# Improving Automatic Video Captioning

PRANAV PRABHU, Virginia Tech, USA

**Abstract**: The video captioning task for machine learning models entails the generation of textual descriptions for a given video. The following paper summarizes the results of our literature review of the video captioning task, including an in-depth description of the problem and its corresponding subtasks, and the problem's significance. This is followed by a detailed overview of three state-of-the-art video captioning models we have identified with increasing complexity – MaMMUT, VALOR, and mPLUG-2. We have considered these models as a baseline. Some simpler models have been executed as experiments, with and without adversarial examples, the details of which are outlined as a qualitative and quantitative analysis, analyzing their strengths and where they can still be improved.

**ACM Reference Format:**
Pranav Prabhu. 2018. Improving Automatic Video Captioning. *Proc. ACM Meas. Anal. Comput. Syst.* 37, 4, Article 111 (August 2018), 10 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Video captioning describes the autonomous generation of textual descriptions for a video. Captions can come in a variety of forms, such as subtitles which explicitly describe what is occurring during a scene, or broader descriptions which summarize an entire segment of the video. One approach to this problem is to use the entire video segment as the input for a machine learning model which will then generate the caption as the corresponding output. Alternatively, the problem can be broken down into subtasks utilizing the audio and visual components of the video separately to generate outputs that can be combined into a final caption. For example, frames from the video can be used as inputs to image captioning and segmentation tasks to identify key components in the video. On the other hand, the audio can be used alongside natural language processing models to use dialogue to enhance video captioning. The current state-of-the-art models often incorporate each of these into a multi-modal approach to generate the most accurate video captions possible.

Accurate video captioning techniques have significant ramifications for how video media is consumed. The user experience of watching a video is improved greatly by accurate and context-aware captions, as it allows for a much more enjoyable and informative viewing, especially in sound sensitive environments. This is particularly crucial to those with hearing impairments, as automatic video captioning promotes equal access to media consumption. Additionally, this level of accessibility is sometimes enforced legally such as through the Americans with Disabilities Act in the United States, and thus automatic video captioning can help creators and distributors with compliance without the need for a dedicated team for caption generation. Furthermore, the inclusivity of video captioning can be applied to non-native speakers, who may rely on captioning to assist in language learning, thus making educational content more accessible. Lastly, captions can play a key role in content indexing, and accurate captioning can assist users in searching for relevant information within a video. The overall improvement to the user experience that video captions provide demonstrates the overwhelming need for machine learning models that can accurately generate these captions.

Author's address: Pranav Prabhu, Virginia Tech, Blacksburg, USA, ppranav02@vt.edu.

ACM 2476-1249/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

## 2 BACKGROUND

The field of video captioning has witnessed a lot of significant advancements in recent years due to models adapting different sophisticated approaches to the multimodal nature of videos. The first model we looked at was the MaMMUT model. This model uses a two-pass approach in the first pass, it focuses on contrastive tasks by using an image encoder and then in the second pass, it generates text (e.g., captions) based on the learned representations using a language decoder. Using the two-pass system the MaMMUT model is able to map regions in images and video frames to corresponding words in captions [8].

