

UCLA

UCLA Electronic Theses and Dissertations

Title

Learning Counterfactual Reasoning By Answering Counterfactual Questions From Videos

Permalink

<https://escholarship.org/uc/item/6nh2k5pp>

Author

Hu, Qingyuan

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Learning Counterfactual Reasoning

By Answering Counterfactual Questions From Videos

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Computer Science

by

Qingyuan Hu

2023

© Copyright by
Qingyuan Hu
2023

ABSTRACT OF THE THESIS

Learning Counterfactual Reasoning By Answering Counterfactual Questions From Videos

by

Qingyuan Hu

Master of Science in Computer Science

University of California, Los Angeles, 2023

Professor Nanyun Peng, Chair

Multimodal counterfactual reasoning is a vital yet challenging ability for AI systems. It involves predicting the outcomes of hypothetical circumstances based on vision and language inputs, which enables AI models to learn from failures and explore hypothetical scenarios. Despite its importance, there are only a few datasets targeting the counterfactual reasoning abilities of multimodal models. Among them, they only cover reasoning over synthetic environments or specific types of events (e.g. traffic collisions), making them hard to reliably benchmark the model generalization ability in diverse real-world scenarios and reasoning dimensions. To overcome these limitations, we develop a video question answering dataset, ACQUIRED: it consists of 3.9K annotated videos, encompassing a wide range of event types and incorporating both first and third-person viewpoints, which ensures a focus on real-world diversity. In addition, each video is annotated with questions that span three distinct dimensions of reasoning, including physical, social, and temporal, which can comprehensively evaluate the model counterfactual abilities along multiple aspects. We benchmark our dataset against several state-of-the-art language-only and multimodal models and ex-

perimental results demonstrate a significant performance gap ($>13\%$) between models and humans. The findings suggest that multimodal counterfactual reasoning remains an open challenge and ACQUIRED is a comprehensive and reliable benchmark for inspiring future research in this direction.

The thesis of Qingyuan Hu is approved.

Kai-wei Chang

Baharan Mirzasoaleiman

Nanyun Peng, Committee Chair

University of California, Los Angeles

2023

To my loving Mom and Dad, and my grandparents.
Their guidance, belief in my abilities and unconditional supports have been my constant
motivation.

To all my family members, near and far,
who have celebrated all my achievements.

To my dear boy friend Ke Huo,
whose unwavering love, support, and understanding have been the cornerstone of my
academic journey in the past 6 years.

To my exceptional undergraduate Professor Julia Rayz,
who introduced me to the world of NLP and shaped my academic and research trajectory.

To each friend who has accompanied me on this journey,
your camaraderie has enriched my experiences and made the path to this thesis more
memorable.

TABLE OF CONTENTS

1	Introduction	1
1.1	Motivation	1
1.2	Contributions	2
1.3	Thesis Statement	4
2	Background	5
2.1	Visual Question Answering Datasets	5
2.2	Visual Understanding Models	7
2.3	Causal and Counterfactual Reasoning	7
3	The ACQUIRED Dataset	8
3.1	Dataset Design & Collection	8
3.1.1	Problem Definition	8
3.1.2	Commonsense Dimensions	8
3.1.3	Video Resources & Sampling	10
3.1.4	Collection Workflow	11
3.1.5	Quality Validation	12
3.2	Dataset Statistics	13
3.3	More Details of The Dataset	14
3.3.1	Dataset Splits	15
3.3.2	Word Distributions	15
4	Experiments & Results	21

4.1	Experimental Setup	22
4.2	Experimental Results	22
4.3	Discussion	23
4.3.1	Multimodal Models Performance	23
4.3.2	ChatGPT and GPT-4	25
4.3.3	Commensense Dimensions	25
4.3.4	Viewpoints	25
4.3.5	Human Performance	26
4.4	More on GPT Baselines	26
5	Limitations & Future Work	29
6	Conclusion	31
	References	32

LIST OF FIGURES

1.1	The ACQUIRED dataset is a video question answering (QA) dataset that specifically focuses on <i>counterfactual reasoning</i> on diverse real-world events. Our dataset concerns three types of commonsense reasoning dimensions: physical, social, and temporal, and encompasses videos from both third-person (upper) and first-person (lower) viewpoints. Each question is curated with a correct and a distractor answer. Each answer is by itself individually judgeable, and hence our dataset can be approached in either binary True/False or multiple-choice setting. .	3
3.1	Data collection workflow.	10
3.2	MTurk Annotation User Interface: (a) We ask workers to follow the indicated instruction. All the blue-colored text bars on the top of the page are expandable. Workers can click to expand them for detailed instructions of the annotation task. (b) We design an user-friendly and interactive annotation tool where annotators and simply input their annotations and get an instant feedback from our model.	17
3.3	Top-40 frequent word-types in the dataset.	18
3.4	Top-40 frequent word-types in Oops! part of ACQUIRED.	18
3.5	Top-40 frequent word-types in Ego4d part of ACQUIRED.	19
3.6	Top-40 frequent word-types in CLEVERER.	19
3.7	Top-40 frequent word-types in TrafficQA.	20

LIST OF TABLES

2.1	Comparisons of different visual question answering datasets. ACQUIRED is the first to feature all the dimensions.	6
3.1	Sample data points of the ACQUIRED dataset.	9
3.2	Deployed model fooling rates during collection.	13
3.3	General statistics of the two video domains.	14
3.4	Verb-token ratio (total # verb-types / total # tokens) of CLEVRER, trafficQA and ACQUIRED	15
3.5	Noun-token ratio (total # noun-types / total # tokens) of CLEVRER, TrafficQA and ACQUIRED	16
4.1	Exemplar inputs and outputs in T/F setting: 10 frames of the original video clips are subsampled and fed into VL-Adaptor.	24
4.2	Model benchmarking performance on our ACQUIRED dataset.	28

ACKNOWLEDGMENTS

I would like to first express my heartfelt appreciation to the members of my research team: Te-Lin Wu, Yu Hou, Nischal Reddy Chandra, Ziyi Dou and J.R Bronker. Furthermore, I would like to thank my advisor Prof. Nanyun Peng for giving me research opportunities and mentorship. Special thanks to Guanxuan Xu for helping out with my thesis writing. Furthermore, I extend my gratitude to Dr. Marjorie Freedman, Dr. Ralph Weischede for mentorship and guidance on this project. Last but not least, I appreciate all members of my thesis committee, Prof. Nanyun Peng, Prof. Baharan Mirzasoaleiman, and Prof. Kai-wei Chang for reviewing and advising my thesis.

CHAPTER 1

Introduction

1.1 Motivation

Multimodal counterfactual reasoning refers to the ability to imagine and reason about what might have happened if certain conditions were different from what actually occurred based on vision and language inputs. It involves mentally simulating alternative scenarios and evaluating their potential outcomes. This cognitive process plays a crucial role in human intelligence, as it allows us to understand causality, make predictions, and learn from past experiences. For AI models, developing the capacity for counterfactual reasoning is a significant area of research and a challenging task. By enabling AI models to engage in counterfactual reasoning, we can enhance their understanding of causal relationships and their ability to assess the impact of interventions or changes in conditions.

However, despite the significance of counterfactual reasoning, it remains a relatively unexplored area of research with only limited studies focusing on this aspect. To assess the overall reasoning capabilities of models, several visual question answering datasets have been proposed on both images [AAL15, JHV17] and videos [YGL20, XHL21]. These datasets require reasoning skills such as commonsense reasoning, extracting human/object-to-object relations, and inferring physical properties.

