Exposing the Limits of Video-Text Models through Contrast Sets

Anonymous ACL submission

Abstract

Recent video-text models can retrieve relevant videos based on text with a high accuracy, but to what extent do they comprehend the semantics of the text? Can they discriminate between similar entities and actions? To answer this, we propose an evaluation framework that probes video-text models with hard negatives. We automatically build *contrast sets*, where true textual descriptions are manipulated in ways that change their semantics while maintaining plausibility. Specifically, we leverage a pre-trained language model and a set of heuristics to create verb and person entity focused contrast sets. We apply these in the multiple choice video-to-text classification setting. We test the robustness of recent methods on the proposed automatic contrast sets, and compare them to additionally collected human-generated counterparts, to assess their effectiveness. We see that model performance suffers across all methods, erasing the gap between recent CLIP-based methods vs. the earlier methods.

1 Introduction

Relating video and text modalities is one of the important goals in vision and language. Video is a complex signal where people and objects act and interact with each other through space and time. Thus correctly associating a textual description and a video requires understanding of entities, their actions and much more, making it a hard problem.

One of the popular ways of training and evaluating video-text models is via cross-modal matching. Often the task is formulated as a retrieval problem, where the goal is to select the correct match among many (e.g. thousand) candidates, and distractors are picked randomly (Yu et al., 2018). Another way is via multiple-choice prediction, where the goal is to pick the true match out of several (e.g. 5) candidates (Torabi et al., 2016). The latter allows for more controlled choice of negatives, which are typically selected from other videos. Commonly, the retrieval setting is used during training to avoid capturing any specific multiple-choice patterns or biases, while both are used for evaluation.

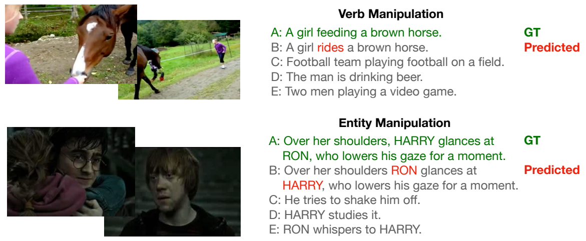


Figure 1: Samples of our video-to-text tasks on the MSR-VTT (Xu et al., 2016) and LSMDC dataset (Rohrbach et al., 2017; Park et al., 2020). A hard negative option is added by manipulating verb (top) and entity (bottom) in the ground truth sentence. Two SOTA methods MMT (Gabeur et al., 2020) and CLIP4CLIP (Luo et al., 2021) incorrectly choose the manipulated sentence (option B) in *both* these cases.

Recent methods that leverage the large-scale CLIP model (Radford et al., 2021) show significant improvement in cross-modal matching, specifically, in the retrieval setting (Fang et al., 2021; Luo et al., 2021). They outperform the prior state-of-the-art methods, often based on the Multimodal Transformer design (Miech et al., 2020; Gabeur et al., 2020; Lei et al., 2021). However, we know that often model performance is "over-estimated" due to the lack of challenging samples in evaluation. For instance, Gardner et al. (2020) show that model performance on several NLP tasks and one image-text task is much lower on *contrast sets*, which are test samples with small perturbation done by human experts in a way that changes the gold label.

In this work, we are investigating whether the video-text models also struggle in an evaluation framework that probes them with hard negatives. Instead of using human-designed contrast sets that are not easily scalable, we propose an automated pipeline that can generate contrast sets via verb and human entity manipulation. Our manipulations are carefully designed to preserve fluency but change semantics of the textual descriptions, making them invalid for a given video. We focus on *entities* and *verbs* to evaluate if the model can truly understand “who did what" in a video. Inspired by (Li et al., 2020; Morris et al., 2020), we leverage a generative T5 language model (Raffel et al., 2020) to manipulate the verb phrase and use heuristics to swap person entities. Note that our pipeline does not require a trained video-text model in the loop.

We apply our automatic manipulations to two popular video-text benchmarks, MSR-VTT (Xu et al., 2016) and LSMDC (Rohrbach et al., 2017). We additionally collect human generated contrast sets to compare with our automatic ones. To make sure that our automatic negatives are of high quality, we also confirm that humans can successfully select the correct description for a given video with our hard negatives. Finally, we benchmark several video-text models on our contrast sets. We find that all methods degrade in performance with the introduction of hard negatives in the multiple-choice setting (Figure 1). This includes the recent CLIP-based works that demonstrated large gains in the retrieval setting. This shows that all methods have difficulty discriminating between entities and verbs when the remaining context is unchanged. We observe that model performance drops especially on cases such as verb antonym swaps, where fine-grained action understanding is important.