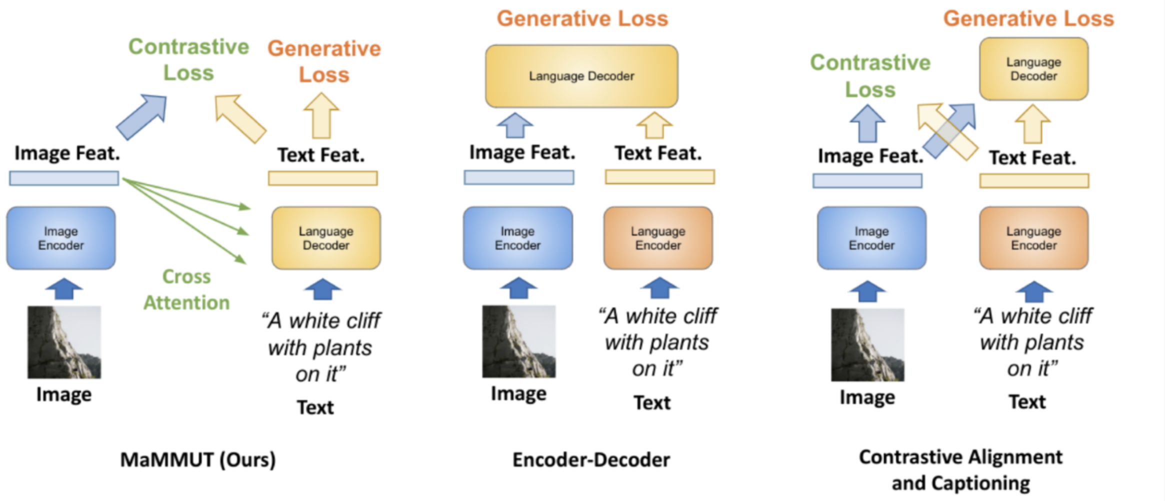

Fig. 1. MaMMUT Model

The second model we looked at was the VALOR model. Similarly to the MaMMUT model the VALOR model uses an image encoder and language decoder but it adds to the MaMMUT architecture by also looking at audio. VALOR jointly models' information from vision, audio, and language modalities. It leverages features from all three domains to enhance multi-modal understanding. The fusion mechanism allows VALOR to capture cross-modal interactions effectively. By creating a shared embedding space, VALOR aims to bridge the semantic gap between video clips and their corresponding textual descriptions [2]. A problem with the VALOR model is that its reliance on contrastive learning may limit its ability to capture subtle nuances in video-text relationships, potentially hindering its performance in scenarios with complex multimodal interactions.

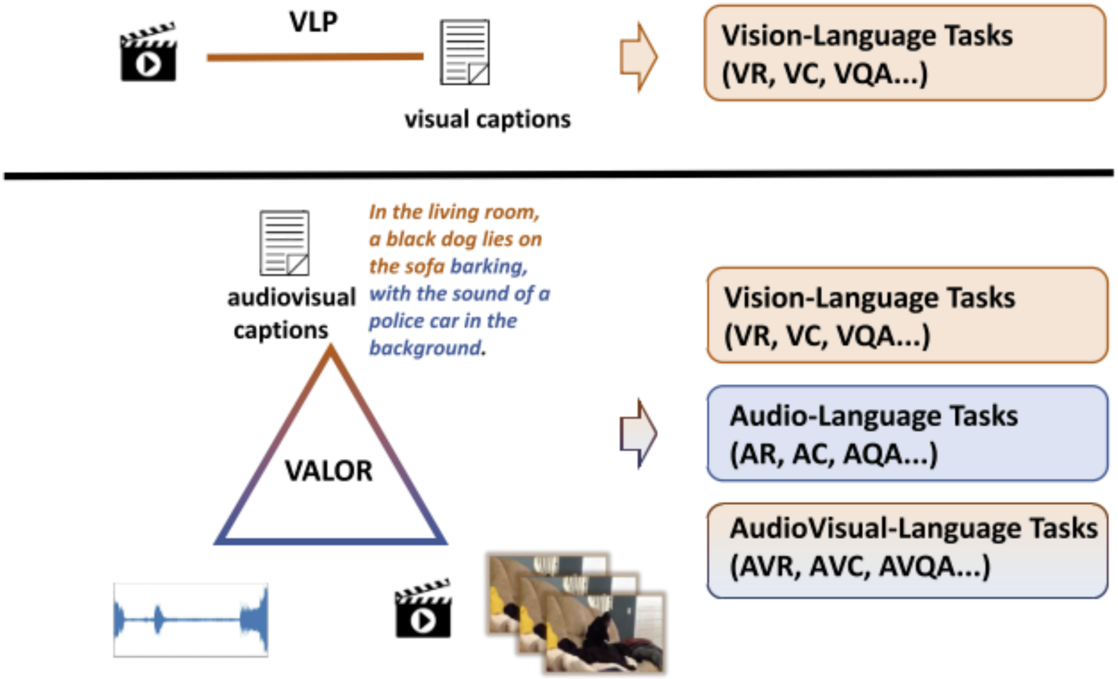
VLP
visual captions
Vision-Language Tasks (VR, VC, VQA...)
audiovisual captions
In the living room, a black dog lies on the sofa barking, with the sound of a police car in the background.
VALOR
Vision-Language Tasks (VR, VC, VQA...)
Audio-Language Tasks (AR, AC, AQA...)
AudioVisual-Language Tasks (AVR, AVC, AVQA...)

