10] etc. These models are tested on three standard video retrieval benchmarks (§4.2) such as MSRVTT [52], MSVD [4] and DiDeMo [1].

Our experiments (§5.1) reveal that objects & attributes are the most crucial to video retrieval followed by actions and syntax. Among video retrieval models, CLIP based models have a better compositional and syntactic understanding when compared with pretrained video models. We further perform detailed studies to fully judge how retrieval models perceive each of the components. We find that (§5.2) video retrieval models have a poor understanding of relationship between objects and its attributes. However, they are extremely sensitive to incorrect object references in the captions. Our studies on action understanding (§5.3) disclose that models have poor sense of action negation and replacing them with incorrect actions lead to slight decrease in video retrieval performance. Finally, we discover (§5.4) that models perform significantly better even without the right syntax. In summary, our contributions in this paper are as follows:

- Ours is the first work to comprehensively investigate the compositional and syntactic understanding of video retrieval models.
- For this study, we propose a set of 10 tasks dealing with different aspects of compositionality and syntax.
- We perform this analysis on a broad range of 12 state-of-the-art models and generalize the findings to the video retrieval task.
- We establish that video retrieval models exhibit distinct and contrasting behaviours for interpreting various elements in the text captions.

2. Related Work

2.1. Video retrieval

In the recent years, there has been a tremendous improvement on the task of video-retrieval. This is mainly due to two reasons (i) with the adaption of transformer based models to vision tasks like image classification [8, 19, 32] (ii) with the availabilty of large scale video-text datasets like HowTo100M [37], WebVid-2M [2] and YT180M [54]. Frozen-in-Time [2] is a dual-stream transformer model pre-trained on WebVid-2M and Conceptual captions-3M [44] datasets and fine-tuned for downstream video retrieval. A prompt based novel pre-training task [28] is proposed to effectively align visual and text features during large scale video-text pre-training. A new pre-training approach Masked Visual-token Modeling (MVM) [11] is presented to better model the temporal dependencies among videos for video-retrieval. To incorporate the rich semantic features of the videos, a novel pretext task Multiple Choice Questions (MCQ) is put forward in which the model is trained to answer questions about the video.

In a parallel direction, image features pre-trained on large amounts of image-text data have been adopted for the task of video retrieval. CLIP4Clip [33] is an end-to-end trainable video retrieval model based on CLIP [41] architecture in which frame features are extracted using clip image encoder and the temporal modelling is performed using a transformer encoder. A two-stage framework CLIP2Video [10] is proposed to enhance interaction among video features and video-text features for video-retrieval. Madasu et at. [36] used off-the-shelf multi-lingual data to enhance the performance of video-retrieval. All these video-retrieval models haven't been tested for syntactic and compositional understanding. To the best of our knowledge, ours is the first work to comprehensively explore syntactic and compositional understanding of video retrieval models.

2.2. Syntactics

Transformer based language models [7, 20, 53] have achieved state-of-the-art results on most natural language understanding tasks [50, 51]. Hence, there has been a growing interest to explore the morphological capabilities of these models [14, 40, 43, 49, 55]. Since all the video retrieval models use pre-trained language models for encoding text captions, we build upon those works and investigate their syntactic understanding.

2.3. Compositionality

Although vision-language models pretrained on large amounts of data achieved state-of-the-results there has been a growing interest to understand the working of these models [6, 23, 23, 27, 38, 47]. These works mainly focus on compositional knowledge these models by proposing

new benchmarks. Winoground [48] dataset was introduced in which a pair of text captions contain the same set of words but pertain to different images. The models are then tested for image and caption match. Another benchmark CREPE [13] was put forward to evaluate two aspects of compositionality: systematicity and productivity. This benchmark contains unseen compounds and atoms in the test split to evaluate the models' generalization. Parcalabescu et al. [39] proposed VALSE dataset to measure visio-linguistic capabilities of pretrained vision and language models. AGQA-Decomp [13] is a new benchmark to measure compositional consistency for the task of Video Question Answering. All these works proposed new benchmarks for compositional reasoning in image-language models. Contrary to these, our work focuses on measuring compositionality of video retrieval models using the standard datasets and doesn't require a new benchmark. Moreover, our experiments are evaluated on 12 models which are significantly higher than the frequency of models used in these works.

