Navigating Hallucinations for Reasoning of Unintentional Activities

Anonymous ACL submission

Abstract

In this work we present a novel task of understanding unintentional human activities in videos. We formalize this problem as a reasoning task under zero-shot scenario, where given a video of an unintentional activity we want to know why it transitioned from intentional to unintentional. We first evaluate the effectiveness of current state-of-the-art Large Multimodal Models on this reasoning task and observe that they suffer from hallucination. We further propose a novel prompting technique, termed as Dream of Thoughts (DoT), which allows the model to navigate through hallucinated thoughts to achieve better reasoning. To evaluate the performance on this task, we also introduce three different specialized metrics designed to quantify the models reasoning capability. We perform our experiments on two different datasets, OOPs and UCF-Crimes, and our findings show that DOT prompting technique is able to outperform standard prompting, while minimizing hallucinations.

1 Introduction

Automatic understanding of human activities in videos is a challenging problem with a lot of real-world applications in domains such as healthcare, security, robotics, and elderly assistance. In past few years, we have seen an impressive progress in recognizing intentional human activities in videos [14]. However, human beings are prone to making mistakes and activities can be unintentional in real-world scenarios. Recognizing unintentional activities is important [9], but it is also important to understand the reasoning behind their occurrence. This can be useful for correcting mistakes and any damage control. Motivated by this, in this work we focus on finding the reasoning behind unintentional activities in videos.

Recently developed multimodal foundation models have shown impressive capabilities across a range of tasks with strong generalization capabilities for zero-shot scenarios [2, 16–18, 23, 42, 47]. We first study the reasoning abilities of existing Large Multimodal Models (LMMs) using prompting to determine the intentionality of actions as we transition to unintentional states. Our analysis reveal that conventional prompting techniques suffer from hallucinations and does not perform well in reasoning about the transition into unintentional activities. We also noticed that even when model is able to identify that the transition from intentional to unintentional has occurred it frequently provided very generic reasons without using the visual context to the fullest extent. Although chain of thoughts [38] prompting provides a framework to obtain specific reasons not just generic ones, it also suffers from hallucinations when trying to reason over unintentional activities.

To mitigate the effect of hallucinations and improve the reasoning over unintentional activities, we propose a multi-step solution. Our solution relies on two key observations; 1) if we let a model hallucinate multiple times, some of the responses might be correct, and 2) multiple-choice questions helps guide the model to find the right answer. We build upon these observations and propose a novel approach termed Dream of Thought (DoT) style prompting. We use the models hallucinations and present to the model as multiple choices and let the model navigate through these choices and provide correct reasoning.

We experiment with two different datasets, OOPs [9] and UCF-Crimes [32], where OOPs focus on unintentional activities in daily life and UCF-Crimes focus on anomalous activities. With extensive evaluations we demonstrate the effectiveness of DoT prompting over simple prompting and chain of thoughts prompting. We make the following contributions in this work,

- We present a novel problem that focuses on reasoning about the transition of an activity from intentional to unintentional.

- We study the capability of existing LMMs and prompting techniques for this task and also provide a novel Dream of Thoughts (DoT) reasoning-based mechanism which outperforms existing methods.
- We provide three different evaluation protocols, rm_{MCQ}, rm_{LLM}, and rm_{FIB}, for response matching (rm) which quantifies the reasoning capability of models for this task.

2 Related works

Large generative models The field of large language models (LLMs) has significantly evolved in recent years, with advanced models like GPT [5], LLaMA [33], ChatGPT [30], and BARD [11]. These models excel at generalizing across various tasks.Emerging Large Multimodal models, derived from these foundational LLMs, are now being explored for vision tasks. Examples include MiniGPT [47], Open Flamingo [2], BLiPv2 [16], and LLaVA [18] in the image domain, and Video LLaMA [42], Video Chat [23], and Video ChatGPT [17] in the video domain. We use these state of the art Large Multimodal Models to study the prposed new task.
Prompting techniques The emergence of large language models (LLMs) and multimodal models has led to the development of techniques to enhance their zero-shot abilities. Notable advancements include the Chain of Thought (COT) prompting by Wei et al [37], Automatic Chain of Thoughts [44] and the Self-Consistent Chain of Thought [35] Zhang et al. [45] further evolved this concept into the Multimodal Chain of Thought, which incorporates both textual and visual data. Wang et al. [35] refined the original CoT approach using the self-consistency criteria. Yao et al. [41] and Long [22] further proposed through the Tree of Thought. The Graph of Thought by Liu et al. [21] expanded on these ideas. Incorporating examples for few-shot learning scenarios has also been shown to improve LLM performance [5, 33] which have been further enhanced upon by [15, 20, 31, 46]. We analyze these existing techniques capabilities to induce reasoning abilities in LMM's and compare with our proposed method.
Reasoning abilities of LLM's Web et al. [36] showed that models like GPT-3.5 and GPT-4 have considerable analogical reasoning abilities, while Liu et al. [19] highlighted their limitations with out-of-distribution data and complex tasks. Malkinski et al. [26] analyzed deep models of analytical reasoning on Raven's Progressive Matrices [36]. The Visual Question Answering (VQA) field has seen significant contributions from studies like [43], [25], [13], and [3], enhancing VQA solutions. Research by Xue et al. [39], Hafner et al. [12], Finn et al. [10], Chang et al. [7], Burda et al. [6], Babaeizadeh et al. [4], and Agrawal et al. [1] has been pivotal in advancing how deep models understand dynamic visuals. To the best of our knowledge LMM's ability to reason over unintentional videos has not been addressed in existing works. **Hallucination in LLM's**: Hallucination in foundational models refers to the creation of inconsistent responses. Mckenna et al. [27] investigated the origins of hallucinations in LLMs, while Yao et al. [40] drew comparisons between these hallucinations and adversarial examples. Wang et al. [34] extended this research to Large Vision Models, examining hallucinations in the visual domain. To address hallucination challenges, Dhuliawala et al. [8] and Manakul et al. [24] introduced self checking and self verification to generate consistent responses. In this work, we use hallucinations to improve the models reasoning capability with the help of multi-step navigation.

