NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

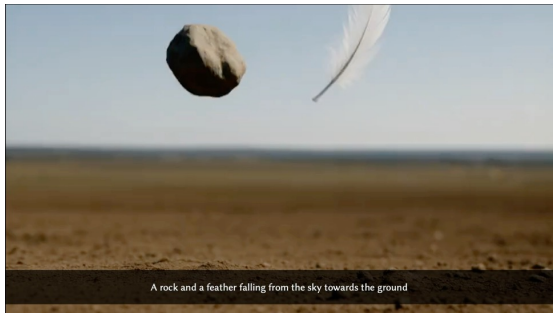
Multimodal Reasoning for Human-Centric Generative Models

Ziwei Liu 刘子纬
Nanyang Technological University

Multimodal Reasoning

Within-Shot *Event Reasoning*

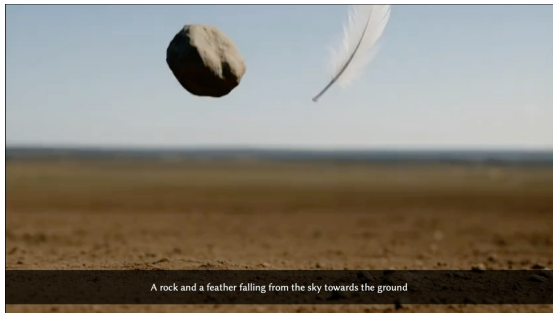
How humans understand causal logic



Multimodal Reasoning

Within-Shot *Event Reasoning*

How humans understand causal logic



→

Cross-Shot *Cinematic Reasoning*

How filmmakers compose stories

→ Toward **story-level human-like reasoning** in generative models

Multimodal Reasoning

Within-Shot *Event Reasoning*

VChain: models event plausibility


A rock and a feather falling from the sky towards the ground

Within-Shot Event Reasoning

VChain: Chain-of-Visual-Thought for Reasoning in Video Generation

Ziqi Huang, Ning Yu, Gordon Chen, Haonan Qiu, Paul Debevec, Ziwei Liu

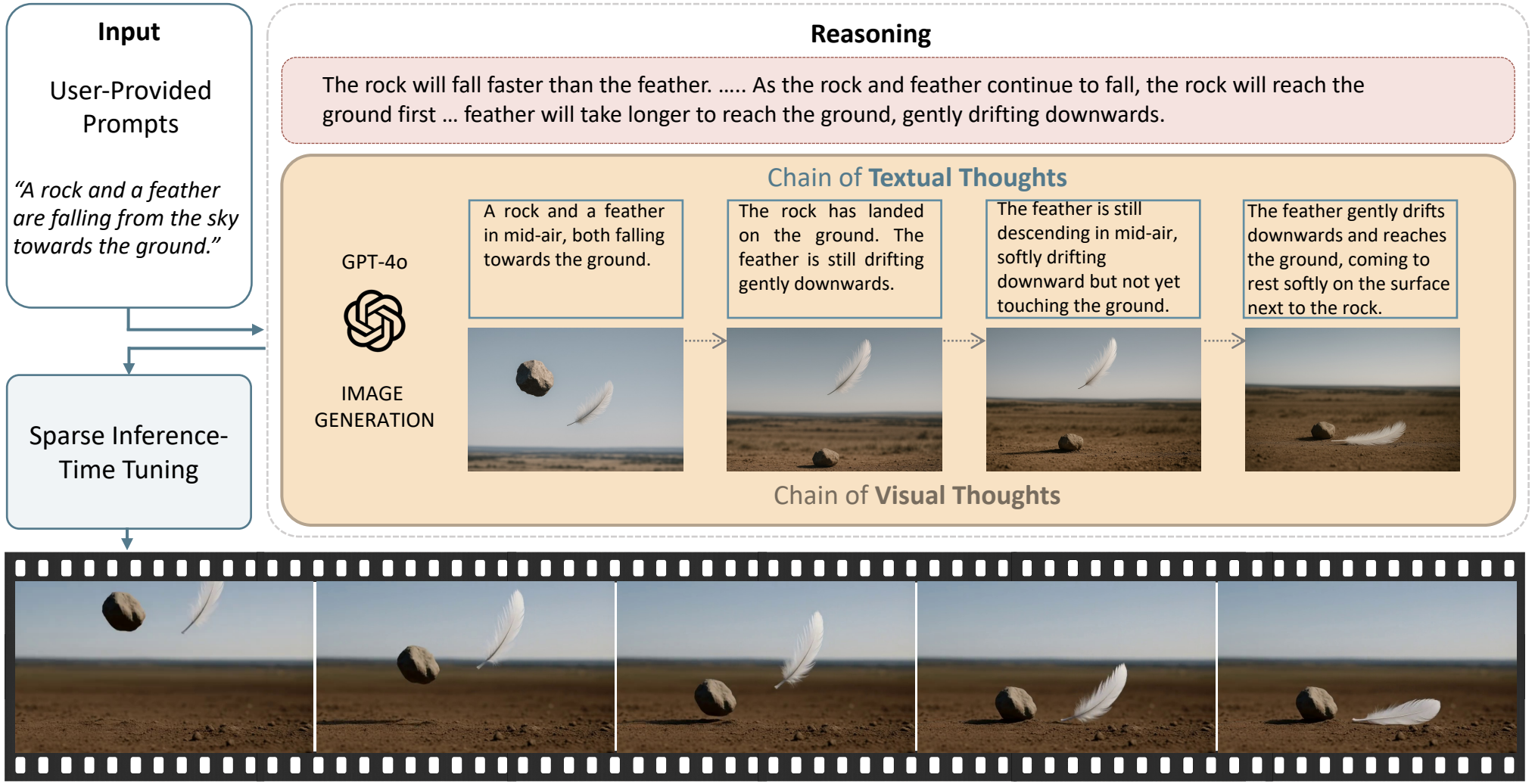
Challenges

- Video Generation Models
  - ✅ Produce smooth and visually appealing clips
  - ❌ Capture causal dynamics, particularly when an interactive trigger initiates a chain of visually grounded consequences (e.g., flipping a light switch or knocking over a glass).

Challenges

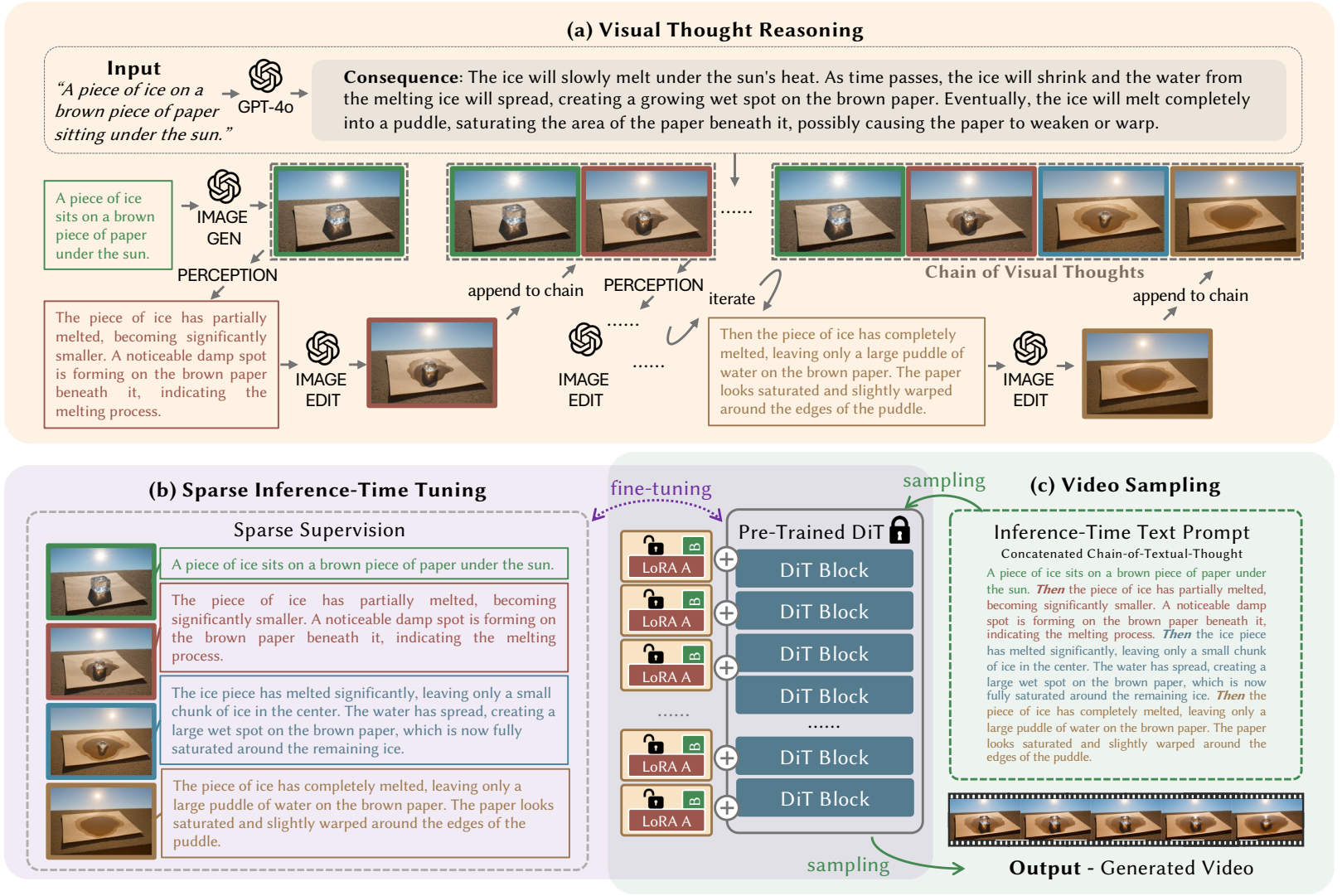
- Video Generation Models
  - ✅ Produce smooth and visually appealing clips
  - ❌ Capture causal dynamics, particularly when an interactive trigger initiates a chain of visually grounded consequences (e.g., flipping a light switch or knocking over a glass).
- In contrast, MLLMs (*e.g.*, GPT-4o)
  - ✅ Visual reasoning, future state prediction, and understanding causal implications
  - ❌ Cannot render videos or high-quality images

