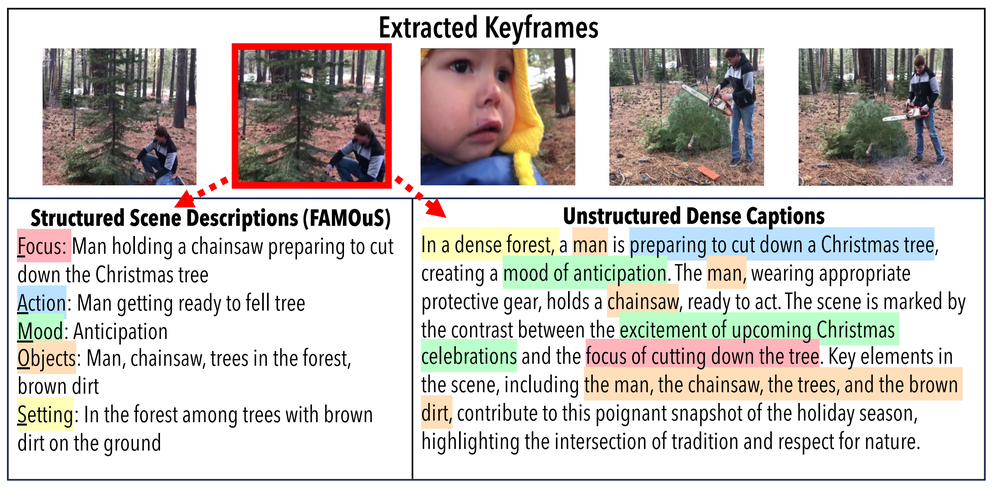
Let's think Frame by Frame with VIP: A Video Infilling and Prediction Dataset for Evaluating Video Chain-of-Thought

Vaishnavi Himakunthala*, Andy Ouyang*, Daniel Rose*, Ryan He*,
Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, William Wang
University of California, Santa Barbara

Motivation

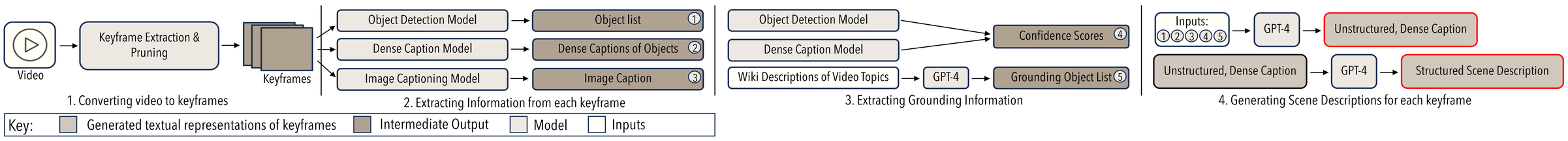
- Investigate video reasoning (VR) abilities of vision-language (VL) systems, the next logical step after text and image
- Evaluate multi-hop, multi-frame reasoning on general real-life videos, missing in VR datasets
- Automate data creation to minimize redundancy and costs, for scalability of VR research

VIP at a Glance



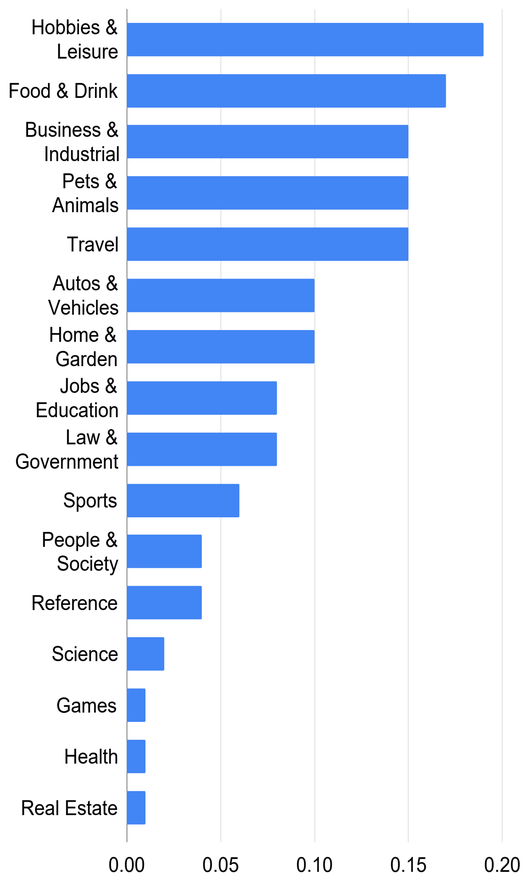
- An **inference time dataset** to assess models' video reasoning abilities
- Contains video keyframes from a wide variety of realistic videos
- Uses two textual representations for keyframes, including a new format of keyframe captioning – FAMOuS
- Introduces two new tasks, Video Infilling and Video Prediction

VIP Data Collection Pipeline



VIP Dataset

VIP's Video Categories



Distribution of VIP's real-world video domains.

