香港中文大學
The Chinese University of Hong Kong

PhoStream: Benchmarking Real-World Streaming for Omnimodal Assistants in Mobile Scenarios

CUHK MMLab
Lu Xudong
luxudong@link.cuhk.edu.hk

Real-World Streaming for Omnimodal Assistants

香港中文大學
The Chinese University of Hong Kong

The first mobile-centric streaming benchmark unifying on-screen and off-screen scenarios for Omni Models

MLLMs have demonstrated strong capabilities in offline audio-visual understanding, yet their potential to function as mobile assistants in continuous real-world streams remains underexplored.

PhoStream is the first mobile-centric streaming benchmark that integrates on-screen and off-screen scenarios to evaluate video, audio, and temporal reasoning. It is constructed through an Automated Generative Pipeline supported by rigorous human verification, while models are evaluated using a realistic Online Inference Pipeline with LLM-as-a-Judge evaluation for open-ended responses.

Question (09:08): The host just announced that the winner of the Roll-Up Challenge gets to stay up late. Wait until the challenge is over and the winner is declared, and tell me, who wins?

Answer (10:38): After the one-minute Roll-Up Challenge ends, the host asks for the rep counts. Christian, the son, completed 35 roll-ups, while his dad, Charles, completed 28, making Christian the winner of the challenge.

Gemini 3 Pro (10:38): Christian wins. (PartlyCorrect, 5)

Doubao-Seed-1.6 (:): " ". (NoResponse, 0)

Doubao-Seed-1.8 (09:08): The winner is the young girl in the purple outfit (the daughter). (EarlyResponse, 0)

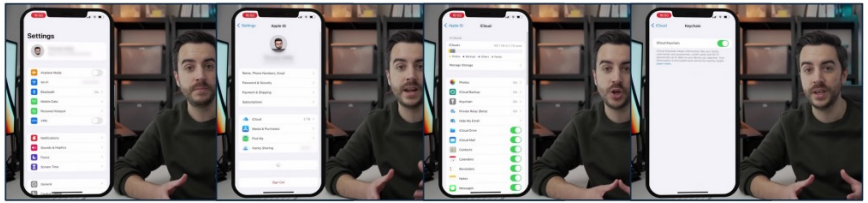
Qwen3-Omni-30B-A3B (09:08): The boy wins. (EarlyResponse, 0)

Qwen3-VL-30B-A3B (09:08): The boy in the blue shirt wins. (EarlyResponse, 0)

Dispider (:): " ". (NoResponse, 0)

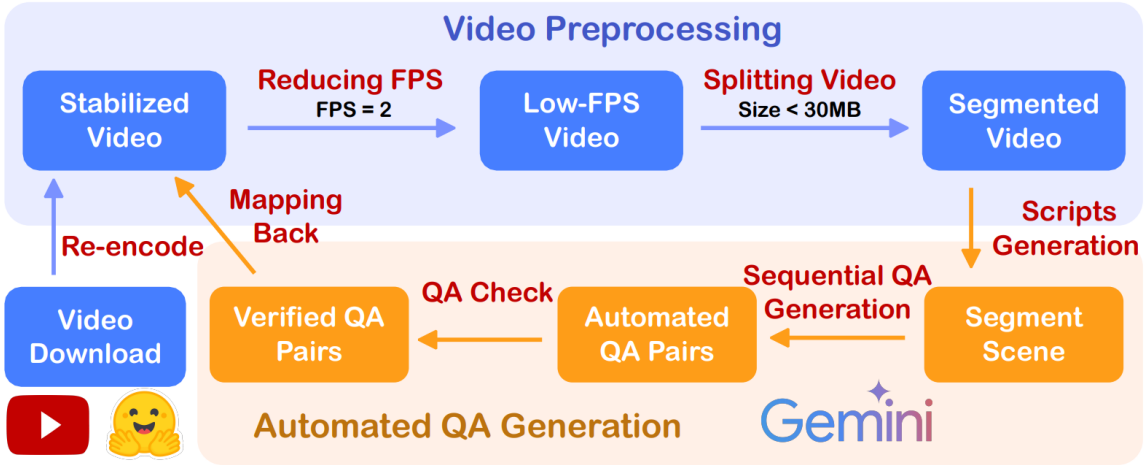
PhoStream – Automated Generative Pipeline

香港中文大學
The Chinese University of Hong Kong



Question (01:20): When he demonstrates enabling the Keychain feature, what is the exact navigation path he takes through the settings menu?