3. Compositional and Semantic Understanding

In this section, we first define syntax and compositionality and subsequently establish the evaluation protocol for syntactic and compositional understanding in video retrieval models. For this evaluation, we augment the existing text captions and create new datasets that assess their syntactic and compositional understanding. We explain this protocol using an example test caption (Q) *"a guy wearing a red shirt drives a car while talking"* from the MSRVTT dataset. Table 1 summarizes different augmentation methods used for the proposed study.

3.1. Compositionality in videos

A video is composed of multiple objects & attributes interacting with each other in a similar or different fashion. To retrieve a video, corresponding text caption is passed as an input to the video retrieval model. This text caption typically consists of objects & attributes and interactions (actions) unique to that particular video. The video retrieval model parses the input caption and computes the matching scores with all the videos. Finally, the video with the highest matching score is the predicted ground truth video. Therefore a video retrieval model should be able to understand each of the objects & attributes and actions present in the caption. This is called compositionality in the visual world. To evaluate the compositional understanding in video retrieval models, we mainly focus on their ability to parse objects & attributes and actions. Next we discuss the evaluation protocol to measure compositionality in VR models.

3.1.1 Object & Attribute knowledge

Object & Attribute removal ($Q_{objattrrem}$): In this setup, we remove all the objects & attributes in the original caption Q and the resulting caption is *"wearing a drives a while talking"*. Here guy, red shirt and car are the objects & attributes.
Object shift ($Q_{objshift}$): To test the VR models ability to relate objects with their attributes, we shift the places of objects in the captions. The modified caption is *"a shirt wearing a red car drives a guy while talking"*.
Object replacement (Q_{objrep}): We evaluate the VR models sensitivity to objects by randomly replacing the objects with an entirely different objects. The replaced caption is *"a surf wearing a red mars drives a guy channel while talking"*.
Object partial ($Q_{objpartial}$): In this setup, the VR models are given access to just 50% of the objects in the caption. This is to understand if the models perform any shortcuts while retrieving videos. Eg: *"a wearing a red drives a car while talking"*. Next, we introduce the tasks for evaluating action knowledge in VR models.

3.1.2 Action knowledge

Action removal ($Q_{actrrem}$): The actions present in the original captions are eliminated. The modified caption is *"a red shirt a car while"* as the actions wearing, drives are removed. This is to understand the influence of actions on the video retrieval performance.
Action negation (Q_{actneg}): A negation is added to all the actions in the captions resulting in the new caption *"a guy **not** wearing a red shirt **not** drives a car while **not** talking"* This tests the VR models ability to comprehend negation in the captions.
Action replacement (Q_{actrep}): In this setup, the actions are randomly replaced with a different set of actions. The replaced actions are neither antonyms nor synonyms. It checks if the models truly recognize the meaning of the action words. Next we present the evaluation protocol for syntactic understanding of VR models.

3.2. Syntactic understanding

In the previous section we elucidated the components for compositional reasoning in videos namely objects & attributes and actions. These components are bind together by syntax there by forming a meaningful caption. Let's consider a part of the example described previously "a guy wearing a red shirt drives a car", if the word "car" and "guy" are interchanged the resulting caption will be "a car wearing a red shirt drives a guy" which is not meaningful. Consequently, syntax also play a crucial role in video retrieval performance along with the compositonality. Subsequently, we put forward the evaluation protocol to measure syntactic

Notation	Caption type	Example
Q	Original caption	a guy wearing a red shirt drives a car while talking
$Q_{objattrrem}$	Object & Attribute removal	a wearing a drives a while talking
$Q_{objshift}$	Object shift	a shirt wearing a red car drives a guy while talking
Q_{objrep}	Object replacement	a surf wearing a red mars drives a channel while talking
$Q_{objpartial}$	Object partial	a wearing a red drives a car while talking
Q_{actrem}	Action removal	a guy is a red shirt a car while
Q_{actneg}	Action negation	a guy not wearing a red shirt not drives a car while not talking
Q_{actrep}	Action replacement	a guy removing a red shirt flying a car while sleeping
Q_{synrem}	Syntax removal	guy wearing red shirt drives car talking
Q_{shuf}	Word order shuffle	talking red shirt drives while car a guy a wearing a
Q_{rev}	Word order reverse	talking while car a drives shirt red a wearing guy a

Table 1. Table shows the types of perturbations applied to the text captions. The example text caption is taken from the MSRVTT [52] dataset. Red color denotes the change from the original text caption.

understanding in video retrieval models.