One specific dataset in the realm of counterfactual reasoning is CLEVRER [YGL20], which generates synthetic videos and associated questions in a controlled environment, featuring simulated object motion and rendered video frames. This dataset allows for evaluating

models using descriptive, explanatory, predictive, and counterfactual questions, covering a wide range of reasoning scenarios. However, the data generation process in CLEVRER is overly synthetic, limiting its usefulness in assessing models’ counterfactual reasoning abilities in realistic contexts. To address this limitation, TrafficQA [XHL21] focuses on real-world traffic event cognition and reasoning in videos, specifically targeting scenarios like traffic accidents. It leverages crowdsourcing to gather diverse types of questions, including fundamental comprehension, counterfactual inference, and event forecasting. Nevertheless, because TrafficQA concentrates solely on traffic events, it fails to encompass other real-life events, resulting in a substantial domain gap between TrafficQA and general video datasets such as Kinetics [KCS17, SCN20] and YouTube [AKL16, ZLL22].

1.2 Contributions

In this work, we construct a benchmark that can evaluate the counterfactual reasoning abilities of visual models on various kinds of real-world events. To this end, we introduce ACQUIRED¹ that covers multiple dimensions of counterfactual reasoning and includes videos of both egocentric and exocentric views. Specifically, based on videos in both Oops [ECV20] and Ego4D [GWB22], we crowd-source 11K questions over 3.9K videos targeting physical, temporal, and social counterfactual reasoning. Both the Oops and Ego4D datasets consist of human activities and interactions in numerous settings, making them ideal sources for curating video question answering datasets. In addition, many videos contain unintentional human actions (e.g., the person accidentally falling down the ladder in Figure 1.1), which naturally enables people to come up with diverse *what-if* questions.

Inspired by [SWH21], we adopt a similar methodology for gathering counterfactual questions. Each question consists of a pair of answers, with one being the correct response and the other serving as a distractor. Importantly, the distractor answer represents a *minimal*

¹Abbreviation of: **A**nswering **C**ounterfactual **Q**uestions **I**n **R**eal-Life **V**ideos

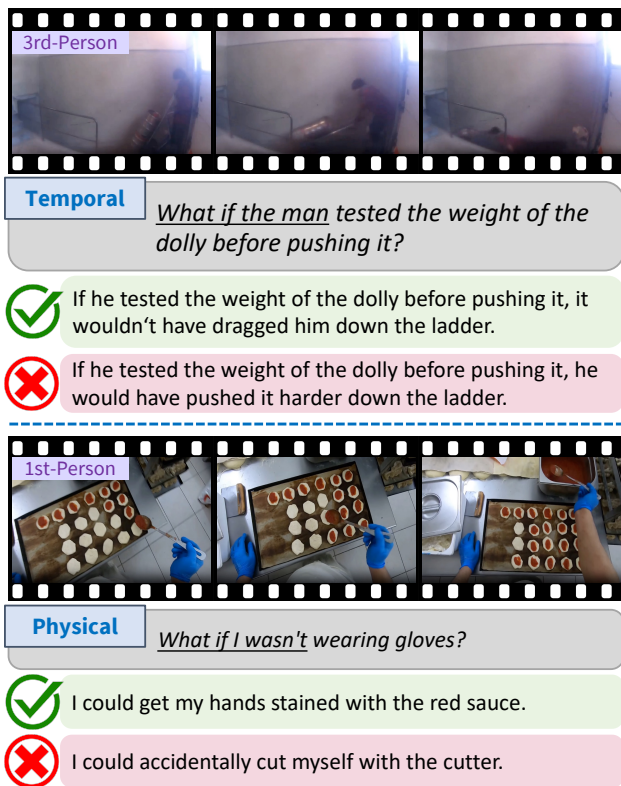


Figure 1.1: **The ACQUIRED dataset** is a video question answering (QA) dataset that specifically focuses on *counterfactual reasoning* on diverse real-world events. Our dataset concerns three types of common-sense reasoning dimensions: physical, social, and temporal, and encompasses videos from both third-person (upper) and first-person (lower) viewpoints. Each question is curated with a correct and a distractor answer. Each answer is by itself individually judgeable, and hence our dataset can be approached in either binary True/False or multiple-choice setting.

contrastive counterpart to the correct answer. As we can see from examples in Figure 1.1, the design of using complementary pairs requires the model to understand the subtle differences between different options, which ensures that the model exhibits an intuitive grasp of counterfactual reasoning. In addition, having one distractor for each question allows for testing models in either True/False or multiple-choice setting.

We extensively evaluate numerous strong language models such as GPT-4, as well as state-of-the-art video-language models such as VALOR on our ACQUIRED dataset. The experimental results suggest that models struggle to effectively utilize the video contexts

and perform counterfactual reasoning, with multimodal models achieving only comparable and sometimes inferior performance than language-only models. Moreover, the significant gap between the human and model ($>13\%$) performance highlights the challenging nature of our task and room for improvements in visual counterfactual reasoning.

1.3 Thesis Statement

This work provides the first sizable video question answering dataset that focuses on typed reasoning and specifically counterfactual reasoning in a diverse set of events, where the reasoning can span across physical, social, and temporal aspects, and the videos include both third-person and ego-centric viewpoints.

CHAPTER 2

Background

This chapter offers an overview of three lines of relevant research to this work: visual question answering, visual understanding models, and counterfactual reasoning.

2.1 Visual Question Answering Datasets

In 2.1, we list several representative visual QA datasets as well as their key features. The Visual Question Answering (VQA) dataset [AAL15] is one of the pioneering works in this direction and has been a standard benchmark for evaluating the reasoning ability of image-language models [GKS17]. Follow-up datasets such as CLEVR [JHV17] and GQA [HM19] automatically construct compositional questions over real or synthetic images and perform the evaluation in a systematic way. To further evaluate the commonsense reasoning ability of models, VCR [ZBF19] crowd-sources commonsense question-answer pairs associated with rationales over static images extracted from movies. Video question answering is more challenging than image question answering and is gaining increasing attention from the research community, leading to several video QA datasets being constructed [LYB20, TZS16, XZX17, JSY17, MHJ17, LYB18]. Among them, CLEVRER [YGL20] improves upon CLEVR and uses programmatically generated videos capturing collisions of synthetic objects to evaluate the model reasoning abilities along multiple dimensions. Social-IQ [ZCL19] and TrafficQA [XHL21] employ videos depicting real-world events, wherein Social-IQ primarily emphasizes human social interactions, while TrafficQA focuses on traffic events and accidents.

Dataset	Visual Source	Question Source	Reasoning Domain			Counterfactual
			Physical	Temporal	Social	
<i>Image QA datasets</i>						
VQA [AAL15]	Diverse Real-world Event	Human	✓	✗	✗	✗
CLEVR [JHV17]	Synthetic Object	Automatic	✓	✗	✗	✗
GQA [HM19]	Diverse Real-world Event	Automatic	✓	✗	✗	✗
VCR [ZBF19]	Movie	Human	✓	✗	✓	✗
<i>Video QA datasets</i>						
CLEVRER [YGL20]	Synthetic Object Collision	Automatic	✓	✓	✗	✓
VLEP [LYB20]	TV & YouTube	Human	✓	✗	✗	✗
MovieQA [TZS16]	Movie	Human	✓	✓	✓	✗
MSRVTT-QA [XZX17]	Diverse Real-world Event	Automatic	✓	✗	✗	✗
TGIF-QA [JSY17]	Tumblr GIF	Automatic & Human	✓	✓	✗	✗
MarioQA [MHJ17]	Gameplay Video	Automatic	✓	✓	✗	✗
TVQA [LYB18]	TV	Human	✓	✓	✗	✗
Social-IQ [ZCL19]	YouTube	Human	✗	✗	✓	✗
TrafficQA [XHL21]	Traffic Event	Human	✓	✓	✗	✓
ACQUIRED	Diverse Real-world Event	Human	✓	✓	✓	✓

Table 2.1: **Comparisons** of different visual question answering datasets. ACQUIRED is the first to feature all the dimensions.