3 Method

Problem statement We focus on understanding the transition from intentional to unintentional activities in videos under zero-shot setting. Given a model $p()$ which takes a prompt $\mathcal{P}$ and a video $\mathbf{V}$ with n frames as input, the objective is to identify the reasoning $\mathbf{R}$ behind the activity's transition from intentional to unintentional in the video.

3.1 Background and motivation

The Chain of Thought (COT) prompting [38] method has been shown to enhance the reasoning abilities of LLMs in large-parameter models. Our preliminary experiments indicate that Large Video Language models face specific challenges due to hallucinations as well as lack of ability to infer relationships between events, which seems to be affecting inference and causal understanding.

While studying these issues, we observe that repeated trials substantially provide accurate responses occasionally, approximately achieving one correct response out of every few attempts with the CoT prompt. Moreover, in [28, 29] the authors show that humans also interpret problem-solving in a combinatorial manner, using some heuristics

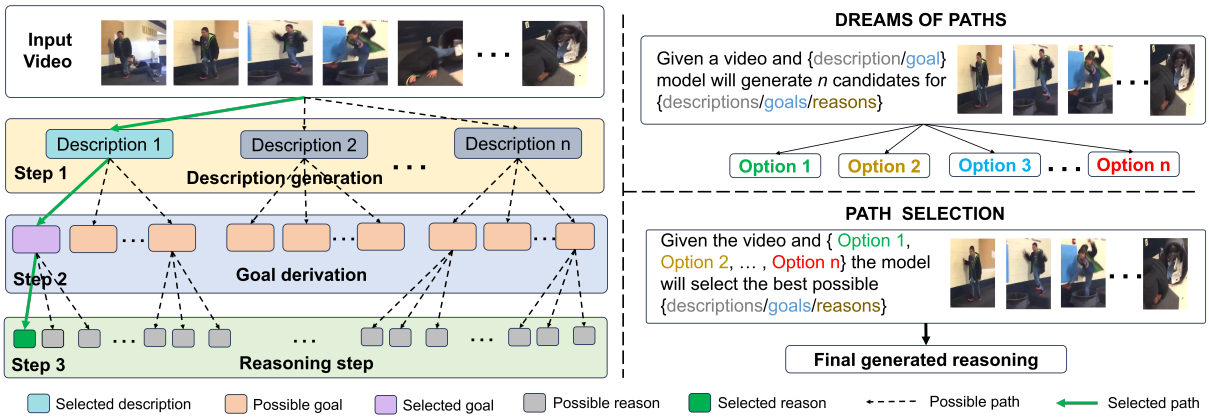

Figure 1: ***Overview of the proposed Dream of Thoughts framework:*** The left figure shows an overview of the three-step process with all the possible paths generated by the Large Video Language Model using the video and provided prompts. The right figure describes the Dream of Paths mechanism for generating thoughts to cover the most probable options and the Path Selection mechanism for navigating through the best possible options.

to decide from various possibilities. The possibilities at each stage are generated by our prior experience in solving problems, which also generates the plan to solve the problem. Motivated by this, we introduce a multi-step prompting strategy which exploits models hallucinations and attempt to navigate through those hallucinated responses to achieve better reasoning.

3.2 Proposed approach

We introduce Dream of Thought (DoT) prompting to improve the models' ability to generate the correct response by filtering through multiple responses. It is a multi-step process which consists of three steps to obtain cues consisting of essential components to obtain the reason. Specifically, we first obtain a description of the video and using this as the cue, we generate the goal of the intentional activity in the video, which enables us to reason why the intentional activity is failing. An overview of the proposed approach is shown in Figure 1.

At each step, DoT generates a range of possible answers (Dreams of Paths) to a given question. We then employ a Multiple Choice Question (MCQ)-style prompt for effective selection of the most appropriate response (Path Selection) to the specific video. This strategy capitalizes on the models' generative capability to provide diverse options, with the MCQ prompt acting as a filter to select the most appropriate output. Similar strategy has been explored in Tree of Thoughts (ToT) [41] mechanism but there are some key differences; 1) ToT requires a scoring mechanism to select the best possible option in each step, whereas, we pose this as MCQ for the model itself, and 2) our proposed DoT utilize cues from different steps as a context for next steps, whereas ToT treats each step as a partial path with no such motive.