2 Related Work

Defending and generating adversarial examples (Jia et al., 2019; Jin et al., 2020) have been mostly explored in NLP since the reign of pre-trained language models (LMs) (Devlin et al., 2019). Li et al. (2020); Garg and Ramakrishnan (2020); Morris et al. (2020) show that substituting words in a sentence with masked LMs (Devlin et al., 2019; Liu et al., 2019) can successfully mislead the classification and entailment model predictions to be incorrect. Template-based (McCoy et al., 2019; Glockner et al., 2018) and manually crafted (Gardner et al., 2020) perturbations on evaluation datasets have also been studied for textual entailment.

Language-based adversarial examples can be collected to study the robustness of vision-language models as well. Shekhar et al. (2017) introduces FOIL-COCO dataset to evaluate the vision-language model's decision when associating images with both correct and "foil" captions. Hendricks and Nematzadeh (2021) show that vision-language Transformers are worse at verb understanding than nouns. New versions of the VQA dataset (Antol et al., 2015) are proposed to study robustness of VQA models (Shah et al., 2019; Li et al., 2021). Our work is different in that we use pre-trained LMs to introduce perturbations and evaluate robustness of video-language models.

3 Designing Contrast Sets

In this section we present our approach to automatically constructing *text-based* contrast sets for video-language tasks. Suppose we are given a video V_i and description s_i. Contrast sets $\hat{C}_i = \{\hat{s_1}, ..., \hat{s_i}\}$ are designed such that $\hat{s}_i$ is semantically inconsistent with V_i and yet models incorrectly select $\hat{s}_i$ over s_i in a video-to-text multiple-choice setting. While there are different ways to create valid $\hat{C}_i$, we investigate manipulating 1) *person entities* and 2) *verb phrases* in the original descriptions. Qualitative examples of $\hat{C}_i$ are shown in the Appendix.

3.1 Contrast sets for Person Entities

First, we investigate swapping the name or *identity* of a person. The LSMDC dataset (Rohrbach et al., 2017; Park et al., 2020) includes movie descriptions with character identities (e.g. *Harry Potter*), and a list of characters present in each movie along with their gender. We replace each character's identity with one from the same movie and with the same gender, to prevent the language statistics alone from detecting the swapped IDs.

For the MSR-VTT dataset (Xu et al., 2016) we do not have the identities, however 80% of videos have gender cues in the descriptions. Thus the contrast sets are created by swapping the gender of a person mentioned in a sentence and the corresponding pronouns (e.g., *A woman is pushing her stroller* $\rightarrow$ *A man is pushing his stroller*). This is done with a template that maps gender-sensitive words and pronouns to their counterparts (see Appendix).

3.2 Contrast Sets for Verb Phrases using Language Models

The above rule-based strategies cannot be directly translated to create contrast sets for verb phrases: 1) a substitute verb phrase is not guaranteed to be inconsistent with a video, and 2) the sentence may look unnatural and no longer be textually plausible. Based on their success in adversarial attack generation (Li et al., 2020; Garg and Ramakrishnan, 2020; Morris et al., 2020), we instead leverage pre-trained language models (LMs) to automatically manipulate the verb phrases.

We identify verb phrases in a sentence using Spacy (Honnibal and Montani, 2017), replace them with a mask token [MASK], and select top K phrases that best fit the mask token using probability scores from a LM. Different from prior work (Li et al., 2020), we use T5-base model (Raffel et al., 2020) instead of masked language models (Devlin et al., 2019; Liu et al., 2019) to easily support generating multi-word candidates. We additionally finetune T5 to learn verb phrases in the downstream training data with unsupervised denoising objective (Raffel et al., 2020). This is done to mitigate the distribution shift between ground truth and generated descriptions.

We then filter the K sentence candidates with the following criteria: 1) There is no verb in the sentence. 2) Verbs are rare or unseen in training descriptions. 3) The sentence has a high perplexity obtained by GPT2-XL (Radford et al., 2019) to ensure grammaticality and plausibility (Morris et al., 2020). Lastly, we check that the semantics of a candidate is *inconsistent* with the original sentence. This is when *a)* the candidate verb is an antonym[1] of original verb, or *b)* the word embedding (Mrkšić et al., 2016) of candidate and original verb and their sentence encodings (Reimers and Gurevych, 2019) both have low cosine similarity scores.

3.3 Human-Generated Verb Contrast Sets

Are language models capable of generating contrast sets of good quality? To answer this question, we follow the original contrast sets work (Gardner et al., 2020), and create negatives manually to see if the performance on machine and human generated contrast sets is similar. We use the Amazon Mechanical Turk (AMT) platform and ask workers to modify a verb phrase such that a sentence becomes inconsistent with a video (see Appendix).

