VideoCon: Robust Video-Language Alignment via Contrast Captions

Hritik Bansal[1] Yonatan Bitton[2] Idan Szpektor[2*] Kai-Wei Chang[1*] Aditya Grover[1*]
[1]UCLA [2]Google Research
{hbansal, kwchang, adityag}@cs.ucla.edu
{yonatanbitton,szpektor}@google.com

arXiv:2311.10111v1 [cs.CV] 15 Nov 2023

Abstract

Despite being (pre)trained on a massive amount of data, state-of-the-art video-language alignment models are not robust to semantically-plausible contrastive changes in the video captions. Our work addresses this by identifying a broad spectrum of contrast misalignments, such as replacing entities, actions, and flipping event order, which alignment models should be robust against. To this end, we introduce the VideoCon, a video-language alignment dataset constructed by a large language model that generates plausible contrast video captions and explanations for differences between original and contrast video captions. Then, a generative video-language model is finetuned with VideoCon to assess video-language entailment and generate explanations. Our VideoCon-based alignment model significantly outperforms current models. It exhibits a 12-point increase in AUC for the video-language alignment task on human-generated contrast captions. Finally, our model sets new state of the art zero-shot performance in temporally-extensive video-language tasks such as text-to-video retrieval (SSv2-Temporal) and video question answering (ATP-Hard). Moreover, our model shows superior performance on novel videos and human-crafted captions and explanations. Our code and data are available at https://github.com/Hritikbansal/videocon.

1. Introduction

Semantically aligning data points from diverse modalities is a long-standing goal of AI. We focus on video-language alignment, which is challenging due to the complexities involved in understanding of entities, their relationships, and temporal order of the depicted events [17]. Recent models such as VideoCLIP [55], ImageBind [14] learn a shared embedding space. Similarly, generative models such as Flamingo [1], mPLUG-Owl-Video [61] can provide a classification label (e.g., yes/no) when queried about video-language alignment.

Despite large-scale pretraining, prior work [5, 36, 38, 51] highlights that video-language alignment models are not robust to semantically plausible manipulations to an original aligned caption in the form of contrast captions, such as from 'dog runs away *before* it eats food' to 'dog runs away *after* it eats food'. Such pitfalls in robustness questions the trustworthiness of alignment models for large-scale deployment. To mitigate these shortcomings, one possible solution is to scale video-language pairs more for increased diversity during pretraining. However, this is challenging due to the difficulties in sourcing new, high-quality and permissible content, as well as the requirement for substantial storage capacity. Several works [11, 13, 16] have shown that naively training models on web-scale data has diminishing returns on downstream tasks, and emphasize the importance of data quality. Furthermore, the recent studies [28, 62] demonstrate that applying a contrastive objective to the pretraining datasets does not encourage the model to grasp the fine-grained details within image/region-caption data.

To this end, we take a scalable, active strategy to gather high-quality data that is deliberately enriched with the attributes that we want to instill in alignment models. We create a novel dataset, **VideoCon**, to improve the robustness of models. Specifically, the dataset consists of a variety of semantically plausible video-language misalignments in contrast captions. These misalignments include altering *objects (entities)*, *actions*, *attributes*, *relations*, *counts*, *event orders*, and introducing *hallucinations* (Figure 2). To construct VideoCon, a large language model (PaLM-2 API) takes video-caption pairs as input and generates high-quality contrast captions for a given misalignment type. To make our dataset temporally-challenging, we skipped "easy" video-caption pairs whose alignment could be inferred based on a single frame (image) understanding [9, 26] (§3.1). In addition, the LLM generates natural language explanations (NLEs) [42] to the differences between original and altered captions, which are used for further robust training. We performed human verification on a sample of VideoCon and found that it is of high-quality. Finally,

*Equal Advising.

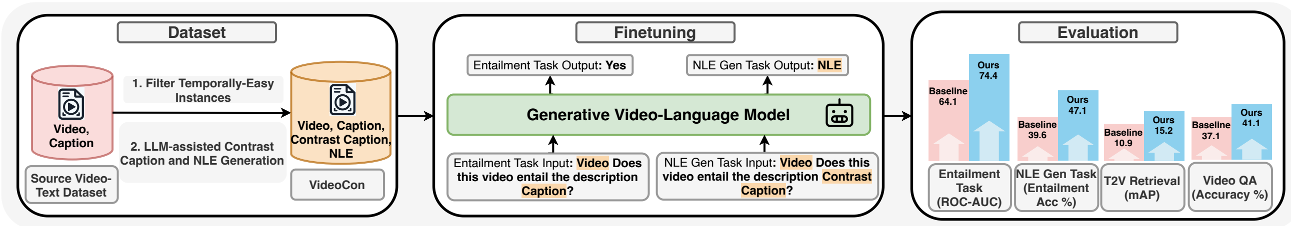


Figure 1. **Overview of our VideoCon approach.** First, aligned video-language pairs are filtered to retain temporally-challenging instances. Then contrast captions and natural language explanations (NLE) are generated by an LLM to create the VideoCon dataset. Second, a video-language alignment model is finetuned with VideoCon on the alignment and NLE tasks. Finally, the finetuned model is evaluated against the baseline model. Our results show that it outperforms the baseline, achieving SOTA results on downstream tasks.

to evaluate the model's generalization capabilities, we collect human-generated contrast captions and NLEs for the videos sourced from external datasets that did not contribute to VideoCon's development.

