Seeing the Unseen: Visual Metaphor Captioning for Videos

Anonymous ARR submission

Abstract

Metaphors are a common communication tool used in our day-to-day life. The detection and generation of metaphors in textual form have been studied extensively but metaphors in other forms have been under-explored. Recent studies have shown that Vision-Language (VL) models cannot understand visual metaphors in memes and adverts. As of now, no probing studies have been done that involve complex language phenomena like metaphors with videos. Hence, we introduce a new VL task of describing the metaphors present in the videos in our work. To facilitate this novel task, we construct and release two datasets- a manually created dataset with **705** videos and **2115** human-written captions and a synthetic dataset of **90886** MSCOCO images with synthetically generated metaphor captions. We propose a novel video metaphor captioning system: GIT-LLaVA, which uses a frozen video captioning model augmented by a Large Language Model (LLM) for generating metaphors, as a strong baseline. We perform a comprehensive analysis of SOTA video language models on this task. We publish our datasets and benchmark results for our novel task to enable further research.

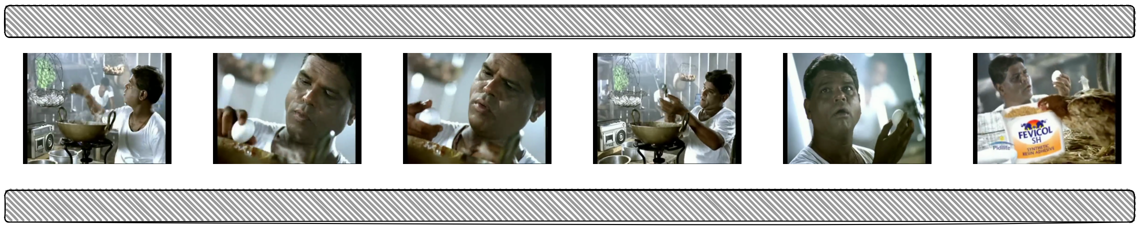
1 Introduction

Metaphors are the most commonly used form of figurative language in literature (Kreuz and Roberts, 1993). Metaphors are a tool to colour the imagination of the reader by introducing unknown concepts in comparison to familiar concepts, thereby allowing them to be understood easily and powerfully. This trope is used in various creative fields like advertisements (Hussain et al., 2017) to convey information more effectively that includes modalities like text, images, and audio. Figure 1 shows an example of using an image to creatively convey an idea. Metaphors are also used in video advertisements. Figure 2 shows a few examples of how metaphors are used in video advertisements to bring emphasis to the product being advertised.

Figure 1: An example of a creative advertisement that uses visual metaphors. The sugar-free nature of lollipop is highlighted by showing ants avoiding them.

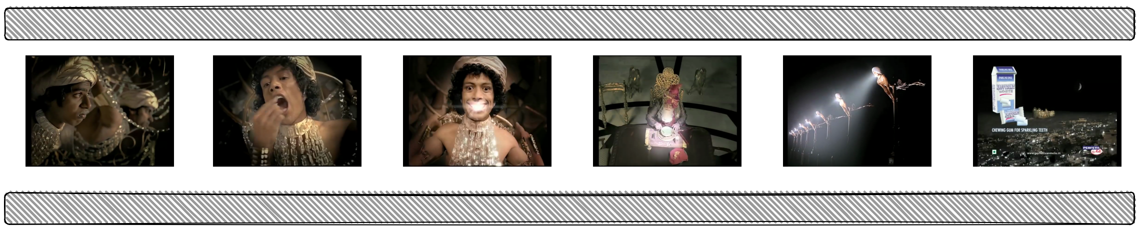
Figurative languages in textual form have been well-studied in literature (Abulaish et al., 2020). With the advent of powerful AI assistants like ChatGPT and BARD and tools that are built on top of them, it is possible to interact with these AI systems through images and audio. Hence it becomes important to build and test models to work with complex language phenomena like metaphors in multiple modalities. Recent works on Visual metaphors (Yosef et al. 2023, Chakrabarty et al. 2023) focus on understanding metaphors present in images and generating images from prompts with metaphors. They show that it is challenging to deal with metaphors presented visually.

Recently, chat assistants that can answer questions related to videos have shown good promise on standard video datasets (Zhang et al. 2023; Li et al. 2023b; Maaz et al. 2023). However, they struggle to understand videos that contain metaphors. To this effect, we build and release a novel video metaphor captioning model built on top of the LLaVA (Liu et al., 2023) model that is trained



Video Explanation: The egg is so strong that is unbreakable with a hammer. The reason is that it was laid by a hen that was fed food from a Fevicol (a glue) box.

Video Metaphor Caption: The adhesion of glue is as strong as an unbreakable egg



Video Explanation: The advertisement happens in a world where humans are used as light towers. The chewing gum makes teeth so white that humans can be used as a light source.

Video Metaphor Caption: The teeth is as white as a light source

Figure 2: Examples of metaphors used in videos to convey ideas creatively along with their explanation

to understand metaphors in videos along with the datasets used to train the model.

Our contributions are

1. A novel low-resource Vision-Language model (GIT video model followed by Vicuna LLM) pretrained and fine-tuned for video metaphor understanding, a task hitherto unattempted (Section: 4).

2. Release of two datasets:

   (a) A benchmark dataset with 705 videos comprising 2115 manually written captions (Section: 3).

   (b) A synthetic dataset consisting of $90,886$ images from the MSCOCO dataset with synthetically generated metaphor captions, built for pretraining (Section: 3.3).