Fig. 2. VALOR Model

The last model we looked at was the mPLUG-2 model which adds another layer of complexity by making it able to perform all of the following tasks Text Classification, Image Captioning, Image Classification, and Video Question Answering either in conjunction with one another or separately depending on the users' needs. By leveraging this modular design, mPLUG-2 aims to benefit from modality collaboration while mitigating challenges posed by diverse modalities [1]. The mPLUG-2 model is the most advanced model we looked at, as it combines so many utilities into one tool.

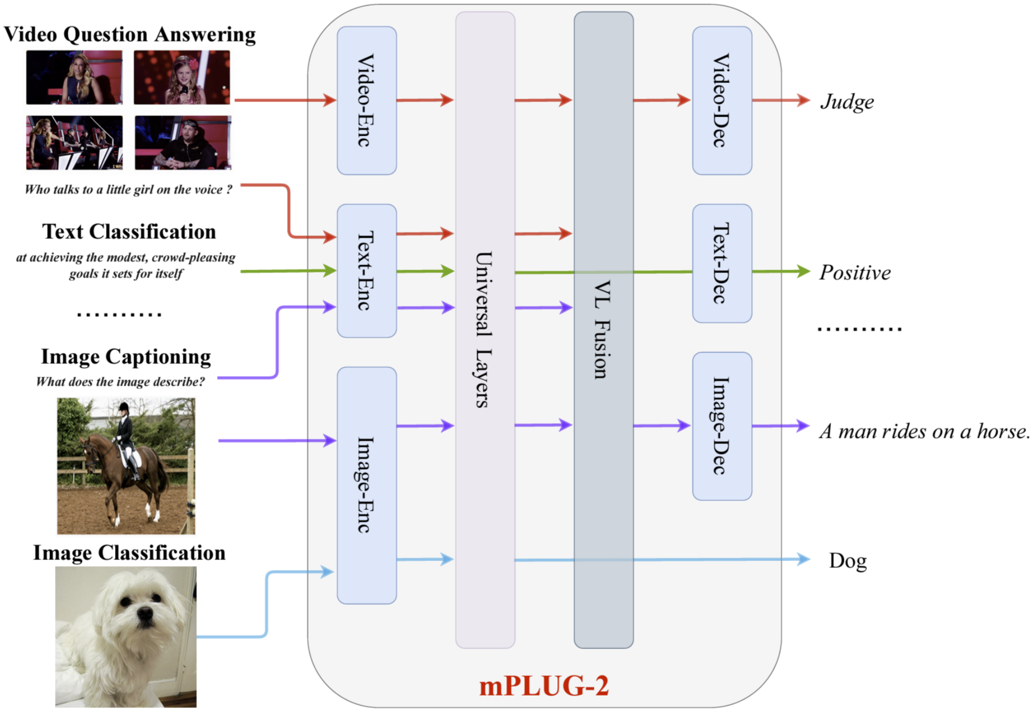

Fig. 3. mPLUG-2 Model

# 3 PROPOSED METHOD

## 3.1 Machine Learning Methodologies

For our project, we would like to run multiple existing machine learning models that take videos and caption them. Originially, we had planned to run MaMMUT, VALOR, and mPLUG-2 but these models proved to be too intensive on compute resources so we have opted to run simpler models and compare the results of the models. The two models are a Convolutional Neural Network with Long Short-Term Memory, or CNN-LSTM, and the LLaVA model.