We finetuned a generative video-language model (mPLUG-Owl-Video) on the VideoCon dataset. The trained model surpasses existing video-language alignment models by a large margin on the LLM-generated test set for both video-language alignment and NLE generation tasks. Interestingly, we observed that our finetuned model generalizes to unseen videos and human-generated contrast captions and NLEs, and outperforms the baseline models. For instance, our model's ROC-AUC exceeds the baseline model by 12 points on the human-generated contrast captions. This indicates that our model has developed a better understanding of the entities, their interactions, action understanding, as well as the temporal order of the events for robust video-language alignment.

We further assessed the effectiveness of robust training via contrast captions on zero-shot downstream video-language tasks such text-to-video retrieval and video question answering on the temporally-challenging and action-intensive SSv2-Temporal [45] and SSv2-Events [5]. Our model achieves state-of-the-art (SOTA) performance, improving on SSv2-Temporal by 4.3 mAP, SSv2-Events by 3.6 mAP points. In addition, our model also achieves SOTA on temporal and causal video question answering in the ATP-Hard dataset, increasing 4% accuracy. This suggests that equipping a model with the knowledge of contrast captions is highly data-efficient and effective in improving its robustness in comparison to scaling the pretraining data. The complete pipeline is illustrated in Figure 1. The dataset and the model will be released upon acceptance.

2. Video Language Alignment

We are interested in assessing the semantic alignment between the video[1] and text data since it powers many practical applications such as video-text retrieval [57], video generation [7, 47] and video captioning [59]. To this end, [14, 39, 49, 55] designed (image)video-text alignment models that are utilized for evaluating the semantic similarity between the two modalities. However, previous works [5, 36, 38, 51] have questioned their robustness to semantically plausible changes to the video descriptions, termed here *contrast captions*. Our aim is to improve the robustness of video-text alignment models by training on contrast captions with a wide range of misalignments.

Consider a dataset $\mathcal{D} = \{(V_i, T_i, C_i, E_i)\}$ where V_i is a video, T_i is an aligned caption, C_i is a contrast caption which is a perturbation of T_i but misaligns with V_i, and E_i is a natural language explanation for the misalignment between V_i and C_i. We consider two video-language alignment tasks: (a) video-language entailment, (b) natural language explanation.

Video-Language Entailment (VLE) casts video-text alignment as a Visual Entailment (VE) task. VE was originally defined for images as premises and texts as hypothesis [53, 54]. We extend VE definition also for videos as premises, under which a classification model $A_{vle}(V,T)$ predicts whether a video V entails a text T.

Natural Language Explanation (NLE) requires a model, $A_{nle}(V,C)$, to generate an open-ended explanation for the discrepancy between a video V and a non-entailing caption C.

In this paper, we address both VLE and NLE tasks under a multitask setting in which a single video-language generative model generates the binary label for entailment and the open-ended explanation.

3. VideoCon: Contrast Captions Generation for Robust Video-Language Alignment

Our research goal is to measure the impact of a comprehensive dataset on increasing the robustness of video-text alignment models. To this end, we first collect video-caption

[1] Like prior works [32, 55], we use only the video frames (the visual channel) without the soundtrack (the audio channel).

Figure 2. **Overview of the VideoCon data generation process from top to bottom.** Specifically, we prompt a large language model (PaLM-2) with the original caption that is grounded in the video, and the intended type of misalignment within the contrast caption. We consider *seven* kinds of misalignments including object, action, attribute, counting, spatial relation, hallucination, and event order flip. We provide a generated contrast caption and the corresponding natural language explanation for each misalignment type.

pairs where the caption cannot be derived from a single frame of video. We then categorize a wide range of semantically plausible manipulations of video captions. Using an LLM for large-scale computation, contrast captions and related explanations are generated for the defined categories, constructing the VideoCon dataset. Finally, we extend VideoCon to include human-created contrast captions as held-out evaluation on unseen videos. We detail the dataset construction steps below.

3.1. Temporally-Challenging Instance Selection

To construct VideoCon, we start with existing datasets that include natural (real) videos and associated high-quality human-written captions: MSR-VTT [57], VaTeX [48], and TEMPO [17]. MSR-VTT and VaTeX consist of 20 captions and 10 captions per video, respectively, while TEMPO consists of a single caption per video. More dataset details are in Appendix §B.

TEMPO is designed to create temporally-challenging instances, while MSR-VTT and VaTeX contain more general video-caption pairs. For MSR-VTT and VaTeX, we filter out instances, where the caption is highly associated with a single frame in the video based on an image-text alignment model. In such cases, a video-text alignment can leverage shortcuts and align the video to its caption without understanding the temporal or causal relations depicted in the video. We want to filter such instances.

To this end, we employ the End-to-End VNLI model [60] to calculate an alignment score $A_{vle}(V, T)$ between a video $V = \{I_1, I_2, \ldots, I_N\}$ and a text T where I_i is a frame from the video sampled at a rate of 1 frame per second. Formally,

$$A_{vle}(V, T) = \max_i(VNLI(I_i, T)) \quad (1)$$

where $VNLI(I_i, T)$ is the image/text entailment score. There are 20 and 10 captions per video in the MSR-VTT and VaTeX datasets, respectively. We retain 5 captions per video from these datasets with the lowest $A_{vle}(V, T)$, and the remaining captions are filtered out. Post-filtering, the percentage of temporally-challenging instances increased from 36.5% to 81.5% in MSR-VTT, and from 42.6% to 71% in VaTeX.

