Diving Deep With Video-Text Models in Representing Motion

Anonymous ACL submission

Abstract

Videos are more informative than images because they capture the dynamic activities in the scene. By representing motion in the videos, we can represent the dynamic activities in the scene. In this work, we create a benchmark. to investigate how existing video-text models respond to different motion descriptions that describe the activity. We evaluate several video-text models on the task of retrieval of motion descriptions, and the results fall far behind human expert performance. We also propose a method of improving motion understanding in video-text models. Our proposed method consistently surpasses other previous methods over two benchmark datasets.

1 Introduction

Since the large-scale use of contrastive learning to represent images and text (Radford et al., 2021), various efforts have been made to build video-text models (Ni et al., 2022; Luo et al., 2022; Fang et al., 2021; Wang et al., 2021) to relate video to text. Videos provide a way to access the dynamic things or motion in the scene that a single image cannot capture. Motion is probably the most crucial component in videos that is not present in images. That is why it is important to model motion in the videos. Motion in the videos could be due to the action involved, or it could be the effect of camera movement like in the egocentric video, or it could be a combination of both camera motion and action. While there is no single definition of motion in actions, we can define motion in general as the movement of actor or movement of actor and object in the world, and the goal of the action usually plays a pivotal role.

Unlike image-text models, video-text models are more challenging to design because of the additional temporal dimension. In this work, we study how existing video-text models perceive motion due to the action. The task for which we instigate video-text models is, given a video, we need to retrieve the correct description of motion in the video.

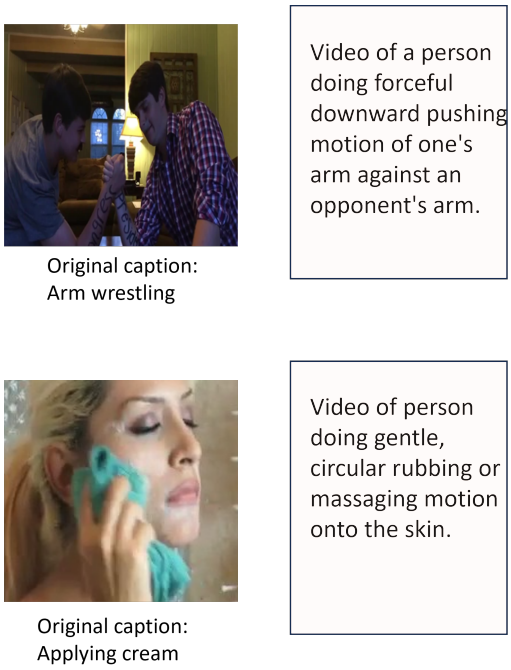

Figure 1: **Qualitative examples of the additional descriptions of verbs :** The left side shows the video clip and original caption. The right side box includes the video's motion description generated by GPT-4.

The challenge is the lack of datasets that explicitly describe the motion in a video. Howto100M dataset (Miech et al., 2019), Spoken Moments in Time (Monfort et al., 2021) are all popular video-text datasets used to train video-text models, but they don't explicitly give motion descriptions. Since motion is very much interlinked to actions, we use human action datasets like Kinetics-400 (Kay et al., 2017), UCF-101 (Soomro et al., 2012), and HMDB-51 (Kuehne et al., 2011) to circumvent the lack of quality annotations of motion descriptions. The advantage of action datasets is that for every action label, we can obtain the correspond-

ing characteristic motion of the action by using large language models like the GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), which, to a large degree, produce accurate descriptions of motion for each action. In this work, we use GPT4 to obtain motion descriptions of all the actions in the datasets. Figure 1 shows some examples of the descriptions generated by GPT4 and corresponding videos and original captions.

We evaluate several video-text models on the motion description retrieval task on HMDB-51 and UCF-101 datasets. We also compare it against the human-estimated performance and show that all models fall far behind human performance, raising questions about the design of video-text models. We then propose a method to incorporate motion description in video-text models given the constraints of lack of detailed annotations of motion in videos. Our proposed approach incorporating motion description improves several other video-text models by over 10% accuracy on UCF-101 and 3% HMDB-51 dataset.