4 Experiments

4.1 Datasets and Multiple Choice Design

MSR-VTT (Xu et al., 2016) is composed of 10K YouTube videos each paired with 20 natural descriptions and is typically evaluated on retrieval performance with 1000 video text pairs as candidates in the test set. The multiple choice version (Yu et al., 2018) has 2,990 test videos as queries, and a positive caption with 4 random captions from other videos as 5 answer options. We label this split as the *Random MC*. We design another MC problem by replacing one negative option with one from our contrast sets. In particular, *Gender MC* swaps gender in an original sentence; *Verb*$_{LM}$ *MC* and *Verb*$_H$ *MC* include verb-based negatives generated by our approach and by humans.

LSMDC (Rohrbach et al., 2017) includes short movie clips and captions. Characters in these captions are labeled as SOMEONE and we cannot construct contrast sets for person-entities. We instead use captions in (Park et al., 2020) that include the character identities. We create a new training/test split using the same movies in training and test so that the test identities have been seen during training. We call this modified dataset **LSMDC-IDs**. Using this set, *Random MC* is newly defined with 4 negative captions drawn randomly from different clips of the same movie. *ID MC* swaps the character IDs (Section 3.1) as negatives, and *Verb MC* includes the verb contrast sets, as before.

4.2 Video-Text Models and Evaluation

We benchmark Transformer (Vaswani et al., 2017) based video-language models in our experiments. Multi Modal Transformer (MMT) (Gabeur et al., 2020) learns the joint representation between text and multiple modalities in videos. CLIP-Straight (Portillo-Quintero et al., 2021) applies frozen CLIP features (Radford et al., 2021) for zero-shot prediction. Inspired by Dzabraev et al. (2021), we also extend MMT to take frozen CLIP features as input, which we denote as MMT-CLIP. CLIP4CLIP (Luo et al., 2021) and CLIP2Video (Fang et al., 2021) directly finetune CLIP with temporal pooler and are the state-of-the-art in retrieval tasks. ViT-B/32 model is used for CLIP experiemnts, see Appendix C for more implementation details. We train the above models with a contrastive loss to learn the joint video-text representation. In MC settings, we mark it as correct, if a ground truth sentence is scored the highest. We also report video-to-text (V $\rightarrow$ T) Recall@1 for retrieval evaluation.

4.3 Results

Table 1 shows results on the MSR-VTT dataset. In video-to-text retrieval, we see a significant gap in performance between the CLIP-finetuned models and all other models; even CLIP-Straight outperforms MMT in this metric. Next, we see that *Random MC* is nearly solved by almost all models. However there is a significant drop in performance

[1] Extracted using VerbNet (Schuler, 2005).

Approach	V → T (R@1)	Random MC	Gender MC	Verb$_{\text{LM}}$ MC	Verb$_{\text{H}}$ MC
MMT	27.0	97.6	**84.0**	83.4	**80.3**
MMT-CLIP	30.8	97.2	**84.0**	80.9	78.3
CLIP-Straight	27.2	91.1	69.6	65.4	64.1
CLIP4CLIP	43.1	**98.4**	82.7	**83.7**	80.2
CLIP2Video	**43.3**	98.3	78.5	81.1	79.0
Human	-	-	-	92.7	94.5

Table 1: Method comparison on **MSR-VTT** dataset. Human is majority vote over 3 judges.

Approach	V → T R@1	Random MC	ID MC	Verb$_{\text{LM}}$ MC	Verb$_{\text{H}}$ MC
MMT	17.7	73.2	65.2	56.2	65.3
MMT-CLIP	23.8	**74.8**	**70.1**	**56.9**	65.8
CLIP-Straight	4.3	53.3	39.8	38.9	42.8
CLIP4CLIP	**25.0**	72.9	69.1	54.1	**66.3**
Human	-	-	-	90.2	93.9

Table 2: Method comparison on **LSMDC-IDs** dataset. Human is majority vote over 3 judges.

across all models when evaluated on contrast-set based MC. Interestingly, the performance gap between MMT and the finetuned CLIP models with high retrieval performance (CLIP4CLIP and CLIP2Video) is gone in this setting, meaning stronger retrieval performance does not guarantee robustness to word-level manipulations. We also observe that models with frozen CLIP features perform better on *Gender MC* than *Verb MC*, and finetuning the CLIP features on video-language task can make the model less sensitive to gender information. Finally, to verify that the automated verb-based contrast sets are valid, we note that: models on *Verb$_{LM}$ MC* perform on par with the human produced ones *Verb$_{H}$ MC*, and humans maintain accuracy greater than 90% on both contrast sets.[2]

Table 2 presents results on the LSMDC-IDs dataset. Overall, it is more challenging than MSR-VTT, as it often requires more fine-grained understanding. Similar to MSR-VTT, models obtain lower performance on contrast-set MC designs. While we see that models found *Verb$_{LM}$ MC* to be more difficult than *Verb$_{H}$ MC*, our automated contrast sets are valid as humans still perform above 90% for both cases. We also notice that the ID swaps are easier than the verb swaps, and CLIP features are helpful in distinguishing character IDs (MMT vs. MMT-CLIP). Table 6 in Appendix shows that model accuracy drops by at least 13.9% when the "negative" IDs appear more frequently in the training data than the original IDs, meaning the models struggle to identify IDs in the long-tail.