3.2. Categories of Contrast Captions

We aim for VideoCon to include a wide range of misalignments in its contrast captions. Overall, VideoCon covers *seven* misalignment types, exemplified in Figure 2. We include replacement of *objects* (entities) and *actions* following the analysis in [36, 38], and replacement of *attributes*, *counts*, *relations*, as well as adding unrelated but plausible information to captions as *hallucinations* following [29, 31, 34]'s study of image/text alignment model brittleness. Since most video-text models rely on pretrained image backbones, they are likely to suffer from similar problems. Finally, following [5]'s analysis that video-text models do not understand temporal order of the events, we include *event order flipping* as misalignment type.

3.3. Data Generation using an LLM

To generate contrast captions and corresponding NLE we first assign one of the seven misalignment types (§3.2) to each caption in the input video-text datasets (§3.1) (details

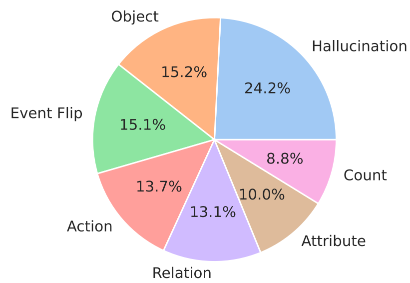


Figure 3. **Distribution of the types of misalignments within the contrast captions of the VideoCon dataset.** We observe that the dataset has good representation for all the kinds of misalignments ranging from 8.8% to 24.2%.

in Appendix §C). Then, given a video V and a misalignment type m, we prompt PaLM-2 API[2] [2] to generate a contrast caption and accompanied explanation (our type-specific prompts are detailed in Appendix §D).

Analyzing the LLM generations, we found that sometimes the output caption C do not contradict the original caption T. For example, a generated contrast caption "*a person riding a car*" does not contradict the original caption "*a person riding a mustang*". To filter such cases, we employ a Natural Language Inference (NLI) model [19] and remove cases in which the contrast caption is assessed as entailed by the original caption $NLI(T, C) > 0.5$. Post-filtering, each tuple (V, T, C, m) is converted to the two instances of video/language entailment task: $A_{en}(V, T) = 1$ and $A_{en}(V, C) = 0$. We present the dataset statistics for the entailment task in Table 1, including train/eval/test splits. In addition, Fig. 3 shows the distribution of misalignment types in the dataset. We observe that VideoCon maintains a high density across the 7 misalignments ranging from 8.8% to 24.2%.

We also found that some generated explanations do not describe the differences between T and C well. For example, the explanation "*two friends are not traveling together*" does not fully describe the discrepancy between "*three friends traveling together*" and "*two friends are traveling together*". To filter these out, generated examples are removed if $NLI(F(T, C), E) < 0.6$ where $F(T, C)$ is the premise comprising the original and contrast captions. Specifically, premise will be 'Expected Caption: T Actual Caption: E' and hypothesis will be 'Difference between Expected and Actual Caption: E'. This filter indicates that the information in the explanation is not entailed by the difference between the two captions. The dataset statistics for the NLE task is presented in Table 1. We refer to the final LLM-generated dataset as VideoCon (**LLM**).

[2] https://developers.generativeai.google/products/palm

	Video-Language Entailment			Natural Language Explanation		
Source	**Train**	**Val**	**Test**	**Train**	**Val**	**Test**
MSR-VTT	38366	478	16538	15888	206	6788
VaTeX	66480	736	8110	30180	345	3636
TEMPO	10712	7098	2708	4165	2739	1073
Total	115558	8312	27356	50233	3290	11497

Table 1. Statistics for the VLE and NLE tasks in VideoCon.

To assess the quality of VideoCon (**LLM**), we perform human evaluation on 500 contrast captions and NLEs (details in Appendix E). The human evaluator found 91% of the contrast captions and 89% of the NLEs to be valid, indicating the high-quality of VideoCon (**LLM**).

3.4. Data Generation using Humans

To study whether a model trained on VideoCon (**LLM**) generalizes to out-of-distribution videos and its performance on human-generated contrast captions, we randomly selected a set of videos from the validation set of ActivityNet [10]. This dataset consists of captions matched with segments in the video, e.g., "*a little boy is climbing on an outside gym*" matched to the first 10 seconds of its related video. We extracted video segments with an associated caption. Human workers[3] on Amazon MTurk were then shown the video segments and associated captions and were asked to create a semantically plausible contrast caption and a corresponding NLE (more details in Appendix §F). *We did not communicate any type of target misalignments to encourage natural diversity of human created contrast captions.*

Overall, we collected 570 tuples $(V, T, C_{human}, E_{human})$ where V is the video, T is the original caption, C_{human} is the human-written contrast caption, and E_{human} is the human-written explanations. We denote this dataset by VideoCon (**Human**). We sample 100 instances from this dataset, and found 93% to be clean. In addition, we observe that many of the human-generated contrast captions perturbing one or more objects (35%) and actions (35%) depicted in the caption. While $8\% - 10\%$ of the contrast captions flip the order of the events and attribute of the objects. As this dataset is largely unfiltered, it contains a mix of temporally-easy and challenging instances. We also constructed a more temporally-challenging subset of 290 instances, denoted VideoCon (**Human-Hard**), by filtering out tuples in which $A_{vle}(V, T) < 0.5$ (Eq. (1)), as in §3.1.

