Unhackable Temporal Hacking for Scalable Video MLLMs

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, Wenbing Tao

StepFun ICLR

Project Page: Scan here !

WeChat: Connect with me !

Postdoc seeking: Connect with me !

Part 1: Rethink Video MLLM

Key Questions

- Anti-scaling Law in Video MLLMs: Why more video-text data and larger video llm size lead to worse video understanding?
- Potential Risks of existing Video-Language Modeling: Will current video-language modeling paradigm, which primarily utilizes video-text pairs, potentially introduce risks to model optimization, such as shortcut learning?
- RL perspective: Is it possible to describe the video mllm process from the lens of reinforcement learning (RL)?

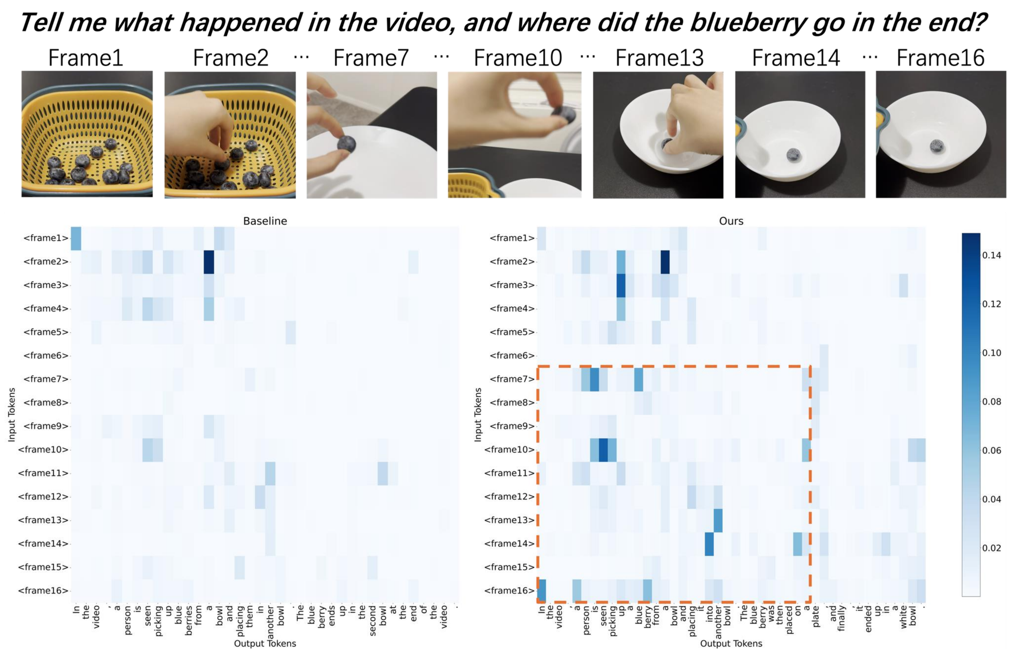


Figure: Attention map visualization illustrates which specific frames the model's output focuses on.

Part 2: Temporal Hacking Theory

What is temporal hacking?