3. Strong baselines which are the SoTA benchmarks for the task of "Video metaphor captioning" (Table: 1).

4. A new metric- Average Concept Distance (ACD) for evaluating the quality of metaphors generated by the model (Section: 6).

1.1 Problem Statement

Input: Video
Output: Caption describing the metaphor.

Video metaphor captioning is the task of describing the metaphor in the video. Given a video 'v', the model generates a single line description of the following format: 'Primary concept' is as 'property' as 'secondary concept'. The model should hence identify the object being compared, the object it is being compared to, the property that links both, and put them all together as a caption.

1.2 Motivation

Vision and Language (VL) models have shown great performance in standard Image-Text and Video-Text tasks (Gan et al., 2022). They however still struggle with tasks that require deeper understanding like metaphors in images (Akula et al., 2022). While concurrent works focus on understanding visual metaphors in images, no such work has been done on understanding metaphors in videos.

Understanding and describing metaphors present in the video is a very challenging task, as established in our work. Hence it could be used as a benchmark to test larger models on their video understanding capabilities in the future. Our framework of using a video captioning model for obtaining video representation can be adapted to other low-resource domain-specific tasks in the future.

1.3 Background

Lakoff (1993) describes metaphor as a mapping between a source and target domain through shared properties. For example, consider the sentence "*The development has hit a wall*". Here, hitting a wall denotes that the development has been halted. The target domain is halting and the source domain is wall and the property of wall is used to describe halting.

Metaphors and similes can be simplified to a syntax of A is B, where A is being compared to B. We use this simple syntax inspired from Akula et al. (2022). A is denoted as the primary concept and B is referred to as the secondary concept. For example, in the sentence "*The blanket is as white as snow*", the primary concept is the blanket and it is compared to the secondary concept snow. The property that links them is their colour. Following prior work, we use the following template to describe the metaphors present in the videos: *Primary Concept* is as *property* a *Secondary Concept*

2 Related Work

Recently, significant efforts have been made to understand metaphors to detect and generate them. Many sentence-level and token-level datasets have been released to facilitate the same (Birke and Sarkar 2006; Steen et al. 2010; Tsvetkov et al. 2014; Mohammad et al. 2016; Mohler et al. 2016).

Metaphor Detection is the task of classifying if the given sentence/token contains a metaphor or not. In recent years, metaphor detection has been explored with the aid of large language models. Choi et al. (2021) used the contextual embeddings from BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) with a late interaction mechanism to make use of linguistic metaphor identification theories. Aghazadeh et al. (2022) probed and analyzed the metaphorical language encoded in the large language models. Su et al. (2020) used a combination of global sentence features and POS information to perform token-level metaphor detection. Badathala et al. (2023) used a multitasking approach to detect hyperbole and metaphors together.

Metaphor generation is the task of generating metaphorical sentences given a literal sentence (Abe et al. 2006, Terai and Nakagawa 2010). Metaphor generation was initially modelled as a template-filling task. Veale (2016) used templates to generate metaphoric tweets. Stowe et al. (2020) used masked language modelling by masking the verbs in the literal sentence and training the model to replace it with its metaphoric counterparts. Stowe et al. (2021) used FrameNet embeddings to generate metaphoric sentences by replacing verbs with metaphoric verbs in literal sentences.

Visual Metaphors: The detection and generation of metaphors in textual form have been explored extensively but the use of metaphors in other modalities like images is not explored until very recently. Akula et al. (2022) introduced a set of tasks related to understanding visual metaphors. They showed that existing Vision-Language models are not good at understanding visual metaphors. Yosef et al. (2023) introduced a multimodal dataset that contains metaphors, similes, and idioms with corresponding images for them. Zhang et al. 2021, Hwang and Shwartz 2023, and Xu et al. 2022 explored the uses of metaphors in memes and released datasets for understanding metaphors in memes. Chakrabarty et al. (2023) explored generating visual metaphors from metaphorical input sentences.

Video Captioning: Video captioning is the task of generating a single-line natural language description of the video. Video-Text models are trained on large-scale paired video and language datasets to align frames to text in the captions. Sun et al. (2019) built on BERT (Devlin et al., 2019) model by learning a joint representation for visual and text tokens for video-text tasks. Lei et al. (2021) proposed CLIPBERT that uses sparse sampling to sample short clips from videos to learn visual representation instead of using the whole video and showed remarkable performance. Luo et al. (2020) is a Unified Video and Language pre-training model for both multimodal understanding and generation built by pretraining the model on 5 diverse objectives. Zellers et al. (2021) uses spatial and temporal objectives during pretraining on a large-scale dataset of videos with transcriptions to align videos to text. The GIT model (Wang et al., 2022) is trained on a large corpus of parallel image-text data. It used a single image encoder and single text decoder and modeled multiple vision-text tasks as a language modeling task. These models however cannot follow instructions which makes it difficult to adapt to newer tasks.

Video Assistants: Recent success in using frozen LLMs with vision encoders for instruction fine-tuning for Image-Text tasks (Li et al. 2023a; Liu et al. 2023) has inspired the use of instruction fine-tuning for videos. Video-LLaMA (Zhang et al., 2023) use frozen visual and audio encoders and projects them to the embedding space of LLMs using Q-formers as in BLIP-2 (Li et al., 2023a). Li et al. (2023b) use information from image, video, and ASR tools along with video embedding to align video frames to text. Video-ChatGPT (Maaz et al., 2023) use CLIP (Radford et al., 2021) as the visual encoder and Vicuna (Zheng et al., 2023) as the LLM and train the model on 100,000 video and instruction pairs. Video-LLaVa (Munasinghe et al., 2023) uses audio signals by transcribing them into text in an LLaVA model-like architecture.