As can be seen in Table 2.1, among all the visual QA datasets, there are only a few that attempt to evaluate the counterfactual reasoning abilities of models. In addition, the existing benchmarks are often limited in terms of the video sources and the question types, making it difficult to evaluate the model performance in a diverse real-world setting. ACQUIRED is the first dataset that can comprehensively evaluate the model counterfactual reasoning abilities spanning three distinct dimensions (i.e., physical, social, and temporal) and cover videos that include a wide range of event types and from different viewpoints.

2.2 Visual Understanding Models

The creation of visual QA benchmarks allows for the development of visual understanding models. Many of the previous works have tried to solve these tasks using compositional approaches and scene graphs [SRB17, HAR17, HM18, PSD18, YWG18, SZL19, GLW20, DCD21]. For example, [HAR17] propose to train a modular network in an end-to-end manner to achieve both effectiveness and interpretability; [HM18] utilize scene graphs and perform differentiable neural operations on the graphs to perform visual reasoning. Inspired by the success in pretraining on Internet-scale data [DCL19], pretraining models on large vision and vision-language tasks and then finetuning them on specific downstream tasks has become a standard in tackling visual understanding tasks [SMV19, LCC20, ZY20, LLZ21, ZLH21, FLG21, ZLL22]. Existing works in this direction generally train models on large vision-language datasets with objectives such as masked language modeling and video-text matching. Despite the great progress in this direction, it is unclear if these models can perform counterfactual reasoning. To address this, we benchmark ACQUIRED against state-of-the-art models and systematically study their performance.

2.3 Causal and Counterfactual Reasoning

Humans can infer how an event would have unfolded differently without experiencing this alternative reality and it has been a long-standing research topic in cognitive psychology [VWB15]. To empower such an important ability to artificial intelligence, researchers have tried to build learning models that can infer causal relations and perform reasoning in various fields [QBH19, YGL20, BNM20, ATP20, YWS21, WFH21]. Our constructed benchmark provides a valuable resource for developing and evaluating visual models with counterfactual reasoning abilities.

CHAPTER 3

The ACQUIRED Dataset

3.1 Dataset Design & Collection

3.1.1 Problem Definition

As illustrated in Figure 1.1 and Table 3.1, each data point in ACQUIRED consists of a video and corresponding annotated question and answer pairs. For each question, we collect **one correct** and **one distractor** answer (which can be a slightly perturbed version of the correct one), where both of which are individually judge-able by themselves respectively. And hence, our dataset can be approached as a binary *True/False (T/F)* prediction task as well as a *multiple-choice (MCQ)* (2 choices in this case) question answering task.

3.1.2 Commonsense Dimensions

We adopt the commonsense knowledge categorization proposed in [SWH21] to collect QAs that focus on the following three dimensions: *physical*, *social*, and *temporal*. The **physical** dimension concerns the knowledge of objects involved in the events and their properties (e.g., shape, size, functionalities, affordances), as well as the motion and location of the events. The **social** dimension looks at human social behaviors, particularly attributes such as personality, emotions, inner interests/intentions, and social activities.¹ The **temporal** dimension regards the aspects of events/activities in their temporal orderings, duration, and

¹As most videos from Ego4D show tasks performed solely by the camera wearer without social interactions, we do not require the social dimension to be annotated.





Sub-sampled Key Video Frames	Question-Answer Pairs
	<p>(Temporal) Q: What if the two persons had swerved to their left before reaching the shore?</p> <p>Correct: They would not have had a beach landing.</p> <p>Wrong: They would have had a beach landing.</p>
	<p>(Social) Q: What if the skier was a stranger to the two people standing still?</p> <p>Correct: The skier does not throw the snowball.</p> <p>Wrong: The skier still throws the snowball.</p>
	<p>(Physical) Q: what if the wheel was in a bike?</p> <p>Correct: He would need to take out the screw before being able to set the wheel on the table</p> <p>Wrong: He would set the whole bike along with the wheel on the table.</p>
	<p>(Physical) Q: What if I let the cutting board lie on the counter?</p> <p>Correct: The cutting board would be dried slower.</p> <p>Wrong: The cutting board would be dried quicker as it occupies a larger area.</p>

Table 3.1: Sample data points of the ACQUIRED dataset.

frequency/speed of motions. Although some questions can be answered using more than one commonsense dimension, we ask the annotators to label with the main one used.

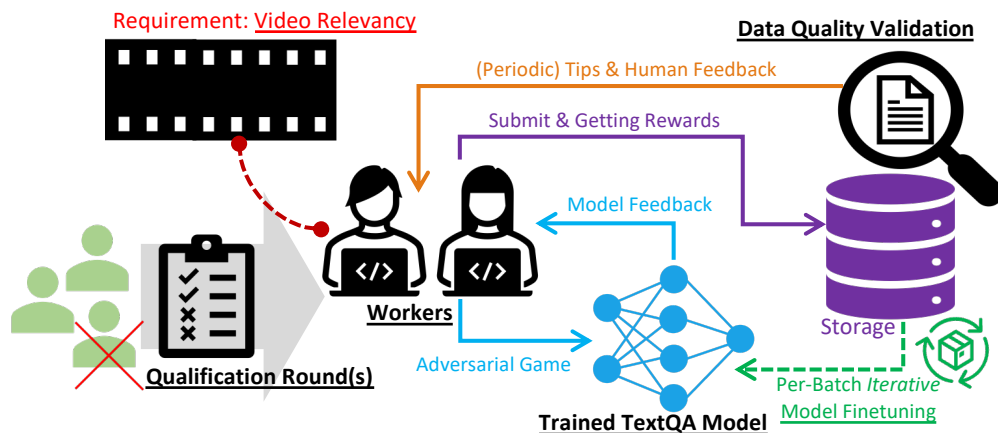


Figure 3.1: Data collection workflow.

3.1.3 Video Resources & Sampling

We utilize the Oops! [ECV20] dataset for third-person view videos and Ego4D [GWB22] for first-person views, where both of which feature text descriptions of the video contents. Oops! concerns predicting the failing (oops) moment of an intended action in a video, and hence is event-rich and a good testbed for reasoning what could the outcomes turned out differently. Ego4D collects videos of humans performing daily activities in the first-person view, which adds a desirable task-knowledge layer on top of its event-richness.

As we are annotating subsets of videos from the aforementioned sources, we have the privilege to encourage a more balanced *key events* distribution from the videos to be annotated. Specifically, we (1) use NLP tools such as semantic role labeling (SRL) to extract key verbs (events) for each video description², and group the videos accordingly, (2) each time sample an event group with a probability inverse proportional to the current launched key event distribution, (3) sample a video from the event group in (2), and repeat until reaching a desired number of videos (to be annotated).

²We use the originally annotated narrations in Ego4D.

