VideoCon: Robust Video-Language Alignment via Contrast Captions

Supplementary Material

A. Detailed Related Work

Foundation Models for Video-Language Understanding. Towards the goal of building general-purpose AI systems, instantiations such as GPT-3 [8], CLIP [55], ALIGN [23] have scaled up self-supervision within single modality (e.g., text) or multiple modalities (e.g., vision-language) by utilizing vast amount of data from the web [20, 44]. Post-training, these models can solve a wide range of downstream tasks through few-shot learning or task-specific finetuning. Similar foundation models have emerged for video-language understanding [1, 4, 49, 55, 56] by pre-training on large amount of video-text pairs scraped from the web [6, 36, 58]. In addition, prior works have either leveraged the pretrained CLIP model for video-language tasks [12, 33, 34] or adopted a socratic approach [50, 63] to employ LLMs (GPT-3) in reasoning over video captions. We highlight that despite the large-scale training of the video-language foundation models [14, 55, 56], they lack robustness to semantically plausible contrast captions (e.g., changing the temporal order of the events) which severely limits their real-world use for alignment applications. We provide a fix to the issue by creating a novel video-centric VideoCon dataset for robust training.

Improving Video-Language Robustness. Prior work [37, 38, 51] highlights that the video-text models cannot comprehend the semantics of the text with focus on manipulating the verb and entities grounded in the video description. At the same time, [5, 51] indicate that the video-text models are not robust to the temporal order of events depicted in the video. To improve the temporal understanding, [5] finetunes a pretrained model with temporal order loss. Despite this, their models do not achieve good zero-shot performance on downstream tasks consistently and is highly dependent on the choice of the finetuning dataset. In our work, we categorize a wide range of plausible misalignments in the contrast captions, 7 in total, and create a temporally-challenging VideoCon dataset by filtering image-temporally-easy instances using a image-text alignment model. Our dataset also covers a wide range of video-text domains covered in MSR-VTT, VaTeX, and TEMPO datasets. Finally, we show that VideoCon enables robust training of the model that achieve state-of-the-art zero-shot performance on various video-language tasks.

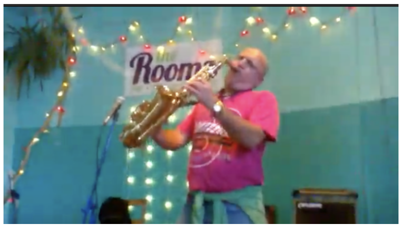
Video-Language Alignment Evaluation. Many traditional applications such as text-to-video retrieval [15, 48, 57] require evaluation of the semantic alignment between the natural language text and raw video. With the rise of creative generative models [40, 41], recent methods [22, 60] have emerged for robust and faithful evaluation of the alignment between the input text and generated image. Similarly, we would soon require robust video-language alignment evaluation to assess the faithfulness of upcoming text-to-video generative models [7, 47]. In this work, we indicate that the existing video-text models such as VideoCLIP and ImageBind are not robust to semantic changes in the video captions, which becomes critical for faithful video-text alignment evaluation. Beyond this, prior work [30, 43] has shown that fine-grained feedback can be useful for evaluating and training better models. In our work, we propose VideoCon and finetune a video-language generative model to perform robust entailment task and provide fine-grained natural language explanations for the observed misalignments between the video and text. As a result, we achieve large performance gains on unseen VideoCon (Human) test set as well as downstream tasks.

B. Details about Video-Language Datasets

MSR-VTT [57] is a large-scale video descriptions dataset covering a wide range of daily life categories ranging from music to cooking. Originally, the dataset contains 10K videos with 20 human-written descriptions for every video. The duration of the video clips in the dataset is between 10-30 seconds. In our work, we filter the videos that are no longer publicly available on Youtube. As a result, we removed 29% of the videos. We utilize the video-text data from MSR-VTT train-val set for VideoCon train-val set, and MSR-VTT test set for VideoCon test set.

VaTeX [48] is large-scale dataset that is focused on enhanced the linguistic complexity and diversity of the video descriptions. The dataset consists of 600 human activities video content from the Kinetics-600 [24]. Originally, the dataset contains 26K videos in the train set and 3K videos in the validation set with 10 human-written descriptions for every video. We used half of the VaTeX training set for VideoCon train-val set and half of the VaTeX validation set for VideoCon test set. Further, we filter the videos that are no longer publicly available on Youtube. As a result, we removed 23% of the videos.

Since MSR-VTT and VaTeX are general-purpose datasets collected from the web, prior work [9, 26] has shown that many of the video-text pairs in these datasets are not temporally-challenging. As shown in Figure 8, a single frame from a VaTeX dataset video shares sufficient semantic information with the video caption, and hence it is not temporally-challenging. The abundance of such instances in the dataset do not encourage the models to develop robust video-language understanding capabilities. Hence, we utilize End-to-End VNLI model [60] to filter temporally-easy instances and make VideoCon temporally-extensive.



a person plays an instrument while wearing a pink shirt

Figure 8. **Illustration of a temporally-easy instance (video-text pair) from the VaTeX dataset.** We observe that the video caption ('a person ... pink shirt') is well-grounded in just a single frame of the video. As a result, the video-text models are not incentivized to develop video-centric understanding (e.g., temporality) while training on such instances.