DoT consists of three main steps, 1) generating description, 2) goal derivation, and 3) reasoning, which make use of Dream of Paths (DoP) and Path Selection. We will first describe Dream of Paths and Path Selection, and then explain the three steps involved in DoT prompting.

Dream of Paths: At each step, we generate n possible options as a solution to the task in corresponding step. The model $p()$ to generate n candidate solutions $x_i \sim p(x_i|V, \dots)$.

Path selection: After obtaining n possible solutions to our problem, we then propose the task as a MCQ form problem where the model has to select one out of n possible solutions: $x \sim p(x|x_1, \dots, x_i, P_s, V)$ using a prompt P_s, "*The list of possible descriptions/goals/reasons for the video are given as (descriptions/goals/reasons). Select the most appropriate descriptions/goals/reasons.*"

Generating description (*D*): In the first step, we generate n concise summaries of the video content using a prompt: $d_i \sim p(d_i|P_d, V)$,where prompt P_d is "*Summarize the video action and infer the list of objects exhaustively, from the relevant visual context to the activity occurring in the video.*". Following this, we engage in the Path Selection step to derive the most accurate description of the video: $d \sim p(d|d_1, d_2, \dots, d_n, V, P_s)$.

Goal derivation (*G*): Using the summary, we derive n possible intended activity to be executed within the context of this video using a prompt:

Algorithm 1 Dream of Thoughts (DoT)

Input: Model $\mathcal{M}$, video V_i
Output: Reasoning R

$P = [P_d, P_g, P_r]$ ▷ Define prompts for reasoning
$c = []$ ▷ Initialize empty list c for storing context
$n = N$ ▷ Set n to number of options to be generated
$P_s = SelectionPriompt$ ▷ Set the selection prompt
for j in P **do**
  $c_i = []$ ▷ Initialize empty list c_i
  for $i = 1$ **to** n **do**
    $c_i \mathrel{+}= model(c \mid P_j, V, c)$ ▷ Update c_i with model output
  end for
  $c \mathrel{+}= model(c \mid c_i, c, V, P_s)$ ▷ Update c with model output
end for
$R = c[-1]$ ▷ Set reason to the last element of c

$g_i \sim p(g_i|d, V, P_g)$, where prompt P_g is given as *"If the summary of the given video is <video summary>, logically infer the most probable intention of the actions being attempted in this video."*. We then perform the Path Selection step to obtain the best possible description for the video: $g \sim p(g|g_1, g_2, g_n, P_s, V, d)$.

Reasoning step (*R*): Utilizing the information pertaining to the intended activity, we generate a set of n probable factors that could have potentially hindered the successful completion of the aforementioned task: $r_i \sim p(r_i|V, g, P_r)$, using a prompt P_r, "*The goal of the intended activity taking place in the given video is described as: (goal), provide a visual description of the event that leads to the failure to perform the activity with the greatest probability.*" This step is again followed by the Path Selection step to obtain the best possible description for the video$r \sim p(r|r_1, r_2, r_n, P_r, V, g)$.

3.3 Evaluation and metrics

We perform comparison of the responses with the ground truth reasons at both high and low level context. For high level context analysis, we aim to match underlying reasons provided by the model with the ground truth reasoning. For this, we introduce the rm_{LLM} metric. For low level contextual analysis we measure how accurately the model can predict specific attributes of the reason such as subject, verb and object. We propose two metrics for this, rm_{MCQ}, and rm_{FIB}. Leveraging keyword-based metrics, we can more precisely assess the presence of hallucinations in these models. Specifically, if the keywords are absent, it suggests that hallucination may have occurred, where the keywords have either been replaced by synonyms or include hallucinatory details not originally present.

1) **Low level context evaluation**: The ground truth encompasses subject, object, and verb components extracted from the ground truth, denoted as s_i for the ith video. Our evaluation revolves around the identification of these "keywords" within the predicted responses. This evaluation is applied when the reasoning task is framed as either a multiple-choice question (MCQ) task, or a fill-in-the-blanks task. We experimented with existing metrics for generated text evaluation such as BLEU and Sacre BLEU, but these metrics were unable to match the responses providing most of the scores close to 0 therefore we do not use these metrics.

1.1) *MCQ evaluation:* For MCQ style task, since we provide the ground truth option as one of the options and rest of the options are unrelated, the presence of keywords in the response provides a reasonable estimate of how correct the answer is and also allows us to judge the accuracy of the output. The rm_{MCQ} accuracy is obtained as,

$$rm_{MCQ} = \Sigma_{i=1}^{N} \mathbf{1}[s_i \in pred_i] \quad (1)$$

where $pred_i$ is the prediction given by the model for the ith video in the dataset. Here N is the total number of samples and $pred_i$ is the prediction provided by the model for the i_{th} video.

1.2) *Fill-in-blank evaluation:* In FIB style task since we are removing one of the possible keywords which has to be completed by the model we evaluate the number for keywords model is able to output correctly. We remove s_i from the ground truth reason gt_i.

$$rm_{FIB} = \Sigma_{i=1}^{N} \Sigma_{x_j \in s_i} \frac{\mathbf{1}[x_j \in pred_i]}{len(s_i)}, \quad (2)$$

Here N is the total number of samples, $pred_i$ is the predicted made by the model for the i_{th} video.