VChain Overview

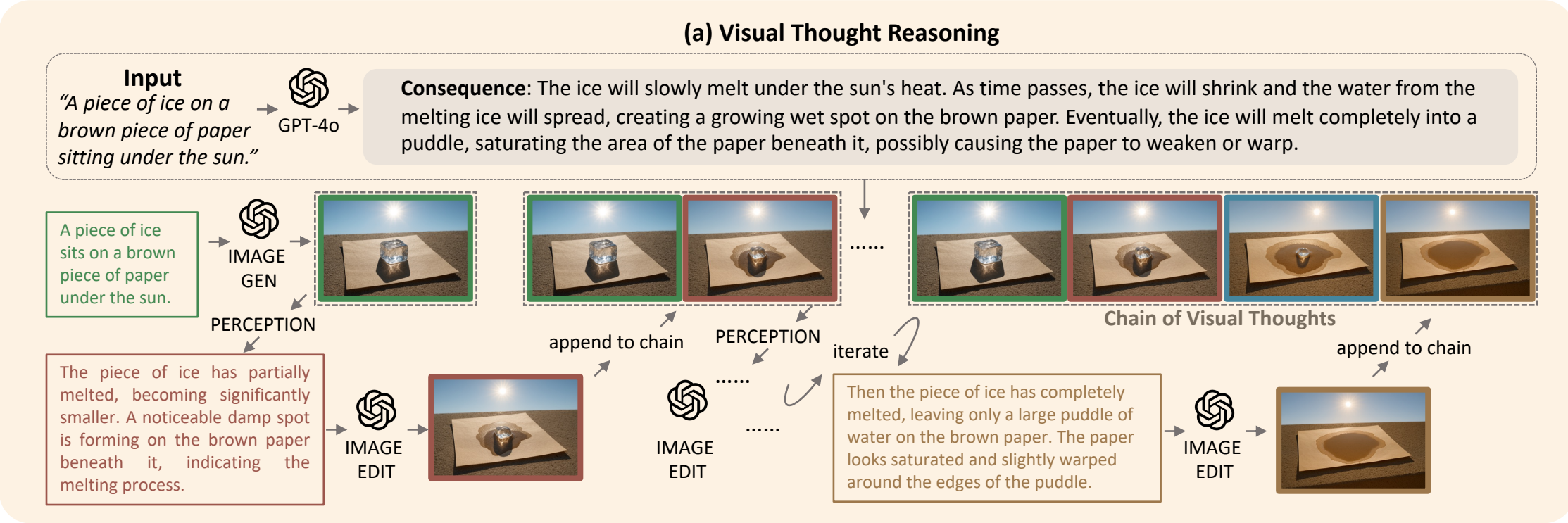


Output - Generated Video

3-Stage Inference-Time Method

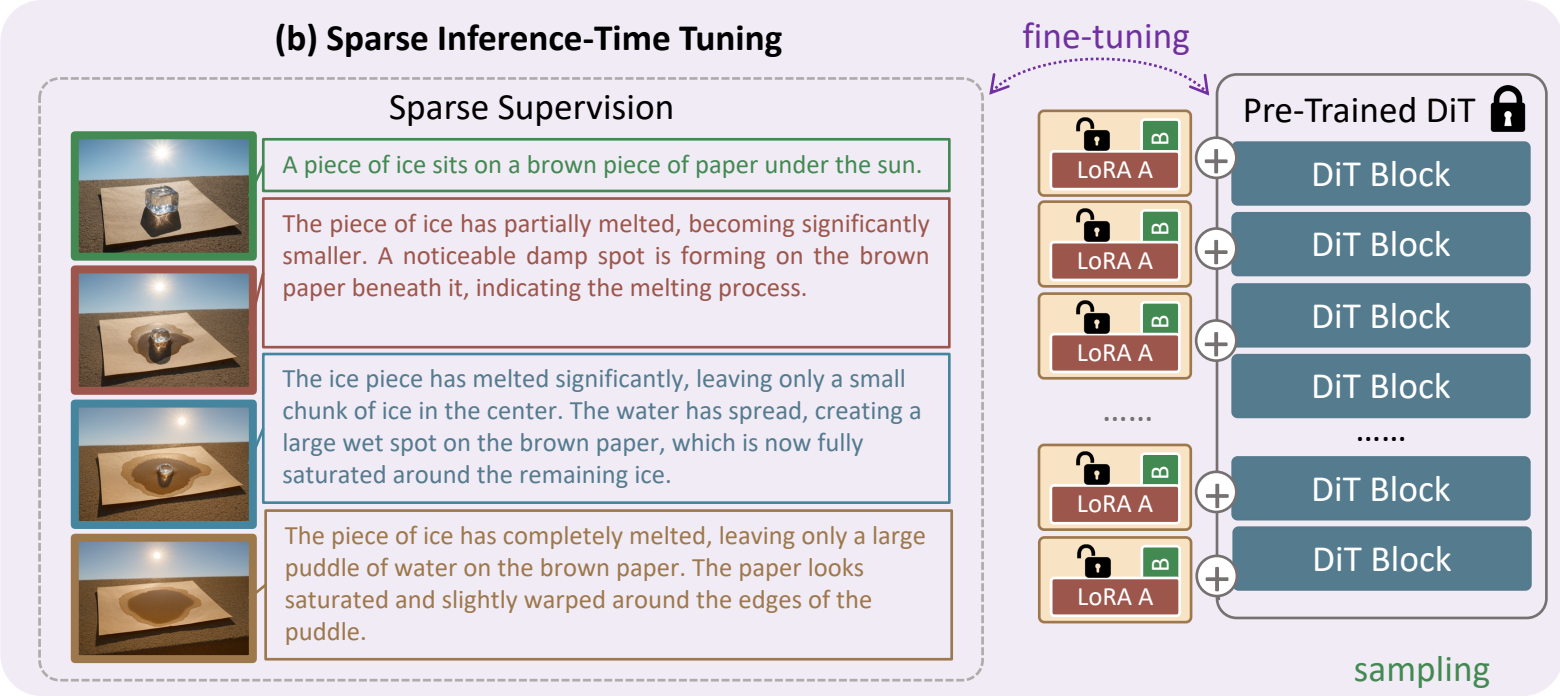


Step 1. Visual Thought Reasoning



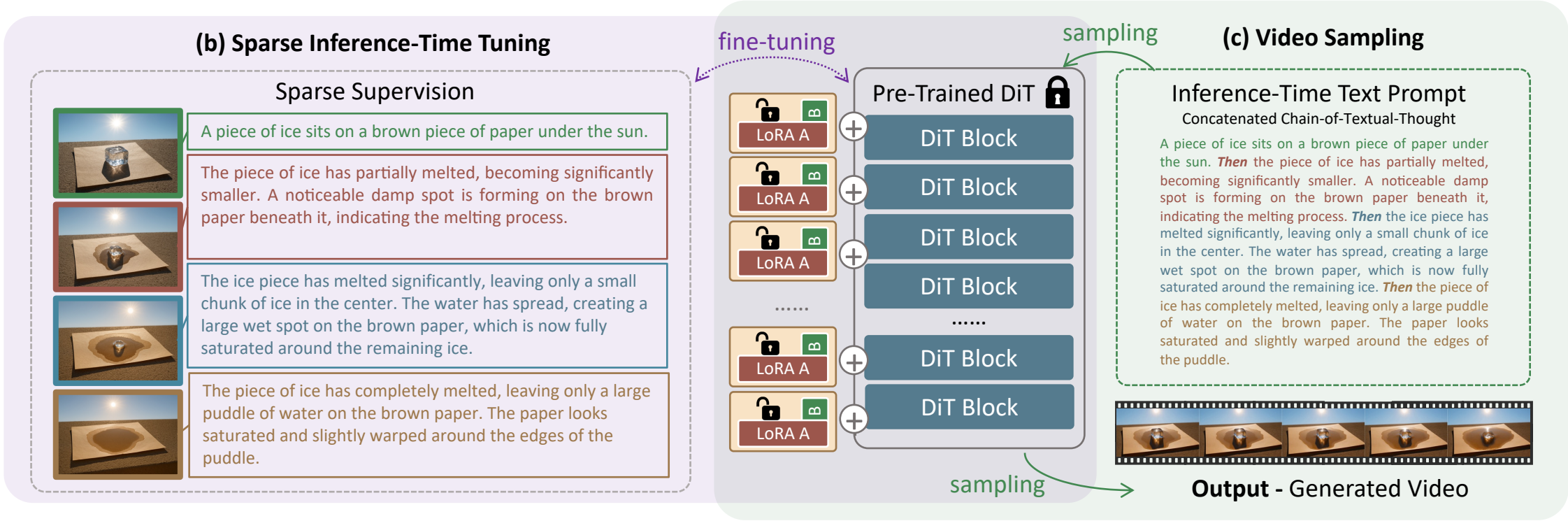
- Leverage the reasoning power of multimodal large language models