TEMPO [17] is an unique temporal reasoning video-text dataset. The dataset is constructed from merging two 5 second segments of the videos in the DiDeMo dataset [3]. TEMPO dataset consists of two versions – template-based (TL) and human-written (HL). In our work, we use the video-captions from the TEMPO (HL) dataset. The VideoCon consists of 11K TEMPO training video-text pairs for its train-val set, and 1.8K TEMPO testing video-text pairs for its testing set.

Overall, VideoCon has 27K and 5K unique videos for training-validation and testing, respectively. In addition, it consists 62K and 13K unique captions for training-validation and testing, respectively.

C. Misalignment Assignment

Here, we assign the type of misalignment within the contrast caption for a given video caption. The video caption and the assigned misalignment is then used to prompt large language model (LLM) to generate the contrast caption.

We consider instances from the datasets (V, T) where V is the video caption and T is the text caption. If the caption contains one of the keywords from Table 6, we assign *relation* misalignment to it. If the caption contains a number ('one' - 'ten'), we assign *count* misalignment to it.

For the instances from TEMPO dataset, the captions are assigned *object*, *action*, *attribute*, *hallucination*, *event order flipping* misalignments with equal probability. For the instances from the MSR-VTT and VaTeX dataset, we identify whether the (V, T) instance is temporally-easy $(V, T)_{\text{easy}}$ or temporally-challenging $(V, T)_{\text{challenging}}$ using the End-to-End VNLI model, as described in §3.1. For the temporally-challenging instances $(V, T)_{\text{challenging}}$, we utilize the PaLM-2 LLM API to identify whether the video caption T describes multiple events Ev. For example, 'a girl walks down a hill and eats icecream' has two events i.e., 'walking down a hill' and 'eating icecream' (Ev = multiple). On the other hand, 'a person moving a toy away from the child' consists only a single event (Ev = single). We assign *event order flipping* misalignment to all the captions from $(V, T)_{\text{challenging}}$. We assign *object*, *action*, *attribute*, and *hallucination* misalignment with equal probability to the captions from $(V, T)_{\text{easy}}$.

'above', 'below', behind', 'in front of', 'top of', 'under', 'inside', 'outside', 'beneath', 'left of', 'right of', 'upwards', 'downwards', 'up', 'down', 'far away', 'towards'

Table 6. The list of keywords that indicate spatial relations between entities in the video captions.

We use Spacy [18] to extract POS tags for the words in the video caption. We ensure that the captions without any adjective, verb, noun parts-of-speech words in the captions are not assigned *attribute*, *verb*, and *object* misalignment, respectively.

D. LLM Prompt

We present the prompts used to generate contrast captions for VideoCon dataset in Figure 9 - 15. We have separate prompts for every misalignment where we provide the task description, guidelines, and a few in-context examples. In our work, we use PaLM-2 LLM API. Specifically, we utilize 'chat-bison@001' with chat parameters temperature = 0.5, max output tokens = 256, top p = 0.95, and top k = 40.

E. Human Annotation for Data Quality

We use the workers from Amazon Mechanical Turk platform to assess the quality of the LLM generated data. We present the screenshot of the annotation interface in Figure 16. Specifically, the annotators are asked to decide whether the contrast captions contradict the original video captions. In addition, we ask the annotators to decide whether the generated natural language explanations correctly describe the discrepancy between the caption and contrast caption. The annotators are first asked to perform a qualification test and then selected for the final annotations. We assign one annotator per annotation instance. The human annotators were paid at $18USD per hour, with the total expenditure of $180 USD.

Your objective is to generate a contradiction sentence using a provided “input sentence” based on a specific “misalignment scenario” called “Object Misalignment”. In this scenario, you should modify a key object in the ”input sentence”.

Please also identify the portion of the “input sentence” you’ve expanded and label this as the “source.” Then, specify the new elements introduced in the “sentence + object misalignment” as the “target”.

Your last task is to provide a “Correct Misalignment” description, clarifying how the “input sentence” is different from the “sentence + object misalignment”.

Key Requirements: - The “sentence + object misalignment” should be plausible and could theoretically occur in real life.

Guidelines:
1. The “sentence + object misalignment” should be clearly distinguishable from the “input sentence”.
2. Your replacements should be creative yet reasonable.
3. Avoid changing gender, color, or race of humans in the sentence.
4. The “Correct Misalignment” should describe how the ”input sentence” diverges from the “sentence + object misalignment”.

Input Sentence: a smartphone and a finger pointing to the bluetooth buttons
Sentence + Object Misalignment: a smartphone and a toe pointing to the bluetooth buttons
Source: “finger”
Target: “toe”
Correct Misalignment: a finger is pointing to the bluetooth buttons instead of a toe

Input Sentence: woman plays a song on the piano
Sentence + Object Misalignment: woman plays a song on the cello
Source: “piano”
Target: “cello”
Correct Misalignment: woman plays a song on the piano instead of cello

Input Sentence: a man is going in the wheel skate
Sentence + Object Misalignment: a man is going in the bicycle
Source: “wheel skate”
Target: “bicycle”
Correct Misalignment: a man is going in the wheel skate instead of the bicycle

Now it’s your turn.