Syntax removal (Q_{synrem}): Our first experiment focuses on the effect of syntax on VR models. We modify the caption by keeping just the objects & attributes, actions and eliminate any meaningful syntax among them. The resulting caption is *"guy wearing red shirt drives car talking"*.

Word order shuffle (Q_{shuf}): In this setup, all the words are shuffled in the caption. This destroys the order of compositionality and syntax. This tests the order sensitivity of VR models.

Word order reverse (Q_{rev}): In this setup, we preserve the word order except that in the reverse order. It evaluates the positional knowledge of video retrieval models. Next, we present the experiment set up for quantifying the compositional and syntactic understanding.

4. Experiments

In this section, we explain the video retrieval models and datasets used for the proposed analysis.

4.1. Models

We experiment with two categories of video retrieval models. The first category of models are pretrained on large scale video-text datasets like WebVid-2.5M [2] and YT-Temporal-180M [54] and fine-tuned for downstream video retrieval datasets. These include Frozen-in-Time (FiT) [2], MCQ [15], MILES [16], VIOLET [11] and MVM [12]. The second category involves models that adapt pretrained image-text features such as CLIP [41] for the task of video retrieval. This category comprise of seven architectures namely TS2NET [31], CLIP4CLip [33], CLIP2Video [10], XCLIP [34], XPOOL [17], EMCL [25] and DiCoSA [26].

4.2. Datasets

We perform the evaluation on three video retrieval datasets: MSRVTT [52], MSVD [4] and DiDeMo [1]. MSRVTT has 10000 videos and each video has multiple captions totalling 200K. We report the results on MSRVTT-9k split (9000 for training and 1000 for testing). MSVD consists of 1970 videos and 80K captions. The training split has 1300 videos and the test split has 670 videos. The captions in these datasets are single sentence. DiDeMo is made up of 10K videos and 40K descriptions. Following [33], we concatenate all the sentences and evaluate as paragraph-to-video retrieval.

4.3. Implementation

We use spacy[1] to identify parts-of-speech for all the words in the caption. We consider nouns, adverbs and adjectives as objects & attributes, verbs as actions and rest of the parts-of-speech as syntax. We use the exact set up used by the state-of-the-art video retrieval models and measure the performance on all the augmented datasets.

5. Results and Discussion

5.1. Objects & Attributes vs Actions vs Syntax: Do all of them matter?

Our aim is to analyze the importance of three components: **objects & attributes**, **actions** and **syntactics** that make up a text query for retrieving videos. Hence, we test the video retrieval models with text captions that have missing objects & attributes ($Q_{objattrrem}$), actions (Q_{actrem}) and syntax (Q_{synrem}). Tables 2, 3 and 4 show the results on MSRVTT, MSVD and DiDeMo datasets respectively. It is evident from the table that there is a drop in video retrieval performance when tested with text captions that don't have actions (Q_{actrem}). The drop is more pronounced among CLIP based models than pretrained video models. This shows that actions play a role for retrieving correct videos. However, we see that the performance drop is not as ex-

[1] https://spacy.io/usage/linguistic-features

		Text-to-Video Retrieval				Video-to-Text Retrieval			
Type	Model	Q	Q_{actrem}	$Q_{objattrrem}$	Q_{synrem}	Q	Q_{actrem}	$Q_{objattrrem}$	Q_{synrem}
Pretrained video	FiT [2]	26.1	22.8	5.2	20	27.9	23.7	5.8	25.7
	MCQ [15]	26	21.9	4.1	20.1	19.4	15.7	3.7	18.6
	MILES [16]	26	21.3	3.3	19.9	17.5	15.2	2.9	17.1
	VIOLET [11]	35.6	29.5	0.1	25	-	-	-	-
	MVM [12]	36.3	31	8.7	33.7	-	-	-	-
CLIP [41]	TS2NET [31]	36	30.6	6	29.3	25.4	21.2	4.3	41.4
	CLIP4Clip [33]	43.4	37	9.7	35.3	43.6	39	10.3	39.7
	CLIP2Video [10]	46	38.8	8.4	35.3	43	38	10	40.8
	XCLIP [34]	46.1	39.8	10.5	35.6	45.4	40.2	11	42.2
	XPOOL [17]	46.9	39.5	7.6	36.4	44.4	39.6	11.1	42
	EMCL [25]	47.8	40.8	8.2	37.4	46.2	39.5	11.6	42.8
	DiCoSA [26]	47.9	41.3	9.1	38.3	45.9	41.2	13.4	43.1

Table 2. The table shows the results on MSRVTT [52] dataset in both text-to-video and video-to-text retrieval settings. Q denotes the performance (R@1 score) on the original unchanged dataset. Q_{actrem}, $Q_{objattrrem}$ and Q_{synrem} is the R@1 score on datasets that have excluded actions, attributes and syntax respectively.