3.1.4 Collection Workflow

We collect our dataset via Amazon Mechanical Turk (MTurk). Each MTurk worker is asked to carefully *watch a given video* for creating the QA pairs. As depicted in Figure 3.1, our dataset collection process comprises four main steps: (1) We design a **qualification questionnaire** focusing on examining one’s understanding of the key concepts in our problem design, i.e., the concept of counterfactuality, the requirement of video relevancy, common sense reasoning dimensions, and what types of QA pairs are more desirable. (2) Once the workers pass the qualification test, they are directed to an interface where a **pretrained (text-only) QA model** is deployed in the loop of the QA creation process. Bonus monetary rewards are given if the deployed model fails to predict correctly the creations. (3) Internal members then conduct a **quality validation** on the created samples and provide customized tips and/or feedback to the workers for potential improvements. (4) Lastly, our deployed model is **iteratively finetuned** on the validated samples after each batch of annotations, which results in a constantly improved model to incentivize more challenging sample creations.

3.1.4.1 Details of Human Annotations

We build a user interface to collect QA pair annotations. In addition, we collect human performance as a benchmarking source.

Our annotation interface (Figure 3.2b) is launched with Mturk tasks (Figure 3.2a). Upon accepting each Mturk HIT, our workers will be directed to the annotation web app and do the rest of the task. Workers will be asked to create QA pairs for different domains and assign their T/F labels. If the QA pair successfully fools the model, a green tick will be shown.

3.1.5 Quality Validation

In order to further ensure the sample quality as well as summarize common mistakes to provide custom human feedback to the annotators, our internal members conduct the second-phase manual sample validation in conjunction with the deployed model results. We cross-validate the annotations among our internal members in the ramping-up phase to ensure quality. We also accumulate detailed guidelines from our manual validation process for providing effective feedback. After scaling up, we continue to validate the annotations via uniform subsampling across each annotator. Our validation criteria are well aligned as can be seen in the high 0.85 Kappa score for commonsense dimension agreements; and 0.91 overlapping ratios for video relevancy.³

3.1.5.1 User Interface

We use an internal validation interface with a question-answering setting to accept or reject a sample. This tool also allows us to fix wrong domain categorization and T/F labels annotated by the workers. Specifically, the validation questions include:

- Should we discard this question group from our dataset (repetitive / not fixable at all)?
- Does this question group need any editing to reduce ambiguity or to further fool the model?
- Check the T/F of the two sentences.
- Select the domain that you think this question group can be categorized into.
- select one of the type that you think this question group can be categorized into.
- To answer this question, do you need to refer to the video?

³We did not use Kappa score for video relevancy because there is an unbalanced "agreed" distribution of "yes" and "no" (22:1) in our validation results for this criteria, which would result in unfair Kappa score.

- Does this question group conform to our question format?

3.2 Dataset Statistics

General Statistics. Table 3.3 summarizes the essential statistics of the collected dataset, where Table 3.3a is for videos obtained from the Oops! [ECV20] dataset whereas Table 3.3b is for videos from Ego4D [GWB22]. The frame-per-second rate (FPS) of videos from either source is mostly 30.

Key Annotated Events. We plot the distributions of most frequent key verbs (for main event types) and nouns (for entities involved in events) in Figure 3.3a and Figure 3.3b, respectively, to have a rough visual inspection of the diversity of the created samples. The key verbs/nouns are firstly determined by the SRL parses of the question and answer sentences (separately considered), and followed by lemmatization. Both plots are summaries of the two video sources, and more plots broken down by video sources and comparisons with existing works are in the Section 3.3.

Deployed Model. Table 3.2 reports the model fooling rates in our collected data across the two data sources. We encourage our annotators to develop QA pairs that can successfully fool our model by setting up monetary rewards and unlimited trials.

Videos From	Avg. Fool Rate (%)	Avg. Fool Accuracy
Oops!	65.44	34.56
Ego4D	52.94	47.06

Table 3.2: **Deployed model fooling rates** during collection.

Type	Counts
Total Unique Videos	2,910
Total Unique QA-Pairs	8,712
Type-Token Ratio	0.0288
Physical / Social / Temporal (%)	34 / 33 / 33

Type	Mean	Std	Max	Min
Tokens in a Question	11.3	3.3	28	5
Tokens in an Answer	8.3	5.6	46	5
Video Frames (Count)	297.8	217.4	3283	74
Video Duration (sec)	10.7	7.5	111.6	3.2

(a) Videos from Oops!

Type	Counts
Total Unique Videos	979
Total Unique QA-Pairs	2,365
Type-Token Ratio	0.0257
Physical / Social / Temporal (%)	77 / 0 / 22

Type	Mean	Std	Max	Min
Tokens in a Question	11.4	3.4	20	6
Tokens in an Answer	9.6	5.7	37	5
Video Frames (Count)	398.6	54.9	572	270
Video Duration (sec)	13.3	1.8	19	9

(b) Videos from Ego4D

Table 3.3: General statistics of the two video domains.

3.3 More Details of The Dataset

Our dataset consists of a mixture of QA pairs collected from two data sources: Ego4d and Oops!. For each dataset split, we create an indexing `.json` file and summarize each QA instance with a video id (index), a domain (physical/social/temporal), a type (counterfactual), a question, a correct answer, a distractor, and a key to the correct answer and a video

link URL. Our official data release will encompass all the aforementioned essential fields.

3.3.1 Dataset Splits

We split our data into train/val/test based on the ratio 0.45/0.05/0.5, with each unique video only appearing in one split.

3.3.2 Word Distributions

Figure 3.4a and Figure 3.4b plot the most frequent verbs (mainly for events) and nouns (mainly for entities) distributions of the Oops! proportion of our dataset, while Figure 3.5a and Figure 3.5b plot the ones of the Ego4D proportions.

Figure 3.6a and Figure 3.6b are distributions of the CLEVRER dataset. Figure 3.7a and Figure 3.7b are distributions of the TrafficQA dataset. It can be seen from these charts, alongside Table 3.4 and Table 3.5 that the event types in both datasets are quite uni-modally towards their original intended domains (which is reasonable), with all four ratios much lower than those of our dataset.

Dataset	verb-token ratio	verb-token ratio
CLEVRER	0.0001	5.14e-6
TrafficQA	0.0053	0.0004
ACQUIRED	0.0687	0.0047

Table 3.4: **Verb-token ratio** (total # verb-types / total # tokens) of CLEVRER, trafficQA and ACQUIRED

Dataset	noun-token ratio	noun-token ratio
CLEVRER	0.0002	2.57e-5
TrafficQA	0.0036	0.0006
ACQUIRED	0.1133	0.0089

Table 3.5: **Noun-token** ratio (total # noun-types / total # tokens) of CLEVRER, TrafficQA and ACQUIRED

OOPS-QA Annotations

* Please Make Sure You Read ALL the Instructions Below Before Doing the HIT!

Task Description
<p>In this HIT, we ask you to create video-related questions, i.e., questions aroused by watching a specific video, as well as the answers to the created questions. Specifically, we ask you to create one true (correct) answer and one false (incorrect or distractor) answer for each question. The questions (along with their answers) will be categorized into physical/social/time domains, with only one type: counterfactual type.</p> <p>Your goal is to write challenging video question-answer pairs that can be answered by humans, however, make sure that the correct answers cannot be inferred without watching the videos.</p> <p>We deploy an agent to detect if the answers can easily be induced from texts alone and if repetitive patterns have occurred multiple times.</p>
Payment
Task Details

Video 1

Intended Goal ver1: A person takes a bowl

Failure ver1: N/A

Intended Goal ver2: A person walks towards the cooker

Failure ver2: N/A



The link here will work only if you accept this HIT.