Input Sentence: <insert caption>
Sentence + Object Misalignment:
Source:
Target:
Correct Misalignment:

Figure 9. PaLM-2 LLM API prompt to generate contrast captions with *Object* misalignment.

F. VideoCon (Human) Data Creation

To assess the generalization performance of our model, we create a human-written dataset in VideoCon. Specifically, we ask the human annotators to create contrast captions and NLE while looking at the video segments taken from ActivityNet validation data [10] and their associated captions. We present the screenshot of the annotation interface in Figure 17. The annotators are **not** instructed to generate any specific kinds of misalignments in their contrast captions, and just asked generate semantically plausible contrast captions and their NLE. The annotators are first asked to perform a qualification test and then selected for the final annotations. We assign one worker per annotation instance. The human annotators were paid at $18USD per hour, with the total expenditure of $260 USD. We present a few examples from the VideoCon (Human) dataset in Figure 18.

G. Finetuning Details

During finetuning, we use low-rank adaptation (LoRA) [21] of the mPLUG-Owl-Video (7B) [5] applied to all the layers of the attention block i.e., query, key, value, output, gate, up, and down projection matrices. We set the LoRA $r = 32$, $\alpha = 32$, and dropout = 0.05. The model is finetuned on the VideoCon (LLM) training set (§3.3) for 2 epochs. The finetuning was performed using Adam [25] optimizer with the linear-warmup of 200 steps followed by cosine decay learning schedule where the maximum learning rate = 10^{-4}. We chose this learning rate after performing a hyperparameter search over $\{10^{-3}, 10^{-4}, 10^{-5}, 2 \times 10^{-5}\}$ based on the validation loss. We utilized 4 A6000 GPUs with the total batch size of 64 and one gradient accumulation step. We finetune our model by utilizing 32 frames in the video. Specifically, we create 32 segments of the video, and sample the middle

[5] https://github.com/X-PLUG/mPLUG-Owl/tree/main/mplug_owl_video

Your objective is to generate a contradiction sentence using a provided “input sentence” based on a specific “misalignment scenario” called “Action Misalignment.” In this scenario, you should modify specific action performed by the object in the “input sentence”.

Please also identify the portion of the “input sentence” you’ve expanded and label this as the “source”. Then, specify the new elements introduced in the ”sentence + action misalignment” as the “target”.

Your last task is to provide a “Correct Misalignment” description, clarifying how the “input sentence” is different from the “sentence + action misalignment”.

Key Requirements:
- The “sentence + action misalignment” should be plausible and could theoretically occur in real life.

Guidelines:
1. The “sentence + action misalignment” should be clearly distinguishable from the “input sentence”.
2. Your replacements should be creative yet reasonable.
3. Avoid changing gender, color, or race of humans in the sentence.
4. The “Correct Misalignment” should describe how the ”input sentence” diverges from the “sentence + action misalignment”.

Input Sentence: a person repairing the car
Sentence + Action Misalignment: a person driving the car
Source: “repairing”
Target: ”driving”
Correct Misalignment: a person is repairing the car instead of the driving it

Input Sentence: a woman is singing
Sentence + Action Misalignment: a woman is yelling
Source: “singing”
Target: “yelling”
Correct Misalignment: a woman is singing instead of yelling

Input Sentence: an animated cartoon of a monster catching a man by the foot and then launching him like a slingshot
Sentence + Action Misalignment: an animated cartoon of a monster throwing a man by the foot and then launching him like a slingshot
Source: “catching a man”
Target: “throwing a man”
Correct Misalignment: a monster is catching a man instead of throwing a man

Input Sentence: a robot is entering a hall talking to a person
Sentence + Action Misalignment: a robot is leaving a hall talking to a person
Source: “entering”
Target: “leaving”
Correct Misalignment: a robot is entering a hall not leaving it

Now it’s your turn.

Input Sentence: <insert caption>
Sentence + Action Misalignment:
Source:
Target:
Correct Misalignment:

Figure 10. PaLM-2 LLM API prompt to generate contrast captions with *Action* misalignment.

frame from each video.

H. Human Agreement for the Generated NLE Automatic Evaluation Methods

Given the potential noise inherent in automated methods based on Q^2 and PaLM-2, we sought to ascertain their efficacy for NLE evaluation. We conducted a comparative analysis between these automated judgments and human judgments on a sample of 500 instances derived from VideoCon (LLM) and VideoCon (Human), as shown in Table 7. We find that both the metrics achieve high ROC-AUC or agreement with the humans, thus, establishing their usefulness for scalable NLE evaluation.

	VideoCon (LLM)	**VideoCon (Human)**
Q^2-Human ROC-AUC	92	89
PaLM-2-Human Agreement	77.40%	72.50%

Table 7. Human agreement analysis to assess the efficacy of the Q^2 and PaLM-2 as entailment evaluators for NLE generation task. We find that both automatic metrics reliably estimate the human judgements for the task. Hence, both of them can be used for scalable NLE evaluation.

I. Details about Downstream Tasks

We provide details about the downstream task datasets and the evaluation setup in §I.1 and §I.2.

Your objective is to generate a contradiction sentence using a provided “input sentence” based on a specific “misalignment scenario” called ”Counting Misalignment”. In this scenario, you should modify the mathematical count of the objects in the “input sentence”.

