A Feint6K: Data Collection

A.1 Data Collection

We develop a Gradio app to collect counterfactually augmented captions from the annotators. For each question, the annotator is first presented with the video and caption(s) from the MSR-VTT [29] or VATEX [23] dataset. Then the annotator is asked to provide a groundtruth text with the matched action in the video, and another five counterfactually augmented texts with novel actions not present in the video. A screenshot of our augmented text annotation web app is demonstrated in Fig. 6.

A.2 Dataset Statistics

We annotate a total of 6,243 videos, each with 5 counterfactually augmented captions. Original videos come from the validation set of the MSR-VTT dataset [29] (947 videos) and the test set of the VATEX dataset [23] (5296 videos).

A.3 Annotation Guidelines

To get annotators familiar with the high-level task, *i.e.*, video-text understanding, and the specific work to accomplish, *i.e.*, annotating counterfactually augmented texts, we provide detailed guidelines about the goal of this project, expected outcomes from the annotations, good and bad practice, *etc*. The full annotation guidelines is available from our project page.

A.4 Notes

Certain actions appear more frequently than others. Although we encourage annotators to design diverse actions that are plausible given the context in the video, annotators may sometimes resort to simpler and more common actions when novel actions are hard to come up with. Actions such as "jump", "climb", and "laugh" appear more frequently than other actions in our annotated counterfactually augmented captions.

Actions in Feint6K explore a much broader space of actions than standard datasets. Actions in standard video-text datasets are limited to common videos available from the internet. As annotators are freely exploring open-set actions that fits the context in the video, actions in Feint6K could explore a much broader space of actions, such as "kick a snowman", "throw up a guitar", or "drop a watermelon".

A.5 Ethics

The data we collect from the human annotators include the counterfactually augmented captions as well as categorical prediction labels when evaluating human performance. Before starting the annotation work, each annotator first sign the consent form acknowledging that: (i) they choose to participate this program voluntarily; (ii) the collected annotations will be used in video-text research projects; and (iii) the collected annotations will be open-sourced and shared with other research groups.

B Limitations

Although we attempt to remove shortcuts from current video-text datasets for a better evaluation of video-text understanding, our evaluation task RCAD is still limited by certain biases in these datasets. As all testing videos are sampled from web-collected data, the object-action pairs would follow a long-tail distribution and video-text models may exploit these biases for a higher benchmark performance. For future work we consider using pretrained text-to-video generation models or video editing models to address the bias issue.

C Shortcuts in Multi-Modal Contrastive Learning

In Sec. 4.1 we compute the change of cosine similarities when videos are unchanged but objects or actions in the captions are manipulated. Results in Fig. 5a demonstrate very different patterns for changes in objects and actions. This gap is attributed to: (i) vision encoder's inability to distinguish between different actions from cross-frame reasoning, and (ii) sensitivity of textual embeddings *w.r.t.* various syntactic categories. To investigate the influence from the textual encoders, we visualize changes in cosine similarities between textual embeddings obtained from (1) LLM2Vec [2], a textual encoder finetuned from LLaMA without multi-modal training, (ii) pretrained textual encoder used in InternVideo, *i.e.*, CLIP textual encoder, and (iii) textual encoder from InternVideo. Results in Fig. 8 show that there is a significant gap between LLM textual encoders and encoders trained with multi-modal contrastive learning. This supports our discussions about shortcuts in Sec. 4.1, where objects become shortcuts in contrastive learning and hinders the models from learning effective action representations.

D Qualitative Comparisons of Caption Generation

As detailed in Sec. 4.2, we consider two caption generation methods in our LLM-teacher to obtain "hard" negative captions for the models to learn from. In Method I, we used a pretrained XLM-RoBERTa [5] model to perform mask filling. In Method II, we leverage the in-context learning capabilities of LLMs

and obtain desirable captions with a LLM-powered chatbot (see Fig. 7). In Fig. 9 we present some qualitative comparisons between captions generated by Method I and II. We find that Method I relies heavily on post-processing such as rule-based filtering (*e.g.*, removing ambiguous words – "he" and "it") or language-based filtering (*e.g.*, removing repeated words derived from the same root – "merged" and "merging"). Meanwhile, Method II is also capable of updating the preposition words according to the change of actions or generating actions composed of multiple words.

E Qualitative Results on Feint6K

In Fig. 10 we present some additional failure cases of InternVideo [24] on retrieval from counterfactually augmented data in our Feint6K dataset. Despite the task being trivial for humans, we show that RCAD is hard for video-text models as it requires complex cross-frame reasoning that current models are weak at. In the "folding/unfolding" example, the model must perceive and reason about changes of the paper over time. In the "using/tangling" example, the model must reason about the interactions between the "person" and the "rope" over a sequence of frames.

The evaluation results we present on RCAD demonstrate that current video-text models still fall far behind human-level understanding of videos and calls for more advanced pretraining strategies. Our LLM-teacher introduces a more effective contrastive learning objective and presents an early step towards this goal.

Improvements from LLM-teacher. We analyze the improvements of LLM-teacher as compared to the InternVideo [24] baseline. We find that with LLM-teacher, our model learns to distinguish actions more effectively (see Fig. 11). However, it still struggles to understand complex human activities (*e.g.*, "jump or crawl on a bounce house") and fine-grained activities (*e.g.*, "apply or throw a lipstick").