2) *Reasoning evaluation:* Finally, we evaluate the response provided by the models and match it with the ground truth answer. We make us of using GPT-3.5 for matching the generated and ground truth reason. This evaluation allows us to compare whether the output contains the event which occurs in the ground truth reason. We evaluate the same video five times and report the average score of each video as the rm_{LLM} and the standard deviation of scores per question as std.

4 Experiments

Datasets We performed our experiments on two different datasets, OOPs [9] and UCF-Crimes [32].

Models	MCQ				FIB			
	w goal		w/o goal		w goal		w/o goal	
	rm_{MCQ}	rm_{LLM}	rm_{MCQ}	rm_{LLM}	rm_{FIB}	rm_{LLM}	rm_{FIB}	rm_{LLM}
Video ChatGPT	0.303	0.667	0.240	0.457	0.352	0.648	0.222	0.519
Video LLaMA	0.105	0.092	0.099	0.054	0.383	0.139	0.167	0.206
Video Chat	0.315	0.204	0.278	0.067	0.337	0.226	0.215	0.214
Video LLaMAv2	0.134	0.072	0.040	0.067	0.184	0.059	0.293	0.214

Table 1: ***Reasoning capability of existing models:*** Performance evaluation of existing models on multiple-choice questions (MCQ) and fill-in-the-blank (FIB) style prompting. We analyze both scenarios, prompts with and without goals. MCQ setup consist of four questions, 1 ground truth, 2 random and '*None of the above*'.

OOPs: We conduct detailed experimental analysis using the validation subset of the OOPs dataset. This subset comprises 3,500 YouTube videos, each portraying a variety of failures in diverse real-world scenarios. Along with this, the OOPs dataset also contains natural language descriptions for each video. These descriptions provide insights into the original intentions behind the videos and the circumstances leading to the deviation from planned actions. **UCF-Crimes** Further, we also conduct experiments on UCF-Crimes dataset. It consists of long and untrimmed real-world surveillance videos, with 13 realistic anomalies such as fighting, road accident, burglary, robbery, etc. We use the validation set of this dataset to evaluate our approach, where we select only anomalous videos. These videos have length ranging from 1-3 minutes and there are a total of 65 videos in this evaluation set. We provide natural language descriptions for the crime occurring in the videos from this new test set to evaluate our approach.

Baselines and models For the evaluation and benchmark, we utilize the officially released versions of several state-of-the-art models, namely Video ChatGPT [23], Video LLaMA [42], Video Chat [17], and Video LLaMAv2 [42]. Along with these video-based models, we also use image based model, Open Flamingo [2]. These models serve as comprehensive baselines in our analysis. Further, we also evaluate different prompting strategies including standard prompting and CoT prompting. Each of these models is built upon the LLaMA-7b billion language model, endowing them with substantial capabilities in text generation from video inputs. For the proposed DoT approach, we use Video ChatGPT in all our experiments.

4.1 Quantitative results

We first analyze the reasoning capability of existing LMMs for explaining reasoning behind unintentional activities in videos. Here we explore two different prompting setups, 1) multiple choice questions (MCQs), and 2) fill-in-the-blanks. In MCQ style prompting with n = 3 options (more details in supplementary), we presented several options along with ground truth and prompted the model to select the correct reasoning for the failure. This is evaluated using rm_{MCQ} and rm_{LLM} metrics. In the second setup, we use the ground truth reasoning and randomly remove subject, object or verbs from the sentence and prompt the model to fill in the missing words. This is evaluated using rm_{FIB} and rm_{LLM} metrics.

The performance of studied models for MCQ and FIB style prompting is shown in Table 1. For both, we experimented with two variations, one where the goal is also provided along with the prompt and the other where goal is not provided. Video ChatGPT shows consistently better performance on both FIB and MCQ prompts for all three metrics with and without goal. Video LLaMA and LLaMAv2 show significantly worse performance on MCQ as compared to FIB-style prompts on rm_{MCQ}, rm_{FIB} and rm_{LLM}. Video Chat shows similar performance on rm_{MCQ} and rm_{FIB} but rm_{LLM} for FIB is higher in non-goal setting and similar in with goal setting. Based on this analysis, we experimented with mostly Video ChatGPT for proposed DoT prompting technique.

Next, we evaluate the existing and proposed methods for generating the complete reasoning. We evaluate both CoT and DoT prompting for Video ChatGPT as it was the best performing model in our preliminary experiments. This is evaluated using rm_{LLM} metric along with standard deviation in responses std, which attempts to measure degree of hallucinations in the response.

The evaluation for all the models is shown in Table 2 for both OOPs and UCF-Crimes dataset. We can observe that the proposed DoT prompting demonstrate benefits over existing methods surpassing both the standard and CoT prompts. DoT

Figure 2: ***Qualitative evaluations:*** We show some samples for qualitative analysis of the proposed DoT prompting compared with CoT and standard prompting. First row illustrates examples from OOPs dataset and the second row refers to examples sampled from UCF-Crimes dataset.