Step 2. Sparse Inference-Time Tuning



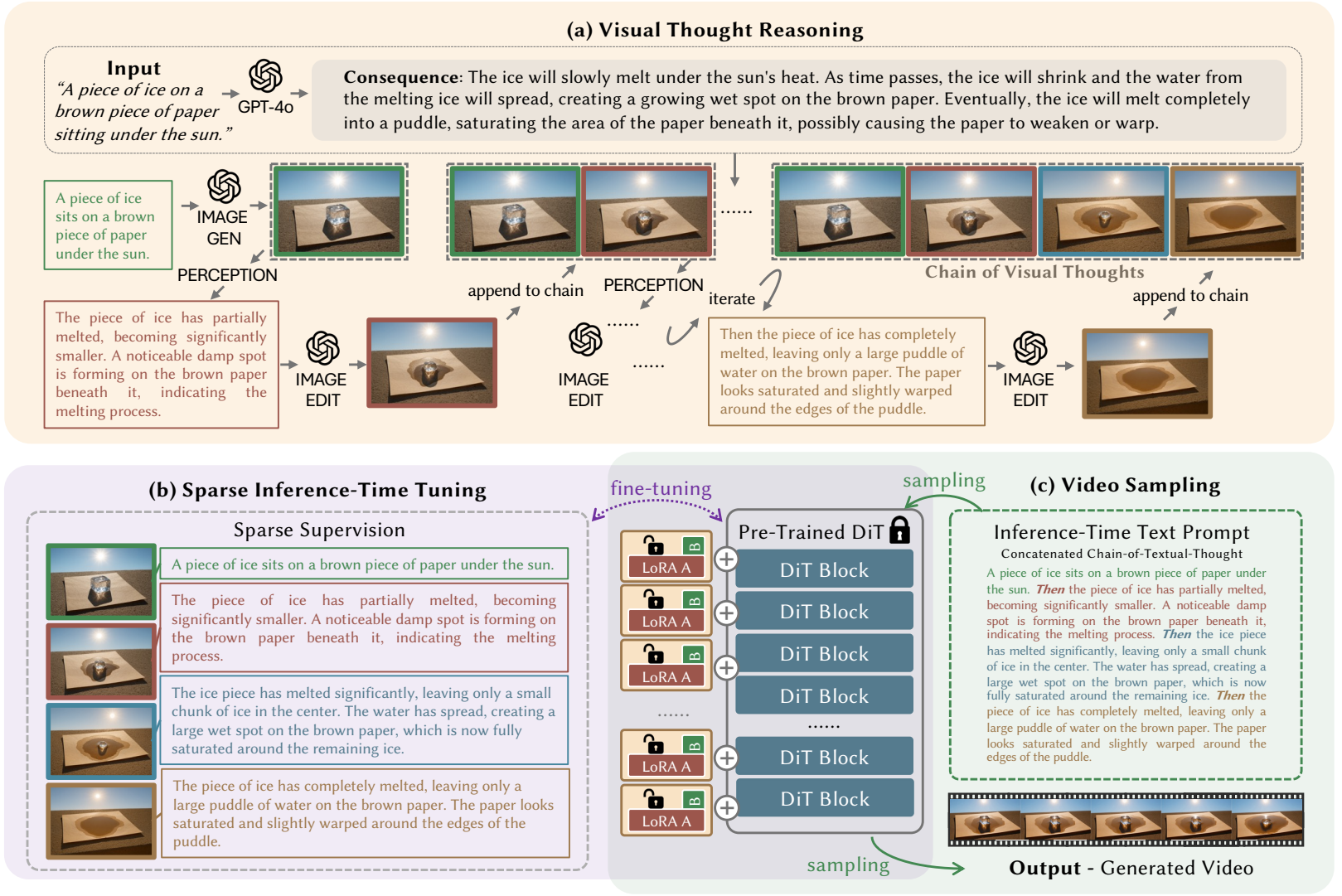
- Self-contained and efficient

Step 3. Video Sampling



- Self-contained and efficient

3-Stage Inference-Time Method



Ablation Study

Input Prompt: *“A rubber duck and a rock fall into a water tank.”*

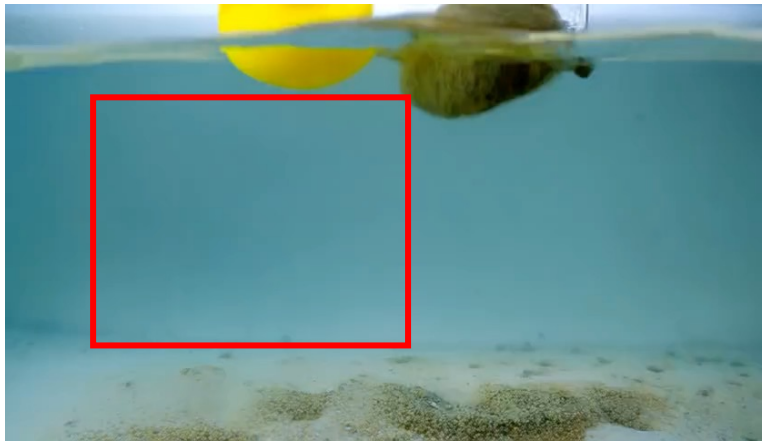
💭 Textual Thoughts

"A rubber duck and a rock are suspended above a water tank. Then the rubber duck and the rock have just entered the water. The water surface shows ripples spreading outwards, and bubbles are rising from the rock as it begins to sink. The rubber duck is floating near the surface, creating small splashes around it, while the rock is visible just below the surface, descending rapidly towards the bottom of the tank. Then the rock has settled at the bottom of the water tank, positioned on its side. Sediments are slightly disturbed around it, creating a faint cloudy area. The rubber duck is floating on the water's surface, slightly off-center, facing forward. Small ripples emanate outwards from both the duck and where the water was disturbed by the rock. The bubbles from the rock have mostly dissipated by now. The scene remains well-lit, capturing the clear distinction between the bright yellow of the duck and the dull gray of the rock."

👁 Visual Thoughts

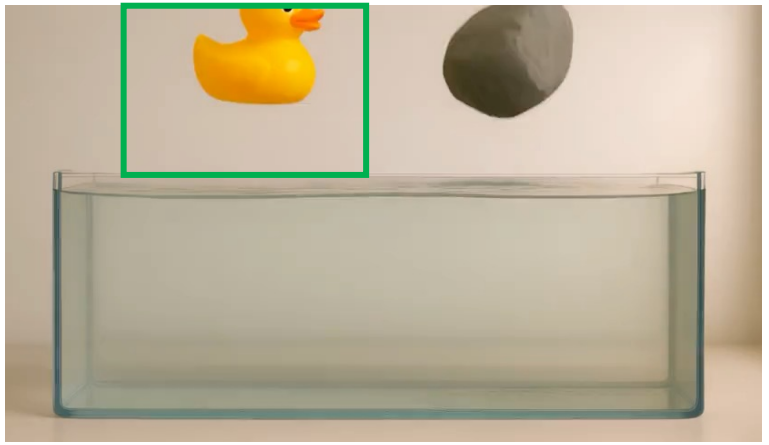
Without Visual Thought

❌ The rubber duck should not be submerged.
It should float on the water's surface.



VChain (Ours)

✅ It's important to see the Visual Thoughts during inference,
to visually see that the rubber duck floats on the water surface rather than sinking.



Baseline Comparisons

Input Prompt: *"A ball is dropped onto a pillow."*

❌ The ball should not bounce back into the air.

T2V

❌ The ball is not supposed to hover over the pillow.

T2V + Prompt Aug

✅ The ball hits the pillow, compresses it slightly, and settles with minimal or no bounce.

VChain (Ours)

Ablation Study

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

👁 Visual Thoughts

Input Prompt:
ice cream cone is left out in the sun."

❌ Although it leverages textual thoughts like "melt" and "cream flowing," it fails to envision the correct visual pattern of typical ice cream melting.

Without Visual Thought

❌ Without tuning, warping artifacts emerge when attempting to bridge spatial misalignments between Visual Thought frames.

Without Sparse Tuning

✅ We use Visual Thoughts to infer how ice cream should visually melt, and Sparse Tuning aligns the video generation prior with these keyframes.

VChain (Ours)

Input Prompt: *“Milk is poured into a cup of black coffee.”*

T2V

T2V + Prompt Aug

Without
Visual Thought

Without
Sparse Tuning

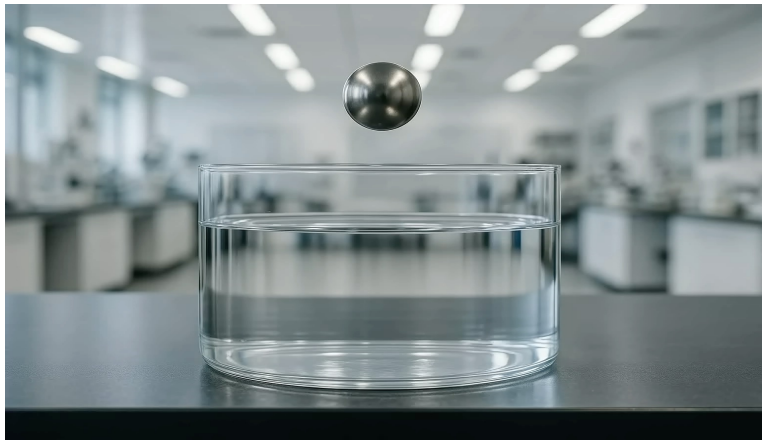
VChain (Ours)

Input Prompt: *“A steel ball is dropped into water.”*

T2V

T2V + Prompt Aug

Without
Visual Thought



Without
Sparse Tuning

VChain (Ours)

Input Prompt: *“POV: You are catching a ball thrown by a friend.”*

T2V

T2V + Prompt Aug

Without
Visual Thought

Without
Sparse Tuning

VChain (Ours)

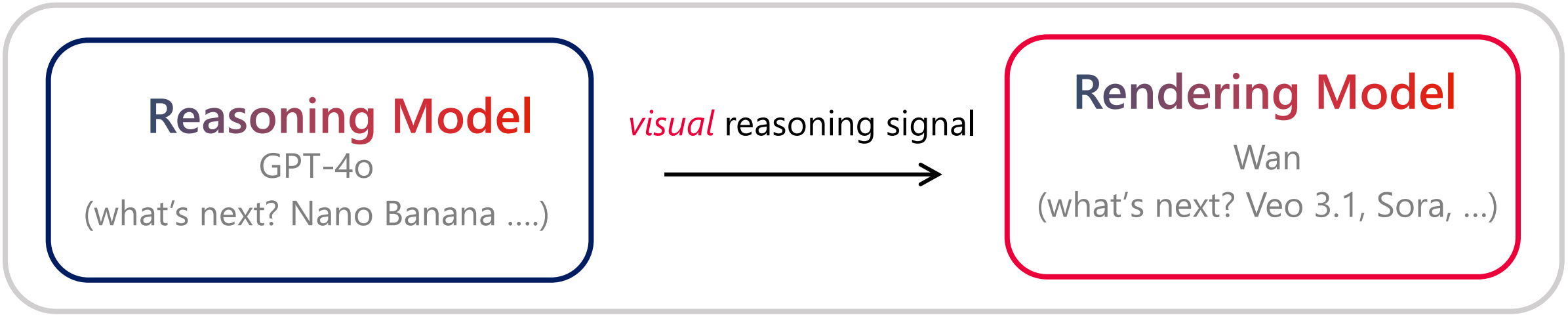
Discussions of VChain

- Inject reasoning signal at inference time
- Self-contained
- Efficient and effective

Discussions of VChain

- Inject reasoning signal at inference time
- Self-contained
- Efficient and effective

VChain is an inference-time tuning framework for reasoning in video generation



Limitations

VChain: Chain-of-Visual-Thought for Reasoning in Video Generation






Ziqi Huang[1] Ning Yu[2]✉† Gordon Chen[1] Haonan Qiu[1] Paul Debevec[2] Ziwei Liu[1]✉

✉Corresponding Authors. †Project Lead.