		Text-to-Video Retrieval				Video-to-Text Retrieval			
Type	Model	Q	Q_{actrem}	$Q_{objattrrem}$	Q_{synrem}	Q	Q_{actrem}	$Q_{objattrrem}$	Q_{synrem}
Pretrained video	FiT [2]	36	32.7	6.9	34.9	36.1	31	7.9	34.9
	MCQ [15]	43.6	36.4	9	42.4	40.3	33.7	9.9	39
	MILES [16]	44	39	8.1	43.9	43.7	37.3	9.6	41.5
	VIOLET [11]	48.3	40.6	10.8	45.8	-	-	-	-
	MVM [12]	49.6	41.5	10.5	45.6	-	-	-	-
CLIP [41]	TS2NET [31]	52.8	38.5	11.8	49.4	51.2	37.1	10.7	48.6
	CLIP4Clip [33]	54.5	42.1	11.9	51.9	51.8	38.5	10.9	50.7
	CLIP2Video [10]	55.8	41.6	11.8	50.6	53.6	40.5	12.5	51.6
	XCLIP [34]	54	39.7	12.4	49.7	54.9	42.6	13.3	48.6
	XPOOL [17]	56.1	47	11.9	53.9	56.6	48	12.4	53.3

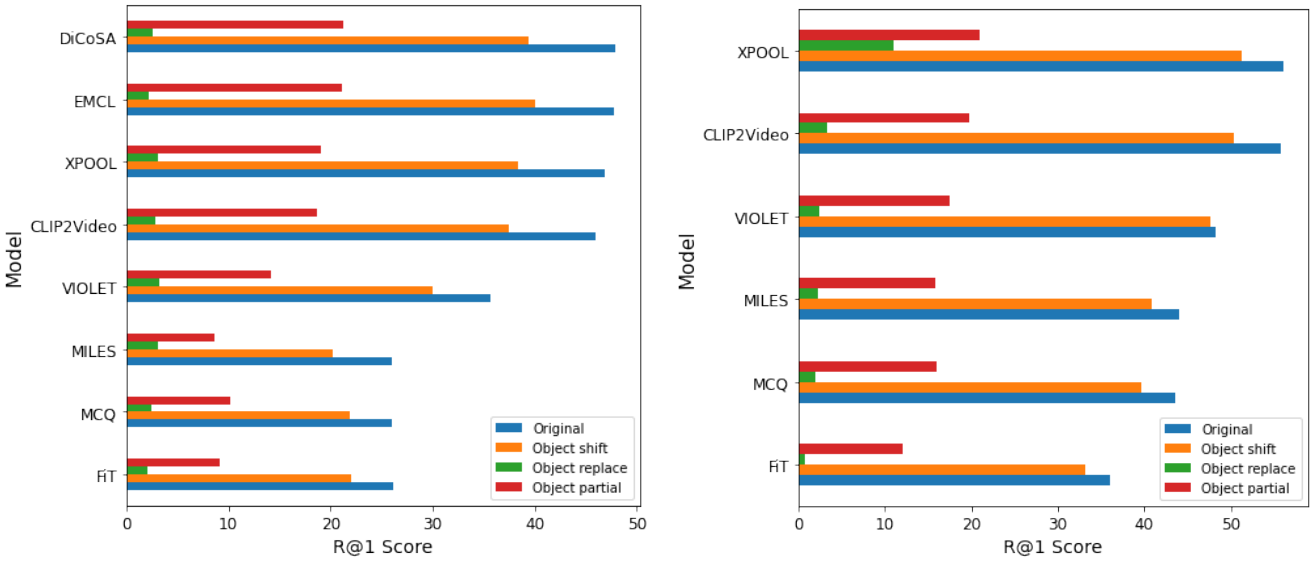
Table 3. The table shows the results on MSVD [4] dataset in both text-to-video and video-to-text retrieval settings. Q denotes the performance (R@1 score) on the original unchanged dataset. Q_{actrem}, $Q_{objattrrem}$ and Q_{synrem} is the R@1 score on datasets that have excluded actions, attributes and syntax respectively.