* And Remember That You Need To Copy Back A Confirmation Code Below!

(a) Human Annotation Instruction

Counterfactual Time

Please provide a counterfactual time question for the given video prompt:

What would happen if the second man came to the first man before he jumped?

Please provide the first response for the above question and whether it is a True/False answer:

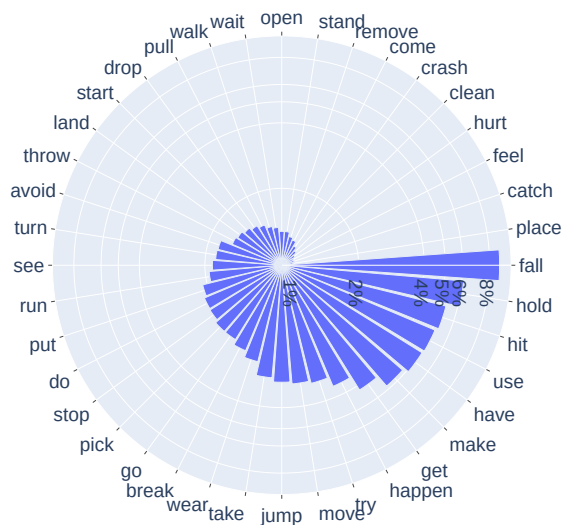
If the second man came to the first man before he jumped, the first man would still have jumped into the puddle. T F

And the other one with an opposite label:

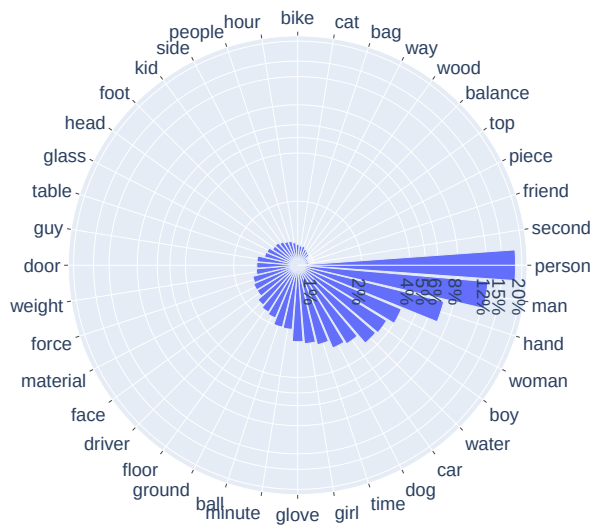
If the second man came to the first man before he jumped, the first man would not have jumped into the puddle. T F

(b) Sample Annotation Interface

Figure 3.2: **MTurk Annotation User Interface:** (a) We ask workers to follow the indicated instruction. All the blue-colored text bars on the top of the page are expandable. Workers can click to expand them for detailed instructions of the annotation task. (b) We design an user-friendly and interactive annotation tool where annotators and simply input their annotations and get an instant feedback from our model.

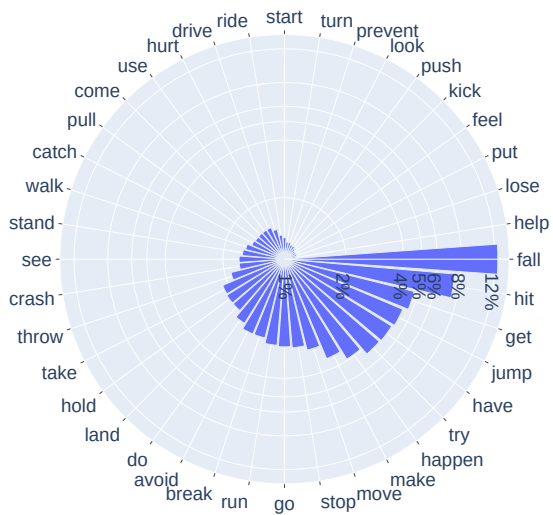


(a) Verbs

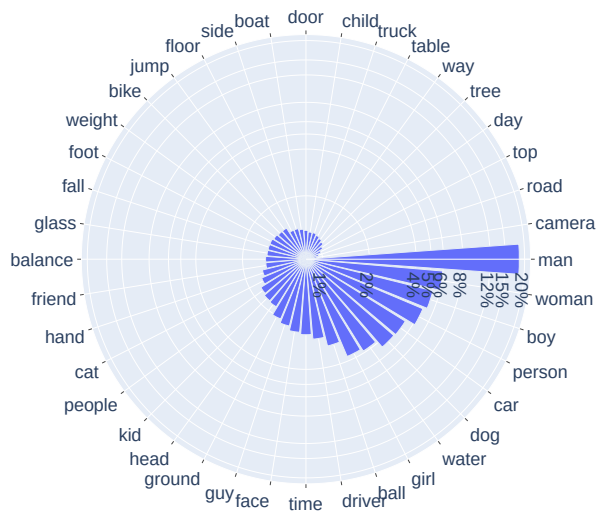


(b) Nouns

Figure 3.3: **Top-40 frequent word-types** in the dataset.

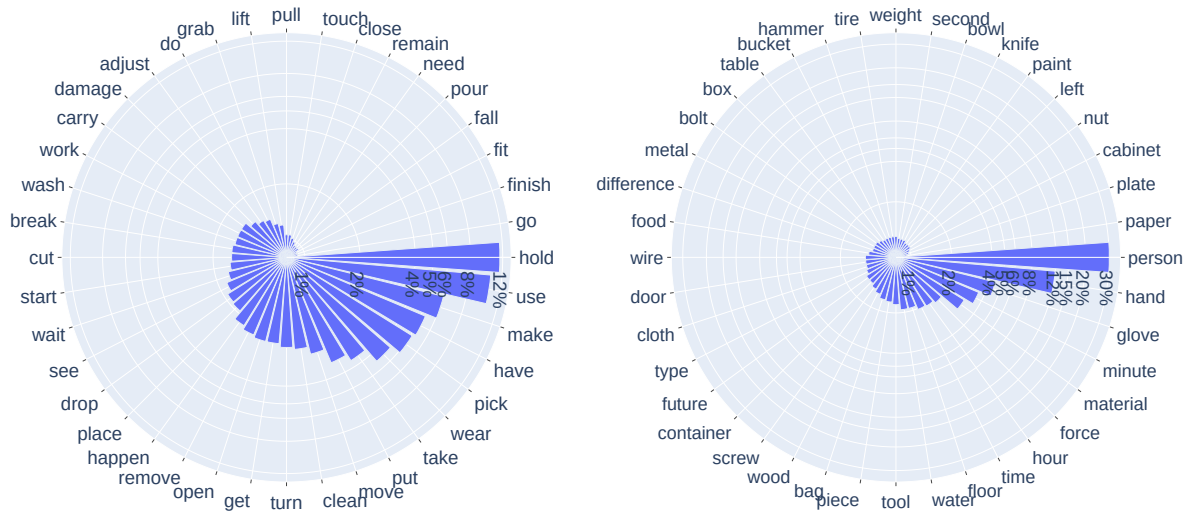


(a) **Top-40 frequent verbs** in Oops!.