All these models are trained on large-scale video and text data. We propose a new model GIT-LLaVA that uses a frozen video foundation model with an LLM that can be fine-tuned with a few hundred videos to perform video metaphor captioning. Also, our work focuses on visual metaphors in videos which has not been explored before.

3 Dataset

No existing datasets have metaphor details available for videos. As advertisements have metaphorical representations in them to convey additional messages to viewers, we choose the Pitt's Ads dataset (Hussain et al., 2017) for constructing our dataset. The Pitt's Ads dataset consists of advertisement images and videos on a wide range of topics. The released dataset contained URLs to $3,477$ videos out of which only 2063 videos are currently available. We annotate these videos with metaphor information for our experiments. Additionally, we also query YouTube with keywords like advertisements, creative advertisements, funny advertisements, etc. using the YouTube Search tool[1]. We filter videos that are less than 2 minutes and add them to our Video Metaphor Captioning (VMC) dataset.

3.1 Annotation Details

We employed three annotators to generate data for our novel task- video metaphor captioning. The annotators were given detailed explanations about metaphors and visual metaphors with examples. They were given two tests with examples consisting of metaphoric and non-metaphoric videos and asked to classify them. The annotators were shortlisted based on their ability to identify metaphors present in the videos. In our final batch of annotators, all three annotators were in the age bracket of 24-30 years. All three annotators are proficient in English with Masters degrees. Each video is annotated by all the three annotators. Other details are discussed in Appendix A.1

[1]https://pypi.org/project/youtube-search-python/

3.2 Dataset Statistics

Interpretation of metaphors present in videos is very subjective and each annotator can understand it differently. We observed that the captions for each video were diverse. We only include videos in our final dataset that are classified as metaphors by all our annotators. This ensures that the VMC dataset has videos that are unambiguously metaphoric.

We employed an additional expert annotator who is a Masters student in English literature and proficient in understanding metaphors to validate the captions written by the three annotators. We also used the *GPT-3.5-turbo* model (Ouyang et al., 2022) to check for grammar and typos in the captions written by our annotators. The annotators were asked to rewrite the captions if any flaw was identified in terms of spelling or grammar. These quality checks ensured the quality of captions present in the dataset.

All videos are accompanied by three captions. Our **Video Metaphor Captioning (VMC) dataset** consists of 705 metaphoric videos with 2115 captions. The train, val, and test split contain 400, 55, and 250 videos each with 1200, 165, and 750 captions respectively.

3.3 Synthetic Dataset Preparation

In addition to the manually annotated dataset, we create and release a synthetically generated dataset for pretraining our model. The manual annotation of videos with metaphor details is both time-consuming and costly. Hence we generate synthetic image and metaphor caption pairs and use them to pretrain the model before finetuning it on videos.

We use images and captions from the MSCOCO dataset (Lin et al., 2014). We prompt *GPT-3.5-turbo* model with the following prompt: "Convert the following image caption to a metaphoric image caption in the following format <primary concept> is as <property> as <secondary concept>. Input: mscoco_caption". For example, we convert the image caption '*A bicycle replica with a clock as the front wheel*' to '*A timepiece is as cyclical as a bicycle's revolution*'. The generated captions were then cleaned to remove captions that did not follow the template in the prompt. The final pretraining dataset consists of 90886 images and corresponding synthetically generated metaphoric captions. Section A.5 discusses the manual evaluation of the quality of generated synthetic captions.

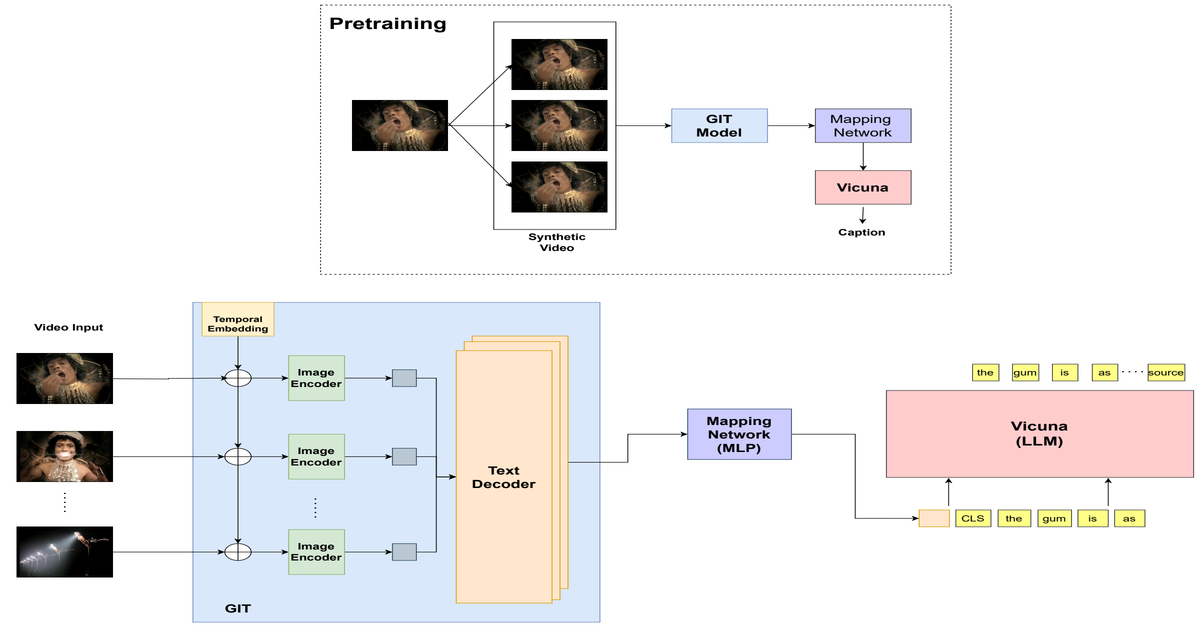


Figure 3: An overview of our Video Metaphor Captioning system, GIT-LLaVA. The text decoder representation of GIT is mapped to the embedding space of Vicuna to generate metaphor captions.

