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 no studies have been done on understanding metaphors in videos, 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 **741** videos and **1142** 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) to generate captions. We build our model on top of the LLaVA model with the GIT model as the encoder and map its decoder to the LLM (Vicuna) using a lightweight mapping network. We show that this allows the video captioning model to develop the ability to understand video metaphors. We publish our datasets and benchmark results for our new 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

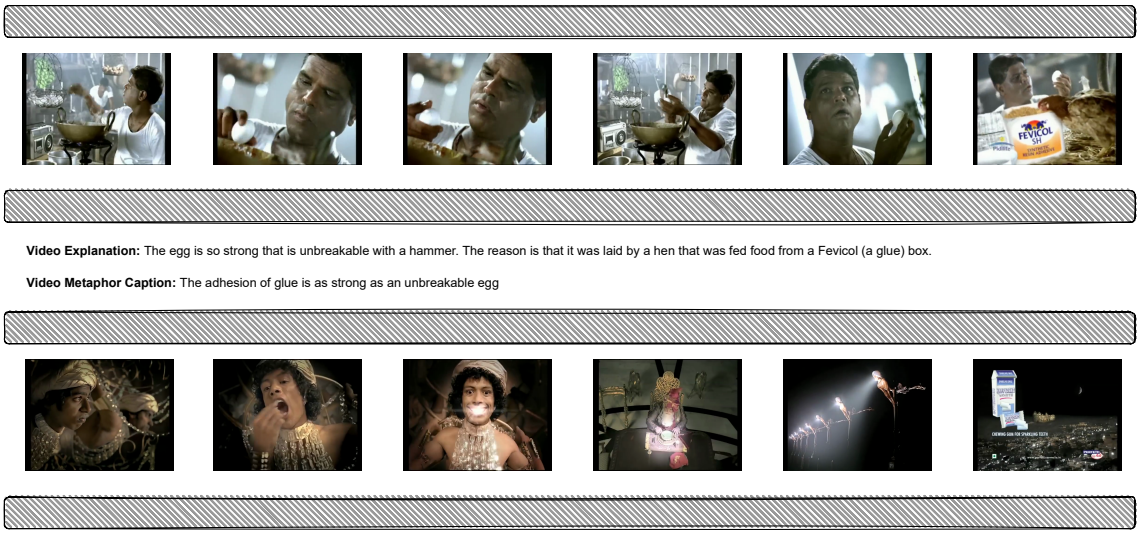

Figure 2: Examples of metaphors used in videos to convey ideas creatively along with their explanation

metaphor captioning model built on top of the LLaVA (Liu et al., 2023) model that is trained to understand metaphors in videos along with the datasets used to train the model.

Our contributions are

1. A novel 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 741 videos comprising 1142 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. Benchmark results for the task of "Video metaphor captioning" (Table: 2).

4. A new metric- Average Concept Similarity (ACS) 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 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 vi-

sual 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.

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, two annotators were in the age bracket of 24-30 years and one above 50 years. All three annotators are proficient in English with Masters degrees. Each video is annotated by all the three annotators.

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.

Cohen's Kappa (κ)	A	B
B	0.651	
C	0.886	0.601
Fleiss' Kappa (K)		0.712

Table 1: IAA calculations with Fleiss' Kappa and pairwise Cohen's Kappa among the annotators

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.

3.2 Dataset Statistics and Annotation Validation

Interpretation of metaphors present in videos is very subjective and each annotator can understand it differently. We observed multiple valid hypotheses for classifying a video as a metaphor or not. We report the Inter Annotator Agreements between our annotators in Table 1. The agreement between annotators is substantial as both Fleiss' Kappa and pairwise Cohen's Kappa are above 0.6 for all cases.

We employed an additional annotator who is a Masters student and proficient in English 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 grammar-corrected caption is then verified by the final annotator before being added to the final dataset.