[1]*Nanyang Technological University* [2]*Eyeline Labs*

Multimodal Reasoning

Within-Shot *Event Reasoning*

VChain: models event plausibility

Multimodal Reasoning

Within-Shot *Event Reasoning* → Cross-Shot *Cinematic Reasoning*

VChain: models event plausibility

Cut2Next: models coherent shot transitions

Cross-Shot Cinematic Reasoning

Cut2Next: Generating Next Shot via In-Context Tuning

Jingwen He, Hongbo Liu, Jiajun Li, Ziqi Huang, Yu Qiao, Wanli Ouyang, Ziwei Liu

The Challenge: From Stunning Shots to Coherent Stories

How do we ensure narrative continuity and cinematic flow?

+ ? →

×

Shot A.
Stunning, but Isolated.

Typical Output: Inconsistent Shot

Our Solution: Introducing Next Shot Generation (NSG)

+

Cut2Next Model

✅

Shot A

Shot A

Cut2Next in Action: Mastering Cinematic Cuts

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Shot/Reverse Shot

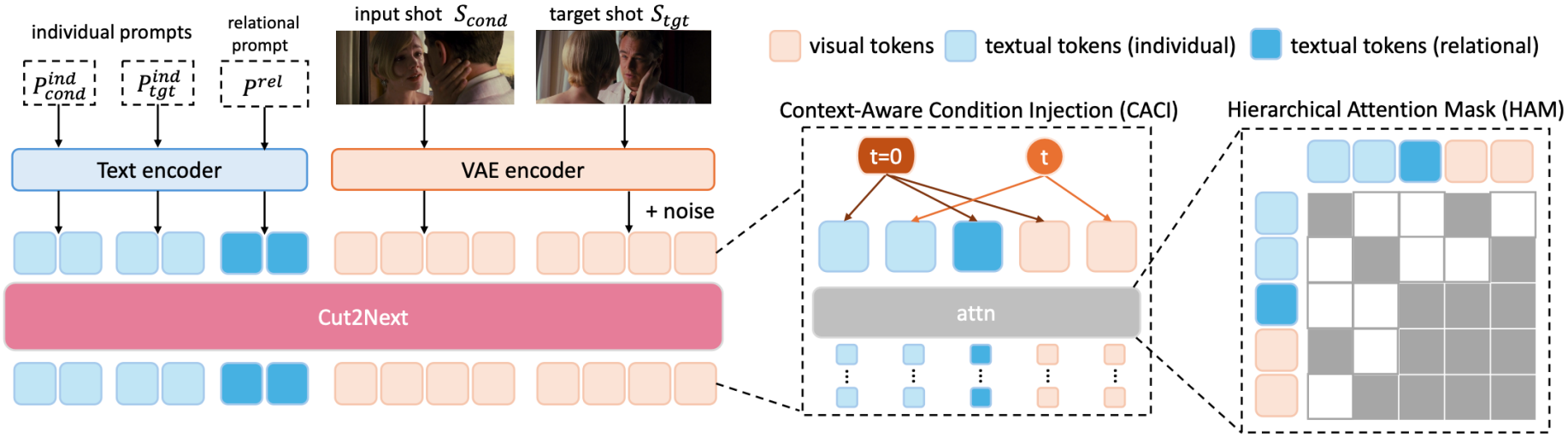
Multi-Angle

Cutaway

Cut-In

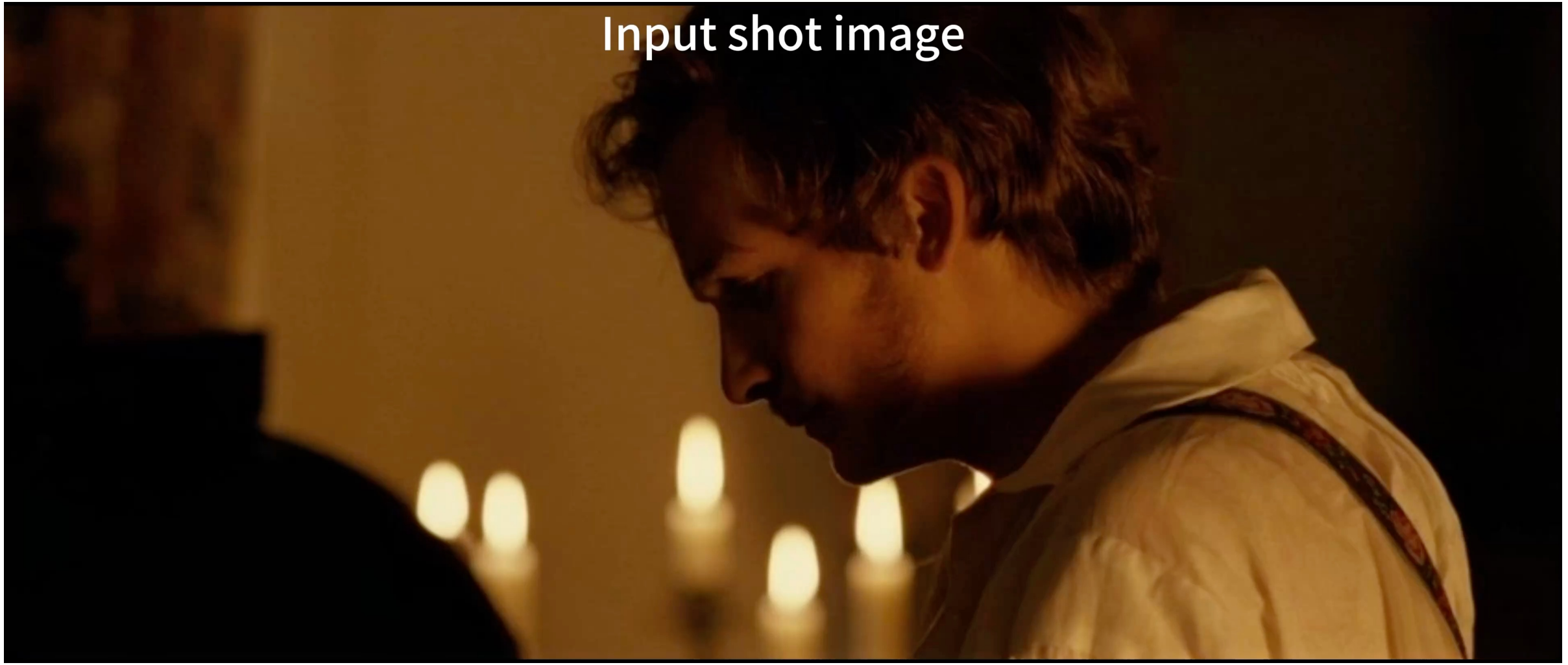
Method Overview

- High-quality datasets: RawCuts & CuratedCuts
- Hierarchical Prompts
- Core Model: Cut2Next (built on FLUX.1)
  - Key Modules: CACI & HAM (No extra parameters)