VIP vs. Video Reasoning Datasets

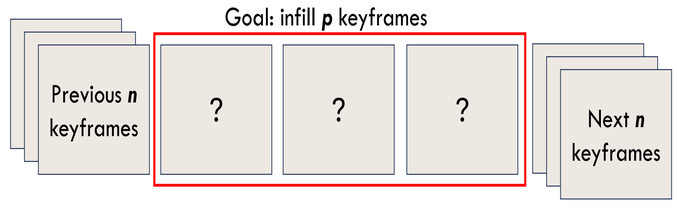
Dataset	Frame	Structured	Domain	Vid. Len.	Cap. Len.	Test Samples
MSR-VTT	✗	✗	Open	20.7s	9.6	3K
YouCook2	✗	✗	Cooking	5.26m	8.8	2K
ActyNet-Cap	✗	✗	Open	2m	13.5	5K
HowTo100M	✗	✗	Instructional	18s	4	24K
VATEX	✗	✗	Open	10s	15.2	6K
VideoStory	✓	✗	Events	12.6m	12.1	16
WebVid-2M	✗	✗	Open	18s	12	5K
VIP	✓	✓	Open	3.6m	114.2	1.5K

VIP Prediction Task Example

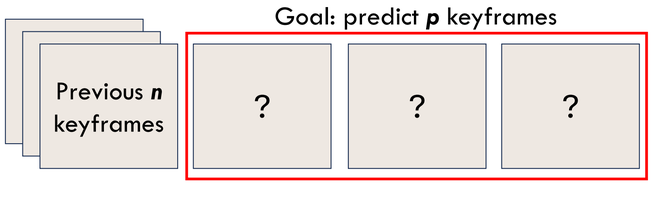
Original Keyframe	Ground Truth	GPT-4's Prediction
	Focus: Woman and young boy Action: Engaged in a learning activity, woman helping the boy with schoolwork Mood: Warm, nurturing, supportive Objects: Brown wooden table, piece of paper, pencil, white shirt, white t-shirt Setting: Comfortable learning space	Focus: Woman ~~back to laptop~~ Action: Woman ~~turning back to her laptop and continues typing~~ Mood: Focused, determined Objects: ~~Laptop~~, table, ~~black metal chairs~~, posters, black computer monitor, ~~TV~~ Setting: Workspace or study area

Task Definition

Video Infilling – Given n previous and next keyframes' descriptions, predict the *p* masked keyframes' descriptions



Video Prediction – Given n previous keyframes' descriptions, predict the *p* next keyframes' descriptions



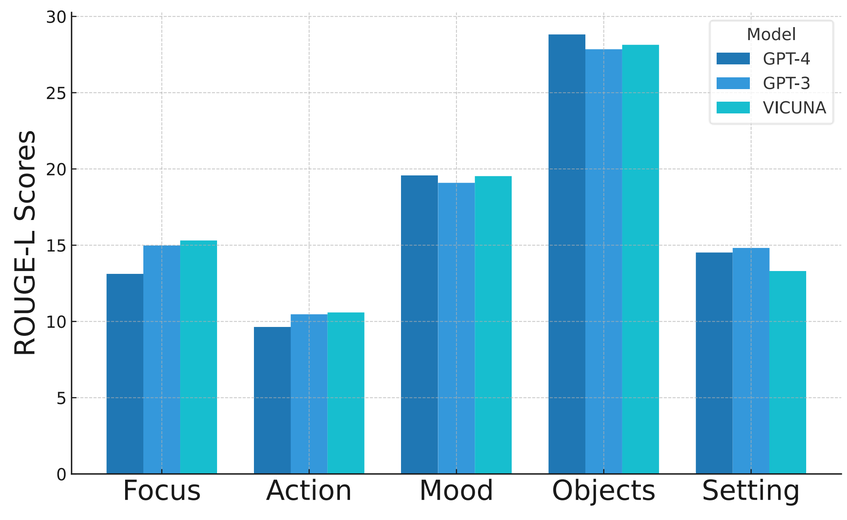
Benchmarking

Models Show Significant Room for Improvement on VIP Tasks

Metric	Model	Infilling-1		Infilling-2		Prediction-1		Prediction-2	
		FAMOuS	Dense	FAMOuS	Dense	FAMOuS	Dense	FAMOuS	Dense
$ROUGE_L$	GPT-4	17.34	24.92	18.44	25.25	15.52	23.31	16.66	25.22
	GPT-3	17.83	26.26	19.34	25.50	16.39	28.43	17.20	25.96
	VICUNA	17.37	25.34	18.85	26.69	15.86	23.75	16.59	25.88
BERTSCORE	GPT-4	18.61	24.81	19.66	25.67	16.24	20.57	17.24	24.79
	GPT-3	18.20	26.24	19.56	23.10	16.60	29.96	17.24	23.47
	VICUNA	17.67	24.07	18.98	28.14	15.80	19.76	16.68	24.34
SENTENCEBERT	GPT-4	53.05	58.53	53.87	58.22	50.57	53.55	51.54	57.06
	GPT-3	52.95	58.54	54.57	55.69	51.04	59.83	51.81	53.99
	VICUNA	52.19	54.66	53.33	58.80	50.40	51.86	50.96	54.86

- Low scores demonstrate difficulty of our tasks on existing LLMs
- Marginal boost in performance with additional context frames
- Filler words/verbosity in dense captions lead similarity metrics to score them higher than FAMOuS descriptions
- Models perform better on the infilling task compared to the prediction task, as bidirectional context is a less complex task

Prediction Task on FAMOuS Categories Reveals Opportunities for Enhancing Dynamic Components



- Low scores on dynamic components (Action, Focus), motivating other strategies for dynamic video reasoning
- Slightly stronger performance for basic components (Objects, Mood), likely due to a consistent background