A video can contain 1 to 3 captions. Our final dataset- the **Video Metaphor Captioning (VMC) dataset** consists of 741 metaphoric videos with 1142 captions. The train, val, and test split contain 590, 70, and 81 videos each with 895, 112, and 135 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 a time consuming and costly process. In our video metaphor

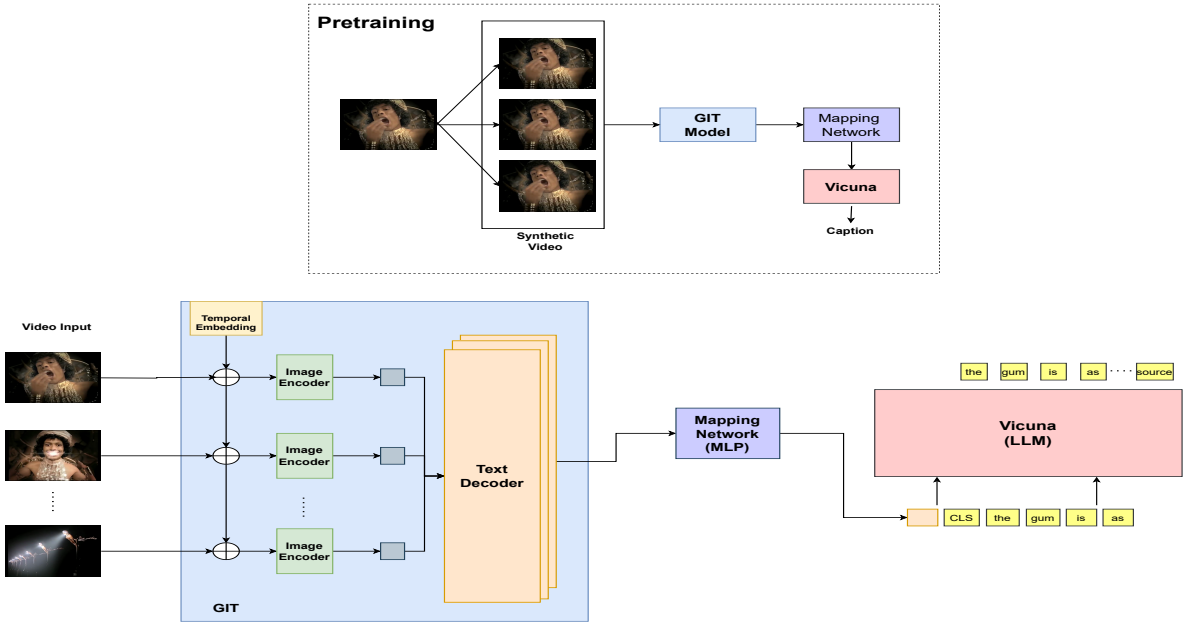

Figure 3: An overview of our Video Metaphor Captioning system, GIT-LLaVA. The text encoder representation of GIT is mapped to the embedding space of Vicuna to generate metaphor captions.

captioning pipeline, we map the text decoder output of the video captioning model to the embedding space of the LLMs. Thus, to train the mapping network it is sufficient if the video captioning model (GIT) can generate a valid caption and a ground truth metaphor caption is present, such that the mapping network can learn the transformation. We simulate this process by feeding images to the GIT model and training it with synthetically generated metaphor captions.

We use images and captions from the popular 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 which were used to pretrain the model.

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.

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 to map the video representation(H_R) to the embedding space of the LLM, similar to the LLaVA model (Liu et al., 2023). We also use task-specific instruction (X_{inst}) as input and the model is trained to generate the answer as output (X_{ans}).

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

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

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

where 'W' denotes the weights of the MLP network and θ represents the parameters of the LLM, X_i denotes the current token predicted. The LLM is trained with this language modeling objective.

We use the LLaVA-13B-V1.5 (Liu et al., 2023) model 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 of output states of GIT 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.

5 Experiments

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 fine-tuning 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 Vicunna-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.

5.2 Video Metaphor Captioning

The model is fine-tuned for video metaphor captioning on our manually annotated dataset. The model is fine-tuned for 5 epochs with early stopping based on the validation set.

Frame Selection:
We explore two frame selection strategies for our model. In our analysis of the dataset, it was found that video advertisements typically consist of a three-act structure like movies. The first act introduces either the primary or secondary concept, the second act discusses the properties and the third act reveals the metaphor. Hence, we split the video into three equal parts and sampled an equal number of frames from each part.