Dataset	OOPs		UCF-Crimes	
Model	rm_{LLM}	std	rm_{LLM}	std
Open Flamingo	0.154	0.128	0.035	0.047
Video LLaMA	0.026	0.048	0.075	0.072
Video Chat	0.064	0.156	0.082	0.143
Video LLaMA2	0.053	0.089	0.081	0.089
Video ChatGPT	0.242	0.217	0.247	0.171
CoT	0.236	0.182	0.271	0.182
DoT	0.279	0.199	0.291	0.160

Table 2: ***Performance evaluation:*** A comparison of existing methods with proposed DoT prompting on OOPs and UCF-Crimes dataset. We show both rm_{LLM} and standard deviation (std) across five trials. CoT refers to Chain of Thoughts and DoT refers to the proposed prompting strategy using VideoChatGPT model.

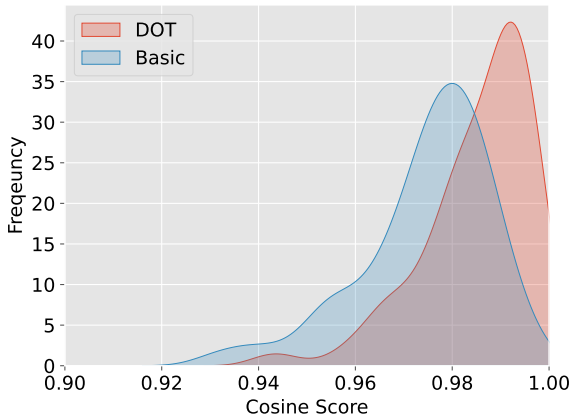

Figure 3: Distribution of cosine similarity between ground-truth and the DoT as well as basic prompt.

outperforms Basic prompts by $\sim 4\%$ Furthermore, Video ChatGPT outperforms Video LLaMA, Video LLaMAv2, and Video Chat models when subjected to basic prompts. Similar results can be observed for UCF-Crimes dataset.

Analyzing hallucinations: We provide insights into the standard deviation of scores across individual questions. High standard deviation implies inconsistent answers and substantial model hallucinations. Conversely, a low standard deviation, coupled with low accuracy, suggests consistent but incorrect responses, while a low standard deviation with high accuracy indicates consistent and correct answers. From Table 2 we can observe that DoT has lower std score than basic prompts by ~ 0.02 whereas it is comparable to that of COT, whereas CoT maintains low uncertainty but struggles to consistently achieve high scores when compared to DoT. Additionally, in Figure 3 we can see that the outputs obtained from DoT prompt display a consistently higher cosine similarity score to ground truth reason as compared to the output obtained from standard prompts (More details in supplementary).

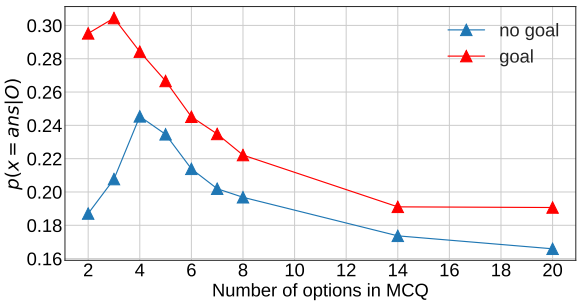

Figure 4: ***Effect of number of options:*** Variation of $p(x = ans|O)$ on reasoning task proposed as MCQ style query, with varying number of present in a MCQ question, where $p(x = ans|O) = 1 if rm_{mcq} >= 0.8$ else $p(x = ans|O) = 0$. Here O refers to the options presented in the MCQ.

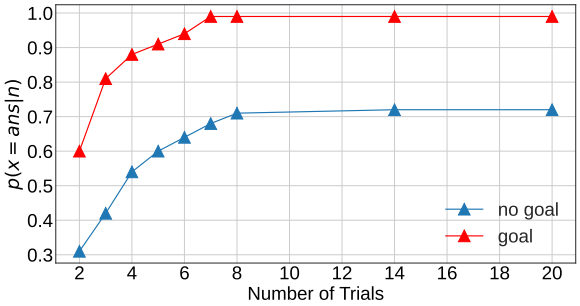

Figure 5: ***Analyzing number of trials:*** Variation of $p(ans \in x|n)$ on reasoning task proposed as MCQ style query, with n is the number of times prompt has been evaluated using LMM and x is set of n outputs obtained using LMM.

4.2 Qualitative Results

We present qualitative results on the OOPs and UCF-Crimes dataset in Figure 2. We can observe that DoT prompting is generating better reasoning for action failures as well reasoning behind the the activity being anomalous in videos, compared to Standard and CoT prompting. The DoT method is better aligned with ground truth reasoning, showcasing its capability across diverse activities such as typing, shooting an air gun. These activities highlight different success scenarios: ongoing success in working, and instant success in air gun shooting. It also demonstrates its effectiveness to identify a wide range of crimes like arson and vandalism showcasing its generalizability.

4.3 Ablation studies

We conduct ablation studies to assess the impact of prompt variations on both accuracy and the presence of hallucinations these ablations studies aid in evaluating the efficacy of each individual step

	with goal		w/o goal	
Model	rm_{LLM}	std	rm_{LLM}	std
Video ChatGPT	0.621	0.213	0.242	0.217
Video LLaMA	0.337	0.261	0.026	0.048
Video Chat	0.205	0.301	0.064	0.156
Video LLaMA2	0.033	0.032	0.053	0.089

Table 3: ***Effect of goal:*** Performance comparison of models on reasoning with provided goals.