4. Experimental Setup

We next describe our evaluation setting for measuring the impact of VideoCon on video-text alignment modeling.

[3] A shortlist that passed our qualification test.

4.1. Finetuning with VideoCon

Our goal in constructing VideoCon (**LLM**) is to improve robustness of video-text alignment models by fine-tuning on this dataset. To this end, we start with the mPLUG-Owl-Video model [61], denoted *Owl-Base*. Its building blocks are CLIP [39] as visual encoder and LLaMA-7B [46] as text encoder/decoder and it was pretrained on VideoChat [27].

> **<u>Entailment Task</u>**:
>
> **Given: V** (Video), **T** (Caption), **C** (Contrast Caption)
>
> **Instruction (I):** *[V] Does this video entail the description [**T**]?*
> **Response (R):** ***Yes***
>
> **Instruction (I):** *[V] Does this video entail the description [**C**]?*
> **Response (R):** ***No***

Figure 4. Entailment task prompt for finetuning.

> **<u>Natural Language Explanation Generation Task</u>**:
>
> **Given: V** (Video), **C** (Contrast Caption), **E** (NLE)
>
> **Instruction (I):** *[V] What is the misalignment between this video and the description [**C**]?*
> **Response (R):** *[**E**]*

Figure 5. NLE generation task prompt for finetuning.

To fine-tune *Owl-Base* on VideoCon (**LLM**), its $\{V, T, C, E\}$[4] tuples were converted into two types of multimodal instruction-response pairs, one for the VLE task (I_{vle}, R) (Fig. 4) and one for the NLE task (I_{nle}, R) (Fig. 5). We then train *Owl-Base* on all instruction pairs from both the tasks with maximum likelihood loss, resulting in a single model *Owl-Con*.

4.2. VideoCon Evaluation Metrics

To evaluate the performance of the *Owl-Con* on video-text alignment we generate *Owl-Con* response to prompt I_{vle} for video V and text $Y \in \{T, C\}$. We then calculate the probability of generating responses s_y =*Owl-Con* (‘Yes’$|I_{vle}(V, Y)$) and s_n =*Owl-Con* (‘No’$|I_{vle}(V, Y)$), and based on these scores the probability for class ‘Yes’: $P_{yes}(V, Y) = \frac{s_y}{s_y + s_n}$. Finally, we compute the ROC-AUC score for $P_{yes}(V, Y)$ over the VideoCon (**LLM**) eval set, with $\{V, T\}$ as label 1 and $\{V, C\}$ as label 0.

To evaluate *Owl-Con* on the NLE task, we prompt it with instruction I_{nle} instantiated on $\{V, C\}$ pairs from the VideoCon (**LLM**) eval set. We compare the generated explanation $\hat{E}$ to the ground truth E by measuring entailment probability $NLI(E, \hat{E})$. In our experiments, we experiment with two NLI automatic metrics: (a) Q^2 score [19], and (b) PaLM-2 API. We performed human evaluation to measure the agreement between the automatic metrics and the human-rating. We found that both metrics achieve high agreement with human assessment (Appendix §H).

4.3. Video-Text Downstream Tasks

We complement the VideoCon intrinsic evaluation over the testset with an extrinsic evaluation over two temporal and action difficult downstream tasks.

We evaluate alignment model performance for *text2video retrieval* over SSv2-Temporal [45] and SSv2-Events [5] datasets. We consider the SSv2-Template captions instead of the label captions since they remove the object-centric bias in model evaluation [26]. We compute input-text/candidate-video alignment score, rank videos and report *mean Average Precision* (mAP). We evaluate alignment model performance for *video question answering* over the ATP-Hard [9] dataset. We cast each question/candidate-answer pair as an imperative statement using PaLM-2 API, measure alignment to the input video and report *Accuracy*. More details on the downstream datasets and the evaluation setup are in Appendix §I.

4.4. Baselines

For the video-text alignment text, we compare *Owl-Con* with the following baselines: (a) End-to-End VNLI as zero-shot *atemporal* model since it does not have access to the temporal order of the video frames, (b) VideoCLIP [55], (c) ImageBind [14], (d) *Owl-Base*, and (e) *Owl-Rand*: *Owl-Base* fine-tuned on VideoCon tuples $\{V, T, \hat{C}, E\}$ where $\hat{C}$ is randomly selected from other captions in the dataset. *Owl-Rand* would indicate if there is merit in the contrast, hard-negative captions in VideoCon. We include additional baselines TACT [5] and VFC [36] for evaluating on the downstream tasks (§5.3).

5. Experiments

We present our intrinsic (VideoCon eval set) and extrinsic (downstream tasks) evaluation results, showing the benefits of VideoCon for robust video-language alignment.

5.1. Performance on VideoCon Entailment Task

We present the ROC-AUC scores of the tested models in Table 2. From the table we see that the baseline models find the VideoCon testset difficult, as reflected by low AUC scores (e.g. *Owl-Base*- 57.2), close to random. Even training on VideoCon train instances, but with “easy” negatives (*Owl-Rand*- 59.7), hardly improves the base models. A significant improvement is achieved with the VNLI-specific model (67), showing that the entailment task is not inherently represented in generic video-language aligned training

[4] V: video, T: original caption, C: contrast caption, E: explanation.