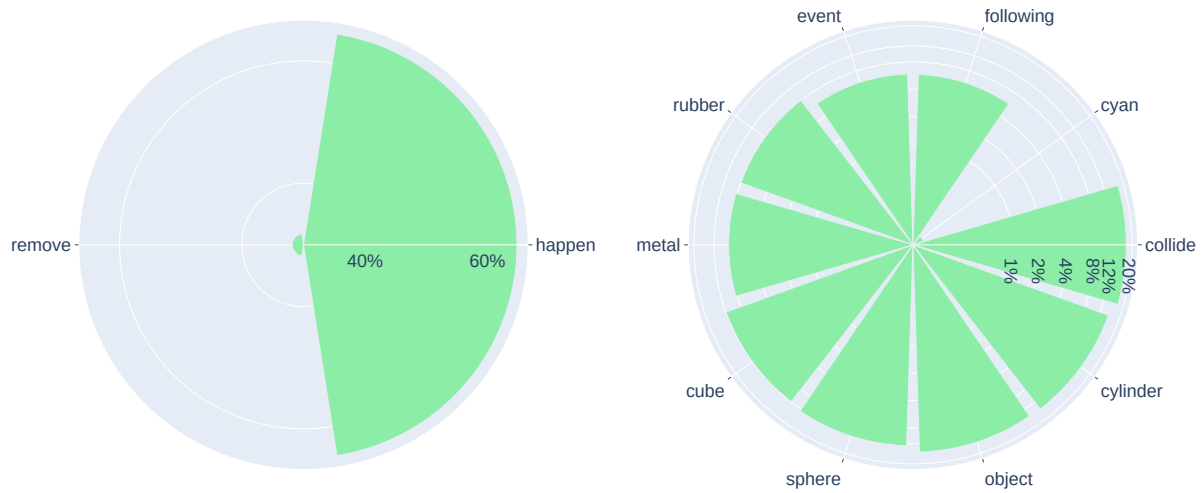


(b) **Top-40 frequent nouns** in Oops!.

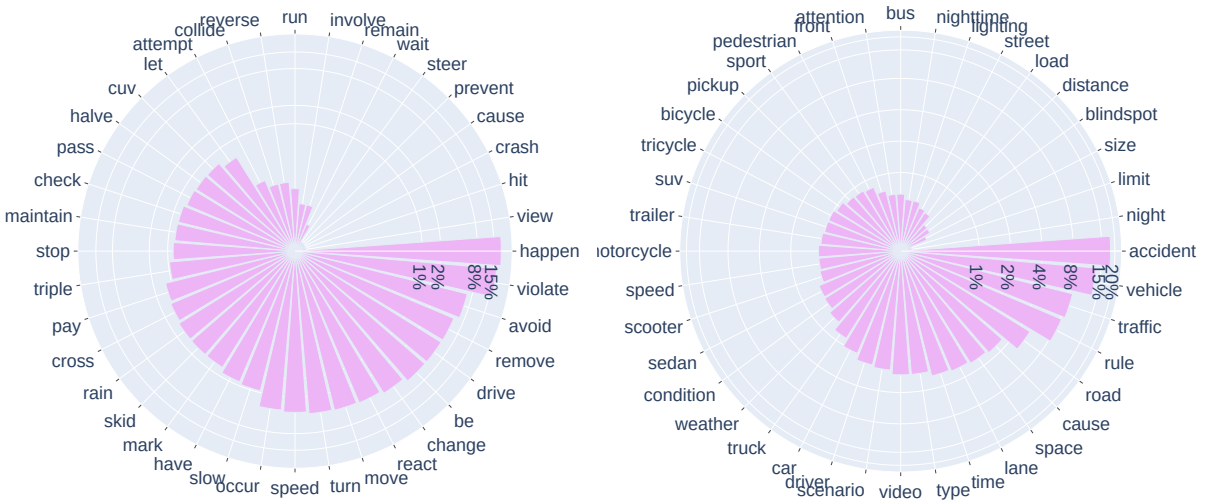
Figure 3.4: **Top-40 frequent word-types** in Oops! part of ACQUIRED.



(a) **Top-40 frequent verbs** in Ego4d. (b) **Top-40 frequent nouns** in Ego4d.
 Figure 3.5: **Top-40 frequent word-types** in Ego4d part of ACQUIRED.



(a) **Top-40 frequent verbs** in CLEVRER. (b) **Top-40 frequent nouns** in CLEVRER.
 Figure 3.6: **Top-40 frequent word-types** in CLEVRER.



(a) Top-40 frequent verbs in TrafficQA. (b) Top-40 frequent nouns in TrafficQA.

Figure 3.7: Top-40 frequent word-types in TrafficQA.

CHAPTER 4

Experiments & Results

We benchmark our dataset with both state-of-the-art language-only and vision-language models. Specifically, we perform experiments with language-only models: DeBERTa [HLG21], UnifiedQA [KMK20] and multimodal models: VIOLET [FLG21], VALOR [CHG23], and VL-Adapter [SCB22] on our dataset. VIOLET is a video-language model that has three components, including a video encoder (Swin Transformer-base [LNC22]), a language encoder (BERT-base [DCL19]), and a cross-modal transformer module that performs cross-modal fusion. VALOR is a recently proposed multimodal model that can take video, language, as well as audio as inputs. VL-Adapter uses a pretrained vision encoder (e.g. CLIP [RKH21]) to extract vision features and feed the vision features as well as text tokens to a pretrained language model (e.g. T5 [RSR20]) so that the model can take both vision and language information.

Inspired by the superior performance of the recent large language models, i.e., the GPT model from OpenAI, we also evaluate its zero-shot performance on the textual parts of our dataset. Specifically, we consider both ChatGPT [Ope23a] and GPT-4 [Ope23b]. In addition, we further include a version of GPT models that can condition on pre-annotated descriptions describing the general contents of the videos, to serve as the pseudo visual (and situated) contexts of the questions.

4.1 Experimental Setup

Data Splits. For our official (to-be-released) dataset, we follow a 45 – 5 – 50 ratio and randomly split the train-development-test datasets. The train split is mainly to adapt models to our QA task settings as well as the counterfactual reasoning style. We ensure that there are no overlaps between videos of different sets and the Oops! and Ego4D videos are equally distributed in each of the splits.

Evaluation Metrics. Models are evaluated by a simple accuracy metric, for both *T/F* and *MCQ* settings. We also further ablate the model performance along the commonsense dimensions and/or viewpoints, for a more detailed performance breakdown and analysis. We also include the pairwise accuracy in the T/F setting following [SWH21], where the model is considered correct if both individual judgments are correct in each pair.

4.2 Experimental Results

Table 4.2 reports benchmark performance. The best-performing multimodal model (VL-Adapter) performs slightly better than its text-only counterparts, UnifiedQA-large (i.e., the language encoder of our VL-Adapter). While this shows that visual contexts and multimodality are effective, the performance gap is not substantial; therefore, there is room for improvement, and more effective methods of multimodal inputs are yet to be explored. While text-only UnifiedQA-3B achieves overall better performance in both T/F and MCQ settings, potentially due to its much larger learnable parameter space, its mediocre pairwise accuracy suggests that the model is still inept at robust counterfactual reasoning in the two facets of the same question.

In general, models perform better in the MCQ settings than the T/F ones. This is intuitive because in the MCQ settings, the model is aware that only one of the two given options is correct and only needs to compare them and select the more reasonable option.