2 Benchmark

2.1 Designing motion description benchmark:

We perform experiments on Kinetics-400, UCF-101, and HMDB-51 datasets. For each class in action datasets, we prompt GPT-4 to produce the characteristic motion of it. We noticed from experiments that we needed to provide some example motion descriptions in the prompt to obtain motion descriptions in the format we are interested in. The prompt we used is "*Characteristic motion of air drumming action is: doing rhythmic, exaggerated hand and arm movement. Characteristic motion of abseiling is: doing controlled descent down a vertical drop. Characteristic motion of swing dancing is: doing energetic, bouncy movements with lots of spins, kicks, and lifts. Similarly, provide Characteristic motion of actionX.*".

We noticed many of the actions in UCF-101 and HMDB-51 datasets have objects involved, and the retrieval task is easier due to the presence of an object in the generated motion description. For this reason, we also created another set of motion descriptions where we replaced the names of objects in the motion description with the generic word "object," which made the task slightly challenging.

3 Proposed Method

In this section, we describe our proposed video-text model that incorporates motion description. A typical vision-language model like CLIP consists of a text and visual encoder. It is usually trained using contrastive learning loss, or its variations (Miech et al., 2020; Momeni et al., 2023a). Due to contrastive learning, representations from the same video caption pair are forced to match and pushed away from other pairs of videos and captions. The quality of representations learned from the video-text model (Momeni et al., 2023b) will depend on the captions used during pre-training. Since no large video-caption dataset containing motion descriptions exists, we can't directly train a video-text model using contrastive learning.

We propose a method that utilizes the rich linguistical understanding of motion in actions from GPT-4 and the generalizability of the CLIP model on new images and text descriptions. Our approach has two parts: Generating the required motion descriptions of actions using GPT-4 described in section 2 and training video-text models to utilize these motion descriptions described in section3.2.

3.1 Problem Setting:

The model is trained on kinetics-400 as source dataset D_s and tested on target datasets D_t : UCF101 and HMDB51. The source dataset D_s consists of videos x_s and labels L_s belonging to classes C_s. The target dataset D_t consists of videos x_t and labels L_t belonging to classes C_t . We use zero-shot setting such that $C_s \cap C_t = \emptyset$. Let the generated motion descriptions from GPT-4 for L_s and L_t be M_s and M_t respectively.

3.2 Architecture Setting:

Figure 2 shows the overview of our approach. We freeze the text encoder from CLIP and fine-tune the visual encoder to learn the motion representation.

3.2.1 Training:

Motion descriptions M_s are passed through frozen CLIP text encoder to obtain the class prototypes of L_s. Our intuition is that these class prototypes can be approximated as classifier weights of the supervised video classifier. Our intuition is from the work of (Nukrai et al., 2022; Liang et al., 2022), which shows that text embeddings from CLIP and vision embeddings from CLIP are very similar and fall within a ball of small radius. The obtained CLIP embeddings of motion descriptions from the