Models	VideoCon (LLM) Test	VideoCon (Human)	VideoCon (Human-Hard)
Random	50.0	50.0	50.0
VideoCLIP [55]	53.2	47.3	47.5
ImageBind (Video-Text) [14]	57.1	65.2	63.0
Owl-Base [61]	57.2	66.8	64.1
Owl-Rand	59.7	68.9	65.5
End-to-End VNLI [60]	67.0	72.4	65.0
Owl-Con (**Ours**)	**84.6**	**78.3**	**74.4**

Table 2. ROC-AUC scores of the tested models for the entailment task on VideoCon test sets.

	VideoCon (LLM)		VideoCon (Human)	
Models	Q^2 **entailment**	**PaLM-2 entailment acc.** (%)	Q^2 **entailment**	**PaLM-2 entailment acc.**(%)
Owl-Base	0.19	36.8	0.23	39.6
Owl-Con (**Ours**)	**0.50**	**65.4**	**0.32**	**47.1**

Table 3. Performance of the tested models on the NLE generation task, measured via entailment metrics.

sets and requires specific training. Yet, the best performance is achieved by training on VideoCon, which addresses the diversity in plausible misalignments and includes "difficult" training examples, reaching 84.6 AUC. This demonstrates the merit of VideoCon for improving video-language alignment robustness. We show qualitative examples for the model predictions in §6.2.

When evaluating on out-of-domain (OOD) data around video types and misalignment distribution, we again see that training with VideoCon offers significant improvement to alignment detection, outperforming all baselines, albeit with smaller relative gains: 17% and 16% improvement compared to *Owl-Base* on (Human) and (Human-Hard) respectively compared to 48% on (LLM) test. In future work, we plan to further diversify the misalignments VideoCon covers to further improve its benefits on OOD cases.

We notice that the performance of the VNLI atemporal model is better than existing video-language alignment models. It might be attributed to its training with contrast captions in [60]. It further highlights that the existing video-language models are not robust in comparison to a atemporal probe on video-language alignment evaluation, corroborating the findings from [9, 26].

5.2. Performance on NLE Generation Task

Table 3 presents the performance of the tested models against the ground-truth on the NLE task, depicting average Q^2 score and PaLM-2 entailment accuracy. The results show that on in-domain VideoCon, *Owl-Con* outperforms *Owl-Base* by an impressive 263% and 178% relative increase on Q^2 score and PaLM-2 accuracy respectively. This indicates the finetuned model can accurately generate NLE that match well with the ground-truth NLE. This indicates that our model can generate accurate NLE for a wide range of misalignments in the video captions, which makes it useful for dense video-language alignment evaluation.

On out-of-domain VideoCon, the improvement is more moderate but still high: 40% and 20% relative increase on Q^2 and PaLM-2 respectively. This is probably due to the more diverse ways humans express explanations compared to LLM prompting. In future work we plan to further address linguistic diversity in explanations for more robust generation and evaluation.

5.3. Performance on Video-Text Downstream Tasks

Models	SSv2-Temporal mAP	SSv2-Events mAP
Random	7.3	3.3
VideoCLIP	9.8	6.4
ImageBind (video-language)	10.5	5.5
Owl-Base	10.9	6.8
TACT [5]	-	7.8
Owl-Rand	12.1	9.9
End-to-End VNLI [60]	14.6	10.4
Owl-Con (**Ours**)	**15.2**	**11.4**

Table 4. Mean Average Precision (mAP) scores for the tested models in the zero-shot text-to-video retrieval tasks.

We next present our results on the two downstream tasks, Text2Video Retrieval and Video Question Answering. Starting with the retrieval task, we report mean Average Precision (mAP) of the tested models on the SSv2-Temporal and SSv2-Events datasets in Table 4. The benefits of training with additional examples tailored for temporal video-language alignment is already evident in the performance of *Owl-Rand*, which improves over the previous SSv2-Events SOTA - TACT with a relative increase of 27%.

However, when training on harder negative contrastive instances, *Owl-Con* achieves a significant improvement,

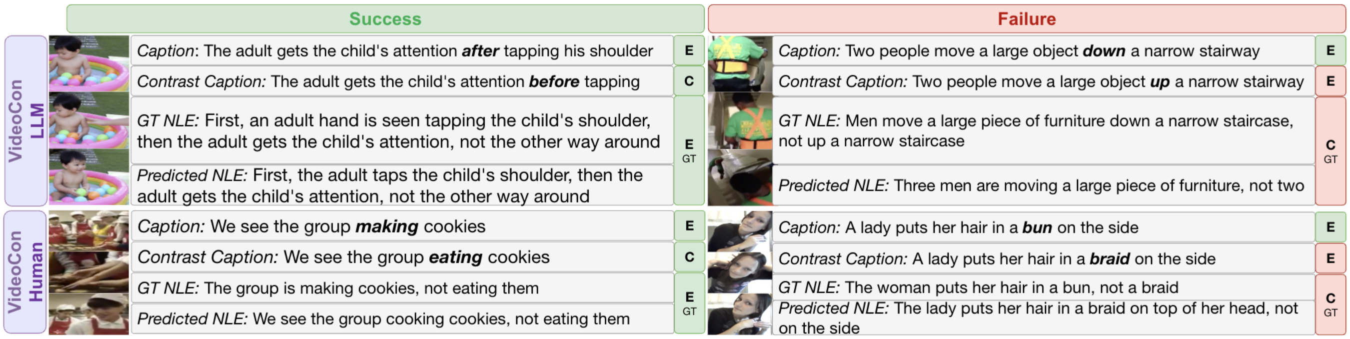


Figure 6. **Qualitative examples for the success (green) and failure (red) modes of our model.** In every example, we present a few video frames in an temporal order from top to bottom, its associated caption, contrast caption, ground-truth NLE from the datasets. Additionally, we present the predicted NLE from our model. The small boxes at the end of caption cells indicate whether our model consider that caption to be grounded in the video. **E** and **C** indicates that the model predicts the caption to entail and contradict to the video, respectively. **E-GT** and **C-GT** indicates the predicted NLE entails and contradicts the ground-truth (GT) NLE, respectively.