Approach	SentBERT Sim.	SentBERT Diff.	CLIP-Text Sim.	CLIP-Text Diff.
CLIP-Straight	55.4	76.0	55.6	71.0
MMT	70.8	93.5	72.1	89.1
CLIP4CLIP	71.8	94.3	68.9	91.9
Human	92.7	93.5	92.2	94.3

Table 3: Model accuracy on *Verb$_{LM}$ MC* in **MSR-VTT**. We select the subsets with the highest and lowest 15% (Sim. and Diff.) semantic similarity with the original sentence. Similarity scores are calculated using: SentBERT in (Reimers and Gurevych, 2019) and zero shot CLIP (Radford et al., 2021) text embedding.

Does Semantic Proximity of Verb Contrast Sets Affect Model Accuracy? To answer this, we use off-the-shelf sentence embeddings to measure the semantic proximity b.w. original and hard negative sentences, and select the subsets of the data with the highest and lowest 15% according to these scores (see examples in the Appendix). In Table 3, we see that models can achieve accuracy greater than 93% on semantically different examples (Diff.) as measured by SentBERT, i.e., on par with humans. However for contrast sets with high semantic similarity (Sim.), model performance is much lower, while human performance is not affected (e.g. CLIP4CLIP drops to 71.8% and humans maintain 92.7% accuracy on SentBERT Sim.). We found that many contrast sets in this subset include *antonyms* of the original verbs (e.g. *pulling vs. pushing*).[3] Distinguishing such antonyms requires fine-grained understanding of actions, which SOTA video-language models fail to demonstrate.

5 Conclusion

We present a pipeline to build automatic contrast sets for video and language tasks, focused on manipulating person entities and verb phrases. We show that models struggle on contrast sets compared to random negatives, and stronger retrieval models do not show better robustness to hard negatives. For verb contrast sets, we find that model performance is strongly correlated with semantic proximity, unlike humans. We leave it as future work to use automatic contrast sets in training to improve model robustness, and designing contrast sets for different concepts/parts of speech.

[2]We report majority vote over 3 human judges.

[3]Recall from Section 3 that we do not apply similarity threshold for antonyms.

6 Ethical Considerations

Our goal is to diagnose performance of video-language models on hard negative samples w.r.t. verbs and person entities. Overall, we envision positive impact from this work, as it aims to expose limitations of the existing models. Some of our entity swaps focus on apparent gender (as described by humans in the video-text datasets), but we do not predict biological sex or gender identity. We construct our verb-focused contrast sets automatically, using a large generative language model, thus potentially some biases present in such a model could propagate into our hard negative samples. Practitioners who wish to use our contrast sets should be mindful of such sources of bias.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Maksim Dzabraev, Maksim Kalashnikov, Stepan Alekseevich Komkov, and Aleksandr Petiushko. 2021. Mdmmt: Multidomain multimodal transformer for video retrieval. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3349–3358.

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

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. *ArXiv*, abs/2004.01970.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. 2017. Cnn architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *EMNLP/IJCNLP (1)*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. *ArXiv*, abs/2106.00245.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. In *The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings* *of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

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. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886.

Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.

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. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *EMNLP*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of HLT-NAACL*.

Jae Sung Park, Trevor Darrell, and Anna Rohrbach. 2020. Identity-aware multi-sentence video description. In *European Conference on Computer Vision*, pages 360–378. Springer.

Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. 2021. A straightforward framework for video retrieval using clip.

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 *ICML*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

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.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6642–6651.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265.

Atousa Torabi, Niket Tandon, and Leon Sigal. 2016. Learning language-visual embedding for movie understanding with natural-language. *arXiv:1609.08124*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487.

Male Nouns	Female Nouns
man → woman	woman → man
men → women	women → men, guys
boy → girl	girl → boy, guy
boys → girls	girls → boys, guys
guy → woman, girl	lady → man, guy
guys → women, girls, ladies	ladies → men, guys

Table 4: List of gender sensitive words mapped to a different gender. Note, that singular and plural form is maintained.

A Contrast Set Construction

Here, we provide more details on construction of each contrast set.

A.1 Gender Contrast Sets

Table 4 shows the mapping of gender-sensitive words. We use these rules to swap only a single word in the sentence. This is to guarantee that swapping gender leads to different semantics (e.g. *man and woman walk together → woman and man walk together* both apply to the same video if all words are swapped). If there are more than one possible mappings, we randomly sample one from a uniform distribution. Lastly, we swap all gender-sensitive pronouns that have the same gender as original noun. These contrast sets are used for the MSR-VTT dataset (Xu et al., 2016).