pected. Videos are time-series image frames which can have same attributes. In those scenarios, actions help in differentiating those videos. We see this effect when the R@1 is lower among pretrained video models and higher among CLIP based models. When the performance is lower, actions do not play a significant role in video retrieval and hence the videos can be retrieved without them in the text caption. On the contrary if R@1 score is higher, we see a notable decline. This is due to the robust video representations of CLIP based models as compared to pretrained video models. CLIP based models accurately encode video representations but, when the differentiating factor among videos i.e actions are missing in text captions leads to retrieval of incorrect videos. In short caption length datasets like MSRVTT and MSVD, we notice a significant drop in performance as compared with DiDeMo which is a paragraph (>1 sentences) dataset. This is because text captions in DiDeMo contains detailed description of the videos and hence, missing actions didn't lead to drop in performance as compared to MSRVTT and MSVD. It demonstrates that actions are not essential in paragraph-video retrieval.

Next, we analyze the performance of video retrieval models tested with text captions without syntactics (Q_{synrem}). From the table, it is clear that there is a reduction in R@1 without the syntax in the text captions. It validates that syntactics are necessary for retrieving correct ground truth videos. For MSRVTT, we observe that models tested without syntax under-perform compared to actions in the text captions and the average difference in performance is 2%. The reverse is true for MSVD and DiDeMo datasets

Type	Model	Text-to-Video Retrieval				Video-to-Text Retrieval			
		Q	Q_{actrem}	$Q_{objattrrem}$	Q_{synrem}	Q	Q_{actrem}	$Q_{objattrrem}$	Q_{synrem}
Pretrained video	FiT [2]	29.2	28	4.4	27.4	28.2	27.1	5.6	27
	MCQ [15]	24.6	22.3	5.6	22.8	23.8	21.2	5.2	21.4
	MILES [16]	28	24.4	3.9	24	22.6	22.4	4.7	22.2
	VIOLET [11]	24.8	23.9	4.5	26	-	-	-	-
	MVM [12]	24.8	23.9	4.5	26	-	-	-	-
CLIP [41]	CLIP4Clip [33]	42.6	25.4	6.7	37.7	41.4	19	7.9	38
	XCLIP [34]	43.2	39.8	8.7	41.5	45.6	41.2	10.3	40.2
	XPOOL [17]	43.7	40.3	8.4	40.1	43.7	40.4	8.9	39.5
	EMCL [25]	46.3	40.2	6.9	41.7	44.8	42.1	9.3	42.3
	DiCoSA [26]	45.4	41.5	7.9	41.1	45.1	41.8	9.5	41.2

Table 4. The table shows the results on DiDeMo [1] dataset in both text-to-video and video-to-text retrieval settings. Q denotes the performance (R@1 score) on the original unchanged dataset. Q_{actrem}, $Q_{objattrrem}$ and Q_{synrem} is the R@1 score on datasets that have excluded actions, attributes and syntax respectively.