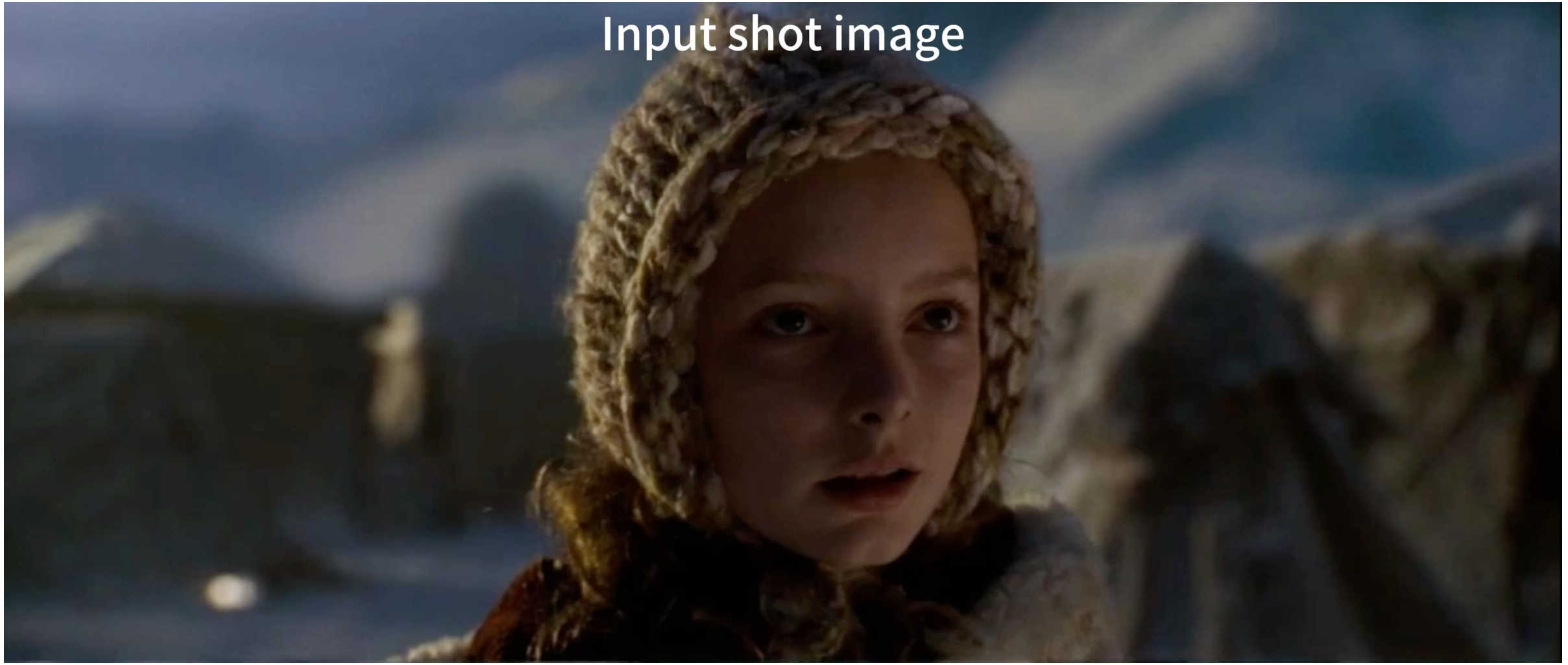
Visual Results

Visual Results



Visual Results

Visual Results



Visual Results

Multimodal Reasoning

Within-Shot *Event Reasoning* → Cross-Shot *Cinematic Reasoning*

VChain: models event plausibility

Cut2Next: models coherent shot transitions

Multimodal Reasoning

Within-Shot *Event Reasoning* → Cross-Shot *Cinematic Reasoning*

VChain: models event plausibility

VBench-2.0: evaluates event reasoning

Cut2Next: models coherent shot transitions

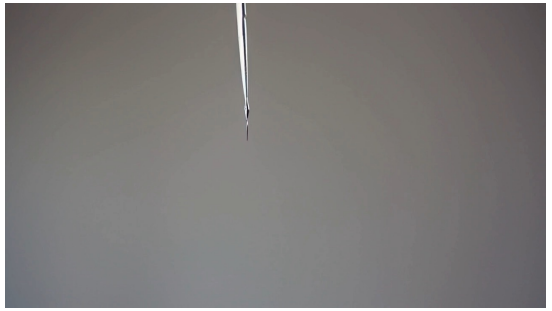
Within-Shot Event Reasoning
Understanding & Evaluation

VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, Ziwei Liu

Problem

- Disregard of Intrinsic Faithfulness
  - **Look well != Real**
    - Wrong Physics
    - Fake Motion
    - Low-quality Motion
    - Entity Sudden Fusion/Division



Wrong Physics

Fake Motion

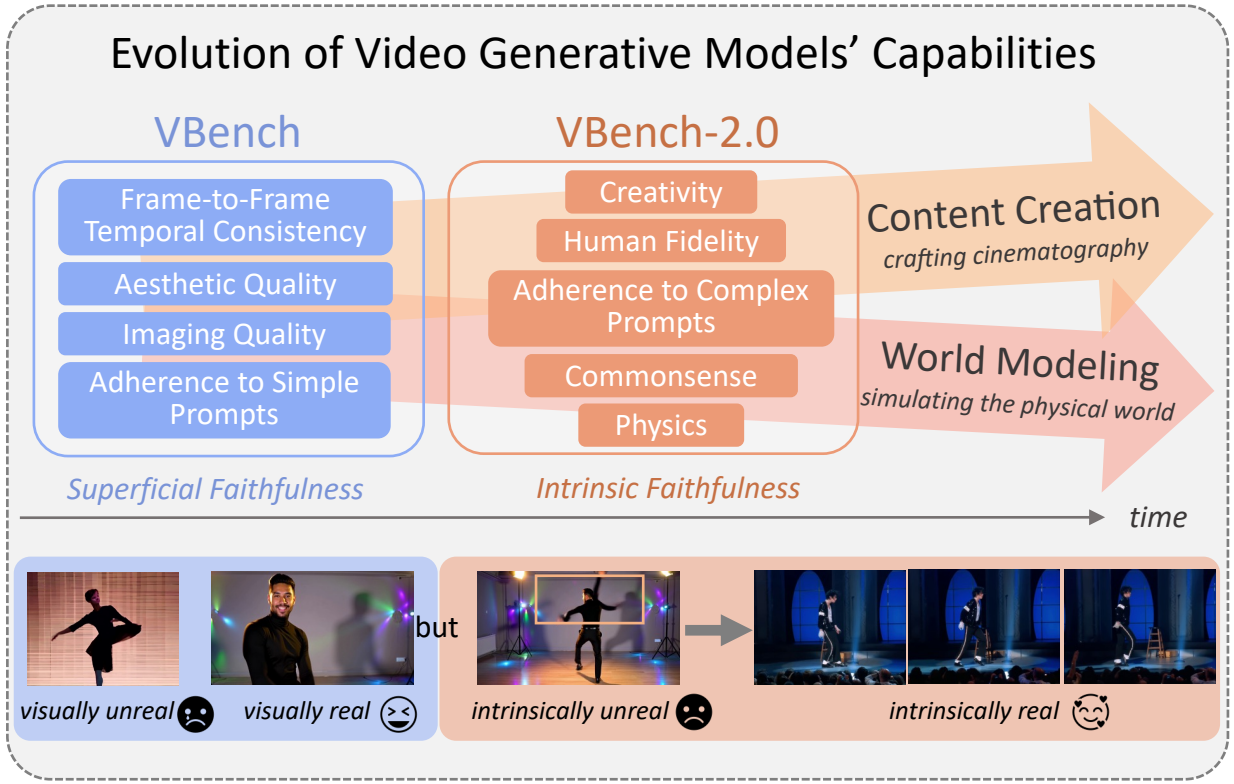
- Inaccurate/Unreproducible Evaluation
  - **Tuning with Scarce Short-Generated Videos**
  - **Using Close-source Models (gpt)**