4.3 Discussion

4.3.1 Multimodal Models Performance

It appears that ChatGPT and GPT-4 fed in with video description perform better than the vanilla versions which indicates that the visual context is important. However, our results show that the multimodal models perform slightly worse than UnifiedQA and DeBERTa. We have a few hypotheses on the performance drop. First, we don't control the domain of the dataset for the pretraining. It's possible that there are nontrivial domain gaps across pretraining datasets and our dataset. At the end of the day, both UnifiedQA and DeBERTa are QA models, so probably not too surprising to contain more commonsense and can perform QA better. Second, because of the limited number of frames we can feed to the models, we uniformly sampled a number of frames from the video clips. These subsampled frames may suffice the need for simple inference such as activity recognition. But the performance of reasoning could be affected negatively if the subsampling misses key events which results in incomplete context. Third, our dataset focuses on counterfactual questions which may emphasize more on the reasoning capability depending on fine-grained visual understanding and potentially some imagination. Yet the selected models, such as VALOR, VIOLET, were trained more on alignment, grounding, and simple factual/forward (i.e., non-counterfactual) reasoning. Table 4.1 shows an example of how videos and sentences are fed into different models. In this particular tricky setting, VL-Adaptor performs better than the other three language-only models, which supports the hypothesis that visual context is a crucial part of reasoning.

Model	Input	Output	Correct?
VL-Adaptor	<p>What if the girl had been wearing a helmet? If the girl had been wearing a helmet, she might have had a concussion.</p> 	True	Yes
ChatGPT	<p>The answer to the question "What if the girl had been wearing a helmet? " is "If the girl had been wearing a helmet, she might have had a concussion." True or False?</p>	False	No
desc. Chat GPT	<p>The video is about "A young girl is riding her sled lying down on her stomach." The answer to the question "What if the girl had been wearing a helmet? " is "If the girl had been wearing a helmet, she might have had a concussion." True or False?</p>	False	No
UnifiedQA-3B	<p>The answer to the question "What if the girl had been wearing a helmet? " is "If the girl had been wearing a helmet, she might have had a concussion." True or False?</p>	False	No

Table 4.1: Exemplar inputs and outputs in T/F setting: 10 frames of the original video clips are subsampled and fed into VL-Adaptor.

4.3.2 ChatGPT and GPT-4

In the case of ChatGPT, its MCQ setting accuracy is lower than that of the T/F setting compared to others. We suspect that ChatGPT might have a weaker reasoning ability compared with GPT4. We observe that often ChatGPT refuses to give an answer in the MCQ settings because of insufficient conditions while it leans towards false when it was asked the same question in a T/F setting.

Perhaps surprisingly, despite the remarkable capabilities of the GPT series, they do not perform as impressively, even when provided with descriptions transcribing the major visual events in the videos. This suggests that the annotators in our curation task indeed closely examine many visual details in order to create more challenging samples.

4.3.3 Commonsense Dimensions

The rightmost parts of Table 4.2 report the performance breakdown along commonsense reasoning dimensions. We observe a general trend: most of the models perform better in physical and social dimensions compared to the temporal dimension; the physical dimension generally exhibits the highest performance. That observation implies that, even after being finetuned on our dataset, the models still fall short of capturing temporal commonsense as opposed to the other two kinds of knowledge. This can also be hypothetically attributed to the fact that the pretraining data for the language models encapsulate more physical and/or human social knowledge.

4.3.4 Viewpoints

We take the best-performing multimodal model (VL-Adapter) and ablate its performance along different video viewpoints. We find that, despite being pretrained mostly on third-person viewpoint videos, the generalization ability of the models towards first-person viewpoints is sufficiently good. However, as the videos from Ego4D are not intended to explicitly

contain failed actions from the camera wearers, it could be more challenging for our annotators to construct diverse and subtle counterfactual questions as compared to the videos from Oops!. Nevertheless, we argue that the counterfactual reasoning ability of the models should be equally crucial regardless of video viewpoint, and our dataset can inspire relevant research serving as a first-of-its-kind counterfactual video QA encapsulating videos from varying viewpoints.

4.3.5 Human Performance

We randomly sub-sample 500 videos to estimate human performance: these are reported in the last two rows of Table 4.2. The human performance highlights a significant gap above all the model results, especially for the MCQ settings. We hope future modeling endeavors can close the gap in visual counterfactual reasoning.

4.4 More on GPT Baselines

The prompts engineered for both ChatGPT and GPT-4 baselines are shown below:

```
1 # T/F setting without video description
2 GPT Prompt = The answer to {Question} is {Correct/wrong Answer}, True or
   False?
3
4 # T/F setting with video description
5 GPT Prompt = The video is about {Video description}, the answer to {
   Question} is {Correct/wrong Answer}, True or False?
6
7 # MCQ setting without video description
8 GPT Prompt = Which of the following is the correct answer to {Question}? (
   a) answer 1 (b) answer 2
9
10 # MCQ setting with video description
```

11 GPT Prompt = The video is about {Video description}. Which of the following is the correct answer to {Question}? (a) answer 1 (b) answer 2

Modality	Model	QA-Format	Viewpoints	Overall Accuracy \uparrow (%)	Dimension Breakdowns			
					Physical	Social	Temporal	
Text-Only	DeBERTa-V3	T/F	—	70.12	70.61	70.32	69.19	
		MCQ	—	70.35	72.10	68.62	69.01	
	UnifiedQA-base	T/F	—	68.93	70.22	69.32	66.33	
		MCQ	—	67.63	68.53	69.01	65.13	
	UnifiedQA-large	T/F	—	69.59	71.00	69.88	67.18	
		MCQ	—	70.38	71.57	71.83	67.38	
	UnifiedQA-3B	T/F	—	70.49	70.58	72.20	68.99	
		T/F (Pair.)	—	54.91	55.31	56.21	53.26	
		MCQ	—	73.40	73.36	75.80	71.60	
	Vanilla ChatGPT	T/F	—	52.80	51.36	48.06	54.04	
	Desc.-ChatGPT	T/F	—	55.20	50.82	52.90	52.48	
		MCQ	—	42.40	36.96	43.22	47.83	
	Vanilla GPT-4	T/F	—	53.80	53.89	53.16	54.32	
	Desc.-GPT-4	T/F	—	56.20	55.00	58.23	55.56	
MCQ		—	60.80	61.41	55.48	65.22		
Multimodal	VIOLET	T/F	All	66.15	70.20	64.45	60.24	
		T/F (Pair.)	All	48.25	54.03	44.60	40.63	
		MCQ	All	69.33	70.20	70.23	67.19	
	VALOR	T/F	All	63.83	66.54	62.50	60.02	
		T/F (Pair.)	All	43.00	46.51	42.46	37.26	
		MCQ	All	55.06	58.28	51.76	51.69	
	VL-Adapter			All	68.75	71.56	67.94	64.40
			T/F	3rd	66.32	66.01	67.90	65.07
				1st	72.63	75.49	—	62.82
		T/F (Pair.)		All	51.19	54.27	49.56	47.74
				3rd	47.82	47.60	49.50	46.40
				1st	60.40	62.23	-	53.44
		MCQ		All	71.53	72.70	70.39	70.25
				3rd	69.13	67.63	70.35	69.48
			1st	75.34	76.29	—	72.05	
Human	T/F	All	83.60	81.82	100	77.27		
Performance	MCQ	All	92.59	90.91	100	90.91		

Table 4.2: Model benchmarking performance on our ACQUIRED dataset.