(a) MSRVTT (b) MSVD

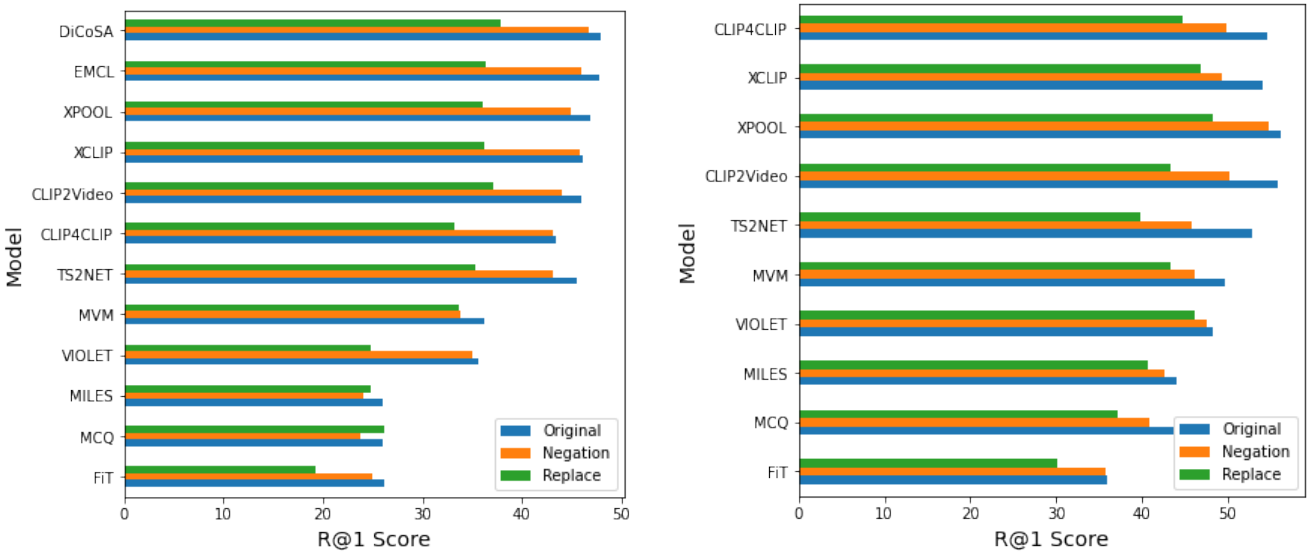
Figure 1. We perform ablation studies on the role of objects & attributes in video retrieval. The video retrieval models are evaluated on three tasks namely: Object shift ($Q_{objshift}$), Object replacement (Q_{objrep}) and Object partial ($Q_{objpartial}$). Results show that swapping of objects has minor effect on performance followed by masking 50% objects. The highest drop is seen when the objects are randomly replaced. These ablation studies are performed on MSRVTT [52] and MSVD [4] datasets

.

where there is a huge difference of 9%. In addition, we also notice that CLIP based models are more sensitive to syntax than pretrained video models. Finally, we evaluate the video retrieval models on text captions in the absence of objects & attributes ($Q_{objattrrem}$). As seen from the results, these models perform poorly (a drop in 20%) which underscores the significance of objects & attributes. We also notice that $Q_{attrrem}$ trails Q_{actrem} and Q_{synrem} by a huge margin. This difference is more striking among CLIP based models as opposed to pretrained video models.

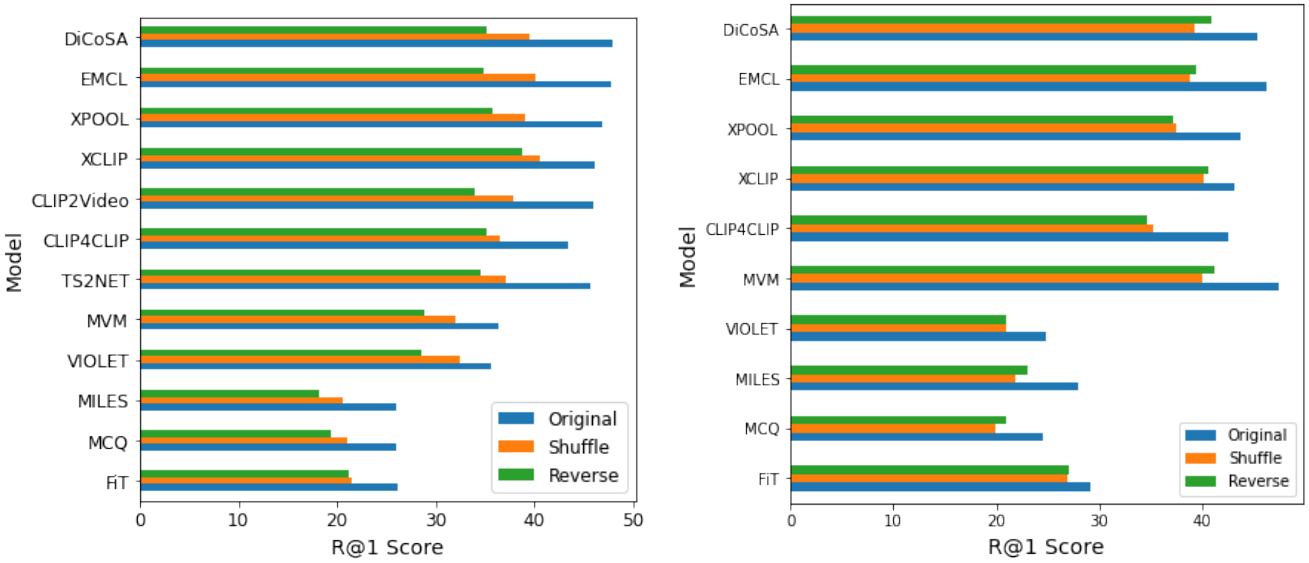
5.2. What role do Objects & Attributes play in video retrieval?