Please also identify the portion of the “input sentence” you’ve expanded and label this as the ”source”. Then, specify the new elements introduced in the “sentence + counting misalignment” as the “target”.

Your last task is to provide a “Correct Misalignment” description, clarifying how the “input sentence” is different from the “sentence + counting misalignment”.

Key Requirements:
- The “sentence + counting misalignment” should be plausible and could theoretically occur in real life.
- Only focus on the counts of the objects; do not replace or remove any existing objects, actions or attributes in the ”input sentence.”

Guidelines:
1. The “sentence + counting misalignment” should be clearly distinguishable from the “input sentence”.
2. Avoid changing gender, color, or race of humans in the sentence.
3. The “Correct Misalignment” should describe how the “input sentence” diverges from the “sentence + counting misalignment”.

Input Sentence: a man is entering a room with three surgeons
Sentence + Counting Misalignment: a man is entering a room with one surgeon
Source: “three surgeons”
Target: “one surgeon”
Correct Misalignment: the man enters the room with three surgeons instead of one surgeon

Input Sentence: three girls singing on stage on the voice
Sentence + Counting Misalignment: six girls singing on stage on the voice
Source: “three girls”
Target: “six girls”
Correct Misalignment: three girls are singing on the voice instead of six girls

Input Sentence: a video showcasing 6 different peoples reactions to a certain video the video seemed family oriented
Sentence + Counting Misalignment: a video showcasing 2 different peoples reactions to a certain video the video seemed family oriented
Source: “6 different peoples reactions”
Target: “4 different peoples reactions”
Correct Misalignment: six different people were showcasing their reactions to a video instead of four different people

Now it’s your turn.

Input Sentence: <insert caption>
Sentence + Counting Misalignment:
Source:
Target:
Correct Misalignment:

Figure 11. PaLM-2 LLM API prompt to generate contrast captions with *Count* misalignment.

I.1. Text to Video Retrieval

We perform text-to-video retrieval evaluation on *Something-Something* (SSv2) dataset [15, 26] that covers a wide range of 174 daily actions and around 100K videos. Originally, the dataset captions are presented in two forms: *label* and *template*. In our work, we utilize *SSv2-template* since it removes the bias in the evaluation due to object recognition instead of temporal modeling.

Following this, [45] came up with a list of 18 actions (classes) that require models to capture rich temporal-information in the video (e.g., ‘Moving away from [something] with your camera’). Each class contains 12 videos associated with it. We call this dataset as **SSv2-Temporal** consisting of 216 (18×12) candidate videos for every text query (action).

In addition, [5] create a subset called **SSv2-Events** with 49 actions (classes) that consist two verbs in the action templates that are indicative of multiple events in the video (e.g., ‘Poking [something] so that it spins around’). Overall, this dataset consists 2888 (49×12) candidate videos for every text query (action).

We use the video-text alignment models to rank each video for every action-specific text query. We report the mean average precision (mAP) performance of the models based on the ranking. We want a robust video-language model to achieve high mAP scores on this dataset.

I.2. Video QA

We assess the VideoQA performance of the video-language alignment models on *ATP-Hard* dataset [9]. It is a causal-temporal split [6] of the Next-QA validation dataset [52] [7]. It

[6] https://stanfordvl.github.io/atp-revisit-video-lang//assets/atp-hard-ct4.txt

[7] https://github.com/doc-doc/NExT-QA/blob/main/dataset/nextqa/val.csv

```
Your objective is to generate a contradiction sentence using a provided “input sentence” based on a specific “misalignment scenario” called ”Attribute Misalignment”. In this scenario, you should modify an attribute of an object in the “input sentence”.

Please also identify the portion of the “input sentence” you’ve expanded and label this as the “source.” Then, specify the new elements introduced in the “sentence + attribute misalignment” as the “target”.

Your last task is to provide a “Correct Misalignment” description, clarifying how the “input sentence” is different from the “sentence + attribute misalignment”.

Key Requirements:
- The “sentence + attribute misalignment” should be plausible and could theoretically occur in real life.

Guidelines:
1. The “sentence + attribute misalignment” should be clearly distinguishable from the “input sentence.”
2. Your replacements should be creative yet reasonable.
3. Avoid changing gender, color, or race of humans in the sentence.
4. The “Correct Misalignment” should describe how the “input sentence” diverges from the “sentence + attribute misalignment”.

Input Sentence: man in blue shirt is test driving his new car
Sentence + Attribute Misalignment: man in red shirt is test driving his new car
Source: “blue”
Target: “red”
Correct Misalignment: a man in blue shirt instead of the red shirt

Input Sentence: a group of people playing with giant beach balls
Sentence + Attribute Misalignment: a group of people playing with small beach balls
Source: “giant”
Target: “small”
Correct Misalignment: a group of people playing with giant beach balls instead of the small beach balls

Input Sentence: there is a man with serious face looking cruelly
Sentence + Attribute Misalignment: there is a man with happy face looking kindly
Source: “serious face looking cruelly”
Target: “happy face looking kindly”
Correct Misalignment: a man is with the serious face looking cruelly instead of the happy face looking kindly

Now it’s your turn.

Input Sentence: <insert caption >
Sentence + Attribute Misalignment:
Source:
Target:
Correct Misalignment:
```

Figure 12. PaLM-2 LLM API prompt to generate contrast captions with *Attribute* misalignment.

consists of 2269 instances $(V, Q, \{A_1, A_2, A_3, A_4, A_5\}, A)$ of video V, question Q, and five multiple-choice options $\{A_1, A_2, A_3, A_4, A_5\}$, and a ground-truth answer A.