A.2 Person ID Contrast Sets

The first character ID in a sentence is replaced by a different character ID that appears in the same movie and has the same gender. Among all the candidates, the manipulated ID is sampled from a uniform distribution. The following character IDs in the same sentence have uniform chance of being kept or swapped using the same strategy. These contrast sets are used for the LSMDC-IDs dataset.

A.3 Verb Contrast Sets

Attack Selection We use Spacy to get the POS tags, and find verb phrases that match a list of pre-defined patterns (verb; verb + preposition).

Candidate Generation We use T5 model and performed beam search (beam size = 50) to generate $K = 50$ multi-word candidates.

Candidate Constraints We keep a candidate if the lemmatized verbs [4] in it appeared more than 30 times in the training set. For fluency, we calculate perplexity score of original and manipulated sentence using GPT2-XL (Radford et al., 2019), which we call ppl_o and ppl_m. We calculate the normalized difference of perplexity scores $\text{ppl}_{\text{diff}} = \frac{\text{ppl}_o - \text{ppl}_m}{\text{ppl}_o}$ to remove a candidate that is less plausible than the original. Specifically, candidates are kept if $\text{ppl}_{\text{diff}} < 0.6$, or $\text{ppl}_{\text{diff}} < 1.4 \cap \text{ppl}_m < 750$. Lastly, the semantic inconsistency constraints are satisfied if the word embedding (Mrkšić et al., 2016) of the lemmatized verbs in the candidate and original sentence have cosine similarity score lower than 0.4, and the sentence embeddings (Reimers and Gurevych, 2019) have cosine similarity score lower than 0.8.

B Contrast Set Examples

Random examples of automatically constructed contrast sets using descriptions from MSR-VTT and LSMDC-IDs datasets are shown in Table 5.

We also illustrate the top/bottom 10% (Sim./Diff.) according to SentBERT similarity, as discussed in the main paper. A few examples from each subset are shown in Figure 2.

C Implementation Details

- **MMT** (Gabeur et al., 2020): We use the following features extracted from video[5]: motion from S3D (Xie et al., 2018), audio from VGGish (Hershey et al., 2017), scene embeddings, face, OCR, Speech, and Appearance. We refer to Miech et al. (2018); Gabeur et al. (2020) for more details about the features.

  For MSR-VTT, we use the released checkpoint from their code[6], which is pre-trained on HowTo100M dataset (Miech et al., 2019) and further finetuned on MSR-VTT.

  For LSMDC-IDs which needs re-training, we used their finetuning code for LSMDC dataset (Rohrbach et al., 2017). The model is trained with max margin ranking loss on 1 Nvidia RTX-6000 GPU for 12 hours. Hyperparameter search was done to find margin of 0.05, batch size of 32, and Adam opti-

[4] https://www.nltk.org/_modules/nltk/stem/wordnet.html

[5] https://github.com/albanie/collaborative-experts

[6] https://github.com/gabeur/mmt

Dataset	Original	Person Entity	Verb Phrase
MSRVTT	Two men are doing wrestling.	Two women are doing wrestling.	Two men are dancing.
	A man in black shirt is talking with his two friends.	A woman in black shirt is talking with her two friends.	A man in black shirt is running with his two friends.
LSMDC-ID	His gaze steely, Jenko lowers his gun.	His gaze steely, Schmidt lowers his gun.	His gaze steely, Jenko raises his gun.
	Jenko and Schmidt sit in the rear pew.	Zach and Schmidt sit in the rear pew.	Jenko and Schmidt stand in the rear pew.

Table 5: Examples of person entity and verb phrase hard negatives in MSR-VTT and LSMDC-IDs.