In the previous sections (§5.1), findings from our experimental results suggested that objects & attributes are the most important component in text captions while retrieving videos. To investigate further, we perform additional detailed studies on their importance. In captions there can be multiple objects & attributes and every pair of object & attribute is distinct from the other. Any slight modification



(a) MSRVTT (b) MSVD

Figure 2. Figure shows the performance comparison (R@1 score) of video retrieval models on action ablation studies. The VR models are evaluated on captions with negated actions and replaced actions from MSRVTT [52] and MSVD [4] datasets respectively. These studies illustrate that VR models have incomplete knowledge of negation and also are immune to action replacement in the captions



(a) MSRVTT (b) DiDeMo

Figure 3. Video retrieval performance (R@1) on word order task. We test the models on original (unchanged) captions, captions with shuffled word order and captions with reversed word order for MSRVTT and DiDeMo datasets. We demonstrate that VR models act like bag-of-words and do not require substantial word order information.

in the pairs can totally change their correspondence and thereby the ground truth video and hence, video retrieval models should be able to account for these changes. We perform a test in which we interchange the places of objects while keeping rest of the caption same $Q_{objshift}$. In the second study, we randomly replace objects in the caption Q_{objrep} and evaluate the models on the modified ones. The final ablation involves keeping just half the objects in the captions ($Q_{objpartial}$). This is to assess if VR models adapt any shortcuts and still retrieve correct videos without

the critical information. Figure 1 demonstrates the results for these studies. As shown in the figure, there is a slight deterioration of video retrieval performance when there is a object shift in the caption. The drop is a meagre 5.5% for MSRVTT and 3.6% in case of MSVD dataset. It demonstrates that VR models do not quite fully understand the relationship between object and its attribute. On the other hand if the objects are randomly replaced (Q_{objrep}) with different unrelated objects, there is a massive degradation in R@1 score. In fact, the performance is quite similar to the models performance tested on captions without objects & attributes. These results prove that video retrieval models are extremely sensitive to alteration of objects. Figure 1 shows that there is a noticeable fall in performance when the VR models have access to just 50% of the object data in the captions. The R@1 score lags by 30% in MSRVTT and 22% in MSVD datasets. It reinforces the aforesaid extreme sensitivity nature of the retrieval models towards objects. Furthermore, random object replacement performs far worse than partial objects in the captions. This highlights that factual object description even though 50% is much more crucial than access to the entire caption albeit with incorrect objects.

5.3. Do VR models pay attention to actions?

We demonstrated in the section 5.1 that actions play a role in video retrieval. Now, this raises an important question ***How much attention do VR pay to actions in the captions?*** To investigate this we perform certain ablation studies on the action understanding of VR models in the text captions. We replace the action word with the negation of it (Q_{actneg}) and test the performance of VR models on the newly formed captions. In parallel, the actions in the captions are randomly replaced with different actions (Q_{actrep}) and VR models are evaluated on the altered captions. In Figure 2, we provide the results of VR models tested on captions with negated (Q_{actneg}) and replaced (Q_{actrep}) actions. From the figure, it is evident that action negation (Q_{actneg}) achieves comparable results to Q and there is a slight drop in performance in case of action replacement (Q_{actrep}). Most of the actions are expressed in positive sense in these datasets and this is not always the case. For a fine-grained description of videos, the actions of the static objects can be communicated in a negation form. So, naturally video retrieval models are expected to understand the negation in captions. However, we notice that action negation has similar performance as original captions which demonstrates that VR models lack the capability of action negation sense. Next, we randomly replace the actions with a different action and test the attention of VR models. In an ideal scenario, the performance of these models should drop drastically as the replaced actions do not correspond to that of the ones in ground truth videos. Nevertheless, we see that the R@1 score of action replacement (Q_{actrep}) is only slightly less than original caption Q. In fact the average drop in R@1 is only 6.8% in MSRVTT and 7.5% in MSVD. Hence even though the actions are important in video retrieval, VR models use other influential information such as objects & attributes to retrieve ground truth videos.