4 Our Model

We model video metaphor captioning as a sequence to sequence task. The video representation is obtained through a pretrained video captioning model and prefixed with an instruction sequence to a Large Language Model (LLM). The LLM generates the caption as a sequence of tokens conditioned on the video input and the instruction. Figure 3 illustrates the model architecture.

We sample 'k' frames from the input video 'V', where k depends on the input restrictions of the video captioning model.

$$V_{input} = [f^1, f^2, ..., f^k] \quad (1)$$

where f denotes each frame sampled from the video. The sampled frames are fed to the video captioning model (C) whose decoder output is used as the representation for the video (H_V). We train a simple Multilayer Perceptron (MLP) network with parameters 'W' to map the video representation H_V to the embedding space of the LLM (H_R), similar to the LLaVA model (Liu et al., 2023). We use task-specific instruction (X_{inst}) as input and the model is trained to generate the answer as output (X_{ans}) autoregressively.

$$H_V = C(V_{input}) \quad (2)$$

$$H_R = W.H_V \quad (3)$$

$$X_{ans} = \sum_{i=1}^{n} log P_\theta(X_i | X_{inst}, H_R) \quad (4)$$

where θ represents the parameters of the LLM, X_i denotes the current token being predicted. The LLM is trained with this language modeling objective. We refer to this model as 'GIT-LLaVA'. We also explore a variation of GIT-LLaVA called GIT-LLaVA-X where we split the video into multiple equal-sized clips and obtain video full video representation by summing up the video representation of each clip.

We use the LLaVA-13B-V1.5 (Liu et al., 2023) model-like architecture for our experiments. We use the Generative Image Text Transformer model (GIT) (Wang et al., 2022) as the video captioning model for obtaining the video representation and Vicuna (Zheng et al., 2023) as the LLM. In all our experiments we freeze the weights of the GIT model and only finetune the mapping network and the LLM. Since we train the mapping network to learn the mapping between the GIT decoder state to the embedding space of the LLM, the mapping network maps GIT's understanding of the video in the form of its representation to the LLM's embedding space, allowing the LLM to directly generate output from the video. This also reduces the need to pretrain the model on a huge corpus of Video-Text parallel data which is harder to obtain.

Model	BLEU-4 ↑	Rouge-L ↑	CIDEr ↑	BERT-F1 ↑	ACS↓	ACD↑
Valley	0.9999	14.4012	1.2519	0.4981	0.7716	0.1463
Video-LLaMA	6.6153	33.3710	29.3587	0.5066	**0.2729**	0.3722
GIT	5.8509	42.3980	7.4936	0.6758	0.3952	0.4104
GIT-LLaVA (Ours)	**12.3868**	**49.9178**	**21.6482**	**0.7259**	0.3236	**0.4908**
GIT-LLaVA-X (Ours)	9.3219	48.3053	11.4496	0.7111	0.3493	0.4627

Table 1: Experimental results on our VMC dataset in comparison to other models. ACS and ACD denote the Average Concept Similarity and Average Concept Distance metric weighted by BERTScore respectively. Cosine similarity and distance is computed between the concepts compared in the metaphor caption. The best model is in bold and the next-best model is underlined.

	GIT-LLaVA	GIT-LLaVA-X
Fluency	1.00	1.00
P.C. Con	0.44	0.15
Creativity	0.77	0.81

Table 2: Results of human evaluation of the captions generated by our models. The numbers denote the ratio of captions that are marked as fluent, consistent, and creative respectively to the total number of captions evaluated. P.C. Con denotes the consistency of the primary concept in the caption with the video.

5 Experiments

We pretrain the model on our synthetic image caption dataset and finetune it on the VMC dataset. We discuss the experiment settings for both as follows.

5.1 Pretraining

Our video metaphor captioning system uses a pretrained video captioning model to obtain video representation. The video representation needs to be mapped to the embedding space of the LLM for it to generate fluent captions. Our dataset for video captioning is small and may not be sufficient to learn this mapping. Hence, we initially pretrain the model on a large synthetic data of images and their corresponding metaphor captions.

The images from the MSCOCO dataset are converted to video by repeating the images to form frames of the video. As only the final decoder state representation is being mapped to the LLM embedding space and the video model is frozen, it does not affect the video understanding abilities of our system. This synthetic video is then fed as input to the video captioning model from which the video representations are obtained. The mapping network trained on the synthetic data is used in the finetuning stage where video data is used.

We use the Generative Image-to-Text (GIT) model (Wang et al., 2022) as our video captioning model for obtaining video representation. We use the GIT-large model that is fine-tuned for video captioning on the VaTeX dataset (Wang et al., 2019). We use the Vicuna-13B model (Zheng et al., 2023) as our LLM. We pretrain the model by creating videos consisting of 6 frames of the same image with a batch size of 4. We pretrain the model for 1 epoch on the entire pretraining dataset. In the pretraining stage, both GIT and Vicuna models are frozen and only the parameters of the mapping network are updated.