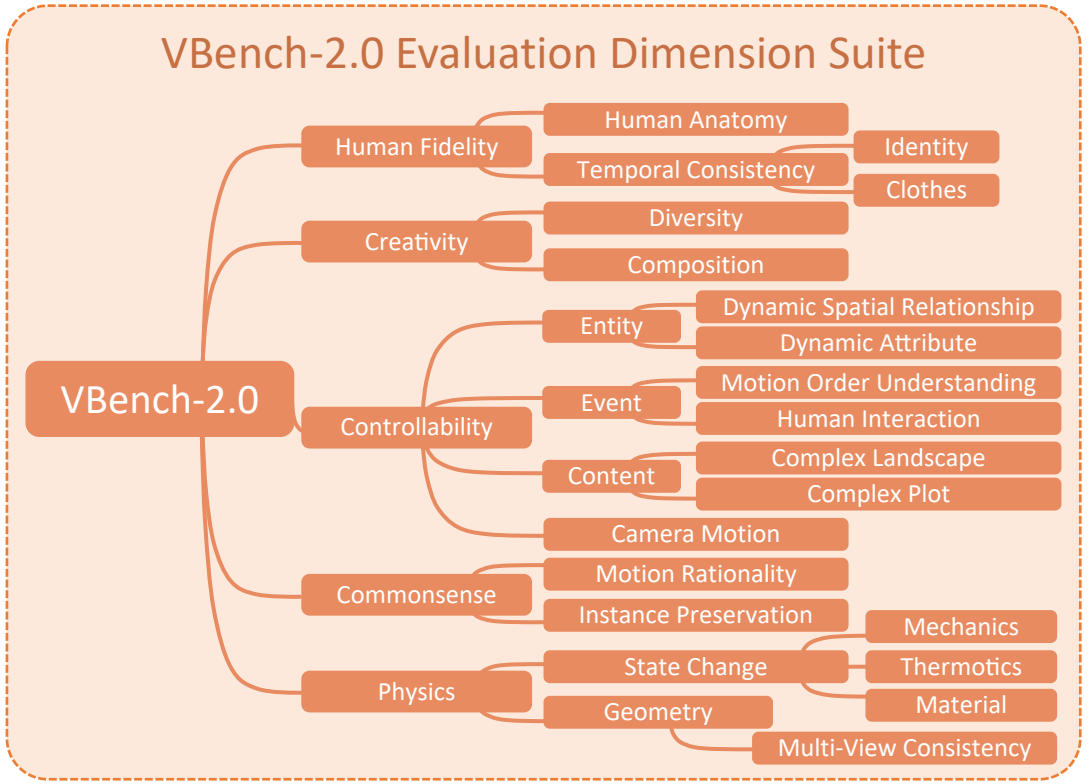
Low-quality Motion

Sudden Fusion/Division

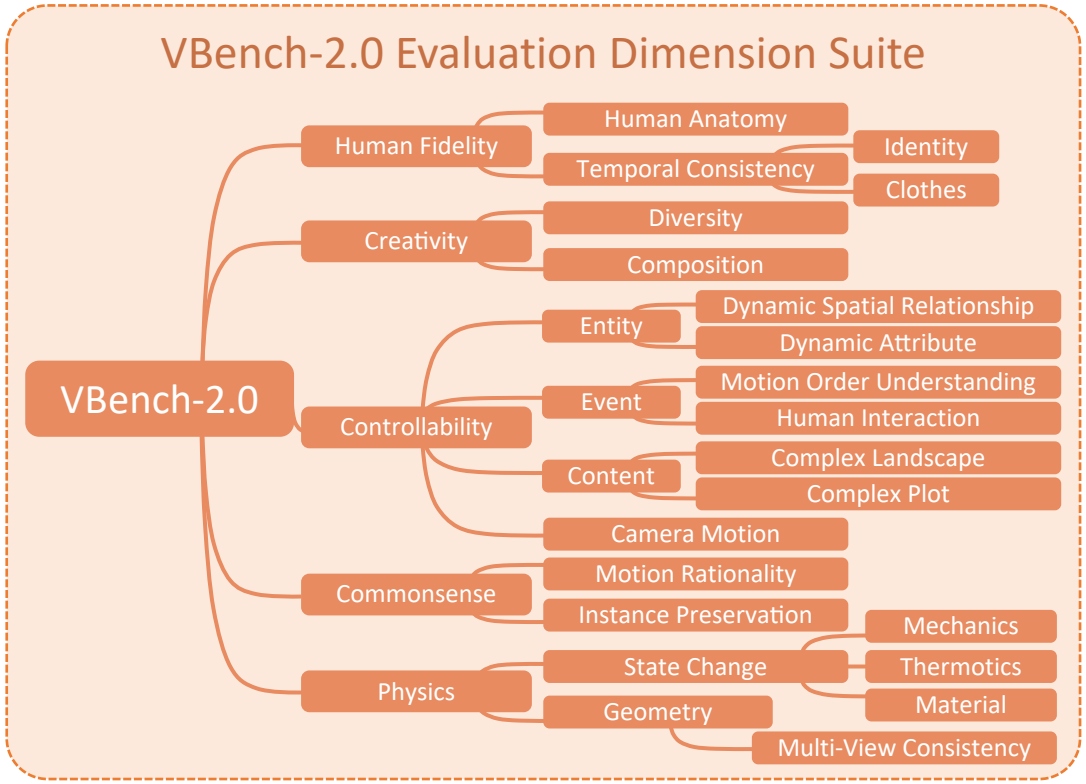
VBench-2.0 For Intrinsic Faithfulness

- Targeting on Next-generation Video Generation Models
  - Content Creation
    - Creativity, Human Fidelity, Prompt Controllability
  - World Modeling
    - Commonsense, Physics, Prompt Controllability



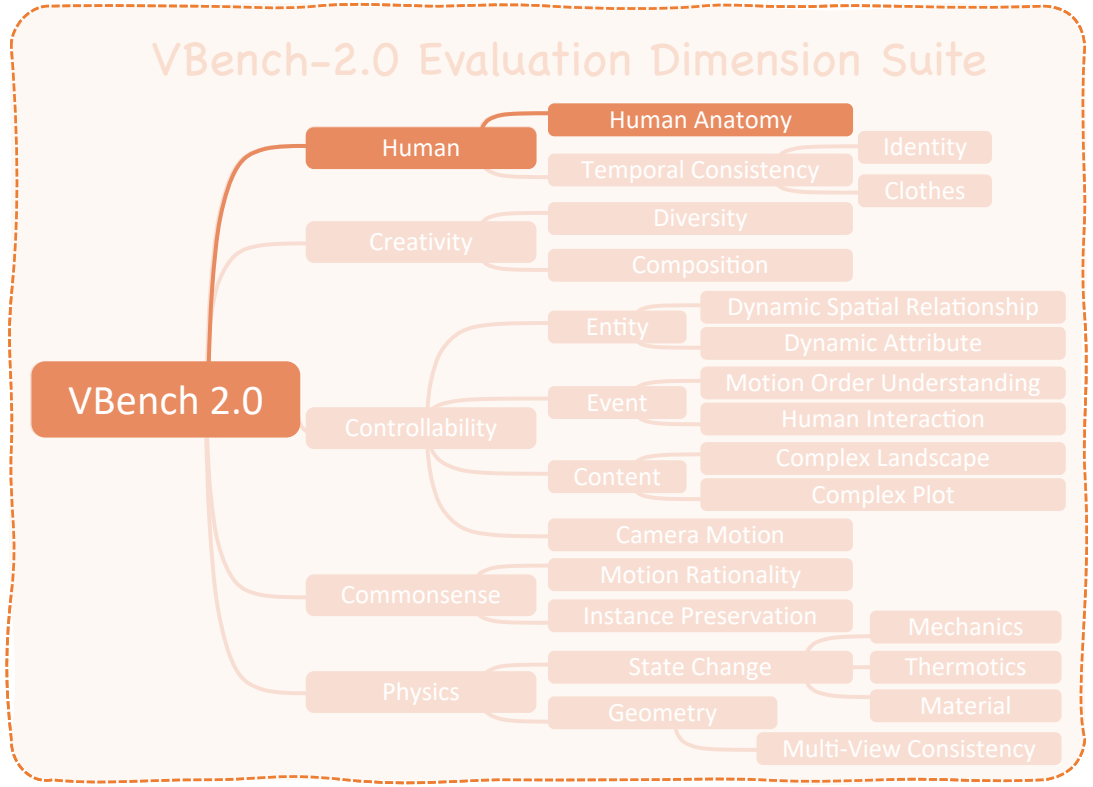


Dimension Suite



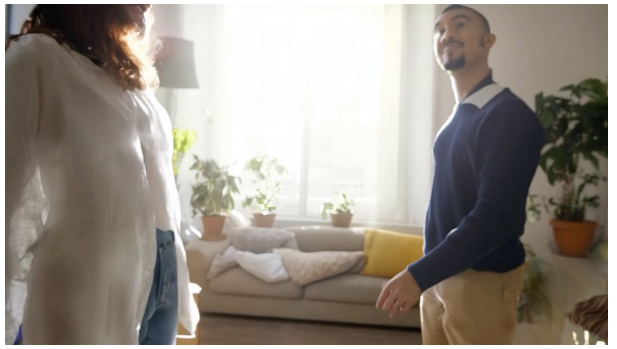
- 18 ability dimensions, hierarchical and disentangled
- Why Multiple Dimensions?
  - Reveal individual model's strengths and weaknesses
  - Different people prioritize each ability dimension differently
- Unique Dimensions
  - Human Anatomy
    - Motion problem in dance, gymnastics
  - Instance Preservation
    - Entity sudden fusion/division
  - Composition
    - Non-realistic compositions
  - Complex Plot
    - Minute-level Long Story

Evaluation Dimension: Human Anatomy



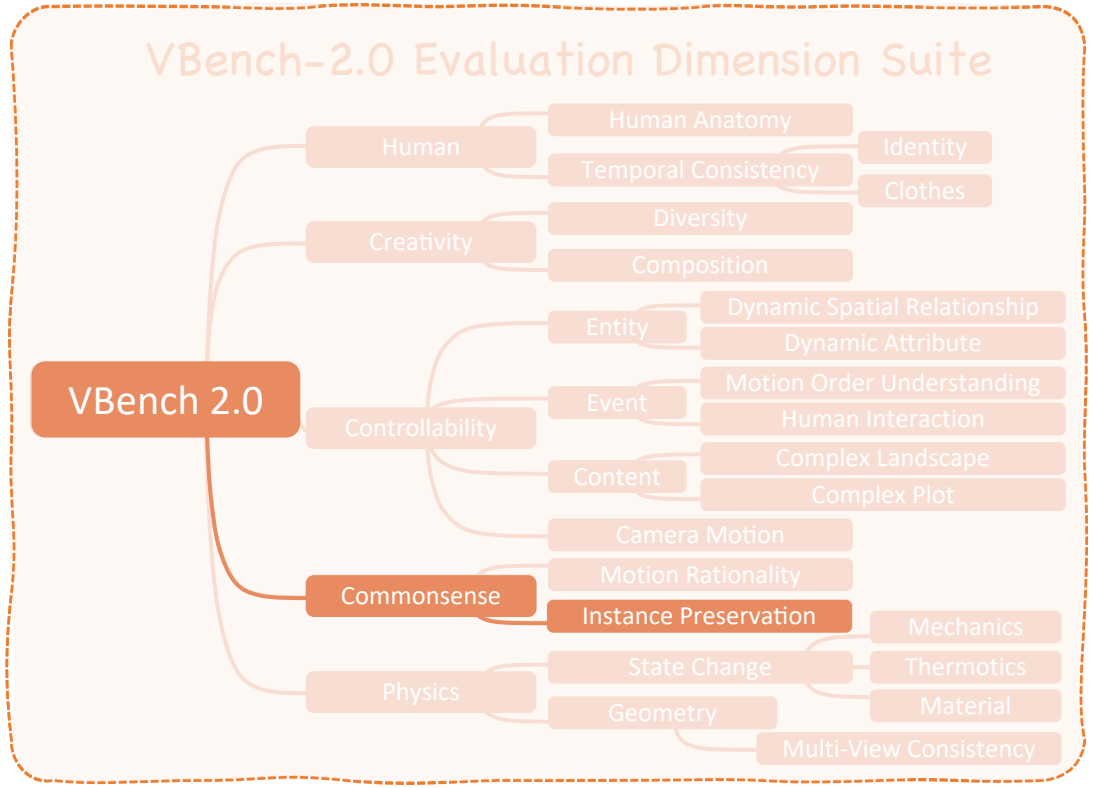
score 91.29%

score 24.21%



Is there any anatomical abnormality in the video?