The GIT-Large model only supports video captioning with 6 frames as input. We experiment with sampling 2 frames in temporal order across the three parts. We also perform additional experiments where 6 frames are sampled from each part, which we call GIT-LLaVA-Extended. The video representation is obtained by considering each part as a video and the final representation is obtained by summing up the representations for each video part. This leads to better metaphor generation as the model can access more frames in the video.

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. We use BFloat16 precision while training the model on 4 A100 GPUs.

5.3 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.

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 8 for 50 epochs.

Video-LLaMA: We use the 13B pretrained model of video-LLaMA that is pretrained on parallel video-text data. We then finetune the vision branch of the model on our VMC dataset.

Valley: Valley is a video-assistant build on top of the LLaVA model. We use the 13B pretrained model of valley and fine-tune it on our VMC

Model	BLEU-1 ↑	Rouge-L ↑	CIDEr ↑	BERT-F1 ↑	ACS ↓
GIT	38.1847	39.9777	32.0064	0.6434	0.3934
Valley	17.6786	18.7736	2.7567	0.5477	0.7910
Video-LLaMA	35.9410	37.1696	**47.6783**	0.5005	0.3130
GIT-LLaVA (Ours)	**42.6690**	**42.7680**	40.9205	0.6534	0.3015
GIT-LLaVA-Extended (Ours)	40.2760	41.9725	26.9294	**0.6542**	**0.2728**

Table 2: Experimental results on our VMC dataset in comparison to other models. ACS denotes the Average Concept Similarity. It represents the average cosine similarity of the concepts compared in the metaphor caption

Model	Fluency ↑	Consistency ↑	Creativity ↑
GIT	0.1142	0.0000	0.2714
Valley	-0.1285	-0.4428	-0.7000
Video-LLaMA	-0.8142	-0.1285	-0.1428
GIT-LLaVA (Ours)	0.3000	0.2000	0.2714
GIT-LLaVA-Extended (Ours)	**0.5285**	**0.3714**	**0.3000**

Table 3: Results of human evaluation of the captions generated by models. Consistency denotes the consistency of the caption with the video. Creativity denotes the quality of the metaphor generated.

dataset by converting it to the data format of valley.

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. In the case of video metaphor captioning, the exact matching of n-grams may not give a clear idea of the performance of the model as it is difficult to generate the exact metaphor in the reference sentences. 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. As no existing metric can be used to evaluate the creativity of metaphors, we introduce a new and intuitive metric called- "Average Concept Similarity" (ACS). It is calculated as follows:

$$\text{ACS} = \frac{\sum_i^n \text{Cosine(PC, SC)}}{n} \quad (5)$$

where PC and SC denote the primary and secondary concepts 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. Sentence Transformers (Reimers and Gurevych, 2019) are used to obtain representations for PC and SC. For captions which do not contain either of 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.

In addition to these automated metrics, we also evaluate and compare the models based on three scores manually given by a set of annotators. We use three metrics for human evaluation- Fluency, Consistency, and Creativity. Fluency denotes how fluent the generated caption is. Consistency denotes the consistency of the generated caption with the video and creativity denotes the quality of metaphor.

7 Results and Analysis

Our models- GIT-LLaVA and GIT-LLava-Extended perform significantly better than other traditional video captioning models despite the smaller scale of pretraining data. Table 2 compares the performance of our models with other baselines. It can be seen that the model performs well on both n-gram overlap-based metrics like BLEU-1, ROUGE-L, and CIDEr and the BERTScore metric. This shows that it generates captions that are semantically similar to the ground truth captions.

Our model achieves the best score (lowest) on our new metric- ACS. It compares the semantic similarity of the primary and secondary concepts used in the metaphor generated. The lower scores confirm that our models generate creative captions in which the comparisons are made to very creative