5.2 Video Metaphor Captioning

The model is fine-tuned for video metaphor captioning on the VMC dataset after the pretraining on the synthetic dataset. The model is fine-tuned for 5 epochs with early stopping on the validation set.

We explore two frame selection strategies for our models. The GIT-Large model only supports video captioning with 6 frames as input. We sample 2 frames in temporal order across the three different parts of the video- start, middle, and end. This ensures that the 6 frames cover the entire span of the video.

We also perform additional experiments where 6 frames are sampled from different parts of the video, which we call GIT-LLaVA-X. The video is split into 4 video clips with equal duration and video representation is obtained for each video clip using the GIT model. The final representation is obtained by summing up the representations for each video clip. Table 3 compares the performances of models with different numbers of video clip splits.

We use a batch size of 4 with an initial learning rate of $2e-5$ with a warmup ratio of 0.03. Cosine Annealing is used as the learning rate scheduler. BFloat16 precision is used while training the model on 4 A100 GPUs.

Baselines: We use the GIT (Wang et al., 2022), Video-LLaMA (Zhang et al., 2023), and Valley

(Luo et al., 2023) as baselines in our experiments. GIT is chosen as the baseline as it is used as our video encoder. Video-LLaMA and Valley have shown promising performance in following instructions in the video setting. They also have diverse vision and language backbones and thus would make for a fair comparison. All the models are finetuned for 50 epochs with early stopping on the validation set. The experiment details for our baselines are discussed in Appendix A.2

6 Evaluation Metrics

We evaluate the performance of our model using a set of automated metrics and human evaluation. The n-gram overlap-based metrics- BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2014) are commonly used to compare the performance of the model in captioning tasks. As discussed in previous works on metaphor generation, the n-gram overlap based metrics cannot capture the quality of generated metaphors. This is because the same information can be conveyed through different comparisons. Hence, we also report BERTScore (Zhang et al., 2019) that compares the semantic similarity of the generated caption and the reference caption.

In the task of video metaphor captioning, the model is trained to generate creative metaphors as output. Previous works rely on manual evaluation to quantify creativity and metaphoricity of the generated captions. As no existing metric can be used to evaluate the creativity of metaphors, we introduce a new and intuitive metric called- "Average Concept Distance" (ACD). It is calculated as:

$$\text{CS} = \text{Cosine}(\text{PC}, \text{SC}) \tag{5}$$

$$\text{ACD} = \frac{\sum_i^n \text{BERTScore(hyp, pred)} * (1 - \text{CS})}{n} \tag{6}$$

where PC and SC denote the primary and secondary concepts in the predicted caption respectively and Cosine denotes the cosine similarity between them. The primary and secondary concepts denote the object of comparison and the object it is being compared to respectively. Average Concept Distance (ACD) is obtained by weighing the cosine distance between the concepts with the BERTScore of the predicted caption. The caption 'The car is as fast as a jeep' is less creative as it makes an obvious comparison while the caption 'The car is as fast as a cheetah' is more creative. This can be captured by the CS metric but a disfluent caption like 'The adsfd is as fast as a cdsak' will also score low on CS and this can be captured by the ACD metric.

S-BERT (Reimers and Gurevych, 2019) (*all-mpnet-base-v2*) is used to obtain representations for PC and SC. For captions that do not contain either PC or SC, the similarity score is set as 1 to penalize the model. Thus the model is evaluated based on how diverse comparison it can make for the object in question. We also discuss the correlation between the proposed ACD metric and human evaluation of metaphor creativity in Section A.6

In addition to these automated metrics, we also manually evaluate and compare the models based on three metrics- Fluency, Primary Concept Consistency, and Creativity.

7 Results and Analysis

We evaluate the models based on both automatic metrics and using human evaluation.

7.1 Automatic Metrics

Table 1 compares the performance of our models with other baselines. Our models- GIT-LLaVA and GIT-LLava-X perform comparable to or better than other traditional video captioning models despite the smaller scale of pretraining data. It can be seen that both GIT-LLaVA and GIT-LLava-X perform well on n-gram overlap-based metrics like BLEU-1, ROUGE-L, and CIDEr and also the BERTScore metric. This shows that it generates captions that are relatively more semantically similar to the ground truth captions than other models.

Figure 4 shows some examples of metaphors generated by our models. Our models achieve the best score (lowest) on the Average Concept Similarity (ACS) metric. It compares the semantic similarity of the primary and secondary concepts used in the metaphor generated. The lower scores confirm that the generated captions are creative with novel comparisons. The ACS values can also be low if the generated captions are not fluent and unrelated words are present in the caption. This was observed in the captions generated by the Video-LLaMA model. The Average Concept Distance (ACD) is used to capture such cases. Relatively high values in ACD indicate that the model generates more consistent and creative captions compared to other baselines. It also indicates that our system is still not perfect as seen by the manual evaluation of generated captions in Table 2. The baseline models