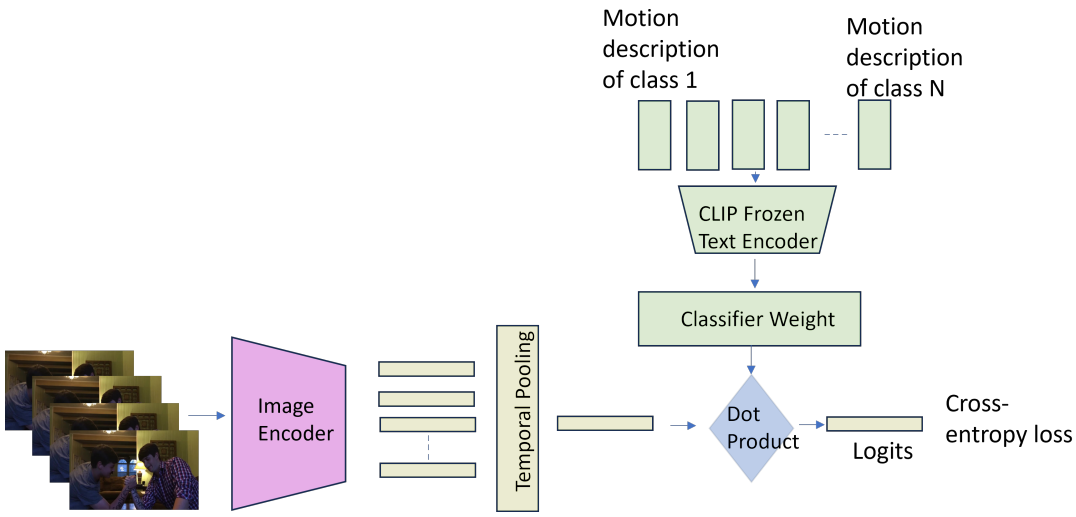

Figure 2: **Schematic representation of our approach encoding motion description in video-text model pipeline :** We integrate motion information as classifier weight in a supervised training paradigm. We finetune the image encoder to integrate the motion information while classifying videos in kinetics400 dataset.

frozen text encoder would approximately translate to visual class prototypes if obtained visually. With that intuition, we use class prototypes from the frozen CLIP text encoder as the classifier weight of a visual classifier.

Given a video v_s from source dataset D_s, T frames are sampled uniformly. The sampled frames are passed through a CLIP pre-trained Image encoder and temporally pooled to obtain a visual feature of the video. The video feature is then taken as a dot product with the transpose of classifier weights W_s to obtain the logits. The model is trained using cross-entropy loss over logits and labels L_s, and the parameters of the CLIP image encoder are updated.

3.2.2 Testing:

The motion descriptions M_t are passed through to obtain classifier weights of W_t. Given a video v_t from target dataset D_t, we obtain the logits indicating the probability that video matches the corresponding motion description M_t.

3.3 Implementation Details:

We use ViT B/16 based CLIP model with 8 frame samples per video. We train the model for 10 epochs on the Kinetics-400 dataset with a learning rate of 0.00005 with a batch of the size of 20 on 4 NVIDIA RTX A5000.

4 Results:

Task: For both the target datasets UCF-101 and HMDB-51, the input is a video and the list of generated motion descriptions for all the classes in the dataset. The video-text model predicts the closest motion description that describes the video. The metric used is the percentage accuracy of correctly predicting the motion description.

Method	Object	Masked Object
Vanilla CLIP	25.92	23.33
Text4Vis	51.23	33.80
Our Method	57.02	44.40
Human estimate	98	98

Table 1: Evaluation of percentage accuracy in motion description retrieval task on UCF-101 dataset.

The models we study are Vanilla CLIP (Radford et al., 2021), Text4Vis (Wu et al., 2023). Vanilla CLIP consists of pre-trained CLIP with temporal averaging of frames. Text4Vis are trained on the Kinetics-400 dataset. More details about the models are given in the Appendix.

Human estimated performance: We also sampled five videos randomly from each class in UCF-101 and HMDB-51. A human expert was asked to select the motion description that correctly describes the video from the list of descriptions generated by GPT4. Human experts are graduate students who have taken computer vision and NLP graduate courses volunteered for this study.

Table 1 reports the performance of various approaches on the UCF101 dataset. Our proposed method beats previous methods by over 6% for motion descriptions containing the names of objects

Method	Object	Masked Object
Vanilla CLIP	25.26	16.67
Text4Vis	34.12	24.93
Our Method	37.53	26.84
Human estimate	97.5	96

Table 2: Evaluation of percentage accuracy in motion description retrieval task on HMDB-51 dataset.

involved and by over 10% for motion descriptions where a generic word object replaces the object's name. We noticed that all the video-text models perform very poorly compared to human-estimated performance. We also see that video-text models have a strong bias toward nouns. When the name of the object involved is changed to a generic object, there is an average of a 10% drop in performance in all methods, indicating the strong bias video-text models have for objects. Table 2 reports the performance of various approaches on the HMDB-51 dataset. Similar trends are exhibited in the HMDB-51 dataset. ,