CHAPTER 5

Limitations & Future Work

We hereby discuss the potential limitations of our work:

(1) Our work focuses on the three commonsense dimensions: physical, social, and temporal. While they likely span the most common types of the reasoning technique, there could be more, e.g., numerical commonsense is not specifically dealt with in this work, nor is non common activities such as fantasies and fictions involved. For future models benchmarked against our dataset, this should be taken account for, i.e., should the models excel at these commonsense dimensions for counterfactual reasoning, we cannot guarantee it is a complete model on all types of reasoning scheme.

(2) The videos used in this work are subsets of readily collected ones from both Oops! [ECV20] and Ego4D [GWB22] mother sets, and hence the event distribution can be bounded by the activities they concern. While we argue that the dataset is, to our best knowledge, first of its kind video QA dataset in terms of diversity and dedication of counterfactual reasoning, the video resources spanning even more diversified situations can be further extended. We will release the manuscripts and our collection tools to help spur future relevant research in such endeavours.

(3) Unlike Oops!, there is not an obvious failed actions occurred in Ego4D, and hence the annotated questions could be confounded by more imagined situations. We argue that the required reasoning technique is essentially the same and the models learn on our dataset should generalize well to situations that actually involve failing actions from egocentric visual contexts. However, we encourage future research to extend the first-person viewpoint (ego-

centric) parts to encompass obvious failing actions to collect just-in-time assistive questions and their corresponding remedial responses.

CHAPTER 6

Conclusion

In this work, we present a novel counterfactual-reasoning-focused video question answering dataset, named ACQUIRED. The dataset provides questions about counterfactual hypotheses over visual events (videos). We collect a correct and a distractor answer for three commonsense reasoning dimensions: physical, social, and temporal. We benchmarked various state-of-the-art language models (including LLMs like GPT) and video-language models on the collected dataset, where the results demonstrate algorithm performance well below human performance ($>13\%$ accuracy).

We hope our studies and the collected ACQUIRED dataset can spur relevant future research, specifically on testing multimodal models' capabilities in counterfactual reasoning, devising assistive AI for remedial and/or cause estimation of observed failures, and more sophisticated visual event understanding and reasoning.

REFERENCES

- [AAL15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. “VQA: Visual question answering.” In *International Conference on Computer Vision (ICCV)*, 2015.
- [AKL16] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. “Youtube-8M: A large-scale video classification benchmark.” *arXiv preprint*, 2016.
- [ATP20] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. “Counterfactual vision and language learning.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [BNM20] Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. “CoPhy: Counterfactual Learning of Physical Dynamics.” In *International Conference on Learning Representations (ICLR)*, 2020.
- [CHG23] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. “VALOR: Vision-audio-language omni-perception pretraining model and dataset.” *arXiv preprint*, 2023.
- [DCD21] Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. “Dynamic visual reasoning by learning differentiable physics models from video and language.” *Neural Information Processing Systems (NeurIPS)*, 2021.
- [DCL19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 4171–4186, 2019.
- [ECV20] Dave Epstein, Boyuan Chen, and Carl Vondrick. “Oops! predicting unintentional action in video.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [FLG21] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. “VIOLET: End-to-end video-language transformers with masked visual-token modeling.” 2021.
- [GKS17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. “Making the v in vqa matter: Elevating the role of image understanding in visual question answering.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

- [GLW20] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. “Multi-modal graph neural network for joint reasoning on vision and scene text.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [GWB22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. “Ego4d: Around the world in 3,000 hours of egocentric video.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [HAR17] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. “Learning to reason: End-to-end module networks for visual question answering.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [HLG21] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. “DeBERTa: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION.” In *ICLR*, 2021.
- [HM18] Drew A Hudson and Christopher D Manning. “Compositional Attention Networks for Machine Reasoning.” In *International Conference on Learning Representations (ICLR)*, 2018.
- [HM19] Drew A Hudson and Christopher D Manning. “GQA: A new dataset for real-world visual reasoning and compositional question answering.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [JHV17] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. “CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [JSY17] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. “TGIF-QA: Toward spatio-temporal reasoning in visual question answering.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [KCS17] 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*, 2017.
- [KMK20] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. “UnifiedQA: Crossing Format Boundaries

- With a Single QA System.” In *Findings of the Association for Computational Linguistics: EMNLP*, 2020.
- [LCC20] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. “HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training.” In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- [LLZ21] Jie Lei, Linjie Li, Luwei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. “Less is more: ClipBERT for video-and-language learning via sparse sampling.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [LNC22] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. “Video swin transformer.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [LYB18] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. “TVQA: Localized, compositional video question answering.” In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- [LYB20] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. “What is More Likely to Happen Next? Video-and-Language Future Event Prediction.” In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- [MHJ17] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. “MarioQA: Answering questions by watching gameplay videos.” In *International Conference on Computer Vision (ICCV)*, 2017.
- [Ope23a] OpenAI. “ChatGPT.”, 2023.
- [Ope23b] OpenAI. “GPT-4 Technical Report.”, 2023.
- [PSD18] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. “FiLM: Visual reasoning with a general conditioning layer.” In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.
- [QBH19] Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. “Counterfactual Story Reasoning and Generation.” In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- [RKH21] 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.” *arXiv preprint arXiv:2103.00020*, 2021.

- [RSR20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. “Exploring the limits of transfer learning with a unified text-to-text transformer.” *Journal of Machine Learning Research (JMLR)*, 2020.
- [SCB22] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. “VL-adapter: Parameter-efficient transfer learning for vision-and-language tasks.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [SCN20] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. “A short note on the kinetics-700-2020 human action dataset.” *arXiv preprint*, 2020.
- [SMV19] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. “VideoBERT: A joint model for video and language representation learning.” In *International Conference on Computer Vision (ICCV)*, 2019.
- [SRB17] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. “A simple neural network module for relational reasoning.” In *Neural Information Processing Systems (NeurIPS)*, 2017.
- [SWH21] Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. “COM2SENSE: A Commonsense Reasoning Benchmark with Complementary Sentences.” In *Association for Computational Linguistics (ACL)*, 2021.
- [SZL19] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. “Explainable and explicit visual reasoning over scene graphs.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [TZS16] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. “MovieQA: Understanding stories in movies through question-answering.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [VWB15] Nicole Van Hoeck, Patrick D Watson, and Aron K Barbey. “Cognitive neuroscience of human counterfactual reasoning.” *Frontiers in human neuroscience*, 2015.
- [WFH21] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. “Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue.” 2021.

- [XHL21] Li Xu, He Huang, and Jun Liu. “SUTD-TrafficQA: A question answering benchmark and an efficient network for video reasoning over traffic events.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [XZX17] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. “Video question answering via gradually refined attention over appearance and motion.” In *ACM International Multimedia Conference (ACM MM)*, 2017.
- [YGL20] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. “CLEVRER: Collision Events for Video Representation and Reasoning.” In *International Conference on Learning Representations (ICLR)*, 2020.
- [YWG18] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. “Neural-symbolic vqa: Disentangling reasoning from vision and language understanding.” In *Neural Information Processing Systems (NeurIPS)*, 2018.
- [YWS21] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. “Counterfactual zero-shot and open-set visual recognition.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [ZBF19] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. “From recognition to cognition: Visual commonsense reasoning.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [ZCL19] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. “Social-IQA: A question answering benchmark for artificial social intelligence.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [ZLH21] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. “Merlot: Multimodal neural script knowledge models.” In *Neural Information Processing Systems (NeurIPS)*, 2021.
- [ZLL22] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohamadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. “Merlot reserve: Neural script knowledge through vision and language and sound.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [ZY20] Linchao Zhu and Yi Yang. “ActBERT: Learning global-local video-text representations.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.