outperforming all baselines, with a relative increase over the best baseline End-to-End VNLI model by 7.5% on SSv2-Temporal and 9.6% on SSv2-Events (46% over TACT), setting new SOTA results. This points at the benefits of exposing the model to temporal examples, such as *actions* and *event-order*.

Models	Accuracy (%)
CLIP	23.8
VideoCLIP	23.4
ImageBind (video-language)	25.4
TACT [5]	27.6
VFC [36]	31.4
Owl-Base	37.1
Owl-Rand	37.2
End-to-End VNLI [60]	39.0
Owl-Con (**Ours**)	**41.1**

Table 5. Accuracy scores for the tested models on the zero-shot video question-answering task on ATP-Hard dataset.

For the Video Question Answering task, we compare the performance of the various models in Table 5. Here too *Owl-Con* achieves SOTA results and outperforms the strongest baseline End-to-End VNLI model with a relative increase of 5.1%. This corroborates the observations in our other experiments, which demonstrate the advantage of the VideoCon datasets, covering various misalignments, especially those pertaining to temporal and causal reasoning over dynamic events. The results also confirm the need for carefully chosen contrastive negative examples, showing that picking negatives at random may mask out the potential benefit of an alignment training set. Finally, the competitive performance of atemporal End-to-End VNLI model on the downstream tasks is surprising and underscores the need for stronger video-language datasets for robust benchmarking.

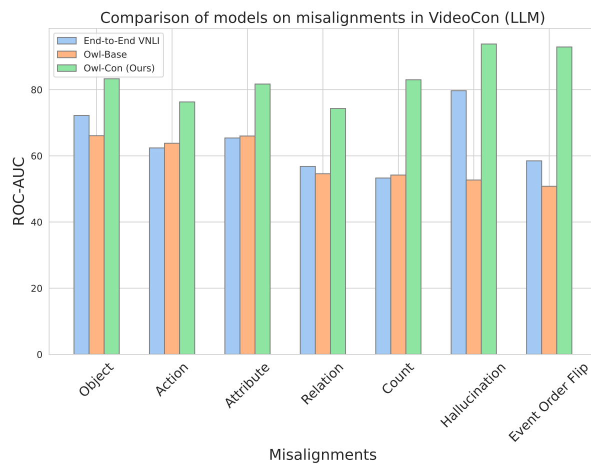


Figure 7. ROC-AUC of End-to-End VNLI, *Owl-Base*, and *Owl-Con* across all types of misalignment in VideoCon (LLM) test set.

6. Analysis

We analyze *Owl-Con* 's performance improvements across the kinds of misalignments in VideoCon. Additionally, we present a few qualitative examples to highlight the success and failure modes of our model.

6.1. Per-misalignment Entailment Results

We compared the ROC-AUC scores of the atemporal End-to-End VNLI, *Owl-Base*, and *Owl-Con* on specific misalignments in the contrast captions from VideoCon (LLM) testset in Figure 7. We observed that *Owl-Con* outperforms the baseline models across all misalignment types. This suggests that our model can reason well about the entities, their relations, and the temporal order of events in the video.

The largest improvement of *Owl-Con* compared to the two baselines is on *event order flip*, indicating that the baselines lack temporal understanding and the VideoCon is efficient in adding this capability to an alignment model. In addition, on *hallucination* both *Owl-Con* and End-to-End VNLI significantly outperform *Owl-Base*, since both models were explicitly exposed to entailment/non-entailment training data. It is surprising to see that while End-to-End VNLI was trained on significantly more entailment data, much of it human-curated, *Owl-Con* outperforms it with only automatically generated data. This could be due to the better encoding of video in *Owl-Con* compared to the atemporal nature of End-to-End VNLI. Finally, the analysis shows other types of atemporal misalignments that are difficult for End-to-End VNLI to sort out, e.g. *counting'* and *relation*, where the training data in VideoCon is useful to improve these capabilities as well. This shows that our approach of detailed analysis of misalignment types of generation of examples for them is effective.

6.2. Qualitative Examples

We highlight a few classification examples of *Owl-Con* in Figure 6. The rows refer to the test source of the instances and the columns refer to the success and failure modes, respectively. In Row1/Column1, we observe that our model provides correct predictions for the entailment between the video and original caption while predicting contradiction for the contrast caption that flips the order of the events i.e., *grabbing attention* and *tapping shoulders*. Interestingly, our model can also provide the accurate NLE when prompted with the video and the contrast caption. This suggests that our model is useful for providing fine-grained details about the video-language alignment. In Row2/Column2, the model confuses 'buns' with 'braids' in hair and gives a wrong NLE that contradicts the ground-truth. This error, due to its inability to distinguish between objects, might be improved by expanding the variety and contrast in the dataset's videos and captions.