The aim of a video QA model is to choose the ground-truth answer from the multiple-choice options. To utilize a video-language alignment model for this task, we first recast the input (Q, A_i) pairs into imperative statements using PaLM-2 LLM API. We present the LLM prompt in Figure 19. For example, Q = ‘what does the white dog do after going to the cushion?’ and A_i = ‘shake its body’ is converted to a statement $S(Q, A_i)$ =‘The white dog shakes its body after going to the cushion’. We use the video-language alignment model to score $S(Q, A_i) \forall i \in \{1, 2, 3, 4, 5\}$. The statement with highest entailment score is considered as the model’s prediction. We report the accuracy on the ATP-Hard dataset.

{A, O}	{A, E, H}	{A, R}	{C, H}	{H, R}
0.24%	0.24%	0.24%	8.29%	11.46%

Table 8. Combination bias in multiple misalignment generation using LLM. O: Object, A: Action, Att: Attribute, C: Counting, R: Relation, H: Hallucination, E: Event Flip.

J. Additional Downstream Tasks

Even though the *Owl-Con* was finetuned for the video-text alignment task, we assess its performance on additional image-to-text and video captioning task.

J.1. Performance on LLaVA-Bench

We evaluate the performance of Owl-Base and Owl-Con on a image chat-related task i.e., LLaVA Bench [31] (Table 9). We find that the performance of Owl-Base outperforms Owl-Con on this task. This indicates that finetuning Owl-

Your objective is to generate a contradiction sentence using a provided “input sentence” based on a specific “misalignment scenario” called “Relation Misalignment”. In this scenario, you should change the relation between the objects in the sentence.

Please also identify the portion of the “input sentence” you’ve expanded and label this as the “source”. Then, specify the new elements introduced in the “sentence + relation misalignment” as the “target”.

Your last task is to provide a “Correct Misalignment” description, clarifying how the “input sentence” is different from the “sentence + relation misalignment”.

Key Requirements:
- The “sentence + relation misalignment” should be plausible and could theoretically occur in real life.
- Relation is a word or group of words used before a noun, pronoun, or noun phrase to show direction, time, place, location, spatial relationships, or to introduce an object. Examples include: “above”, “below”, “inside”, “outside”, “front of”, “behind”, “up”, “down”, “left”, “right” etc.
- Only focus on the relations between the objects; do not replace or remove any existing objects, actions or attributes in the “input sentence”.

Guidelines:
1. The “target” should introduce a contradiction when compared to the ”source,” without being a mere negation.
2. The “sentence + relation misalignment” should be clearly distinguishable from the “input sentence”.
3. Your additions should be creative yet reasonable.
4. Avoid changing gender, color, or race of humans in the sentence.
5. The “Correct Misalignment” should describe how the “input sentence” diverges from the “sentence + relation misalignment”.

Input Sentence: people are dancing and singing outside
Sentence + Relation Misalignment: people are dancing and singing inside the club
Source: “outside”
Target: “inside the club”
Correct Misalignment: people are dancing and singing outside, not inside the club

Input Sentence: a woman talking in front of a camera
Sentence + Relation Misalignment: a woman is talking behind a camera
Source: “in front of a camera”
Target: “behind a camera”
Correct Misalignment: a woman talks in front of a camera, not behind it

Input Sentence: a bowl of grey shrimp is shown above a yellow broth
Sentence + Relation Misalignment: a bowl of grey shrimp is shown below a yellow broth
Source: “above”
Target: “below”
Correct Misalignment: a bowl of grey shrimp is shown above a yellow broth, not below it

Input Sentence: a kid flips over a mattress on a trampoline
Sentence + Relation Misalignment: a kid flips over a mattress under the trampoline
Source: “on a trampoline”
Target: “under the trampoline”
Correct Misalignment: a kid flips the mattress on a trampoline, not under it

Input Sentence: the objects are placed far away from each other
Sentence + Relation Misalignment: the objects are placed close to each other
Source: “far away”
Target: “close”
Correct Misalignment: the objects are placed far away from each other, instead of close to each other

Now it’s your turn.

Input Sentence: <insert caption>
Sentence + Relation Misalignment:
Source:
Target:
Correct Misalignment:

Figure 13. PaLM-2 LLM API prompt to generate contrast captions with *Relation* misalignment.

Con with a specialized video-text entailment data which affects its image chat skills.

J.2. YouCook2 Video Captioning

We compare the performance of the Owl-Base and Owl-Con on the video captioning task on the Youcook2 validation dataset [64] using Rouge-L metric (higher the better). Surprisingly, we find that Owl-Con outperforms Owl-Base despite being trained for entailment and natural language explanation generation task for videos.

Your objective is to generate a contradiction sentence using a provided “input sentence” based on a specific “misalignment scenario” called ”Hallucination Misalignment”. In this scenario, you should add new elements to the sentence without replacing or removing anything that is already there.