Evaluation Dimension: Instance Preservation

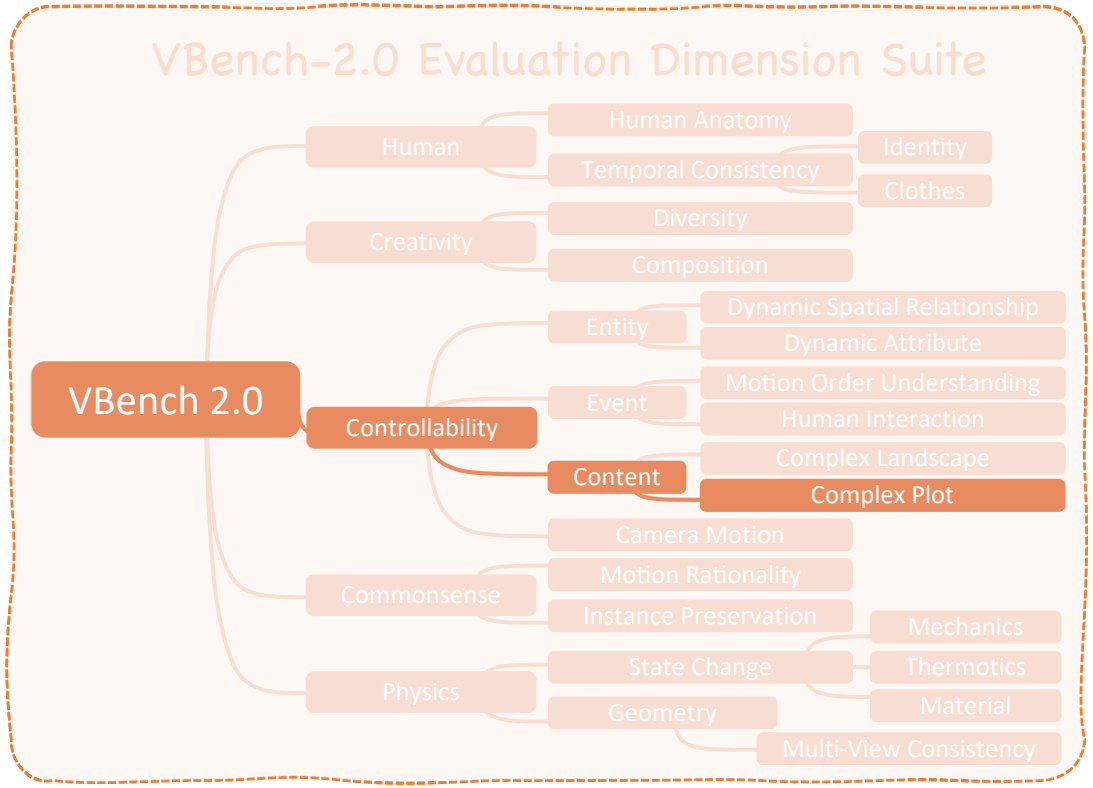


score 100.0% *(better)*

score 65.42%

Does the video show entities suddenly appearing, disappearing, fusing, or dividing?

Evaluation Dimension: Complex Plot



score 40.0% *(better)*

score 0.00%

Does the video follow the plot points one by one?

1.Emily discovers a sleeping sea monster by a shimmering silver lake.

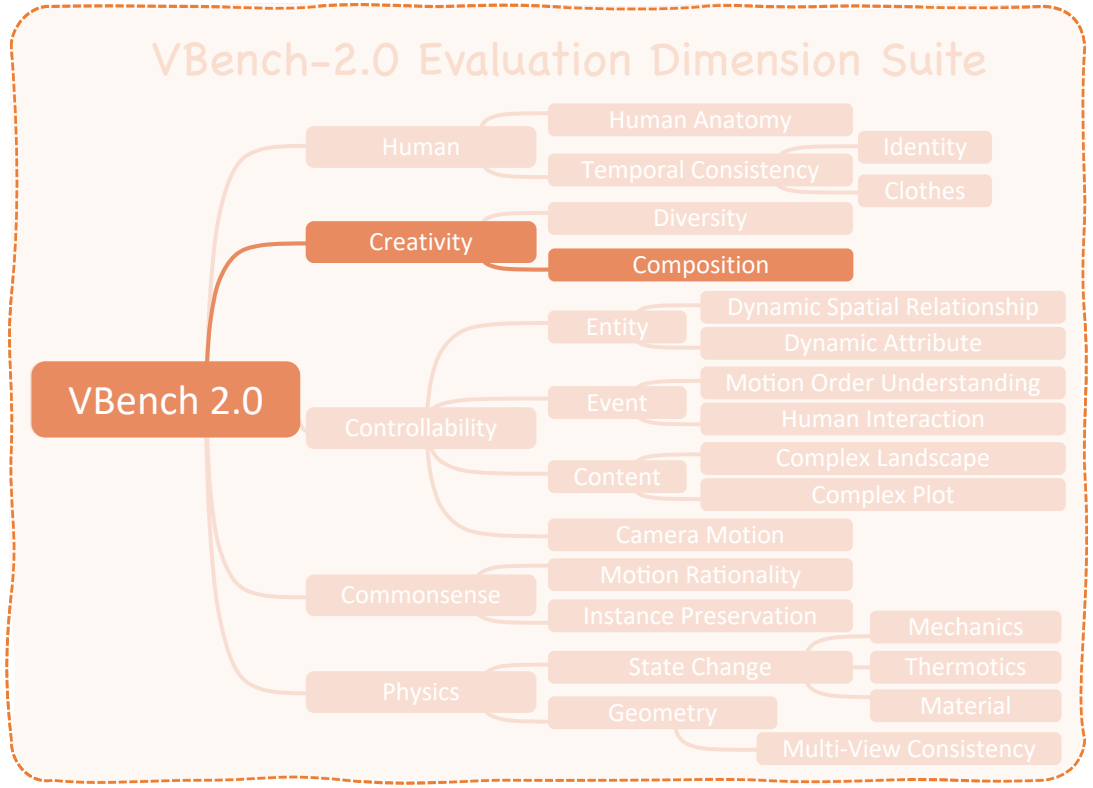
2.She finds a shiny blue gemstone hidden in the monster's scales and awakens it.

3.The sea monster reveals it is a cursed prince and gives Emily a quest to find three royal relics.

4.Emily embarks on an adventure to find the relics and breaks the curse.