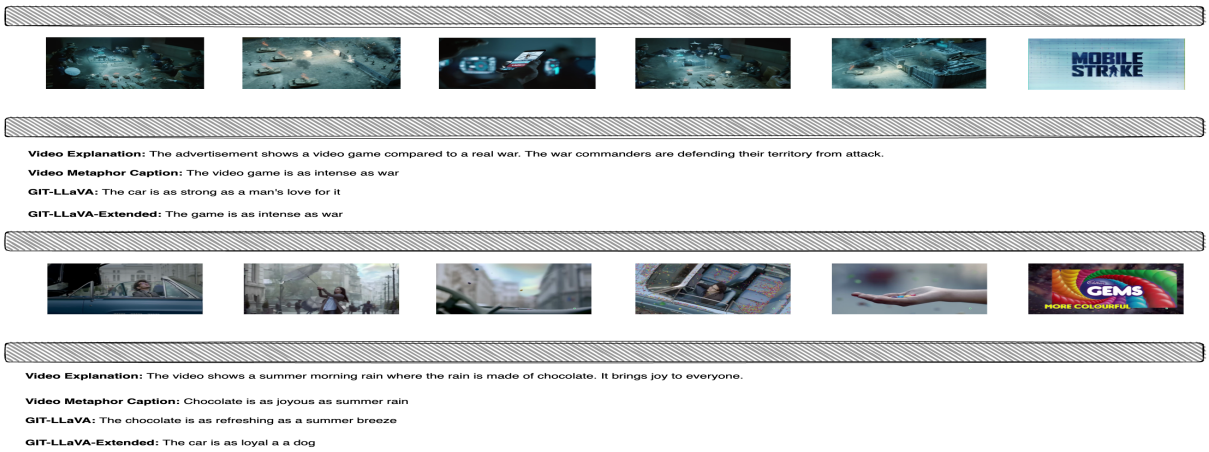

Figure 4: Examples from our manually annotated dataset along with captions predicted by our models.

concepts that are not related to the primary concept. 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 Video-LLaMA model. This is indicated by lower BERTScore and higher ACS values in conjunction. Our models have higher BERT-score as well as lower ACS which indicate that the models generated metaphors that are more relevant to the videos.

It was observed that the Valley model wasn't able to generate quality metaphors even though it was able to generate fluent captions. This is indicated by the poor performance on the ACS metric. The GIT model scored very highly on BERT-F1 but its score was relatively lower on the ACS metric. This shows that our model was able to augment the GIT model to enable it to generate metaphors.

Figure 4 shows examples from our dataset with captions generated by our models. In the first example, it can be seen that the extended model with access to many frames was able to understand that the video was about a game. The GIT-LLaVA model generated a metaphor that focuses on the car used in the video while missing the bigger picture. In the second example, the GIT-LLaVA model describes the breeze seen in the video in the metaphor generated. The extended model is confused by cars appearing in multiple frames leading to describing the car in the metaphor generated. The dataset used in our experiments is small and these problems can be mitigated by training our model on a larger dataset. It was also observed that few captions were repeated in multiple occurrences when the primary concept in the selected frames was similar.

7.1 Human Evaluation

In addition to automated metrics, we also perform human evaluation on 15% of the test set. Table 3 shows the results obtained with human evaluation. We use Best-Worst Scaling (Louviere et al., 2015) to compare models. Four Masters' students who are proficient in English were asked to annotate the captions generated by these five models on three metrics- Fluency, Consistency, and Creativity. The annotators assigned +1 for the best caption, -1 for the worst caption, and 0 for the remaining captions. The mean scores from all annotators are reported in Table 3. The manual evaluation also confirms that our models generate creative captions that are consistent with videos.

8 Summary, Conclusion, and Future Work

In this work, we proposed a novel Vision-Language (VL) task called video metaphor captioning. We constructed and released two new datasets for the task. We proposed a novel VL model that is built on top of the LLaVA model for video metaphor captioning. We showed that by using a frozen video captioning model (GIT) and a lightweight mapping network with LLM, we were able to augment the video captioning model to describe metaphors in the video. We believe that this approach can be extended to different domain-specific tasks with inadequate video data. Our models generated fluent and creative metaphors and it was validated by automatic and human evaluations.

In the future, we plan to adopt stronger models that can also handle audio modality in our video metaphor captioning task.

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-Extended do not have support for audio and cannot understand metaphors introduced through audio. The audio signals can be used to better understand metaphor information 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. 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 with necessary caution in downstream tasks for the same reason.

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 pre-training: 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.

Jordan J. Louviere, Terry Flynn, and Anthony A. J. Marley. 2015. Best-worst scaling: Theory, methods and applications.

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.