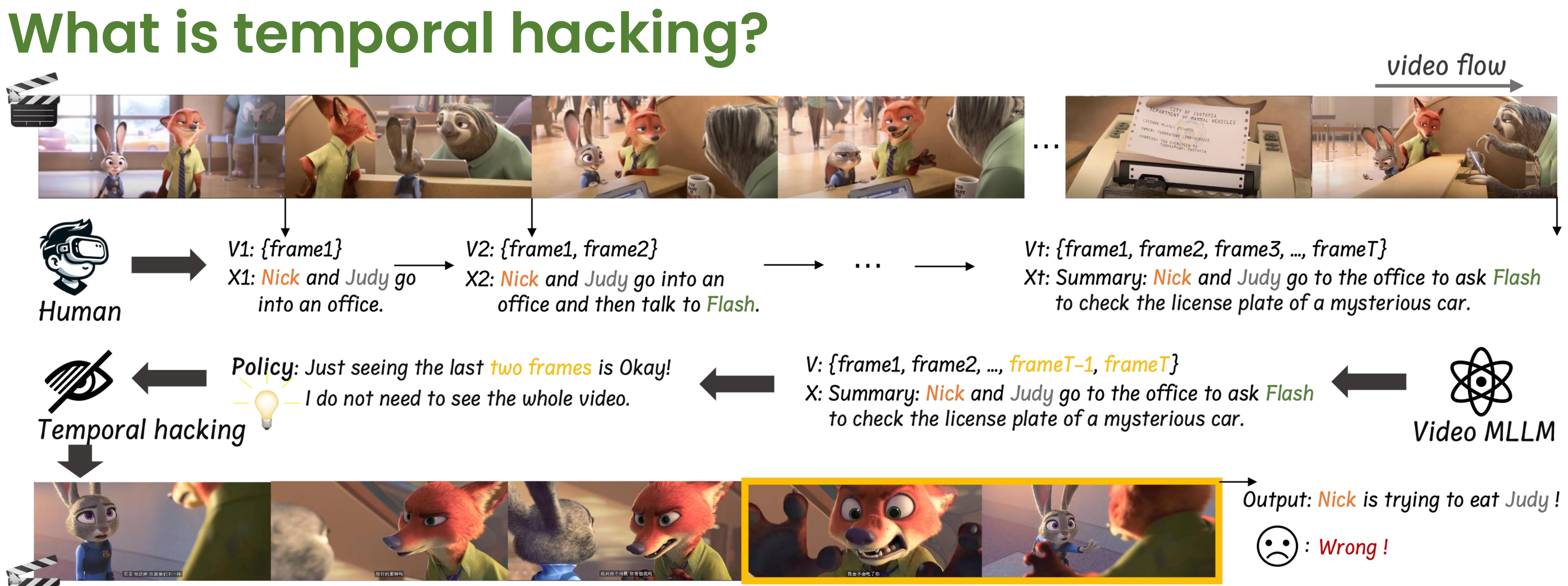


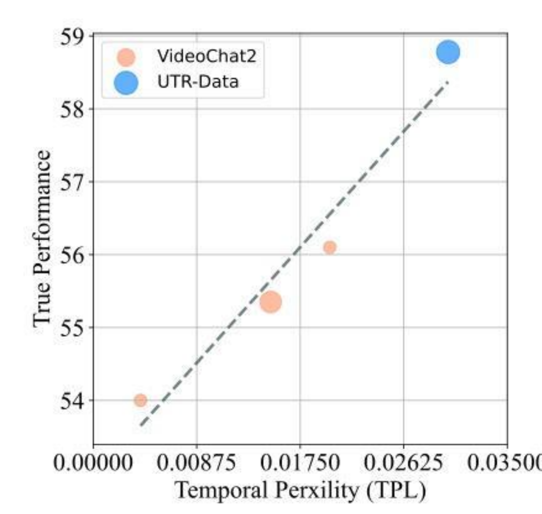
Figure: Temporal hacking illustration.

Theoretical Perspective

	Reward Hacking	Temporal Hacking (Ours)
Objective	$J(\pi)=\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)\right]$,	$J(\pi)=\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=1}^{T}\gamma^t R(V_{1:t},x_t)\right]$
Policy	$\pi^*=\arg\max_{\pi}\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)\right]$,	$\pi^*=\arg\max_{\pi}\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=1}^{T}\gamma^t R(V_{1:t},x_t)\right]$
Hacking	$J(\pi_h)>J(\hat{\pi}),\ \textit{however},\ K(\pi_h)\ll K(\hat{\pi}),$	
ΔR	$\Delta\mathcal{R}=\sum_{t=1,1\le k\le t}^{T}\gamma^t\left(R(V_{1:t},x_t)-R(V_{k:t},x_t)\right)$	

Key Findings

- Misalignment between video frames and text description: At current stage of video-text pair data, textual descriptions struggle to cover each individual frame.
- Frame activation number is limited: We observe that existing video mllms only attend a limited number of frames when conducting video understanding.
- Limited frame observation leads to hallucination more easily: Only observing several frames causes misunderstanding or hallucination more easily.



What causes temporal hacking?