5 Related Work

Various efforts have been made to build video-text models (Luo et al., 2022; Fang et al., 2021) mainly for video-to-text retrieval tasks. These models have developed mechanisms to use CLIP and extend it to video frames. Video-text models (Wang et al., 2021; Ni et al., 2022; Wu et al., 2023; Rasheed et al., 2023; Momeni et al., 2023a) have also been used for general video recognition both in supervised and zero-shot action recognition setting. (Momeni et al., 2023a) introduced verb-focused contrastive training to improve the better verb reasoning by learning with hard negative verb examples. (Park et al., 2022) introduced contrast sets to identify pitfalls in video-text models, recommending the need for fine-grained action understanding to tackle hard negatives in contrast sets. We are the first ones to evaluate video-text models over motion descriptions. Our work is inspired by (Nukrai et al., 2022; Liang et al., 2022) that shows that CLIP text features and CLIP visual features can be roughly exchanged without much loss of performance. This enables us to use CLIP text features as classifier weights in video-supervised classification. We also highlight the limitations in obtaining quality annotations for motion. Our benchmark further suggests that there is evidence that many video-text models have a strong bias to the objects.

6 Conclusion

We introduce a benchmark to understand how motion is understood in video-text models. We highlight the limitations in obtaining quality annotations describing motion in video. We also show that the performance of video-text models for retrieving motion descriptions is poor compared to human expert performance. Our proposed method circumvents some of these issues and improves over other video-text models. We leave it to future work to build better models to capture motion and ignore the biases involved with the object or the scene while designing video-text models.

7 Limitations

We are using image-text data trained CLIP text encoder to represent motion description. This is not the best thing to do as in practice, CLIP text encoder would never have encountered the dynamics of videos while training. However, we hope the method works if we replace this CLIP text encoder with any other video-text model text encoder. Another limitation is that we trained on the Kinetics-400 dataset and tested on the UCF-101 and HMDB-51 datasets. As shown in the results section, the presence of an object or scene can often impact the performance during the retrieval task on these datasets. Furthermore, since we are fine-tuning image encoders on the source dataset, there exists a possibility of overfitting the source dataset, leading to poor transferability on another target dataset.

8 Ethical Considerations

We aim to highlight the neglected aspect of modeling motion in video-text models. We think incorporating motion descriptions and reducing the biases of video-text models to objects and scenes positively impacts the design of video-text models. However, there could be a potential risk introduced in our method, as we rely on GPT-4 to provide us with motion descriptions of actions.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

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

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023a. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15579–15591.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023b. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591.

Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. 2021. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected clip. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4055–4063.

OpenAI. 2023. Gpt-4 technical report.

Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. 2022. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Wenhao Wu, Zhun Sun, and Wanli Ouyang. 2023. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2847–2855.

A Implementation Details

Here, we describe more details about our method and baselines. We report the best results after running experiments on 5 runs.

A.1 Vanilla CLIP:

We use VIT-B/16 pre-trained CLIP text and image encoder and temporally average the frame outputs.

A.2 Text4Vision

We use VIT-B/16 base architecture with 8 frames per video. We use pre-trained weights from https://github.com/whwu95/Text4Vis/tree/main

A.3 Our Model

Our model uses VIT-B/16 pre-trained CLIP text and image encoder. We use eight frame samples per video. CLIP text encoder is kept frozen during the training and CLIP image encoder is fine-tuned. We train the model for 10 epochs on the Kinetics-400 dataset with a learning rate of 0.00005 with a batch of the size of 20 on 4 NVIDIA RTX A5000 for 40 GPU hours. We use a learning warm step of 5 and a weight decay of 0.2. We use the Adam optimizer with the cross-entropy loss for training on the Kinetics dataset with a clip ratio of 0.1.

To obtain motion descriptions, we use GPT-4 API. The entire generated motion descriptions for the three datasets are provided along with this submission.

B Dataset

B.1 Kinetics-400

Kinetics-400 is a large-scale dataset containing 400 classes downloaded from YouTube. It has 240K training videos and 20K validation videos. Some videos are missing if the YouTube user has removed them.

B.2 UCF-101

UCF-101 human action dataset from YouTube videos belonging to 101 classes. It has 13K videos. We report results on full classes on one split provided by authors.

B.3 HMDB-51

It contains approximately 7K videos belonging to 51 classes. We report results on the split provided by the authors.