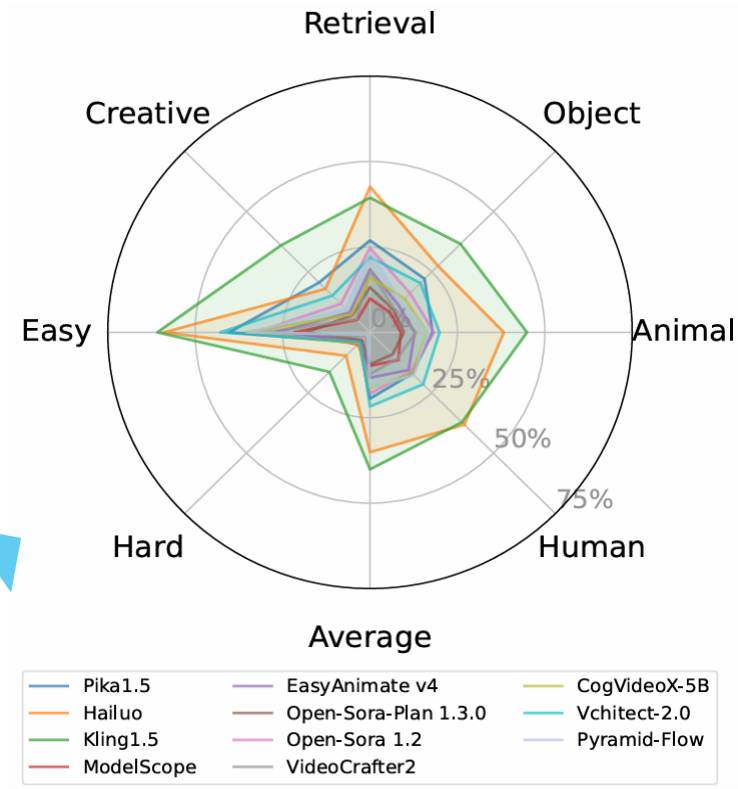
Is Your World Simulator a Good Story Presenter? A Consecutive Events-Based Benchmark for Future Long Video Generation

Yiping Wang[1], Xuehai He[2], Kuan Wang[3], Luyao Ma[4], Jianwei Yang[5], Shuohang Wang[5], Simon Shaolei Du[1], Yelong Shen[5]

[1]University of Washington [2]University of California, Santa Cruz [3]Georgia Institute of Technology [4]University of California, San Diego [5]Microsoft

Why and What

- Current video generation benchmark focuses on **fine-grained** metrics (e.g., aesthetic score, style, spatial/temporal consistency), where top models perform well.
- While if you want to generate a "story" in video, we also care about the **coarse-grained** completion of the story.
- But **top models** fail to generate simple story like **"How to put an elephant in a refrigerator"**, if they contain consecutive events!
- So we create **StoryEval**, a story-oriented benchmark! **None of the 11 top models exceeded an average story completion rate of 50%!**



Pipeline of StoryEval Benchmark

StoryEval Prompt Sets

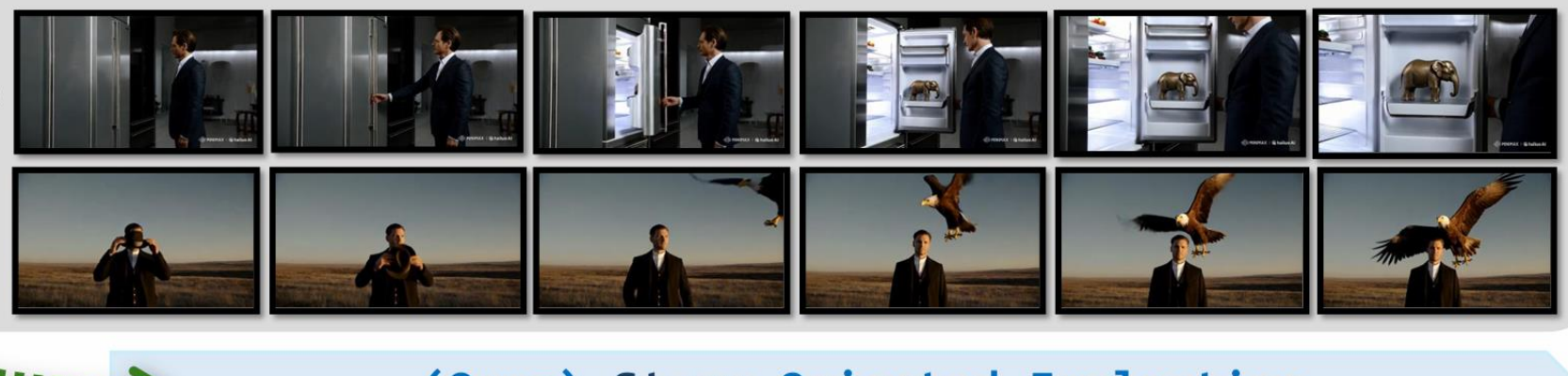
- **423 prompts** covering **7 classes**
- Each shows **a short story** containing **2 – 4 consecutive events**. Examples for 3-event story:
  - "A man opens the refrigerator door, puts the elephant in, and then closes the door."
  - "A man takes off his hat, throws it into the air, and then it is taken by a passing eagle."

Video Generative Models

Closed Source: KLING, Hailuo AI Beta, Pika, ...

Open Source: CogVideo, Open-Sora Plan, EasyAnimate, VideoCrafter2, ModelScope, Vchitect-2.0, Pyramid Flow, OPEN SORA, ...

Prompt → Text-to-Video Generation (Examples from Hailuo) → Video



Video → VLM Verifier: GPT-4o, LLaVA-OneVision

(Previous) Detail-Oriented Evaluation

- **Spatial quality**: aesthetic score, style
- **Temporal quality**: subject/background consistency, temporal flickering, motion smoothness, dynamic degree...
- **Semantics consistency**: video-text alignment on object/color/scene/spatial relationship..
- ...

Many models perform well!

(Ours) Story-Oriented Evaluation

STEP1: Describe the video in detail

STEP2: Analyze and get the **Completion Rate** of the story

(1) ... Overall, based on the frames, the first event is completed satisfactorily. However, the elephant is already inside the refrigerator, and people didn't close the door, so the second and third events are not shown. ==> **Completion List = [1, 0, 0], Completion Rate is only 33.3%**

(2) ... ### Overall Summary: The man takes off his hat, but there are no scenes depicting the man throwing the hat into the air or the eagle taking the hat. ==> **Completion List = [1, 0, 0], Completion Rate is only 33.3%**

All models don't perform well !!!

Property

- Useful supplement to traditional benchmarks (e.g. VBench)
- Rankings given by VLM verifiers align well with those from humans, especially when using unanimous voting.
- **Creative** tasks are in general more challenging.

Comparison

Benchmark	Type	Evaluate Closed Source Models	Prompt Style		Evaluation	
			Retrieval	Creative	Longer Videos	Story Evaluation
FETV	General	✗	✓	✓	✗	✗
EvalCrafter	General	✓	✓	✓	✗	✗
T2VBench	Temporal Dynamics	✓	✗	✗	✗	✗
TC-Bench	Temporal Composition	✗	✓	✓	✗	✗
Chronomagic	Time-Lapse	✓	✓	✗	✓	✗
T2V-Comp	General Composition	✓	✗	✓	✓	✗
VBench	General	✓	✗	✓	✓	✗
StoryEval	Consecutive Events	✓	✓	✓	✓	✓

Scan for interesting **examples** here!