Experimental Perspective

- We introduce Temporal Perplexity (TPL) Score to quantify the severity of temporal assignment. And we discover the relationship between TPL and true performance. Larger TPL indicates a reduced likelihood of reward hacking, thereby leading to superior video comprehension.

TPL Score $\quad \mathcal{T}_{tpl}=-\left(\mathcal{R}_{ppl}(V_{1:T},x_T)-\mathcal{R}_{ppl}(V_{T:T},x_T)\right)$

How to Mitigate Temporal Hacking?

Two Guided Principles

- *Principle I : High frame information density.* The content of the video text should uniquely correspond to as many frames as possible.
- *Principle II : High inter-frame information dynamics.* Descriptions for different frames should be coherent and reflect temporal variations and event progression.

Part 3: Unhackable Temporal Rewarding

- *Spatiotemporal attributes are key to representing unique video frame content.*
- *Bidirectional querying explicitly models spatiotemporal dynamics.*

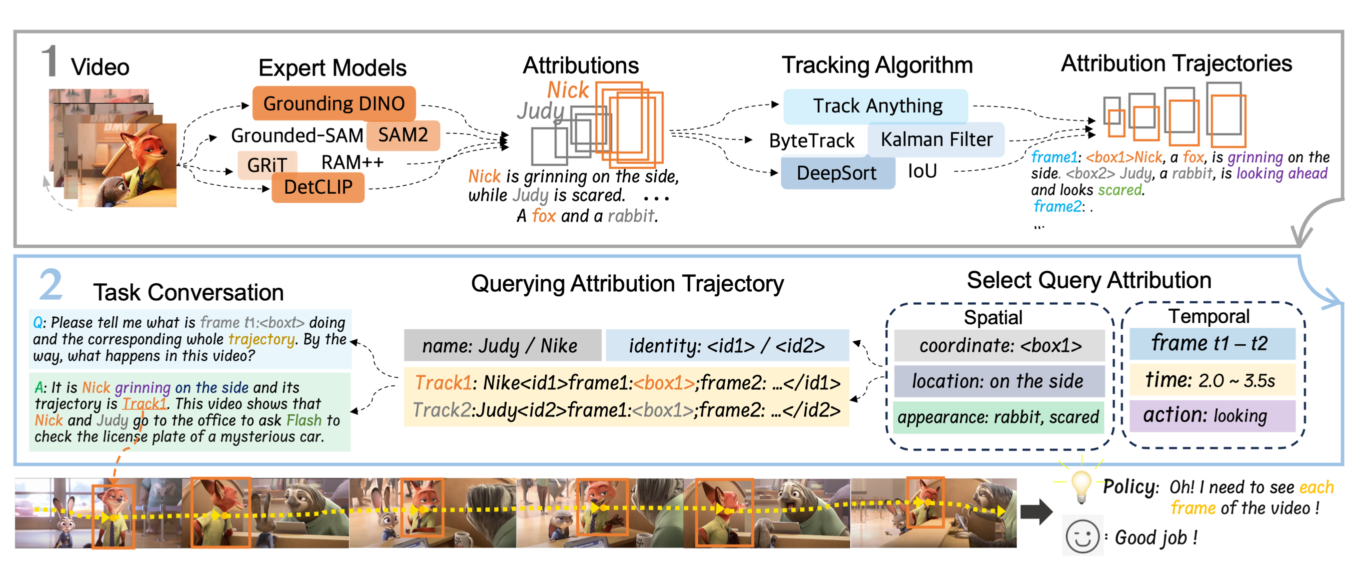


Figure: UTR pipeline. UTR uses expert models to extract spatiotemporal attributes and a tracking algorithm to build trajectories based on confidence levels. It then queries temporal and spatial attributes bidirectionally to generate dialogue data, learning spatiotemporal dynamics.