Please also identify the portion of the “input sentence” you’ve expanded and label this as the “source”. Then, specify the new elements introduced in the “sentence + hallucination” as the “target”.

Your last task is to provide a “Correct Misalignment” description, clarifying how the “input sentence” is different from the “sentence + hallucination”.

Key Requirements:
- The “sentence + hallucination” should be plausible and could theoretically occur in real life.
- Only add elements; do not replace or remove any existing elements in the “input sentence”.

Guidelines:
1. The “target” should introduce a contradiction when compared to the ”source,” without being a mere negation.
2. The “sentence + hallucination” should be clearly distinguishable from the “input sentence”.
3. Your additions should be creative yet reasonable.
4. Avoid changing gender, color, or race of humans in the sentence.
5. The “Correct Misalignment” should describe how the “input sentence” diverges from the “sentence + hallucination”.

Input Sentence: A cola bottle is shown and then it is tossed
Sentence + Hallucination: A cola bottle is shown and then it is tossed along with a frisbee
Source: “tossed”
Target: “tossed along with a frisbee”
Correct Misalignment: There is no frisbee being tossed

Input Sentence: A person is playing a video game where they become aggressive towards a woman robot face
Sentence + Hallucination: A person is playing a video game where they become aggressive and release fireworks towards a woman robot face
Source: “aggressive towards”
Target: “aggressive and release fireworks towards”
Correct Misalignment: The person does not release fireworks at woman robot face

Input Sentence: A man is walking his dog
Sentence + Hallucination: A man is walking his dog while carrying a surfboard
Source: “walking his dog”
Target: “walking his dog while carrying a surfboard”
Correct Misalignment: The man does not carry a surfboard

Input Sentence: Children are playing in the park
Sentence + Hallucination: Children are playing in the park near a giant sculpture
Source: “playing in the park”
Target: “playing in the park near a giant sculpture”
Correct Misalignment: There is no giant sculpture in the park

Input Sentence: A woman is reading a book
Sentence + Hallucination: A woman is reading a book under a parasol
Source: “reading a book”
Target: “reading a book under a parasol”
Correct Misalignment: There is no parasol where the woman is reading a book

Remember: Only add elements; do not replace or remove any existing elements in the “input sentence”. Now it’s your turn.

Input Sentence: <insert caption>
Sentence + Hallucination:
Source:
Target:
Correct Misalignment:

Figure 14. PaLM-2 LLM API prompt to generate contrast captions with *Hallucination* misalignment.

	All	Complex	Conversation	Detail
Owl-Base	69.1	79.0	68.6	51.8
Owl-Con	60.1	67.9	65.0	41.4

Table 9. LLaVA-Bench Evaluation where higher scores are better.

	Owl-Base	**Owl-Con**
Rouge-L	0.08	**0.13**

Table 10. Youcook2 video captioning.

Your objective is to generate a contradiction sentence using a provided "input sentence" based on a specific "misalignment scenario" called "Event Misalignment". In this scenario, you should change the temporal order of the events in the sentence.

Your last task is to provide a "Correct Misalignment" description, clarifying how the "input sentence" is different from the "sentence + event misalignment".

Key Requirements:
- The "sentence + event misalignment" should be plausible and could theoretically occur in real life.
- Only focus on the temporal order; do not replace or remove any existing objects, actions or attributes in the "input sentence".

Guidelines:
1. The "sentence + event misalignment" should be clearly distinguishable from the "input sentence".
2. Your changes should be creative yet reasonable.
3. Avoid changing gender, color, or race of humans in the sentence.
4. The "Correct Misalignment" should describe how the "input sentence" diverges from the "sentence + event misalignment".

Input Sentence: A girl pretends to sneeze and drops something out of her hands and her friend starts to laugh and drops the phone
Sentence + Event Misalignment: A girl drops something out of her hands and then pretends to sneeze and her friend starts to laugh and drops the phone
Correct Misalignment: A girl first sneezes and then drops something out of her hands
Input Sentence: A boy is throwing a ball against a wall and a girl takes the ball and throws it.
Sentence + Event Misalignment: A girl takes the ball and throws it before the boy throws the ball against a wall
Correct Misalignment: A boy is throws the ball against the wall before the girl takes it and throws it

Input Sentence: A small crowd watches as a competitor performs a triple jump, then walks back to the starting mark.
Sentence + Event Misalignment: A small crowd watches a competitor walk to the starting mark, then perform a triple jump
Correct Misalignment: A competitor performs the triple jump before walking back to the starting mark