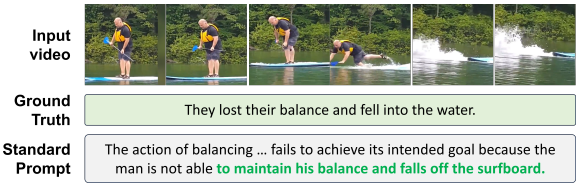

Figure 6: ***Role of visual information:*** We observe some interesting scenarios where the model using a standard prompt with goal of the video provided is able to infer the correct reasoning without any video frames.

within our proposed DoT prompting methodology

Effect of number of options: In MCQ-style question answering, we explore how varying the number of options in MCQs impacts models performance. As shown in Figure 4, we initially observe some gain of 3% and 6% for with and without goal settings respectively which is followed by a noticeable reduction of 12% in the average rm_{MCQ}, when the number of options is increased in both scenarios—with and without a defined goal. We hypothesize that the first increment is due the fact that more options allow the model to generate better options with more probability as shown in Figure 5, but then the performance decreases. This decrease is likely due to the broadening of the model's search space, resulting in more inaccuracies. The score becomes almost constant after 14 options for both with goal and without goal cases.

Effect of goal: Humans demonstrate an impressive ability to comprehend the reasoning behind actions when guided by contextual information. In this experiment, we introduce the goal of the attempted action as a part of the prompt. For this, we construct the prompt as Prompt: *"If the goal of the activity occurring in the video is (goal). Explain the reason behind the failure to achieve the desired goal.*".Analysis of the results, as presented in Table 1 and Table 3, reveals that the inclusion of goal enhances the reasoning capabilities of these models. We can see that the presence of goal increases the rm_{LLM} by 0.4 in Video ChatGPT and by 0.2 $\sim$ 0.3 for Video Chat and Video LLaMA models, whereas Video LLaMAv2 seems to perform worse

Model	rm_{LLM}	std
CoT	0.237	0.182
DoT(w/o des)	0.180	0.153
DoT(w/o goal,des)	0.221	0.182
DoT(rm_{FIB})	0.260	0.183
DoT	0.279	0.199

Table 4: ***Ablation Analysis of the DoT Prompt***.DoT(GPT):final path selection is performed using GPT-3.5. DOT(w/o des) refers to the case when we directly obtain description. Similarly, in DoT(w/o goal, des) we directly obtain goal and description. In DoT(rm_{FIB}) the path selection is performed using rm_{FIB}.

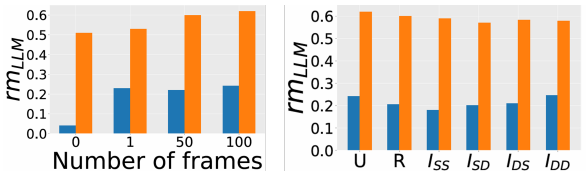

Figure 7: ***Effect of number of frames and sampling strategy:*** The left plot shows the effect of varying the number of sampled frames on rm_{LLM} for reasoning task. In the right plot we show effect of various frame sampling techniques in videos: U(uniform sampling), R(random sampling), I_{SS} (sparse sampling from both intentional and unintentional parts), I_{SD} (sparse from intentional, dense from unintentional), I_{DS} (dense from intentional, sparse from unintentional), and I_{DD} (dense sampling from both intentional and unintentional parts)

in both conditions.

Effect of Dream of Paths: We evaluated the effectiveness of Dream of Paths by modifying the prompt to exclude the Dream of Paths step for both descriptions and goals. The results, as shown in Table 4, reveal that removing this (DoT(w/o des)) leads to a significant decline in performance. This decrease can be attributed to the reliance on inaccurate descriptions for subsequent steps like goal determination and final reasoning, resulting in incorrect overall outcomes. Furthermore, generating a single option for both description and goal (DoT(w/o goal des)) shows marginally better performance compared to DoT(w/o des), yet it falls short of the complete DoT method.

Effect of Path Selection We compared our Path Selection procedure used in against the DoT(rm_{FIB}) approach, where we select the option with the highest rm_{FIB} at each stage, ensuring that the option mentioning the most objects involved in the video is chosen. Our results, as detailed in Table 4, show that using the FIB method, while resulting in a lower std, achieves a slightly lower performance compared to the base DoT by 2%.

4.4 Analysis

Number of video frames: We conduct an analysis on the effect of number of video frames to investigate their impact on models performance. We vary the number of frames, ranging from 0 to 1, 50, and 100 frames. Our observations, as depicted in Figure 7, reveal that the model's performance remains relatively stable concerning the number of frames but experiences a substantial drop when no frames are provided as input. Interestingly, for some scenarios (Figure 6) when merely a goal is provided to the model, it manages to achieve a significantly high rm_{LLM} using only the goal as information about the video, which shows that it utilizes textual conditioning more efficiently than visual modality.

Sampling strategy: Additionally, we explore variations in the frame sampling strategy, ranging from uniform and random sampling to importance sampling. Importance sampling involves selectively sampling frames sparsely or densely from the intentional and unintentional segments of the video. To execute importance sampling, we utilize timestamps provided for intentional and unintentional parts of the video with the OOPs dataset, sampling varying numbers of frames from the start of the video to the beginning of the transition, and from the start of the transition to its end. Our findings, presented in Figure 7, show that sampling strategies do not significantly affect the reasoning capabilities of Video ChatGPT, uniform sampling offers the best overall performance, followed by sampling frames densely from intentional and unintentional parts.

