Learning to Decode Against Compositional Hallucination in Video Multimodal Large Language Models

Wenbin Xing[1], Quanxing Zha[2], Lizheng Zu[3], Mengran Li[1], Ming Li[4], Junchi Yan[5]

[1]Sun Yat-sen University, [2]Huaqiao University, [3]Shenzhen University, [4]School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, [5]Shanghai Jiaotong University

Mail: xingwb@mail2.sysu.edu.cn

WeChat GitHub Huggingface

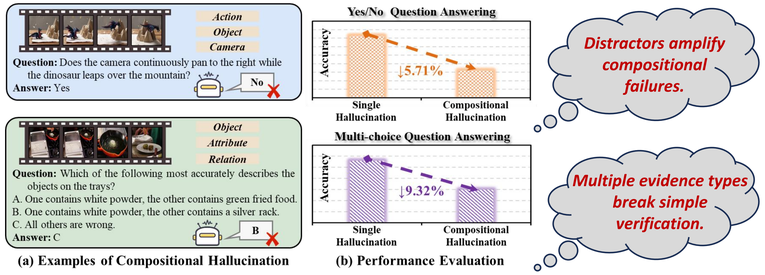
Feel free to contact us! Discussions, and collaborations are warmly welcome.

ICML International Conference On Machine Learning

Motivation

Compositional Hallucinations in VLLMs

- ***Existing video hallucination benchmarks mainly evaluate isolated error types.***
- However, real-world video understanding often ***requires jointly reasoning*** over object, action, relation, temporal order, and camera motion.
- ***We study compositional hallucinations***, where multiple visual evidence types are simultaneously required and removing any one of them makes the answer unsupported or wrong.



(a) Examples of Compositional Hallucination

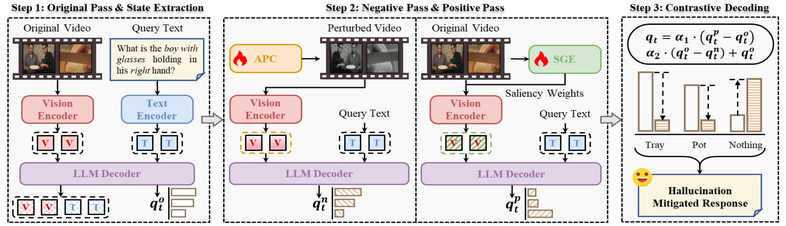
(b) Performance Evaluation

Distractors amplify compositional failures.

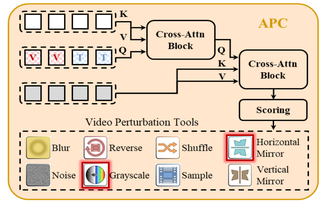
Multiple evidence types break simple verification.

Method

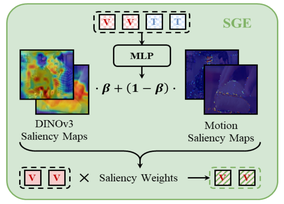
TriCD: Triple-Pathway Contrastive Decoding



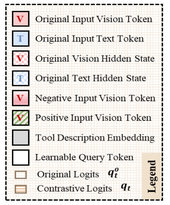
(a) Framework Overview



(b) Adaptive Perturbation Controller (APC)

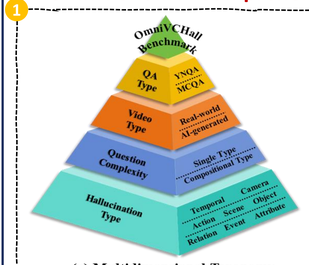


(c) Saliency-Guided Enhancement (SGE)

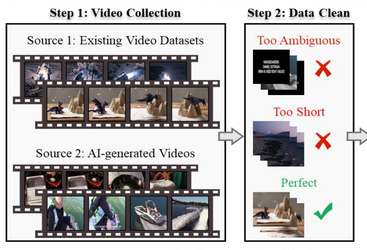


Benchmark

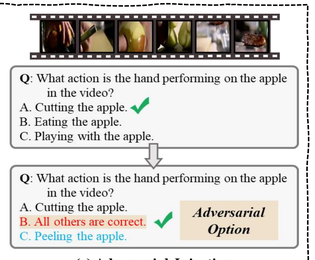
OmniVCHall: A Comprehensive Benchmark for Single and Compositional Video Hallucinations