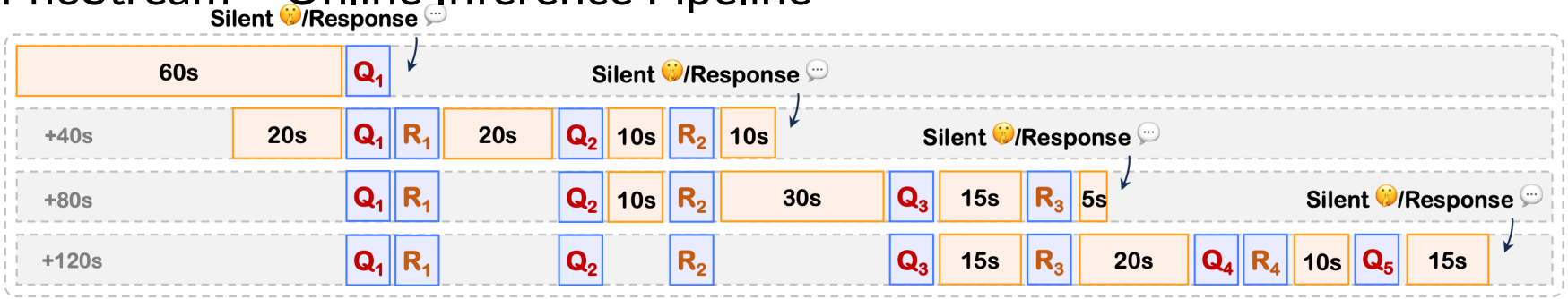
Response (01:28): The user taps the Apple ID banner at the top of the Settings menu (where the specific name is blurred out), then selects 'iCloud', and finally taps 'Keychain'. Finally, the screen shows that the iCloud Keychain toggle is green, indicating that it has been enabled.



- ✓ Converts raw mobile video streams into verified benchmark QA data
- ✓ Uses three stages: video preprocessing, automated QA generation, and multi-step verification
- ✓ Leverages Gemini 3 Pro and temporal verification to ensure QA quality

PhoStream – Online Inference Pipeline

香港中文大學
The Chinese University of Hong Kong



prompt = """You are an expert evaluator judging whether a model's answer provides a reasonable and factually plausible explanation that directly addresses the question, based on the reference answer.

Evaluation Guideline:
- Focus on whether the model gives a coherent reason that logically explains what the question asks. - The answer does not need to reproduce all details from the reference - it only needs to offer a factually grounded and relevant cause. - An answer that captures the essential reason should be considered strong, even if it omits descriptive details. - Accept simplified, rephrased, or high-level reasoning as long as it is consistent with the reference, plausibly explains the phenomenon in the question, and does not contradict known facts. - Do not deduct points for omitting secondary or illustrative details when the core causal logic is present, or for using concise or abstract phrasing. - Only penalize if the explanation is factually wrong, fails to provide a meaningful cause, or is so vague that it does not actually answer the question.

Scoring (integer 0-5):
- 5: Fully accurate and complete explanation. - 4: Correct and logically sufficient explanation; may omit non-essential details but captures the essential reason. - 3: Partially relevant but weakens or misses part of the core causal link. - 2: Tangential or speculative without solid grounding. - 1: Factually incorrect. - 0: No attempt to answer or completely off-topic.

Output Format:
Return a valid JSON object with exactly two keys:
- "explanation": one sentence focusing on whether the answer gives a reasonable and relevant reason for the question
- "score": an integer from 0 to 5

Output only the JSON. No other text, markdown, or commentary.
"""

- ✓ Processes continuous video streams in 1-second intervals with a sliding memory window
- ✓ Issues the query only once at the relevant timestamp, instead of repeatedly querying the model
- ✓ Let the model decide when to respond, enabling evaluation across different temporal reasoning tasks

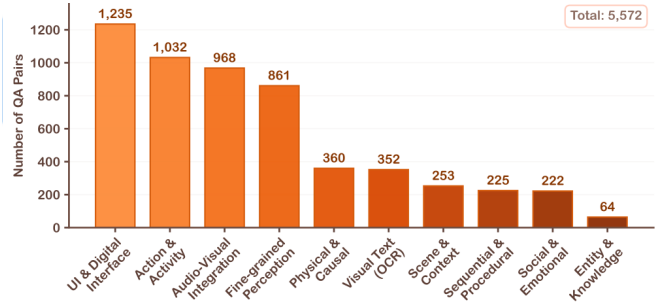
PhoStream – Data Distribution

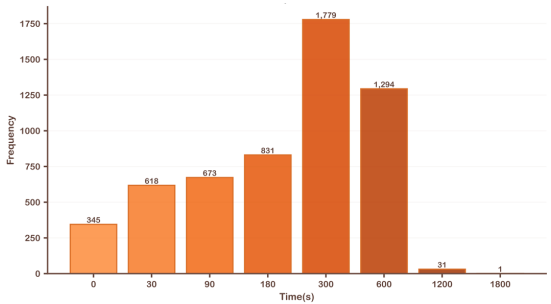
香港中文大學
The Chinese University of Hong Kong