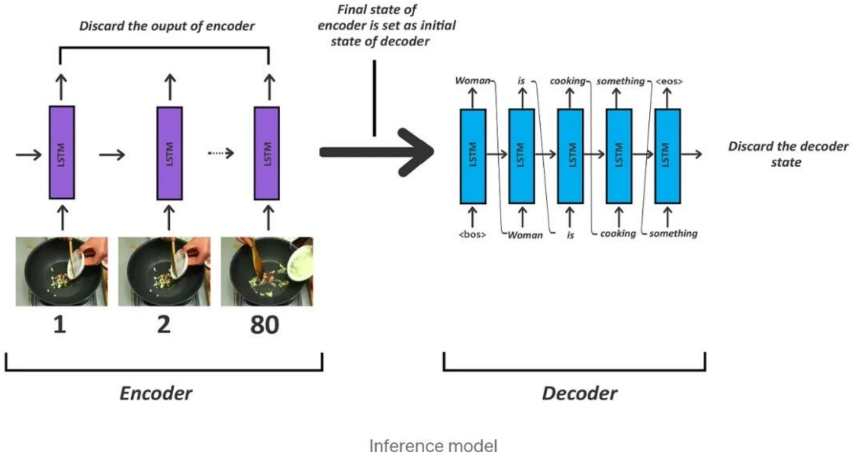
Discard the ouput of encoder
Final state of encoder is set as initial state of decoder
LSTM
LSTM
LSTM
1
2
80
Woman
is
cooking
something
<eos>
Discard the decoder state
<bos>
Woman
is
cooking
something
Encoder
Decoder
Inference model

Fig. 4. CNN-LSTM Model

The CNN-LSTM represents the most naive approach to the video captioning problem. It contains a single encoder phase which outputs a set of features extracted from eighty frames of the video. Afterward, a decoder phase will use this and an initial sentence token to begin generating a sentence, with each token generated being used as the input for the next iteration until a full sentence is generated. The CNN-LSTM used had already been pretrained on MSVD, with notable hyperparameters being 150 epochs, a learning rate of .0007, and 4096 features extracted per frame.

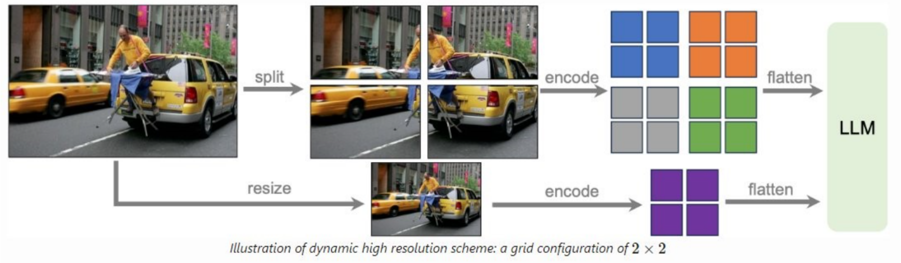
split
encode
flatten
LLM
resize
encode
flatten
Illustration of dynamic high resolution scheme: a grid configuration of 2 × 2

Fig. 5. LLaVA Model

LLaVA expands upon the CNN-LSTM concept by introducing complexity through the splitting of the original frame into four additional sub-frames. These five images are then encoded into a set of features which are flattened and used as input for an LLM which will generate the caption. LLaVA was pretrained and finetuned on MSVD, with notable hyperparameters being only one epoch, a learning rate of .001 for pretraining and .00002 for finetuning, and 2048 features extracted per frame. Due to issues with running the model locally, the testing videos were submitted to an online interface published by LLaVA's team to generate the captions for this experiment. However, the interface is QA based, which would slightly impact the results of the experiment as detailed in the analysis.