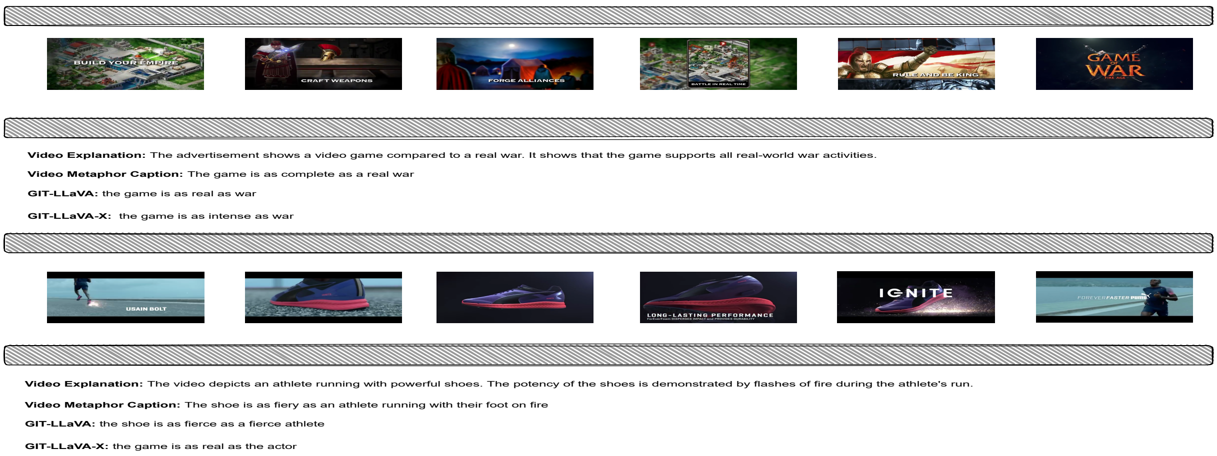


Figure 4: Examples of metaphor captions generated by GIT-LLaVA and GIT-LLaVA-X models.

perform poorly due to the low-resource nature of the task. Overall lower values for ACD further underline the challenge of the task and the inability of the models to generate both creative and consistent metaphoric captions.

7.2 Human Evaluation

In addition to automated metrics, we also perform human evaluation (Table 2) on 100 videos from the test set with outputs generated by all models. Three annotators in the age group of 25-35 were used to annotate these videos. Each annotator annotated 50 videos. 25 videos were common among the annotators. The annotation was done on three metrics-Fluency, Primary Concept Consistency, and Creativity. Primary Concept Consistency denotes if the caption correctly predicted the primary concept in the video. The annotators assigned binary scores for each caption against each of the metrics.

Our models generated mostly fluent captions but were not always consistent with the primary concept of the video. It can seen that by increasing the number of frames, the model hallucinates more and generates less consistent captions. The GIT and Valley models generate fluent captions but they are not consistent and less creative. Video-LLaMA model did not generate any fluent captions. The captions mostly had spelling errors or grammatical errors after finetuning on the VMC dataset due to the nature of the low resource setting. More details about annotator agreement are in Section A.7.

7.3 Error Analysis

The most common case of error is the misprediction of the primary concept in the video as can be seen in Table 2 and Table 4. Figure 6 illustrates a few examples of misprediction. In the first example, the GIT-LLaVA models generate a metaphor about cars when the actual metaphor was about getting a car loan. It was also observed that videos related to shoe brands typically present more about the game and the athletes than about the shoes themselves. This leads to models generating metaphors about people and the game than about the shoes. It was also observed that videos that contain animated objects are confused for advertisements about video games resulting in metaphors being generated about video games. In general, all the models don't seem to have the ability to deeply reason about the video to generate accurate metaphors as shown by the performances on the VMC dataset.

8 Conclusion and Future work

In this work, we proposed a novel Vision-Language (VL) task called video metaphor captioning that probes the language reasoning abilities of the video language models. We constructed and released two new datasets for the proposed task. We also released a new metric to evaluate the creativity of generated metaphor captions. Our VL model that used a frozen video captioning model (GIT) with an LLM decoder (Vicuna) to generate metaphor captions showed that pretraining the model with metaphor knowledge results in relatively better performance in the proposed task. It was observed that all the video language models studied in the work lack a deeper understanding of video and language for a complex task like video metaphor captioning. We believe that our work will enable future research in this direction with our dataset and models being a strong benchmark for progress.

9 Limitations

The scope of our work is only limited to understanding visual metaphors in videos. The models introduced in our work- GIT-LLaVA and GIT-LLaVA-X do not have support for audio and cannot understand metaphors introduced through audio. The audio signals like music and dialogues can be used to better understand metaphor information in videos and we intend to do this in the future.

10 Ethical Considerations

We build our Video Metaphor Captioning (VMC) dataset based on the Pitt's Ads dataset. The original dataset has links to YouTube videos and may contain some videos that propagate biases seen in advertisements. We ensure that no personal information is included in the captions written by our annotators. We also ensure that brand names are replaced with common nouns such that no identifiable information is present in our dataset. Our model uses Vicuna as the decoder and may propagate the biases held by the LLM. We urge the research community to use our models and datasets with necessary caution in downstream tasks for the same reason and use them responsibly.

References

Keiga Abe, Kayo Sakamoto, and Masanori Nakagawa. 2006. A computational model of the metaphor generation process.

Muhammad Abulaish, Ashraf Kamal, and Mohammed J. Zaki. 2020. A survey of figurative language and its computational detection in online social networks. *ACM Trans. Web*, 14(1).

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Annual Meeting of the Association for Computational Linguistics*.

Arjun Reddy Akula, Brenda S. Driscoll, P. Narayana, Soravit Changpinyo, Zhi xuan Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas J. Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. 2022. Metaclue: Towards comprehensive visual metaphors research. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23201–23211.

Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. *ArXiv*, abs/2305.17480.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Annual Meeting of the Association for Computational Linguistics*.

Minjin Choi, Sunkyung Lee, Eun-Kyu Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. *ArXiv*, abs/2104.13615.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. 2022. Vision-language pretraining: Basics, recent advances, and future trends. *Found. Trends Comput. Graph. Vis.*, 14:163–352.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes. *ArXiv*, abs/2305.13703.