✓ Data Source and Statistics

Benchmark	Modality	#Videos	Avg. Dur. (min)	#Ques.	Avg. Q/V	Mobile Centric	Open Ended	Temporal Scope		
								Backward	Instant	Forward
StreamingBench	V, A	900	9.7	4,500	5.0	✗	✗	✓	✓	✓
OVO-Bench	V	644	7.9	2,814	4.4	✗	✗	✓	✓	✓
OmniMMI	V, A	1,121	5.4	2,290	2.0	✗	✓*	✓	✓	✓
ProactiveVideoQA	V, A	1,377	2.1	1,427	1.0	✗	✓	✗	✗	✓
PhoStream (Ours)	**V, A**	578	**13.3**	**5,572**	9.6	✓	✓	✓	✓	✓

✓ Capability and Timestamp distribution





Scenario	Temporal Scope			Total
	Instant	Backward	Forward	
YouTube Vlog	997	629	1,631	**3,257**
Phone Tutorial	417	255	673	**1,345**
Phone Record	150	128	296	**574**
EgoBlind	83	111	202	**396**
Total	**1,647**	**1,123**	**2,802**	5,572

PhoStream – Results

香港中文大學
The Chinese University of Hong Kong

✓ Main evaluation results

Model	Param	Evaluation Score (↑)			Overall (↑)	Forward Task Analysis (%)		
		Instant	Backward	Forward		ER (↓)	NR (↓)	PC (↑)
Proprietary Multimodal Models								
Gemini 3 Pro (Google DeepMind, 2025)	-	**80.83**	**82.19**	16.40	48.70	79.12	**0.11**	20.77
Doubao-Seed-1.6 (ByteDance, 2025)	-	71.28	62.94	**44.26**	56.01	**29.76**	13.38	**56.85**
Doubao-Seed-1.8 (ByteDance, 2026)	-	80.45	77.31	33.38	**56.15**	56.46	2.36	41.18
Open-source Multimodal Models								
Qwen2.5-Omni-7B (Xu et al., 2025a)	7B	67.71	65.20	1.81	34.06	**42.65**	43.50	**13.85**
Qwen3-VL-8B (Bai et al., 2025)	8B	75.22	71.50	**7.18**	**40.25**	85.44	3.50	11.06
Qwen3-VL-30B-A3B (Bai et al., 2025)	30B	73.38	69.46	5.25	38.33	91.33	1.18	7.49
Qwen3-Omni-30B-A3B (Xu et al., 2025b)	30B	**77.18**	**77.24**	1.26	39.02	97.89	**0.07**	2.03
Open-source Multimodal Streaming Models								
Dispider (Qian et al., 2025)	7B	**44.24**	**42.90**	**3.53**	**23.50**	68.52	21.20	10.28
VideoLLM-online-8B (Chen et al., 2024)	8B	24.88	24.72	0.00	12.34	99.54	**0.43**	0.04
MMDuet2 (Wang et al., 2025b)	3B	8.76	8.30	1.39	4.96	**28.55**	59.21	**12.24**

✓ Ablation study on audio modality

Model	Audio	Evaluation Score (↑)			Overall (↑)	Forward Task Analysis (%)		
		Instant	Backward	Forward		ER (↓)	NR (↓)	PC (↑)
Gemini 3 Pro								
Gemini 3 Pro	✓	**80.83**	**82.19**	16.40	**48.70**	79.12	**0.11**	20.77
Gemini 3 Pro	×	77.46	72.84	**18.79**	47.03	**75.34**	0.71	**23.95**
Δ		**+3.37**	**+9.35**	-2.39	**+1.67**	+3.78	**-0.60**	-3.18
Qwen3-Omni-30B-A3B								
Qwen3-Omni-30B-A3B	✓	**77.18**	**77.24**	1.26	**39.02**	97.89	**0.07**	2.03
Qwen3-Omni-30B-A3B	×	74.10	70.92	**1.33**	36.87	**97.68**	0.14	**2.18**
Δ		**+3.08**	**+6.32**	-0.07	**+2.15**	+0.21	**-0.07**	-0.15