5 Conclusion

In this work, we present a novel task regarding understanding of unintentional activities in videos where we formalize it as a zero shot reasoning task. We first analyze the reasoning capabilities of existing LMM models and prompting techniques and then also propose a novel DoT prompting technique which navigates through hallucinations introduced by LLM's to obtain the reasoning. We propose different metrics to quantify the models performance and also analyze hallucinations of the responses. We further demonstrate that the proposed method outperforms existing prompting techniques.

6 Guidelines

6.1 Limitations

In this work we only explore reasoning where the event that causes the action to fail occurs immediately before the actual failure of the action. We do not consider actions which may cause failure of the action at a later moment in time.

6.2 Risks

This research may pose some risk for privacy by being employed extensively for surveillance.

6.3 Licenses

OOPs dataset - Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Video ChatGPT- Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. LLaMA- LLAMA community license agreement UCF-Crimes - Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

6.4 Computation

All experiments we performed using a single V-100 32 GB GPU with each experiment taking around 10 hours.

References

[1] Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. *Advances in neural information processing systems*, 29, 2016.

[2] 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.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2021.

[5] 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.

[6] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

[7] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016.

[8] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models, 2023.

[9] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[10] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE, 2017.

[11] Google. Bard. https://bard.google.com, 2023. Accessed: 2023-11-12.

[12] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

[13] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018.

[14] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, pages 9459–9474. Curran Associates, Inc., 2020.

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[17] 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.

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[19] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023.

[20] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*, 2021.

[21] Zemin Liu, Xingtong Yu, Yuan Fang, and Xinming Zhang. Graphprompt: Unifying pre-training and downstream tasks for graph neural networks. In *Proceedings of the ACM Web Conference 2023*, pages 417–428, 2023.

[22] Jieyi Long. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*, 2023.

[23] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[24] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023.

[25] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14111–14121, 2021.

[26] Mikołaj Małkiński and Jacek Mańdziuk. A review of emerging research directions in abstract visual reasoning. *Information Fusion*, 91:713–736, 2023.

[27] Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks, 2023.

[28] Allen Newell, J. C. Shaw, and Herbert A. Simon. Report on a general problem-solving program. In *IFIP Congress*, 1959.

[29] Allen Newell, Herbert Alexander Simon, et al. *Human problem solving*. Prentice-hall Englewood Cliffs, NJ, 1972.

[30] OpenAI. Chatgpt: Version classic. https://openai.com, 2023. Accessed: 2023-11-12.

[31] Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. Art: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014*, 2023.

[32] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[33] 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.

[34] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. Evaluation and analysis of hallucination in large vision-language models, 2023.

[35] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[36] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837. Curran Associates, Inc., 2022.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[39] Haotian Xue, Antonio Torralba, Joshua Tenenbaum, Daniel Yamins, Yunzhu Li, and Hsiao-Yu Tung. 3d-intphys: Towards more generalized 3d-grounded visual intuitive physics under challenging scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3625–3635, 2023.

[40] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples, 2023.

[41] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[42] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[43] Yifeng Zhang, Shi Chen, and Qi Zhao. Toward multi-granularity decision-making: Explicit visual reasoning with hierarchical knowledge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2573–2583, 2023.

[44] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models, 2022.

[45] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[46] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

[47] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Appendix

A.1 Cosine similarity

To obtain the cosine similarity score for Figure 3 we prompt the model as the **Prompt:** ***"Given the video goal of the activity occurring in the video as <goal> and reason behind its failure as <reason>"*** and take the embedding obtained from the encoder of Video-ChatGPT model. For ground truth encoding we replace <reason> with the ground truth reason similarly for DoT and Basic prompt with reasoning obtained from using repsective prompts.

A.2 LLM Evaluation

We use GPT-3.5 for evaluation using LLM. To obtain the score we prompt GPT-3.5 as **Prompt:** *"You are provided with a question,the correct answer and the predicted answer. The question contains information about the task being attempted to be achieved in the video, along with the context about the objects involved in achieving that goal. The correct answer consists of the reasons behind the failure of achieving that objective and information about the objects present during the failure. Your task is to evaluate the correctness of the predicted answer. Here's how you can accomplish the task:// "____" "INSTRUCTIONS: //" "- Focus on the meaningful match of events between the predicted answer and the correct answer.*
" "- Consider synonyms or paraphrases as valid matches.
" "- Evaluate the correctness and alignment of the predicted answer compared to the correct answer. " ,

"role": "user",
"content":
"Please evaluate the following video-based question-answer pair:
" f"Question: question
" f"Correct Answer: answer
" f"Predicted Answer: pred
" "Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 1, with 1 indicating the highest meaningful match. " "Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where value of 'pred' is a string of 'yes' or 'no' and value of 'score' is in NUMBER, not STRING."
"DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. " "For example, your response should look like this: 'pred': 'yes', 'score': 0.8."
Where the correct reason is the ground truth reason the question is given as ***If the <goal> of the action occurring in the given video infer the reason why the action fails to achieve the intended outcome*** and predicted answer is the answer obtained using the respective prompting technique.