While these two models were not the ones we had originally planned on using, their increasing levels of complexity enabled us to continue to explore the trend of model complexity vs. model accuracy that was present in the original three models.

## 3.2 Technology

To run these models, we will find ones that are trained on the same dataset as the three models detailed above for consistency. This leaves us with multiple choices, given that the three models have been trained and tested using a variety of datasets, so we have chosen MSVD. MSVD, or the Microsoft Research Video Description dataset, is a dataset containing 1550 videos (although some versions of this dataset are extended to 1970 or even over 2000 videos), with a set of reference captions for each. The videos are short clips obtained from YouTube, all hovering around ten seconds in length. For each video, at least ten reference captions were created by human observers who were told to summarize the video in a single sentence [12]. In our experiment, we utilized the 100 videos commonly used a testing split to generate a set of captions to be compared to the MSVD reference captions. While the two models we will be running our experiment on have solely been pre-trained and fine-tuned on this dataset, VALOR and mPLUG-2 have been pre-trained with other data sets as well.

## 3.3 Metrics and Analysis

The system's performance will primarily be gauged using standard evaluation metrics for captioning, such as BLEU, ROUGE-L, and METEOR scores. Bilingual Evaluation Understudy, or BLEU, scores consider matching words, caption length, and precision [11]. BLEU-4, for example, considers the number of matching 4-word sequences within two pieces of text [11]. Recall-Oriented Understudy for Gisting Evaluation, Longest Common Subsequence, or ROUGE-L considers the longest common subsequence of words in the reference and predicted text [11]. Metric for Evaluation of Translation with Explicit Ordering, METEOR, is similar to the previous metrics, but is more lenient, taking into account synonyms and overall sentence structure [11]. These metrics will compare the generated captions with a set of reference captions to evaluate linguistic quality, including aspects of precision, recall, and semantic similarity.

In addition to these quantitative measures, human inspection of the generated captions will play a crucial role in capturing subtleties that automated metrics may overlook. Human evaluators will assess the captions for relevance, coherence, and readability to ensure that the captions not only are technically accurate but also make sense in a real-world context.

By employing these metrics and methods, we can thoroughly evaluate the automatic video captioning system, balancing the need for technical precision with the nuances of human language and interaction.

We planned to analyze the differences in metrics and consider why one model may have had better quality captions than another and in what ways. This way, we can thoroughly categorize the strengths and weaknesses of the various models we are testing.

## 3.4 Adversarial Examples

In addition to the original experiment, we ran experiments with adversarial examples. To create adversarial examples, we utilized existing videos within the MSVD dataset and removed random frames. Then, we generated video captions for these modified videos and compared them to the original captions to see the difference and whether the captions were still reasonable. Since we are using simpler models, we expected the captions to make less sense for the simpler models. However, since LLaVA is more robust, we expected captions would still be of good quality. The metrics from this experiment helped us evaluate the resiliency of the different models I ran.

## 4 RESULTS

The baseline LSTM-CNN [9] model was run on the testing split of the MSVD data set, a portion of the data set containing 100 videos the LSTM had not seen during training. For each of these videos, a caption was generated, allowing us to both qualitatively, as well as generate BLEU, ROUGE, and METEOR scores to compare with published results from the papers of other models.

The baseline LSTM-CNN model was run on the testing split of the MSVD data set, a portion of the data set containing 100 videos the LSTM had not seen during training. For each of these videos, a caption was generated, allowing us to both qualitatively, as well as generate BLEU, ROUGE, and METEOR scores to compare with published results from the papers of other models. The following table shows our resulting metrics with 'A' noting results from when the models were run with adversarial examples. We included VALOR and mPLUG-2 metrics for reference. We were unable to find published MaMMUT values for these metrics.