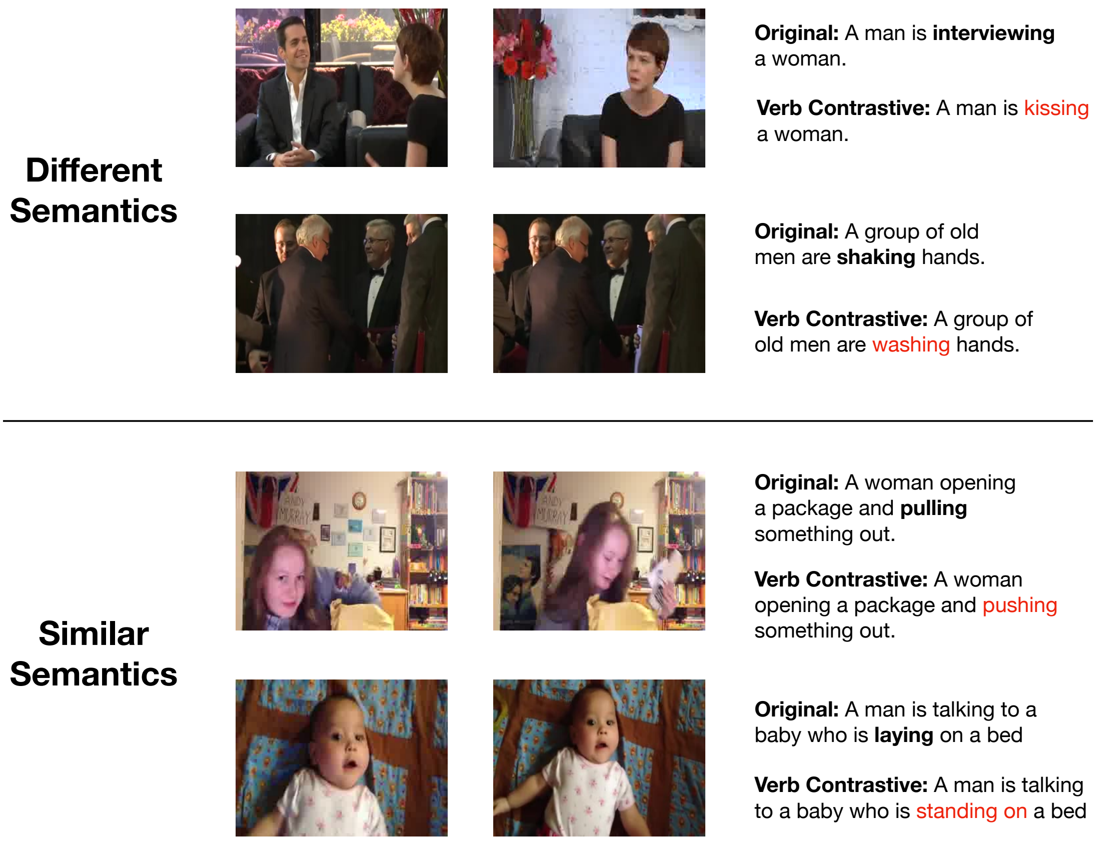


Figure 2: Qualitative example of contrast sets that have different and similar semantics with the original sentence obtained by off the shelf embeddings.

mizer (Kingma and Ba, 2015) with learning rate $5e^{-5}$. The best model was selected by the video-to-text retrieval performance with Recall@1. We found training from scratch performs better than using pre-trained model. This has been also observed by Gabeur et al. (2020) for the LSMDC dataset.

- **MMT-CLIP**: We replace the appearance features in MMT with frozen CLIP ViTB/32 features and train with the same architecture.

- **CLIP-Straight** (Portillo-Quintero et al., 2021): CLIP(ViTB-32) (Radford et al., 2021) features are aggregated via mean pooling to approximate video representation. This video representation and text embedding from CLIP are combined to perform retrieval and MC in a zero shot manner.

- **CLIP4CLIP** (Luo et al., 2021): We use the hyperparameters from the finetuning code[7] to reproduce their results. We use mean pooling for the similarity calculator and CLIP model is initialized with ViTB-32 weights. The model was trained with 4 Nvidia RTX-6000 GPUs for 5 epochs (48 gpu hours). The best model was selected by using Recall@1 in video-to-text retrieval.

[7] https://github.com/ArrowLuo/CLIP4Clip

- **CLIP2Video** (Fang et al., 2021): We used the released checkpoint on MSR-VTT using their code base[8]. This model is not used for LSMDC-IDs because finetuning code was not provided. CLIP model is initialized with ViTB-32 weights.

D Multiple Choice Details

Here we provide more details about our evaluation data. Note, that we use 5 text candidates (1 positive and 4 negative) for all multiple choice (MC) settings.

D.1 MSR-VTT

We use the standard train/val/test split in MSRVTT dataset (Xu et al., 2016).

- Retrieval: 1,000 ground truth video-text pairs in the test set (Yu et al., 2018).
- Random MC: 2,990 videos and all negative options are drawn randomly from other videos (Yu et al., 2018).
- Gender MC: 2,477 video-text instances. Using the original descriptions from Random MC, a single negative is drawn from gender contrast sets to replace one of the options in Random MC (the remaining 3 are kept). Note, that not all videos involved people or contained gender-sensitive words in descriptions, hence some instances are filtered.
- Verb$_{LM}$ MC: 2,554 video-text instances. Constructed using the same strategy as in Gender MC but a single negative is drawn from verb contrast sets generated by language models. Instances are filtered when there are no valid verb contrast sets satisfying constraints in Section A.3.
- Verb$_{H}$ MC: 2,554 video-text instances. We use the instances in Verb$_{LM}$ MC, and a negative is drawn from human designed verb contrast sets.

D.2 LSDMC-IDs

We define a new split using LSMDC descriptions with character IDs (proper names) (Park et al., 2020). Note, that Rohrbach et al. (2017); Park et al. (2020) use development and test sets where videos come from distinct movies than the training data, meaning that IDs in test data are not seen in training. To overcome this issue, we split their *training* descriptions into 80%/10%/10%/ ratio to create new training/validation/test sets that *share* the same movies and identities across splits.