Roger J. Kreuz and Richard M. Roberts. 1993. The empirical study of figurative language in literature. *Poetics*, 22(1):151–169.

George Lakoff. 1993. The contemporary theory of metaphor.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7327–7337.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *ArXiv*, abs/2305.06355.

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.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. 2020. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *ArXiv*, abs/2002.06353.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Ming-Hui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *ArXiv*, abs/2306.07207.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv*, abs/2306.05424.

Saif M. Mohammad, Ekaterina Shutova, and Peter D. Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *International Workshop on Semantic Evaluation*.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, Mubarak Shah, and Fahad Shahbaz Khan. 2023. Pg-video-llava: Pixel grounding large video-language models. *ArXiv*, abs/2311.13435.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

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*. Association for Computational Linguistics.

Gerard Steen, Lettie Dorst, J. Herrmann, Anna Kaal, and Tina Krennmayr. 2010. Metaphor in usage. *Cognitive Linguistics*, 21.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736, Online. Association for Computational Linguistics.

Kevin Stowe, Leonardo Ribeiro, and Iryna Gurevych. 2020. Metaphoric paraphrase generation. *ArXiv*, abs/2002.12854.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rong bo Wang, and Zhi qun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *FIGLANG*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin P. Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472.

Asuka Terai and Masanori Nakagawa. 2010. A computational system of metaphor generation with evaluation mechanism. In *International Conference on Artificial Neural Networks*.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Annual Meeting of the Association for Computational Linguistics*.

Tony Veale. 2016. Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 34–41, San Diego, California. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *ArXiv*, abs/2205.14100.

Xin Eric 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. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4580–4590.

Bo Xu, Ting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. Irfl: Image recognition of figurative language. *ArXiv*, abs/2303.15445.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. In *Neural Information Processing Systems*.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. Multimet: A multimodal dataset for metaphor understanding. In *Annual Meeting of the Association for Computational Linguistics*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv*, abs/2306.02858.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Appendix

A.1 Annotation Details

The VMC dataset consists of three manually written captions for each video. The annotators were asked the following questions for each video:

a) Does this video contain a visual metaphor?
b) Is audio of the video required to understand the metaphor?
c) What part of the video contains the metaphor?
d) What is the primary concept in this video?
e) What is the secondary concept in this video?
f) What is the common property of both concepts?
g) Give a one-line description of the form "*primary_concept*" is as "*property*" as "*secondary_concept*".
h) A free-form description of the video.

Questions a and b are Yes/No questions. The annotators write the time of occurrence of the metaphor in the video for question c. Question g follows the format used for annotation in the MetaCLUE dataset (Akula et al., 2022) for visual metaphor in images. The VMC dataset consists of videos that were marked as metaphors (Quesetion: a) by all three annotators. We instruct annotators to ensure that no identification information is included in the primary and secondary concepts and to use common words in their place. For example, instead of 'The coke is as cool as Messi in the finals', the caption is written as 'The drink is as cool as the football player in the finals'.

A.2 Baselines

We use the GIT (Wang et al., 2022), Video-LLaMA (Zhang et al., 2023), and Valley (Luo et al., 2023) as baselines in our experiments.

GIT: We finetune the GIT model that is already fine-tuned for video captioning on VaTEx dataset on our VMC dataset. The model is fine-tuned with a batch size of 4.

Video-LLaMA: We use the 13B model of video-LLaMA that is pretrained and finetuned on parallel video-text data. We then finetune the vision branch of the model on our VMC dataset with a batch size of 4 and report the performance.

Valley: Valley is a video-assistant build on top of the LLaVA model. We use the Valley-2 7B model that is finetuned on video instruction data. We finetune this model on the VMC dataset with 4 as the batch size.

Model	# of VC	BLEU-4	Rouge-L	CIDEr	BERT-F1	ACS	ACD
GIT-LLaVA	1	12.3868	49.9178	21.6482	0.7259	0.3236	0.4908
GIT-LLaVA-X	2	11.2179	49.2053	16.6004	0.7170	0.3239	0.4847
GIT-LLaVA-X	4	9.3219	48.3053	11.4496	0.7111	0.3493	0.4627
GIT-LLaVA-X	6	7.2926	47.7425	8.2535	0.6998	0.3301	0.4688
GIT-LLaVA-NP	1	0.7171	20.2105	1.8686	0.4243	0.9877	0.0029

Table 3: Ablation study results with different number of video clip segmentations. # of VC denotes the number of video clips. GIT-LLaVA-NP denotes the model that was not pretrained on synthetic data

A.3 Dataset Statistics

VMC dataset consists of 705 videos with 2115 captions. The average duration of the video is 54 seconds, and the average length of the caption is 8.9 words. Figure 5 shows histograms for the distribution of video duration and caption lengths.

A.4 Ablation Study

We perform an ablation study on the number of video clip segments that can be fed as input to the video captioning model. We split the video into 1, 2, 4, and 6 parts and fed the video clips to the GIT model. The final video representation is obtained by summing up the individual clip representations. The models were trained as discussed in Section 5.2. Table 3 shows the results of the ablation study. On comparing the performance of these models with GIT-LLaVA, it can be seen that adding more video clips did not improve the model performance.

We also study the impact of the pretraining stage by directly finetuning the GIT-LLaVA model on the VMC dataset. The poor performance shows that imparting metaphor knowledge in the pretraining stage is essential for model performance as the training data is smaller.