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | METEOR |
|---|---|---|---|---|---|---|
| CNN | 75.09 | 35.56 | 19.58 | 6.47 | 30.31 | 48.03 |
| CNN-A | 74.60 | 33.05 | 15.98 | 3.80 | 30.23 | 46.19 |
| LLaVA | 55.93 | 25.74 | 11.02 | 5.07 | 24.26 | 60.01 |
| LLaVA-A | 57.17 | 26.52 | 11.16 | 4.27 | 23.94 | 60.29 |
| VALOR | X | X | X | 80.57 | 68.0 | 48.00 |
| mPLUG-2 | X | X | X | 70.50 | 85.30 | 48.40 |

We split our analysis along the first and second half of the dataset, as listed in our llava.txt captions file.

### 4.1 Quantitative and Qualitative Analysis for First 50 Video Subset

*4.1.1 Quantitative Analysis.* The quantitative analysis of the video captioning models reveals varied performances across the six models evaluated using the MSVD test dataset of 100 videos. The evaluation metrics include BLEU-1 through BLEU-4, ROUGE-L, and METEOR, which together provide a comprehensive view of each model's capabilities.

The CNN model exhibits strong performance in BLEU-1 with a score of 75.09, indicating high precision in capturing the most common unigrams in captions. However, there's a significant drop in scores as we move to BLEU-4, which suggests that the model struggles with longer sequence matches. Its METEOR score of 48.03, though not the highest, points to a reasonable semantic understanding.

Comparatively, the adversarial version of the CNN model, CNN-A, shows slightly reduced BLEU scores across the board and a marginally lower METEOR score. This minor reduction indicates that the adversarial approach maintains performance while possibly providing robustness against overfitting to the dataset.

LLaVA models have lower BLEU scores, with the highest being BLEU-1 at 55.93, which suggests that it may not capture unigrams as effectively as CNN but still performs reasonably well. Notably, LLaVA models excel in METEOR scoring, with 60.01 and 60.29 for the standard and adversarial versions, respectively. This higher METEOR score indicates a superior capacity to form grammatically and semantically coherent sentences, even though they may not align perfectly with reference captions.

On the other hand, VALOR and mPLUG-2 models, without BLEU-1, BLEU-2, and BLEU-3 scores available, show remarkable BLEU-4 scores at 80.57 and 70.5, respectively, suggesting their advanced ability in capturing longer sequences of text accurately. Their ROUGE-L scores follow suit, particularly for mPLUG-2, which scores an impressive 85.3, highlighting its capability in forming longer, coherent sequences in captions.

*4.1.2 Qualitative Analysis.* The qualitative analysis assesses how well the models generate captions that are coherent, complete, and contextually accurate in describing the video content. LLaVA models, both standard and adversarial, produce captions that, while not always perfectly aligned with the original MSVD dataset captions, convey clear and complete sentences with proper context. For example, the video with the ID ScdUht-pM6s_53_63 reveals LLaVA's strength in generating detailed descriptions that accurately reflect the video content, even though the actions described ("washing dishes" and "peeling potatoes") differ from the original caption ("putting salt on a chicken").

In contrast, the CNN models, while generating accurate unigram content, often fall short in creating complete and contextually rich sentences. The CNN model's caption for the same video ("A man is mixing a") and its adversarial version ("A man is a a on a ") both lack the coherence and detail seen in the LLaVA-generated captions.

The ability of LLaVA to generate captions that are contextually richer and more detailed than those of CNN indicates its potential for real-world applications where comprehensive understanding is crucial. However, the discrepancy between LLaVA's qualitative performance and its lower quantitative scores suggests that there may be a gap between metric evaluations and real-world utility that future research needs to address.

The absence of lower-order BLEU scores for VALOR and mPLUG-2 complicates direct comparison but their high BLEU-4 and ROUGE-L scores suggest these models excel in generating long and coherent sequences that likely match well with human-generated captions. This assumption is supported by their decent METEOR scores, indicating good semantic understanding.

## 4.2 Quantitative and Qualitative Analysis for Second 50 Video Subset

*4.2.1 Quantitative Analysis.* There are a few key trends to note amongst the metrics generated for each of the captions and the metrics published in the papers for VALOR and mPLUG-2.