Part 4: Video-UTR

Experiment

Methods	LLM	Data Scale	MVBench	TempC	VideoMME	MSVD-QA Acc	MSVD-QA Score	MSRVTT-QA Acc	MSRVTT-QA Score	TGIF-QA Acc	TGIF-QA Score	ANet-QA Acc	ANet-QA Score
VideoChat (2023a)	Vicuna-7B	765K	35.5	–	–	56.3	2.8	45.0	2.5	34.4	2.3	–	2.2
VideoChat2 (2024c)	Vicuna-7B	1.9M	51.1	38.5	–	70.0	3.9	54.1	3.3	–	–	49.1	3.3
Video-ChatGPT (2023)	Vicuna-7B	765K	32.7	31.8	–	51.6	2.5	29.6	1.8	–	–	12.4	1.1
Video-LLaVA (2023)	Vicuna-7B	765K	34.1	34.8	39.9	64.9	3.3	49.3	2.8	51.4	3.0	35.2	2.7
VideoLLaMA2 (2024)	LLaMA2-7B	13.4M	54.6	–	46.6	70.9	3.8	–	–	–	–	50.2	3.3
PLLaVA (2024)	LLaMA2-7B	1M	46.6	–	–	**76.6**	**4.1**	**62.0**	3.5	**77.5**	**4.1**	56.3	**3.5**
LLaVA-NeXT-Video (2024c)	Qwen2-7B	860K	54.6	–	33.7	67.8	3.5	–	–	–	–	53.5	3.2
LLaVA-OneVision(2024a)	Qwen2-7B	1.6M	56.7	59.0*	**58.2**	65.3*	3.8*	43.3*	3.0*	52.8*	3.4*	56.6*	3.3*
Video-UTR (Ours)	Qwen2-7B	1.1M	**58.8**	**59.7**	52.6	73.5	**4.1**	58.3	**3.6**	56.4	3.6	55.0	3.2

Methods	LLM	MM-Vet	MMBench	MMMU	MME	$LLaVA^W$	POPE	SEED	AI2D	RealWorldQA
Image-level MLLM										
InstructBLIP (2024)	Vicuna-7B	33.1	36.0	30.6	1137.1	59.8	86.1	53.4	40.6	36.9
Qwen-VL-Chat (2023b)	Qwen-7B	**47.3**	60.6	37.0	1467.8	67.7	74.9	58.2	63.0	49.3
LLaVA-v1.5-7B (2024a)	Vicuna-7B	30.5	64.3	35.7	1510.7	61.8	86.1	58.6	55.5	54.8
LLaVA-v1.5-13B	Vicuna-13B	35.4	67.7	37.0	1531.3	66.1	86.1	61.6	61.1	55.3
ShareGPT4V (2023a)	Vicuna-7B	37.6	68.8	37.2	1567.4	72.6	86.6	69.7	58.0	54.9
LLaVA-NeXT-Img (2024c)	LLaMA3-8B	44.4	72.1	41.7	1551.5	63.1	87.1	–	71.6	60.0
Video-level MLLM										
LLaMA-VID (2023b)	Vicuna-7B	–	66.6	–	1521.4	–	86.0	59.9	–	–
Video-LLaVA (2023)	Vicuna-7B	32.0	60.9	–	–	73.1	84.4	–	–	–
LLaVA-NeXT-Video (2024c)	QWen2-7B	42.9	74.5	42.6	1580.1	**75.9**	88.7	74.6	71.9	60.1
Video-UTR (Ours)	Qwen2-7B	39.6	**76.6**	**43.4**	**1583.6**	69.4	**88.9**	**74.7**	**72.1**	**63.7**

Ablation Setting	Data Scale	MVBench	TGIF-QA	ANet-QA	MMVet	MMBench	POPE
Video-UTR	1.1M	**58.78**	**56.44**	**55.00**	39.59	**76.63**	88.86
- Task Modeling	1.0M	58.45	56.11	54.21	37.33	76.37	**89.29**
- Data Modeling	780K	54.63	54.74	54.15	**42.20**	75.77	89.13
+ More VideoChat2	1.1M	57.65	53.39	53.65	36.56	75.95	88.76

Methods	Overall ↑	Perception ↑	Reasoning ↑
Claude-3.5-Sonnet (2024)	1.35	1.4	1.04
VideoChat2-HD (2024c)	1.23	0.44	1.23
PLLaVA-34B (2024)	1.16	1	1.1
LLaVA-NeXT-Video-34B-HF (2024c)	1.13	0.58	1.03
Video-UTR-7B (Ours)	1.35	1.38	1.24

Table: Performance on MMBench-Video.