7. Related Work

Foundation Models for Video-Language Understanding. 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, 35, 58]. Additionally, prior works have either leveraged the pre-trained CLIP model for video-language tasks [12, 32, 33] 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 semantic changes to the 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 training models on a novel video-centric VideoCon dataset.

Improving Video-Language Robustness. Prior work [36, 38, 51] highlights that the video-text models cannot comprehend the semantics of the text with focus on manipulating the verb, actions, and entities grounded in the video description. 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, and create a temporally-challenging VideoCon dataset. 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 applications such as text-to-video retrieval [15, 48, 57] and text-to-video generation [7, 47] require evaluation of the semantic alignment between the natural language text and raw video. 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 NLE for the observed misalignments between the video and text. In the future, our model can be utilized to enhance alignment through sparse (entailment scores) and dense (fine-grained NLE) feedback [43].

8. Conclusion

We introduced a comprehensive dataset, VideoCon, designed for robust video-text alignment. It features various semantic misalignments and explanations for text-video discrepancies. Through finetuning video-language models on this dataset, we enhanced their performance on complex tasks like text-to-video retrieval and video question answering, achieving state-of-the-art results.

One current limitation and an important future direction is to increase the complexity of the generated contrast captions. Specifically, the model may encounter several misalignments within a single contrast caption. Addressing this issue, the model should be equipped to accurately assign low entailment scores to these contrast captions and consequently generate precise NLEs. An important future direction is to scale VideoCon to larger datasets. Here, we create contrast captions for high-quality captions written by humans for every video, however, the web-scale datasets have low-quality captions that are not well grounded in the video. In this regard, using synthetic data followed by VideoCon-like contrast caption generation can be a plausible approach

[37]. Further, it would be important to scale our VideoCon (Human) dataset more comprehensively to cover a larger set of visual domains (e.g., generated videos), contrast captions and NLE for robust evaluation.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 8

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 4

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 8, 1

[5] Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2503–2516, 2023. 1, 2, 3, 5, 6, 7, 8

[6] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 8, 1

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 8, 1

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[9] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2917–2927, 2022. 1, 5, 6

[10] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 4, 3

[11] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 1

[12] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 8, 1

[13] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 1

[14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1, 2, 5, 6, 8

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 8, 1, 5

[16] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023. 1

[17] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337*, 2018. 1, 3, 2

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

[19] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. Q2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*, 2021. 4, 5

[20] https://commoncrawl.org/. 1

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[22] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 1

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[26] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 1, 5, 6

[27] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 5

[28] Liunian Harold Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. Desco: Learning object recognition with rich language descriptions. *arXiv preprint arXiv:2306.14060*, 2023. 1

[29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3

[30] Weixin Liang, James Zou, and Zhou Yu. Alice: Active learning with contrastive natural language explanations. *arXiv preprint arXiv:2009.10259*, 2020. 8, 1

[31] Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. Evaluation and mitigation of agnosia in multimodal large language models. *arXiv preprint arXiv:2309.04041*, 2023. 3

[32] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2, 8, 1

[33] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 8, 1

[34] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 3

[35] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 8, 1

[36] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591, 2023. 1, 2, 3, 5, 7, 8

[37] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*, 2023. 9

[38] Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586, 2022. 1, 2, 3, 8

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[42] Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8322–8332, 2022. 1

[43] Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*, 2023. 8, 1

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[45] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 535–544, 2021. 2, 5

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 5

[47] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 8, 1

[48] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3, 8, 1

[49] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2, 8, 1

[50] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35: 8483–8497, 2022. 8, 1

[51] Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. *arXiv preprint arXiv:2305.10683*, 2023. 1, 2, 8

[52] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 5

[53] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. *arXiv preprint arXiv:1811.10582*, 2018. 2

[54] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 2

[55] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 1, 2, 5, 6, 8

[56] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*, 2023. 8, 1

[57] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 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, 2016. 2, 3, 8, 1

[58] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8, 1

[59] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. 2

[60] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *arXiv preprint arXiv:2305.10400*, 2023. 3, 6, 7, 1, 2

[61] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1, 5, 6

[62] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022. 1

[63] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 8, 1

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, 35, 58]. In addition, prior works have either leveraged the pretrained CLIP model for video-language tasks [12, 32, 33] 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 [36, 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 $(V, T)_{\text{easy}}$.

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.

'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.

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.

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.

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.

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.