First, both the CNN-LSTM and LLaVA show little to no impact when provided with adversarial data when compared to the original testing data. For example, the ROUGE-L for the CNN drops from 30.31 to 30.23, a -0.26% decrease, while the score for LLaVA drops from 24.26 to 23.94, a -1.32% decrease. Nearly all metrics face at most a 2-point difference, signifying robustness amongst the models against adversarial data.

Next, we notice the CNN outperforms LLaVA in most of the metrics. The CNN begins with a 75.09 BLEU score for 1-grams, or matching 1-word sequences, which drops to 6.47 for 4-grams. This indicates that captions generated by the CNN often feature many of the words present in the reference captions but fail to string them together into meaningful sequences. LLaVA features a similar trend of decreasing BLEU scores for increasing grams, starting at a BLEU-1 of 55.93 which decreases to a BLEU-4 of 4.27. Similarly, LLaVA appears to fail to generate matching multiple word sequences, but the lower BLEU-1 score as well indicates that this may be the result of simply failing to have as many matching words as the reference caption in the first place. As expected, this is also reflected in the ROUGE-L scores, in which the CNN once again outperforms LLaVA with a score of 30.31 vs. LLaVA's 24.26. This indicates that CNN is capable of generating longer matching sequences of words compared to LLaVA. Both the CNN and LLaVA however have BLEU-4 and ROUGE-L scores far lower than the published results for VALOR and mPLUG-2, which have BLEU-4 scores of 80.57 and 70.5 and ROUGE-L scores of 68.0 and 85.3 respectively. This illustrates just how powerful the multimodality approach can be as both VALOR and mPLUG-2 can generate captions with much higher similarity to the reference captions in terms of matching sequences of words.

One thing to note is that these metrics may not be an ideal quantifier for performance given the initial experimental setup. As mentioned, captions for LLaVA were generated using an online QA-based UI, which means the captions were also influenced by the prompt provided to the model, "Generate a brief single sentence caption for this video." As such, the captions for LLaVA may not have maintained the same diction and syntax as the reference captions for the videos, resulting in lower scores despite being semantically similar. This is supported by the METEOR score which evaluates quality based on sentence structure, considering synonyms and stemming to allow for semantically similar captions. LLaVA has a METEOR score of 60.01, while the other models all feature scores around 48, signifying that LLaVA produces better quality captions that are semantically similar to the references. Overall, the disparity between different metrics demonstrates the difficulty in determining the "best" model, although it does appear that increasing model complexity including multi-modality understandably leads to better results.

*4.2.2 Qualitative Analysis.* For simplicity, we chose a representative quote from MSVD when using examples in our analysis.

The CNN-LSTM often captioned an incomplete sentence with poor likelihood of identifying key actions. These incomplete sentences were present in 60% of the captions in this set of 50 captions. However, the key subject was generally referenced. When compared to MSVD captions, the quality of captions from the CNN-LSTM were much lower. For video s1ZABV7AQdA_38_48, MSVD captioned the video as "a crowd of people are chasing a man." For the same video, CNN-LSTM captioned it "a woman is doing on a" and when given an adversarial version of that video captioned it "a woman is making a." As shown here, little difference was found between the quality of captions for the CNN-LSTM when given adversarial examples. This shows the model is resistant to noisy examples.

The LLaVA model output complete sentences for key captions but occasionally misidentified key actions or subjects. For instance, for video u4T76jsPin0_0_11, MSVD captions the video as "a group of men are racing down a track." For the same video, LLaVA captions it as "A group of people run down a track and jump into a pit" and when given the adversarial version LLaVA captions it "A group of people run down a track and jump into a pit." In this scenario, additional details are identified that are not present in the MSVD caption. As shown here, little difference was found between the quality of captions for LLaVA when given adversarial examples. This shows the model is resistant to noisy examples.