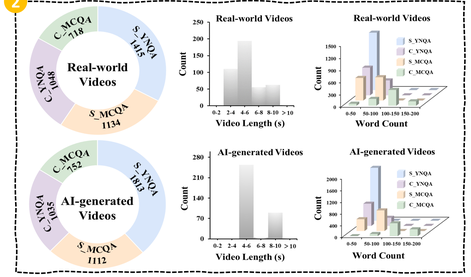
(a) Multidimensional Taxonomy



(b) Construction Pipeline



(c) Adversarial Injection



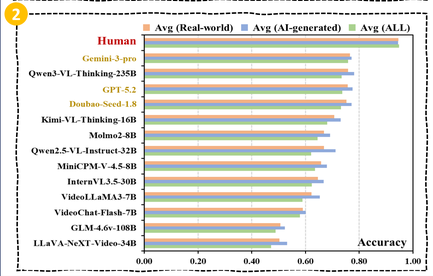
Benchmark	Hallucination Type								QA			Video		
	Object	Scene	Event	Action	Relation	Attribute	Temporal	Camera	YN	MC	Count	Real	Generated	Count
TempCompass (Liu et al., 2024)	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	7,540	✓	✗	410
VidHal (Choong et al., 2024)	✓	✗	✗	✓	✓	✓	✓	✗	✗	✓	1,000	✓	✗	1,000
VideoHallucer (Wang et al., 2024b)	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	1,800	✓	✗	948
EventHallusion (Zhang et al., 2024a)	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	711	✓	✗	400
MHBench (Kong et al., 2025)	✗	✗	✗	✓	✗	✗	✗	✓	✓	✓	3,600	✓	✗	1,200
VidHalluc (Li et al., 2025a)	✗	✓	✗	✓	✗	✗	✓	✗	✓	✓	9,295	✓	✗	5,002
ELV-Halluc (Lu et al., 2025)	✓	✗	✓	✓	✗	✗	✓	✗	✓	✗	4,800	✓	✗	200
OmniVCHall (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	9,027	✓	✓	823

- OmniVCHall contains 823 videos and 9,027 QA pairs, ***covering both real-world and AI-generated videos***.
- OmniVCHall defines ***eight fine-grained hallucination types***: object, scene, event, action, relation, attribute, temporal, and camera.
- OmniVCHall introduces ***adversarial answer options*** such as "All are correct" and "None of the above" to reduce shortcut reasoning.

Experimental Results

Quantitative analysis

Model	Size	Real	Generated	Avg
Human	-	0.95	0.94	0.95
Open-source Models				
VideoChat-Flash (Li et al., 2024c)	2B	0.55	0.57	0.56
VideoLLaMA3 (Zhang et al., 2025)	2B	0.51	0.58	0.54
Qwen3-VL-Instruct (Bai et al., 2025a)	2B	0.59	0.67	0.63
Qwen3-VL-Thinking (Bai et al., 2025a)	2B	0.62	0.69	0.65
Qwen2.5-VL-Instruct (Bai et al., 2025b)	3B	0.52	0.61	0.57
InternVL3.5 (Wang et al., 2025)	4B	0.61	0.66	0.64
MiniCPM-V-4 (Yu et al., 2025)	4B	0.57	0.63	0.60
Qwen3-VL-Instruct (Bai et al., 2025a)	4B	0.64	0.72	0.68
Qwen3-VL-Thinking (Bai et al., 2025a)	4B	0.67	0.73	0.70
Molmo2 (Clark et al., 2026)	4B	0.63	0.68	0.66
LLaVA-NeXT-Video (Zhang et al., 2024b)	7B	0.44	0.49	0.47
VideoChat-Flash (Li et al., 2024c)	7B	0.58	0.60	0.59
VideoLLaMA3 (Zhang et al., 2025)	7B	0.59	0.65	0.62
Qwen2.5-VL-Instruct (Bai et al., 2025b)	7B	0.60	0.69	0.64
InternVL3.5 (Wang et al., 2025)	8B	0.60	0.67	0.64
MiniCPM-V-4.5 (Yu et al., 2025)	8B	0.64	0.68	0.66
Qwen3-VL-Instruct (Bai et al., 2025a)	8B	0.65	0.74	0.70
Qwen3-VL-Thinking (Bai et al., 2025a)	8B	0.67	0.74	0.71
Molmo2 (Clark et al., 2026)	8B	0.65	0.69	0.67
Kimi-VL-Instruct (Team et al., 2025)	16B	0.64	0.72	0.68
Kimi-VL-Thinking (Team et al., 2025)	16B	0.68	0.73	0.71
InternVL3.5 (Wang et al., 2025)	30B	0.62	0.67	0.65
Qwen2.5-VL-Instruct (Bai et al., 2025b)	32B	0.62	0.71	0.67
Qwen3-VL-Instruct (Bai et al., 2025a)	32B	0.69	0.76	0.72
Qwen3-VL-Thinking (Bai et al., 2025a)	32B	0.71	0.77	0.74
LLaVA-NeXT-Video(Zhang et al., 2024b)	34B	0.47	0.53	0.50
GLM-4.5v (Hong et al., 2025)	108B	0.47	0.50	0.49
GLM-4.6v-flash (Hong et al., 2025)	108B	0.44	0.49	0.46
GLM-4.6v (Hong et al., 2025)	108B	0.49	0.52	0.51
Qwen3-VL-Instruct (Bai et al., 2025a)	235B	0.71	0.77	0.74
Qwen3-VL-Thinking (Bai et al., 2025a)	235B	0.73	*0.78*	0.76
Proprietary Models				
GPT-4o (OpenAI, a)	-	0.70	0.73	0.71
GPT-5 (OpenAI, b)	-	0.72	0.75	0.73
GPT-5.2 (OpenAI, c)	-	0.74	0.77	0.76
Gemini-2.5-flash (DeepMind, a)	-	0.71	0.75	0.73
Gemini-2.5-pro (DeepMind, b)	-	0.71	0.74	0.73
Gemini-3-pro (DeepMind, c)	-	*0.76*	0.77	*0.77*
Doubao-Seed-1.6 (Volcengine, a)	-	0.71	0.75	0.73
Doubao-Seed-1.8 (Volcengine, b)	-	0.73	0.77	0.75