A.3 MCQ Style Prompt

: To formulate the MCQ style prompt mentioned in 1 containing n options we first randomly select ground truth reasons behind the failure of actions to obtain n-2 options. In addition to these N-2 options we also provide the ground truth reason for that particular video and None of these option as well. The prompt provided to the model is given as ***The action occurring in the given video fails.You will be given*** $num_options$ ***describing the reasoning behind the failure. The options for this video are given as*** $options_list$***.*** where $num_options$ is the number of options provided in the MCQ style prompt and $options_list$ refers to the list of options provided to the MCQ style prompt.

A.4 FIB style prompt

To formulate the FIB style prompt used in 1 we first use the ground truth reason behind the failure contain a list of s subjects v verbs and o objects. First we randomly remove s , v and $o's$ and replace it with ___. The sentence obtained after it is ***They _____ the _____ too high and ___ a _____ _____ off.*** Finally we prompt the model with ***Given the following video complete the following sentence such that the sentence describes the reasoning behind failure of the intended action in the video. The sentence to be completed is <sentence>. Note: Your task is to complete the given sentence where the blanks are indicated by _____.***

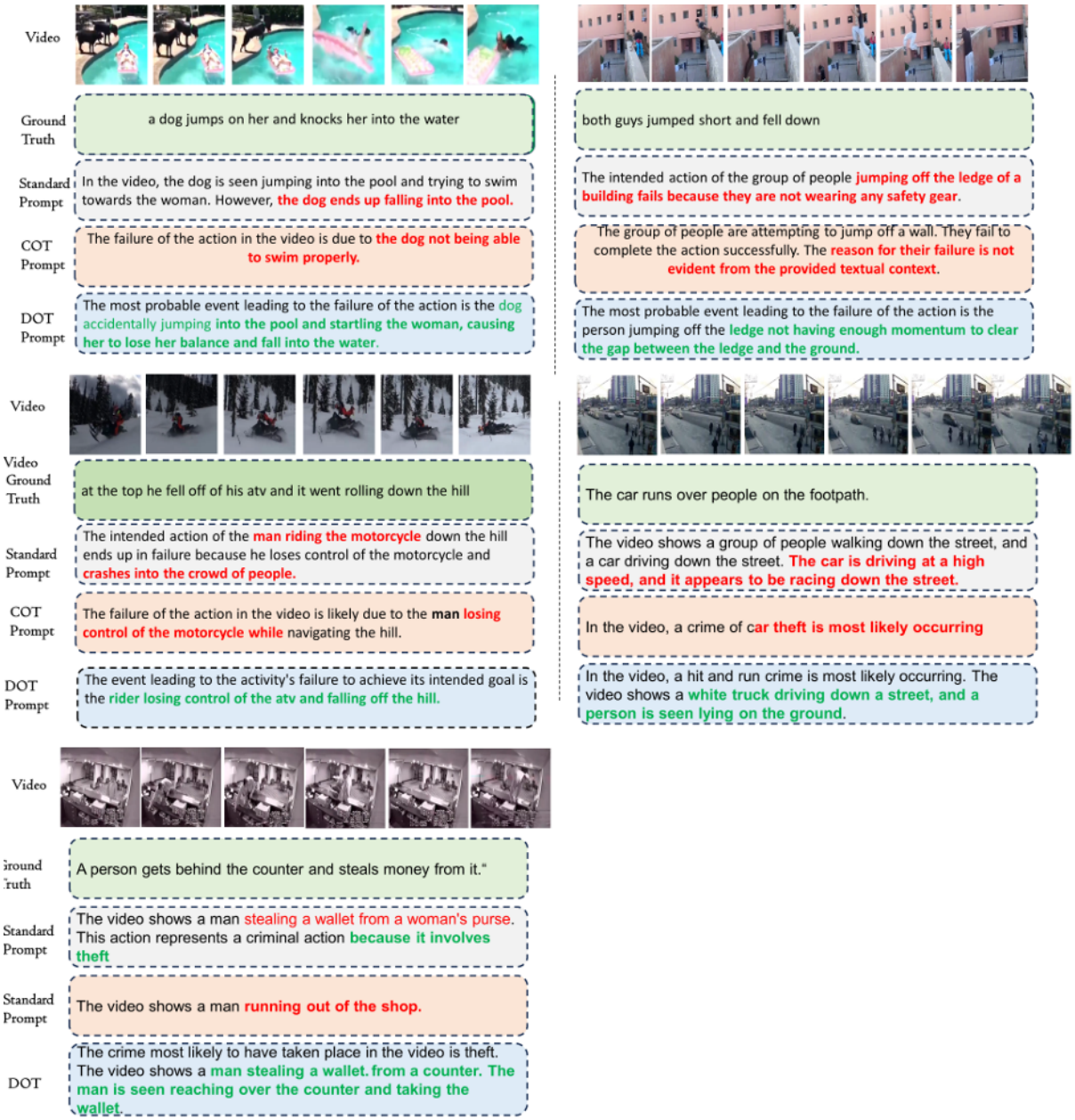

Figure 8: We show some samples for the qualitative results of the proposed DOT prompting compared with COT and standard prompting for UCF-Crimes and OOPs dataset.