When comparing the CNN-LSTM and LLaVA to each other, we notice a clear difference in the quality of captions. Although not consistently accurate, LLaVA's captions are more robust by being longer and more often correctly identifying key attributes of the video while the CNN-LSTM tends to make simple captions that are inaccurate or incomplete. For video UbmZAe5u5FI_132_141, MSVD captions the video as "a woman is removing bones from a fish." For the same video, the CNN-LSTM captions it "a person is being chopped" and LLaVA captions it "A person is seen cutting a fish on a plate and then cutting the inside of the fish." The CNN-LSTM's caption is short and inaccurate. LLaVA's caption is longer and more accurate but is unable to simplify the actions into the concept of deboning a fish. Overall, LLaVA is considerably better at the automatic video captioning task than the CNN-LSTM.

## 5 CONCLUSION

In conclusion, while quantitative metrics are essential for benchmarking, qualitative evaluations provide crucial insights into a model's practical capabilities. Adversarial examples did not provide significantly different metrics. Higher complexity may contribute to higher quality captions as was shown between LLaVA and the CNN-LSTM.

## 6 FUTURE WORK

The future work in this domain will involve not only refining these models to bridge the gap between metric performance and real-world utility but also developing new metrics that can better capture the nuances of human-like caption generation. Additionally, exploring the robustness of state-of-the-art models we mentioned previously, such as MaMMUT, VALOR, and mPLUG-2, with adversarial examples and further testing with more data could help us get a better understanding of the intricacies of these models to better evaluate them. Prompt-based models may be explored for video captioning further through experimental prompt engineering to see if asking different questions results in higher quality captions. Lastly, one could develop methods specifically targeting object recognition and detail inclusion to make existing models more robust. Ideally, these steps will give us a better understanding of the video captioning models that exist so they can be improved.

## REFERENCES

[1] "Papers with Code - mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video," paperswithcode.com. https://paperswithcode.com/paper/mplug-2-a-modularized-multi-modal-foundation.

[2] "Papers with Code - VALOR: Vision-Audio-Language Omni-Perception Pretraining Model and Dataset," paperswithcode.com. https://paperswithcode.com/paper/valor-vision-audio-language-omni-perception.

[3] M. Abdar et al., "A Review of Deep Learning for Video Captioning." [Online]. Available: https://arxiv.org/pdf/2304.11431.pdf

[4] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Wang, "Video Captioning via Hierarchical Reinforcement Learning." [Online]. Available:
https://openaccess.thecvf.com/content_cvpr_2018/papers/Wang_Video_Captioning_via_CVPR_2018_paper.pdf

[5] J. Wang et al., "GIT: A Generative Image-to-text Transformer for Vision and Language."
Available: https://arxiv.org/pdf/2205.14100v5.pdf

[6] "Papers with Code - EnCLAP: Combining Neural Audio Codec and Audio-Text Joint Embedding for Automated Audio Captioning," paperswithcode.com. https://paperswithcode.com/paper/enclap-combining-neural-audio-codec-and-audio.

[7] W. Kuo et al., "MaMMUT: A Simple Architecture for Joint Learning for MultiModal Tasks." Accessed: Feb. 20, 2024. [Online]. Available: https://arxiv.org/pdf/2303.16839v3.pdf

[8] Shreya, "Shreyz-max/Video-Captioning," GitHub, Apr. 01, 2024. https://github.com/Shreyz-max/Video-Captioning.

[9] vsubhashini, "vsubhashini/caption-eval," GitHub, May 28, 2023. https://github.com/vsubhashini/caption-eval.

[10] D. Raj, "Metrics for NLG evaluation," Explorations in Language and Learning, Sep. 16, 2017. https://medium.com/explorations-in-language-and-learning/metrics-for-nlg-evaluation-c89b6a781054.

[11] "MSVD Dataset Corpus," www.kaggle.com. https://www.kaggle.com/datasets/vtrnanh/msvd-dataset-corpus.