- We evaluate ***39 representative VLLMs.***
- Even strong models show a clear gap to human performance: ***the best AI model reaches 0.77 average accuracy, while humans achieve 0.95.***
- ***TriCD consistently improves different VLLM backbones.***

Model	S_YNQA	C_YNQA	S_MCQA	C_MCQA	Avg	-
Qwen3-VL-Instruct-8B	0.75	***0.69***	0.61	***0.53***	0.67	-
+MotionCD	0.76	***0.70***	0.62	***0.54***	0.68	↑1.82%
+TCD	0.77	***0.72***	0.62	***0.55***	0.69	↑3.61%
+DINO-HEAL	0.77	***0.72***	0.62	***0.54***	0.68	↑1.99%
+TriCD	0.79	***0.78***	0.66	***0.64***	0.73	↑9.61%
VideoLLaMA3-7B	0.67	***0.61***	0.56	***0.48***	0.60	-
+MotionCD	0.68	***0.62***	0.57	***0.48***	0.61	↑2.34%
+TCD	0.69	***0.63***	0.57	***0.49***	0.62	↑3.41%
+DINO-HEAL	0.69	***0.63***	0.57	***0.48***	0.62	↑2.92%
+TriCD	0.74	***0.69***	0.63	***0.55***	0.67	↑12.03%
LLaVA-NeXT-Video-7B	0.54	***0.49***	0.38	***0.33***	0.45	-
+MotionCD	0.55	***0.51***	0.38	***0.34***	0.47	↑2.97%
+TCD	0.56	***0.52***	0.39	***0.34***	0.47	↑4.47%
+DINO-HEAL	0.54	***0.51***	0.39	***0.35***	0.47	↑3.22%
+TriCD	0.57	***0.57***	0.41	***0.42***	0.51	↑11.48%
InternVL3.5-8B	0.70	***0.67***	0.54	***0.49***	0.62	-
+MotionCD	0.71	***0.69***	0.55	***0.50***	0.63	↑1.65%
+TCD	0.71	***0.69***	0.55	***0.51***	0.64	↑2.29%
+DINO-HEAL	0.72	***0.68***	0.57	***0.51***	0.64	↑2.75%
+TriCD	0.73	***0.77***	0.58	***0.58***	0.68	↑9.24%
VideoChat-Flash-7B	0.60	***0.57***	0.54	***0.50***	0.56	-
+MotionCD	0.62	***0.59***	0.57	***0.53***	0.59	↑3.94%
+TCD	0.62	***0.58***	0.56	***0.52***	0.58	↑2.84%
+DINO-HEAL	0.62	***0.59***	0.56	***0.52***	0.58	↑3.44%
+TriCD	0.70	***0.66***	0.58	***0.54***	0.62	↑9.76%

Qualitative analysis