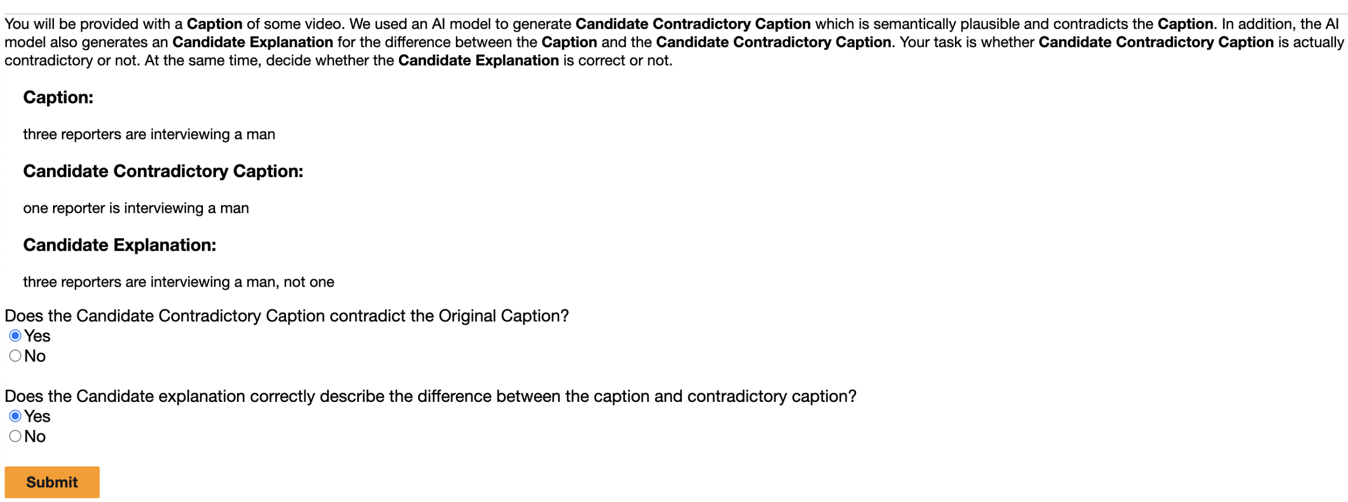


Figure 16. Screenshot of VideoCon data quality assessment interface.

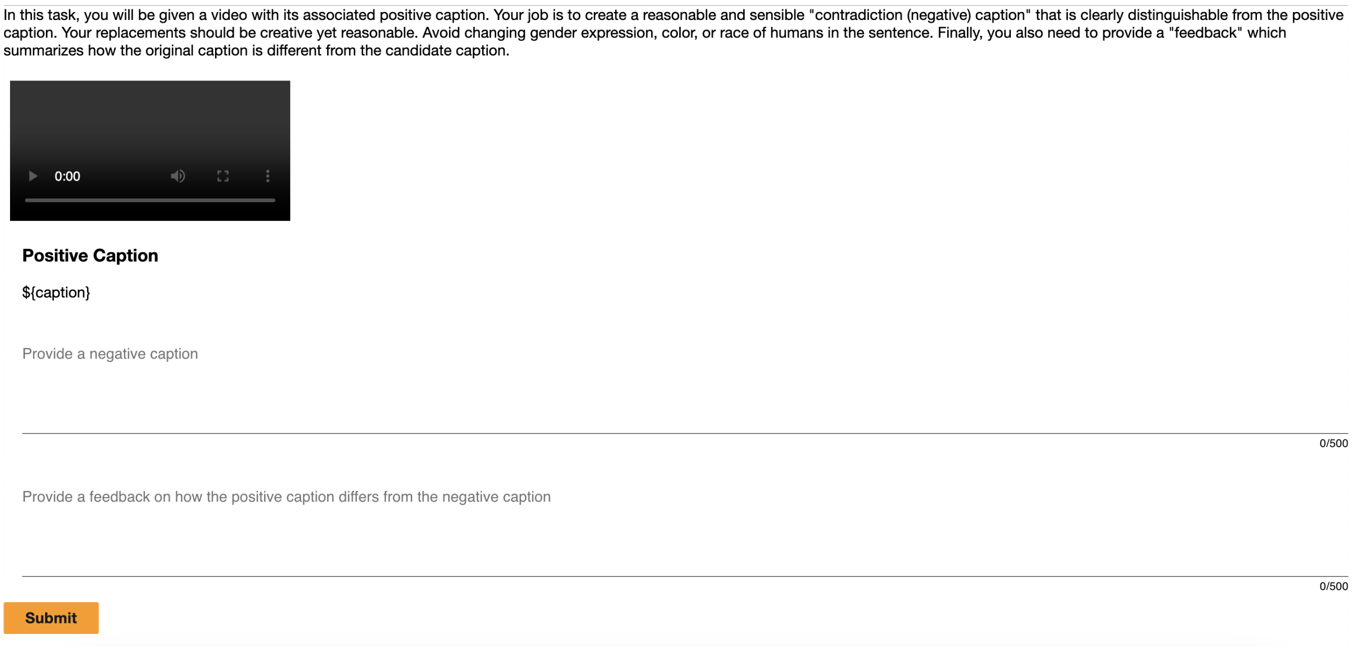


Figure 17. Screenshot of VideoCon (Human) data collection interface.

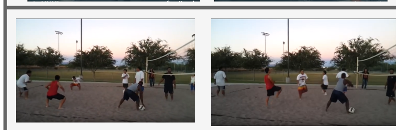
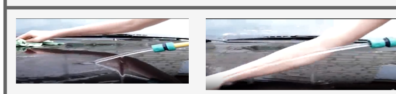
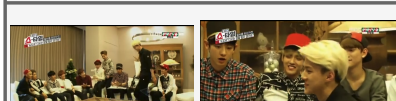
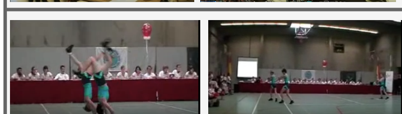
Video Frames	Caption	Human-written Contrast Caption	Human-written NLE
	The lady helps the girl swim	The lady helps the girl dance	The girls are swimming, not dancing
	They fight over the ball, doing ritualistic stunts in between	They fight over the frisbee, doing ritualistic stunts in between	They fight 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.

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.