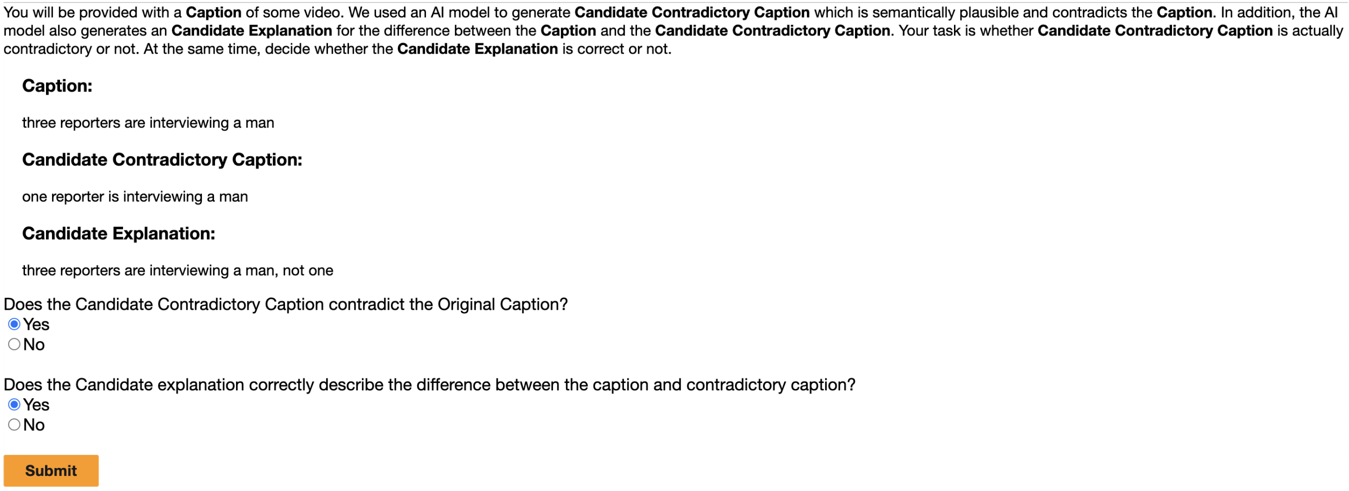
Input Sentence: A man wearing a black t-shirt is holding a cup of food in his right hand. He moves around a piece of food in his left hand to play with the ostrich.
Sentence + Event Misalignment: A man wearing a black t-shirt moves around a piece of food in his left hand to play with the ostrich before holding a cup of food in his right hand.
Correct Misalignment: A man is holding a cup of food before he moves around a piece of food to play with the ostrich

Input Sentence: A person is playing in the doorway, then they begin laughing and grab a doorknob and leave the room.
Sentence + Event Misalignment: A person is playing in the doorway, then they grab a doorknob and leave the room, and then they begin laughing.
Correct Misalignment: They begin laughing before they grabbed the doorknob and leave the room.

Now it's your turn.

Input Sentence: <insert caption>
Sentence + Event Misalignment:
Correct Misalignment:

Figure 15. PaLM-2 LLM API prompt to generate contrast captions with *Event Order Flipping* misalignment.



You will be provided with a **Caption** of some video. We used an AI model to generate **Candidate Contradictory Caption** which is semantically plausible and contradicts the **Caption**. In addition, the AI model also generates an **Candidate Explanation** for the difference between the **Caption** and the **Candidate Contradictory Caption**. Your task is whether **Candidate Contradictory Caption** is actually contradictory or not. At the same time, decide whether the **Candidate Explanation** is correct or not.

Caption:

three reporters are interviewing a man

Candidate Contradictory Caption:

one reporter is interviewing a man

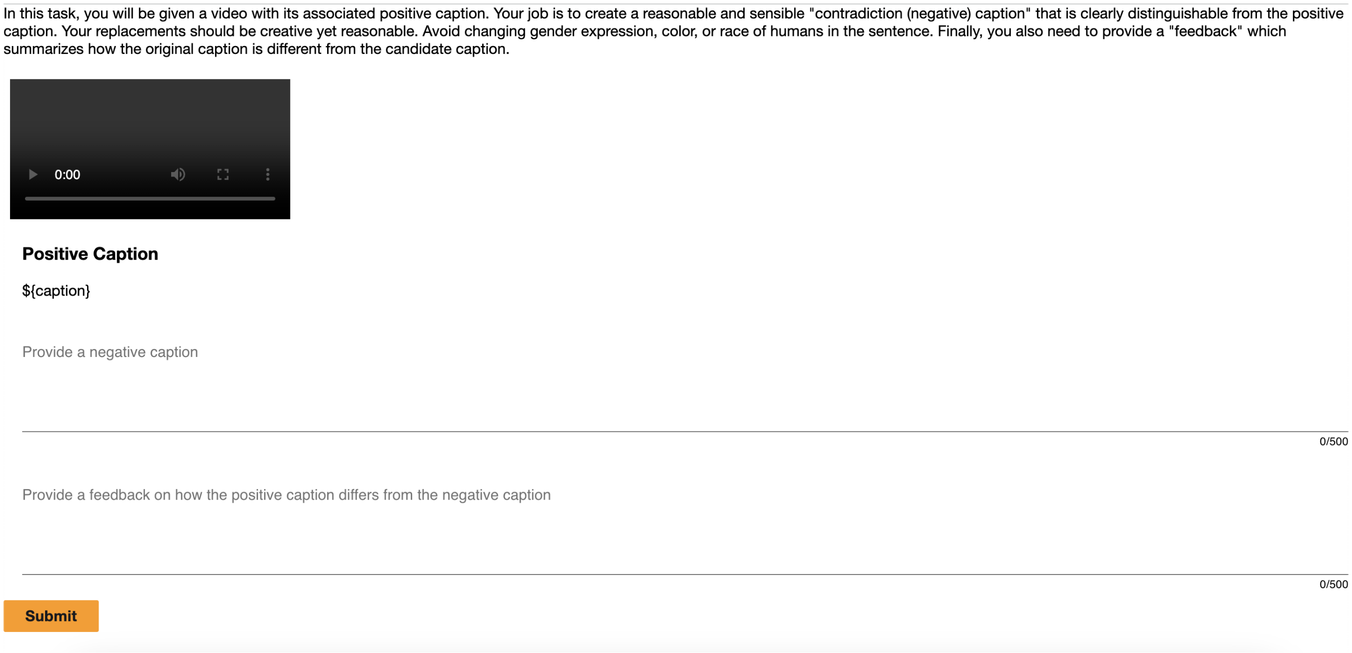
Candidate Explanation:

three reporters are interviewing a man, not one

Does the Candidate Contradictory Caption contradict the Original Caption?
● Yes
○ No

Does the Candidate explanation correctly describe the difference between the caption and contradictory caption?
● Yes
○ No

Figure 16. Screenshot of VideoCon data quality assessment interface.



In this task, you will be given a video with its associated positive caption. Your job is to create a reasonable and sensible "contradiction (negative) caption" that is clearly distinguishable from the positive caption. Your replacements should be creative yet reasonable. Avoid changing gender expression, color, or race of humans in the sentence. Finally, you also need to provide a "feedback" which summarizes how the original caption is different from the candidate caption.

Positive Caption

${caption}

Provide a negative caption

0/500

Provide a feedback on how the positive caption differs from the negative caption

0/500

Figure 17. Screenshot of VideoCon (Human) data collection interface.

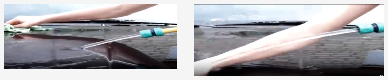
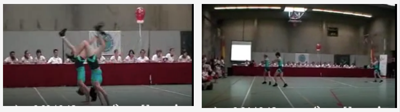
Video Frames	Caption	Human-written Contrast Caption	Human-written NLE
	The lady helps the girl swim	The lady helps the girl dance	The girls are swimming, not dancing
	They fight over the ball, doing ritualistic stunts in between	They fight over the frisbee, doing ritualistic stunts in between	They fight over a ball, not a frisbee
	A video about auto washing is shown	This is a video about auto repair	The video shows auto washing not repairing
	One guy stands up and kneels by the coffee table	Everyone in the room stays seated around the table	At least one person is standing up so not everyone stays seated
	the girls jump and flip in the air, then they start to dance on front a jury	the girls jump and flip in the air, then they bow in front front a jury	The girls dance in front of a jury, not bow in front of them

Figure 18. Example of the instances in the VideoCon (Human) dataset.

K. Why did we alter only one aspect at a time in the contrast captions?

Currently, we alter one aspect to construct the contrast captions. We point that having a *balanced* dataset with multiple changes in a single example is non-trivial due to the LLM's bias towards creating specific combinations. Specifically, we query the LLM to generate new contrast captions with multiple (two or more) misalignments for 1000 video captions. The prompt defined each misalignment with a few in-context examples. We report the least frequent and most frequent combinations of multiple misalignments generated by the LLM in Table 8. We find that most of the contrast captions had hallucination and relation misalignment ({H, R}), while a very few of them had action and object misalignment. While this is an important extension, due to the non-triviality of building a balanced dataset, we will indicate this in the paper as target for future work.

```
You will be provided with a question along with the five multiple choice answers. You need to convert the question and every possible answer to an imperative statement.

Question: how do the two man play the instrument
Choices:
(A) roll the handle
(B) tap their feet
(C) strum the string
(D) hit with sticks
(E) pat with hand
Imperative Statements for every option:
(A) two man play the instrument by rolling the handle
(B) two man play the instrument by tapping their feet
(C) two man play the instrument by strumming the string
(D) two man play the instrument by hitting the sticks
(E) two man play the instrument by patting with hand

Question: how does the man cycling try to sell the watch to the man in the trishaw
Choices:
(A) give him catalogue
(B) show him a video
(C) show him the watch
(D) dismount his bicycle
(E) give him the watch strap
Imperative Statements for every option:
(A) The man cycling tries to sell the watch to the man in the trishaw by giving him the catalogue
(B) The man cycling tries to sell the watch to the man in the trishaw by showing him a video
(C) The man cycling tries to sell the watch to the man in the trishaw by showing him the watch
(D) The man cycling tries to sell the watch to the man in the trishaw by dismounting his bicycle
(E) The man cycling tries to sell the watch to the man in the trishaw by giving him the watch strap

Question: what does the white dog do after going to the cushion
Choices:
(A) drink again
(B) shake its body
(C) smells the black dog
(D) wagging tail
(E) touch lady in blue stripes
Imperative Statements for every option:
(A) white dog drinks again after going to the cushion
(B) white dog shakes its body after going to the cushion
(C) white dog smells the black dog after going to the cushion
(D) white dog wags its tail after going to the cushion
(E) white dog touches the lady in blue stripes after going to the cushion

Now it's your turn.

Question: Q
Choices:
(A) A1
(B) A2
(C) A3
(D) A4
(E) A5
Imperative Statements for every option:
```

Figure 19. Converting the QA pairs into imperative statements for VideoQA dataset.