A.5 Quality Estimation of Synthetic Data

As discussed in Section 3.3, we clean the synthetic dataset by removing the caption that is not of the format: "<primary concept> is as <property> as <secondary concept>". We further evaluate the quality of the generation by manual evaluation. We employed two annotators to annotate the fluency and creativity of the generated captions. The annotators provided binary classification labels for fluency and creativity. The captions were 98.7% fluent and 97.8% creative. This confirms the quality of the synthetic data.

A.6 Average Concept Distance Metric

We compute the correlation of the Average Concept Distance (ACD) metric with the human evaluation of captions. The manual evaluation in Section 7.2 was done for 100 videos with 5 captions for each video. We used the ACD metric evaluator to compare the ACD scores with binary labels provided by the annotators. The ACD scores and binary labels had a Pearson correlation coefficient of 0.403 with p-value $<< 0.0001$. As creativity is a subjective metric, the moderate correlation is very significant.

Metric	Valley	V-LLaMA	GIT
Fluency	0.47	0.00	0.89
PCC	0.07	0.13	0.13
Creativity	0.22	0.01	0.73

Table 4: Results of human evaluation of the captions generated by baseline models. PCC (Primary Concept Consistency) denotes the consistency of the primary concept in the caption with the video.

Cohen's Kappa (κ)	A	B
B	0.6937	
C	0.6345	0.7316
Fleiss' Kappa (K)		0.6844

Table 5: IAA calculations with Fleiss' Kappa and pairwise Cohen's Kappa among the annotators

A.7 Manul Evaluation Results

Table 4 shows the manual evaluation scores obtained for the baseline models. The manual evaluation scores further complement the results obtained with automatic metrics. Table 5 shows the Inter-Annotator Agreement (IAA) between the three annotators for manual evaluation of captions generated by all 5 models. The scores indicate substantial agreement between the reviewers.

A.8 Examples

Figure 6 shows some examples of wrong captions generated by the models.

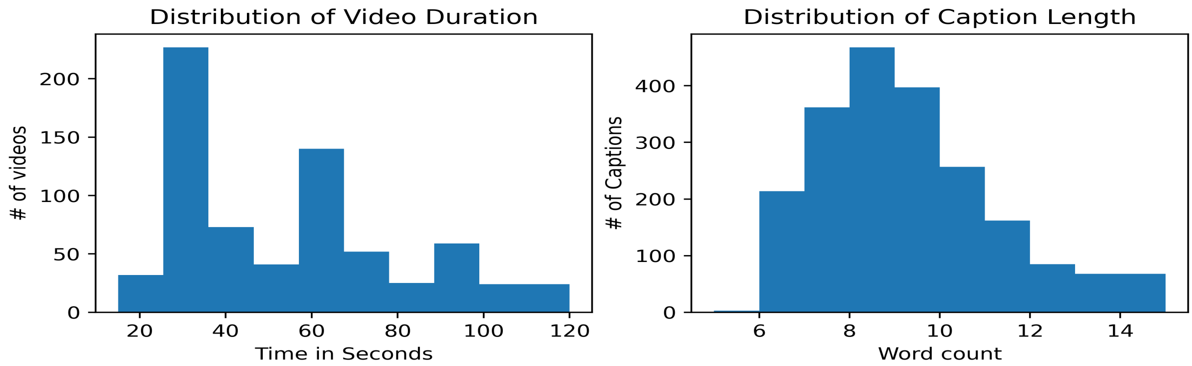


Figure 5: The distribution of video clip duration and caption length in the VMC dataset

Video Explanation: The video depicts a man magically bringing a car out of a magazine. The video suggests that buying a car through a bank loan is as easy as performing a magic trick

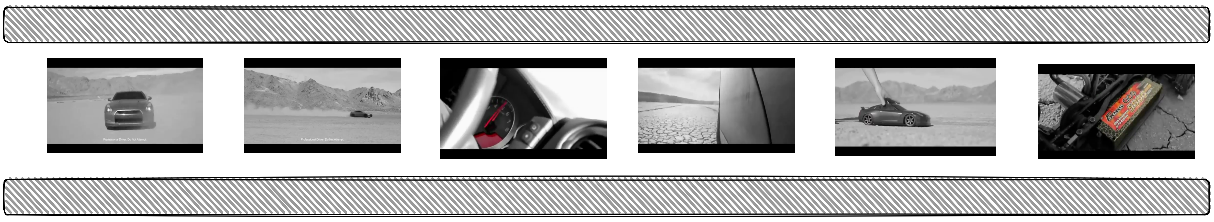
Valley: Your hair looks like a cloud in the sky, and I am the sun shining brightly behind it.

Video-LLaMA: is like importantsecond/ as itsecondary conce

GIT: the drink is as glamorous as the lush life.

GIT-LLaVA: the car is as fierce as a fierce warrior.

GIT-LLaVA-X: the car is as fierce as fire.



Video Explanation: The video depicts a car running at high speed through a desert. It shows that the battery underneath is responsible for the power and performance of the car

Valley: the smartphone's performance is as reliable as a classic car.

Video-LLaMA: the- as fast as a withOtrightarrow

GIT: the chips are as ecstatic as insane.

GIT-LLaVA: the car is as fierce as a racing car

GIT-LLaVA-X: the car is as fierce as a lion.

Video Explanation: The video depicts a bear scratching its back while dancing. The advertisement suggests that the taste of the chocolate feels as unreal as a dancing bear

Valley: The video is like a flower that blooms slowly.

Video-LLaMA: theocolate is as strong as acing be

GIT: the phone is as powerful as men playing rugby

GIT-LLaVA: the milk is as pure as the cow

GIT-LLaVA-X: the game is as intense as war.

Figure 6: Examples of prediction mistakes done by the models on video metaphor captioning