5.4. Does word order of text captions matter?

In the figures 3a and 3b, we present the findings on the word order evaluation. First we observe that models tested on datasets without word order perform worse than the original dataset. The R@1 score is reduced on average by 6.3% and 9.1% on shuffled (Q_s) and reversed (Q_r) MSRVTT captions respectively. On a similar note the performance drops 5.5% on shuffled and 5% on reversed DiDeMo dataset. Additionally, the R@1 decrease is more pronounced on reversed captions than shuffled. This is surprising as the object-action order is preserved in reversed captions in contrast with shuffled. This shows that models adapt bag-of-words approach for syntactic understanding of captions and positioning of object-action order doesn't matter. A possible explanation for this behaviour is: all the video retrieval models use pretrained language models as their text encoder. Recent studies have shown that [35, 46] distributional information is preserved even though the syntactic word order is disturbed and hence, LMs leverage it for hierarchical text understanding. Surprisingly, video retrieval models manifest the same behaviour in caption understanding.

6. Conclusion

In this work, we proposed a comprehensive investigation of compositional and syntactic understanding of video retrieval models. For this study we put forward 10 different tasks to evaluate models reasoning of objects & attributes, actions and syntax for retrieving videos. We experiment with a wide range of 12 state-of-the-art video retrieval models and 3 standard benchmarks. We show that video retrieval performance is heavily impacted by objects & attributes and lightly by syntactics. Furthermore, our results also reveal that word order matter less for video retrieval models. These results shed an important light on the inner workings of video retrieval models. We believe the future works can utilize these findings to design compositional aware video retrieval models.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2, 4, 6

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for

end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2, 4, 5, 6

[3] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927, 2022. 1

[4] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 2, 4, 5, 6, 7

[5] Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos. In *International Conference on Learning Representations*. 1

[6] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10086–10095, 2020. 1, 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 1

[10] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2, 4, 5

[11] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. 2021. 2, 4, 5, 6

[12] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22898–22909, 2023. 4, 5, 6

[13] Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Measuring compositional consistency for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5046–5055, 2022. 1, 3

[14] Aina Garí Soler and Marianna Apidianaki. Let's play monopoly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844, 2021. 2

[15] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 2, 4, 5, 6

[16] Yuying Ge, Yixiao Ge, Xihui Liu, Jinpeng Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Ping Luo. Miles: visual bert pre-training with injected language semantics for video-text retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 691–708. Springer, 2022. 4, 5, 6

[17] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5006–5015, 2022. 4, 5, 6

[18] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 1

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[20] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*. 2

[21] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, 2021. 1

[22] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7366–7375, 2018. 1

[23] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1, 2

[24] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-

temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 1

[25] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. *Advances in Neural Information Processing Systems*, 35:30291–30306, 2022. 4, 5, 6

[26] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. Text-video retrieval with disentangled conceptualization and set-to-set alignment. *arXiv preprint arXiv:2305.12218*, 2023. 4, 5, 6

[27] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 1, 2

[28] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[29] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3032–3041, 2022. 1

[30] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 812–819, 2014. 1

[31] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 319–335. Springer, 2022. 4, 5

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[33] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2, 4, 5, 6

[34] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 4, 5, 6

[35] Avinash Madasu and Shashank Srivastava. What do large language models learn beyond language? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6940–6953, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 8

[36] Avinash Madasu, Estelle Aflalo, Gabriel Ben Melech Stan, Shao-Yen Tseng, Gedas Bertasius, and Vasudev Lal. Improving video retrieval using multilingual knowledge transfer. *arXiv preprint arXiv:2208.11553*, 2022. 2

[37] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2

[38] Anssi Moisio, Mathias Creutz, and Mikko Kurimo. Evaluating morphological generalisation in machine translation by distribution-based compositionality assessment. In *The 24rd Nordic Conference on Computational Linguistics*, 2023. 1, 2

[39] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, 2022. 1, 3

[40] Ellie Pavlick. Semantic structure in deep learning. *Annual Review of Linguistics*, 8:447–471, 2022. 2

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 4, 5, 6

[42] Ishaan Singh Rawal, Shantanu Jaiswal, Basura Fernando, and Cheston Tan. Revealing the illusion of joint multimodal understanding in videoqa models. *arXiv preprint arXiv:2306.08889*, 2023. 1

[43] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2021. 2

[44] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[45] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, 2017. 1

[46] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, 2021. 8

[47] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural

language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019. 2

[48] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1, 3

[49] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, 2020. 2

[50] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2018. 2

[51] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019. 2

[52] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 4, 5, 6, 7

[53] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 2

[54] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 2, 4

[55] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9628–9635, 2020. 2