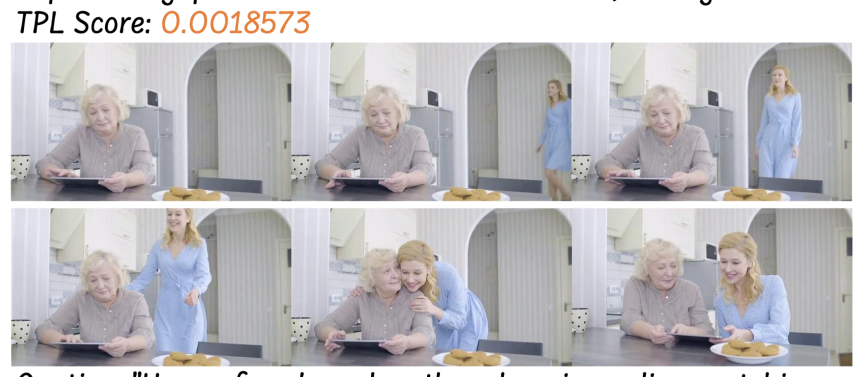
UTR-Data Size	MVBench	TempCompass	VideoMME
0K	54.63	58.88	**53.37**
180K	58.45	58.47	52.30
325K	**58.78**	**59.67**	52.63

Frame Length	MVBench	TempCompass	VideoMME
8	50.93	**56.28**	52.56
24	50.08	56.14	**52.81**
32	**51.40**	56.11	52.07

Table: Scalability ablation study of UTR.

TPL Analysis

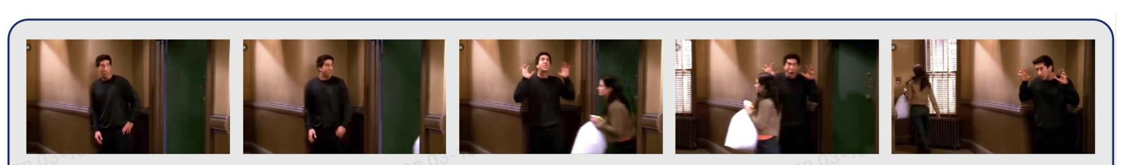
Caption: "Agapanthus flowers at Ile De Brehat, Bretagne France"
TPL Score: 0.0018573



Caption: "Happy female and mother shopping online, watching photos on laptop together"
TPL Score: 0.5056167

Demo case

Successful case: More important details about movie.

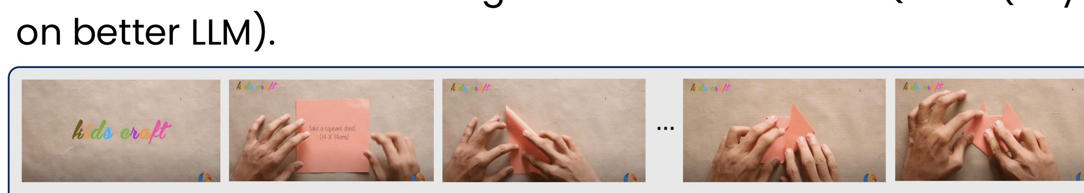


Human: Please describe the video in detail.

LLaVA-NeXT-Video: The man and woman are standing in a hallway, and the man is holding a white bag. They seem to be having a conversation, but it's unclear what they're saying. The woman is wearing a blue shirt, and the man is wearing a black shirt. There is a green door behind them, and a window on the wall. The man appears to be looking at something off-camera, while the woman looks directly at the camera.

Ours: In the video, a man is seen in a hallway and entering a room. He then proceeds to walk towards the woman who is getting out of the room. The man appears to be in a playful mood as he starts making funny faces at the woman, seemingly trying to entertain her. Overall, the video depicts a light-hearted moment between two individuals who seem to be having a good time together.

Failure case: Knowledge-oriented video MCQ task (rely on better LLM).



Question: What is the shape of the paper shown in the video?
Options: **A.** Circle **B.** Rectangle **C. Square** **D.** Triangle
Model output: **D**