[8] https://github.com/CryhanFang/CLIP2Video

	Overall	Rare	Δ
MMT	65.2	48.4	16.8
MMT-CLIP	70.1	56.2	13.9
CLIP4CLIP	69.1	54.2	14.9

Table 6: Accuracy for *ID MC* in **LSMDC-IDs** dataset. We calculate accuracy when the character ID in original sentence is more rare than the swapped ID (column labeld as Rare). Δ is the difference between the two accuracies and we see the best model (MMT-CLIP) has the lowest difference. See Section 4.3 for more details.

- Retrieval: 7,010 ground truth video-text pairs.
- Random MC: 7,010 videos, negative text options drawn randomly from different videos but the same movie.
- ID MC: 7,010 video-text instances. We replace one negative in Random MC with the one from ID contrast sets.
- Verb$_{LM}$ MC: 7,010 video-text instances. We replace one negative in Random MC with one from the language model generated verb contrast sets.
- Verb$_{H}$ MC: 3,500 video-text instances. We replace one negative in Random MC with one from the human designed verb contrast sets (we only crowdsourced 3,500 instances).

E Human Annotation Details

We ran two different human annotations, one to evaluate our Verb$_{\mathrm{LM}}$ MC and another to manually design verb contrast sets. Figures 3 and 4 show the respective HIT UIs. We use Amazon Mechanical Turk interface to get a pool of annotators from native Enlgish speaking countries and with high approval rate, and pay them $15 hour on average which is above a minimum wage.

F Dataset Details

We include additional information on the MSR-VTT (Xu et al., 2016) and LSMDC (Rohrbach et al., 2017) datasets. MSR-VTT contains diverse YouTube videos and corresponding crowdsourced