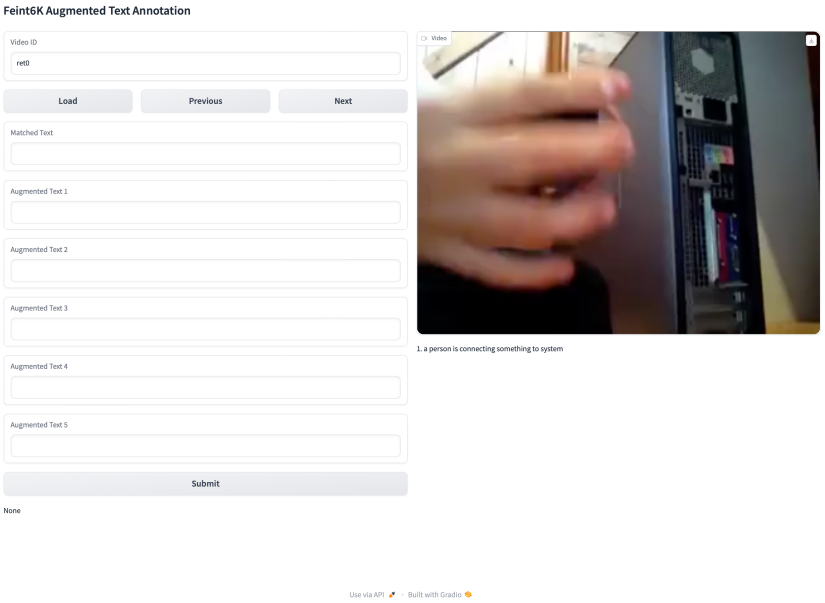

Fig. 6: Screenshot of the Gradio app we develop for caption collection.

Swap an action in the sentence with ten new actions that have very different semantic meaning.

Text: A young lady is buying some vegetables
New text:
A young lady is chopping some vegetables
A young lady is eating some vegetables
A young lady is cutting some vegetables
A young lady is boiling some vegetables
A young lady is washing some vegetables

Text: A person is demonstrating how to open a clam shell
New text:
A person is watching how to open a clam shell
A person is reading how to open a clam shell
A person is demonstrating how to wash a clam shell
A person is demonstrating how to eat a clam shell
A person is demonstrating how to clean a clam shell

Swap an object in the sentence with ten new objects that have very different semantic meaning.

Text: A young lady is buying some vegetables
New text:
A young lady is buying some coffee
A young lady is buying some bread
A young lady is buying some tickets
A young lady is buying some drinks
A young lady is buying some meat

Text: A person is demonstrating how to open a clam shell
New text:
A person is demonstrating how to open a letter
A person is demonstrating how to open a gate
A person is demonstrating how to open a can
A person is demonstrating how to open a container
A person is demonstrating how to open a beer

Fig. 7: We leverage the in-context learning capabilities of LLMs and obtain desirable captions with a LLM-powered chatbot.

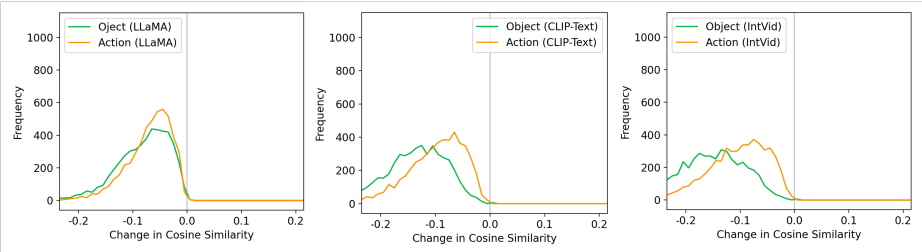

Fig. 8: Comparing changes in cosine similarities using different textual encoders. See discussions in Sec. C.

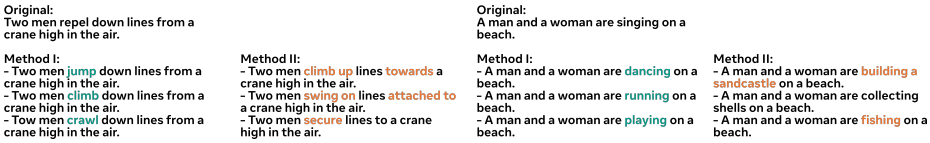

Fig. 9: Qualitative comparisons between captions generated by Method I and Method II. In general we find Method II is capable of exploring a more diverse space of caption and performing multi-word substitutions.

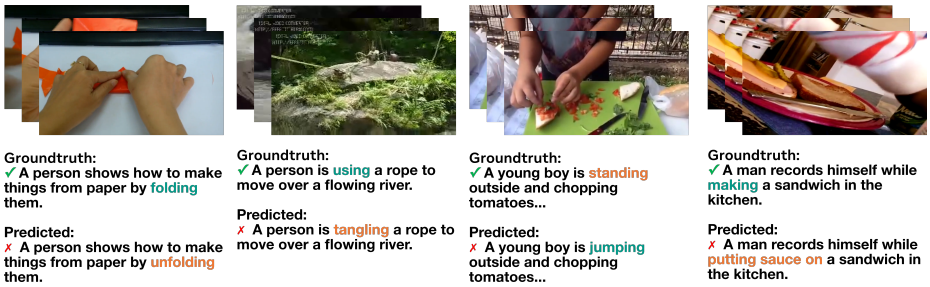

Fig. 10: Failure cases of InternVideo [24] on RCAD in our Feint6K dataset. We show that RCAD is hard for video-text models as it requires complex cross-frame reasoning that current models are weak at.

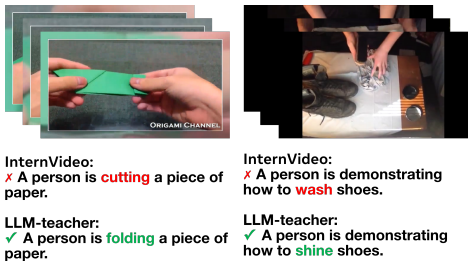

Fig. 11: We analyze the improvements of LLM-teacher as compared to InternVideo [24] baseline. We find that with LLM-teacher, our model learns to distinguish actions more effectively.