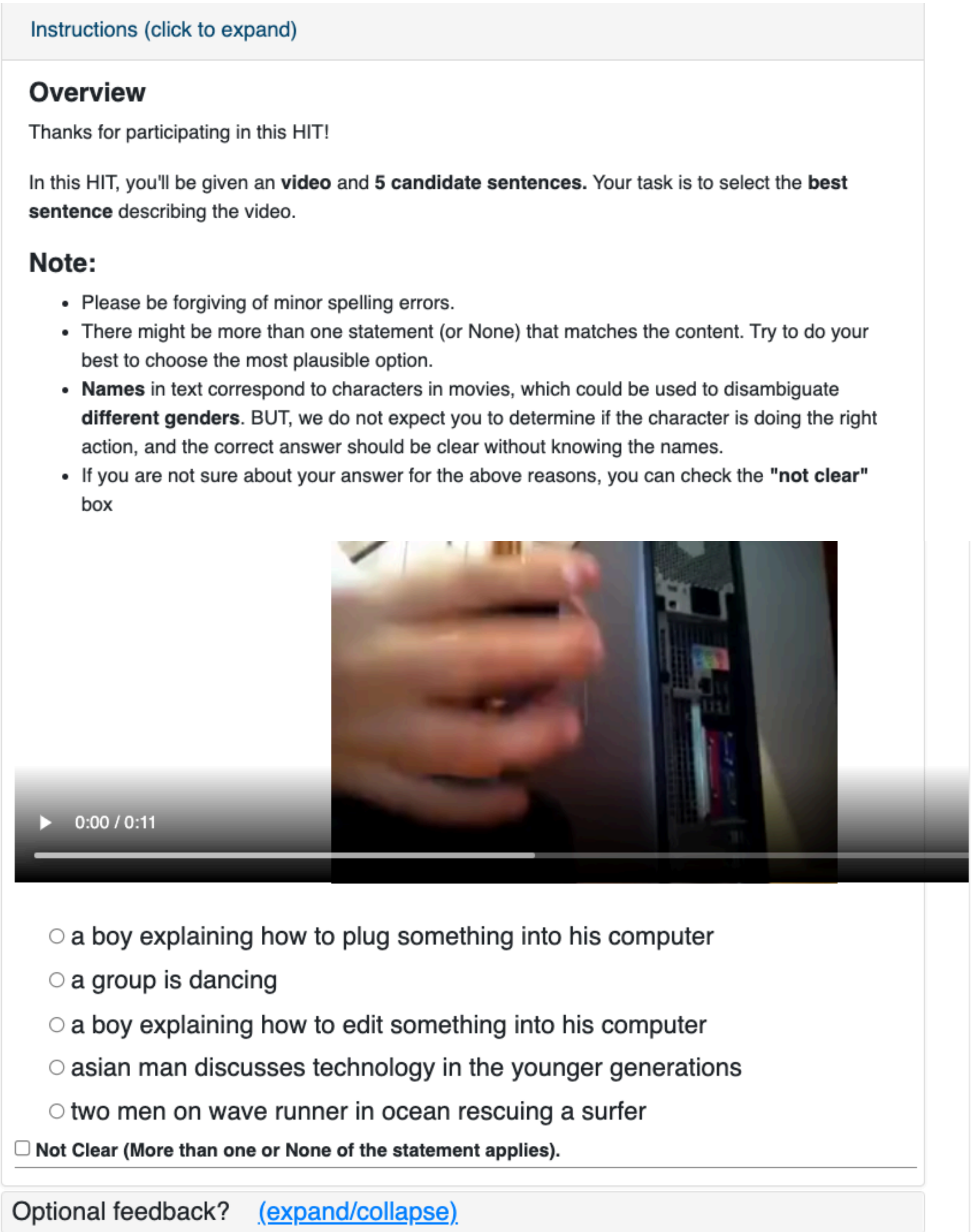


Figure 3: AMT UI for conducting human evaluation in the MC setting with contrast sets.

Instructions (click to expand/collapse)

Overview [Update: 10/25/21]

Thanks for participating in this HIT!

Intro: AI systems have made a great progress in understanding what we see in the media, such as video, using natural language.
One such application is to have a machine go through and find the best video that matches the text description.
But it is still not clear how "good" they are and can understand media in the same level as us.

In this HIT, we are interested if these machines can **detect INCORRRECT details in text** that require more subtle understanding of the video. To do so, we will have to first collect such incorrect descriptions.

Task: You will be given a **video** and a **original sentence** describing the content.
Please **MODIFY** the **HIGHLIGHTED WORD** such that it is **INCORRECT** with respect to what happens in the video.
You are free to change other words, but the highlighted word should ALWAYS BE MODIFIED.

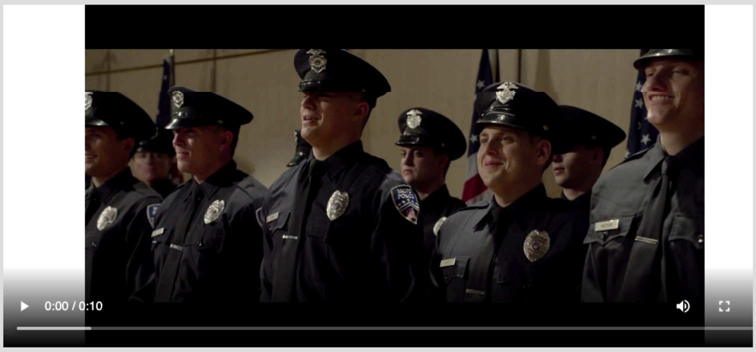
Your written sentence should have the following **PROPERTIES**:

- **(Actually) Incorrect:** Written sentence should include details that are **inconsistent** with the video. While you are trying to write something to fool the machine, the **original sentence should sound more plausible than the modified one**.
  [BAD] Sentence that is NOT Incorrect:
  - Person is **fixing** the computer.
    - **Bad:** Person is **repairing** the machine (try to avoid synonyms).
    - **Good:** Person is **breaking** the computer (**antonyms** are great examples to use).
  - The dancers are **performing** on the stage.
    - **Bad:** The dancers are **dancing** on the stage.
    - **Good:** The dancers are **singing** on the stage (if they are not singing).
- **Plausible:** Written sentence should grammatically make sense and sound plausible. We should not be able to **tell your sentence is incorrect without watching the video.**
  [BAD] Inplausible Examples:
  - A woman **pushing** her stroller.
    - **Bad:** A woman **eating** her stroller.
    - **Good:** A woman **carrying her baby**.
  - A dog is **barking**.
    - **Bad:** A dog is **talking** (usually dogs don't talk in real life).
    - **Good:** A dog is **running towards the owner**. (if dog running is not shown in the video.)

NOTE: You are always welcome to modify multiple words, or even the entire sentence as long as the above properties are met.

More Examples (click to expand/collapse)

HIT:



NOTE:

- **Please look at the examples before you begin!**
- **Please make sure to ALWAYS CHANGE the HIGHLIGHTED word.**
- **You are encouraged to change additional words to make the sentence INCORRECT and still sound PLAUSIBLE (see requirements in instruction).**
- **Please AVOID changing the name of a person.**
- **If a video is not played, please still do your best to write sentence incorrect from image.**

Original Sentence: Jenko **smirks** and Schmidt beams .

Your Incorrect Sentence

Jenko smirks and Schmidt beams.

[Optional] Check to write your sentence!

Your Incorrect Sentence (not required, but if you want to come up with more than one)

Jenko smirks and Schmidt beams.

Figure 4: AMT UI for collecting human-generated verb contrast sets.

descriptions in English language. LSMDC contains movie clips and associated descriptions from scripts or Audio Description, also in English. Both datasets are distributed for research use. The license, personally identifiable information (PII), and consent details of each dataset are in the respective papers. Since LSMDC contains clips from movies, some may contain nudity or violence, etc.