5.The sea monster regains his human form, and they become lifelong friends

Evaluation Dimension: Composition

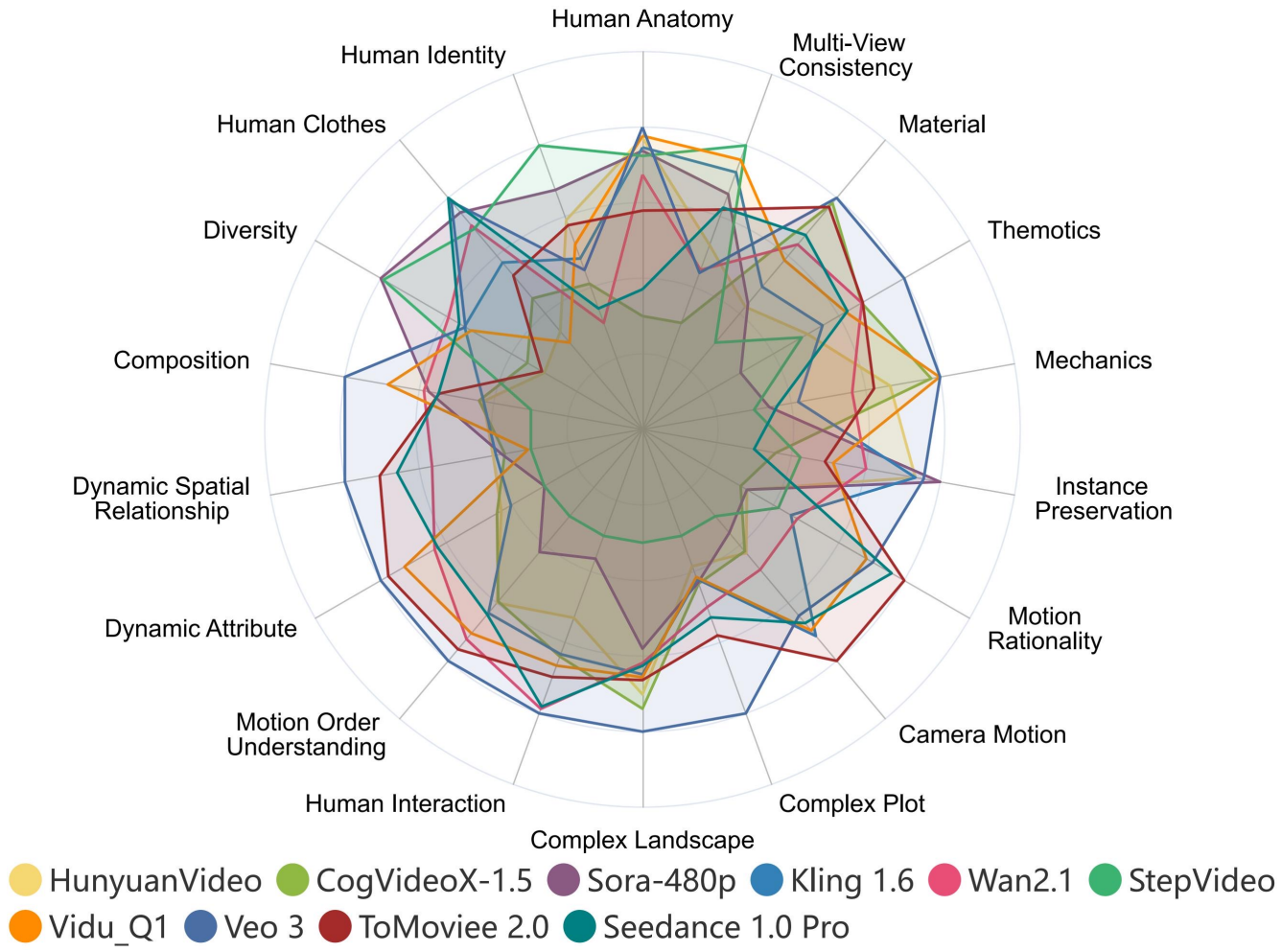


score 100.0% *(better)*

score 0.00%

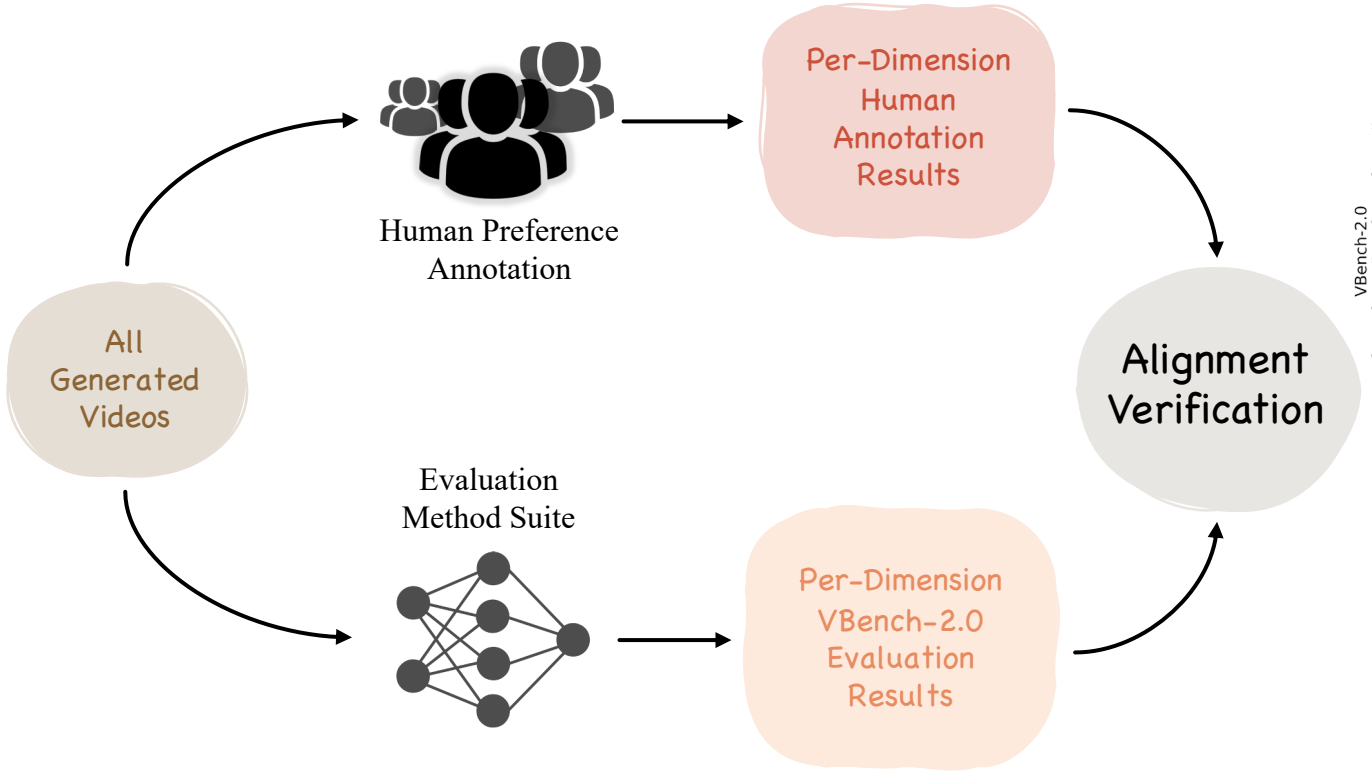
Is the lion performing a handstand?

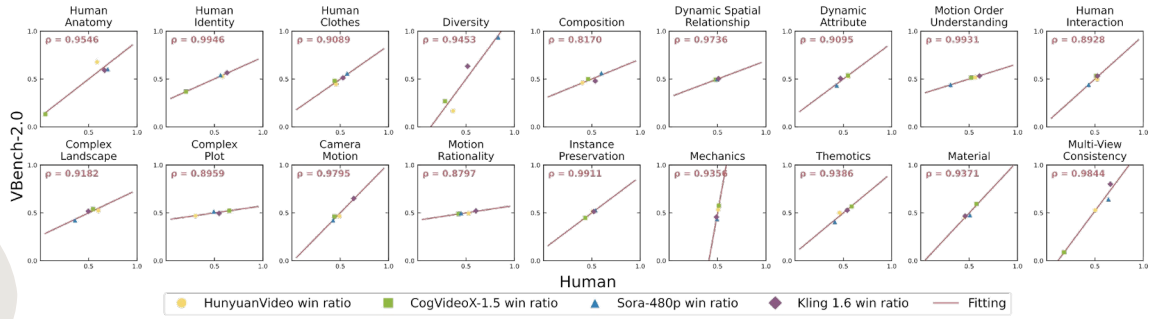
Evaluation Results



- Key Limitations of Recent SOTA Models
  - Generating Complex Plots
  - Controllability in Simple Dynamics
- Prompt Engineering Matter
  - Balancing Controllability and Creativity
  - Partially Compensates for Physical Reasoning Gaps (prompt correct results)
- Superficial Faithfulness vs. Intrinsic Faithfulness: Do Not Miss Out on Any Pillar

Human Alignment of VBench-2.0





VBench-2.0 evaluations across all dimensions closely match human perceptions.

VBench-2.0 Leaderboard

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Evaluation Dimension

- [x] Human Anatomy
- [x] Human Clothes
- [x] Human Identity
- [x] Composition
- [x] Diversity
- [x] Mechanics
- [x] Material
- [x] Thermotics
- [x] Multi-View Consistency
- [x] Dynamic Spatial Relationship
- [x] Dynamic Attribute
- [x] Motion Order Understanding
- [x] Human Interaction
- [x] Complex Landscape
- [x] Complex Plot
- [x] Camera Motion
- [x] Motion Rationality
- [x] Instance Preservation

Model (alphabetical order)	Certification	Sampled by	Evaluated by	Accessibility	Date	Total Score	Creativity Score	Commonsense Score
Veo 3	Gold	VBench Team	VBench Team	API	2025-09-04	66.72%	60.85%	69.48%
Vidu Q1 (2025-04-17)	Silver	ShengShu Team	VBench Team	API	2025-04-21	62.70%	56.54%	65.98%
ToMoviee 2.0	Silver	Wondershare Team	VBench Team	API	2025-09-18	61.78%	45.96%	67.41%
Wan2.1	Gold	VBench Team	VBench Team	Open Source	2025-03-28	60.20%	55.25%	63.98%
Seedance 1.0 Pro (2025-05-28)		Joyme Team	Joyme Team	API	2025-06-26	59.81%	53.04%	64.31%
Kling 1.6	Gold	VBench Team	VBench Team	API	2025-03-28	59.00%	48.58%	65.45%
Sora-480p	Gold	VBench Team	VBench Team	API	2025-03-28	58.38%	60.57%	64.32%
StepVideo	Gold	VBench Team	VBench Team	Open Source	2025-03-28	55.78%	51.26%	60.78%
HunyuanVideo	Gold	VBench Team	VBench Team	Open Source	2025-03-28	55.30%	41.84%	63.44%
CogVideoX-1.5	Gold	VBench Team	VBench Team	Open Source	2025-03-28	53.35%	43.65%	58.18%

Join our Leaderboard
Let's Beat Veo3!

Fully Open-Source

- ***Evaluation Method Suite (code)***
- ***Prompt Suite (text prompts)***
- ***Human Preference Annotations***
- ***Generated Videos (mp4)***

Veo3, Wan2.1, Kling1.6, Vidu-Q1, Sora, HunyuanVideo, ToMoviee 2.0, Seedance 1.0 Pro, StepVideo, CogVideoX (more to be added)

Github

Multimodal Reasoning

Within-Shot *Event Reasoning*

How humans understand causal logic

VChain: models event plausibility

VBench-2.0: evaluates event reasoning

→

Cross-Shot *Cinematic Reasoning*

How filmmakers compose stories

Cut2Next: models coherent shot transitions

Multimodal Reasoning

Within-Shot *Event Reasoning*

How humans understand causal logic

VChain: models event plausibility

VBench-2.0: evaluates event reasoning

→

Cross-Shot *Cinematic Reasoning*

How filmmakers compose stories

Cut2Next: models coherent shot transitions

ShotBench: evaluates cinematic reasoning

Cross-Shot Cinematic Reasoning *Understanding & Evaluation*

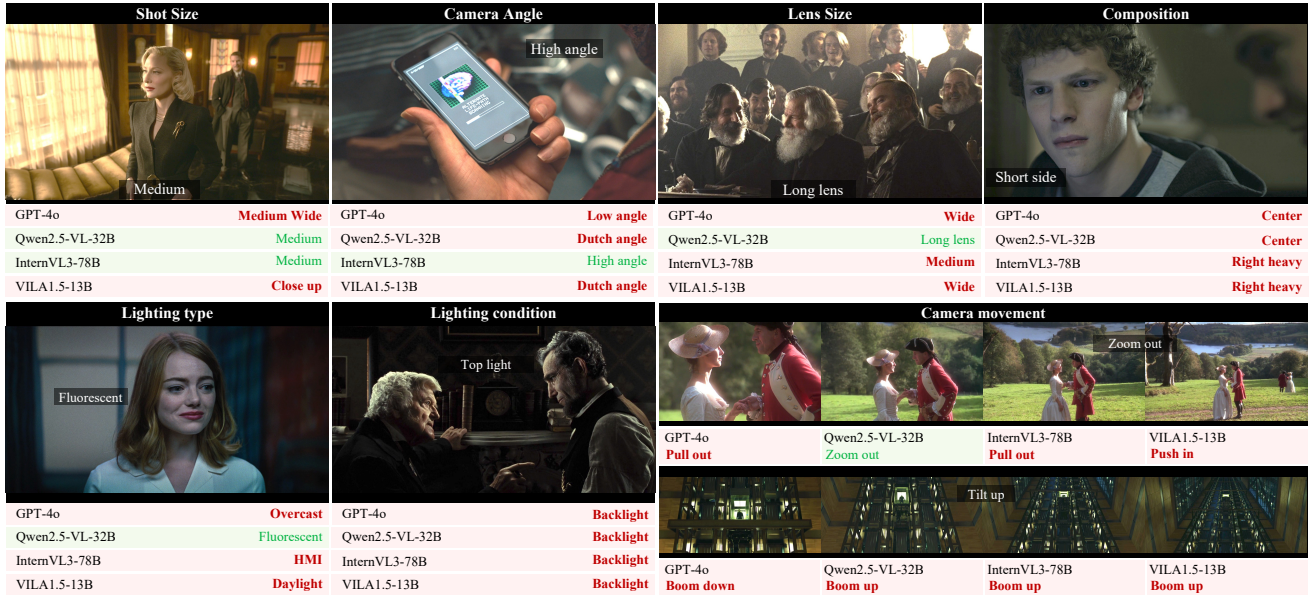
ShotBench: Expert-Level Cinematic Understanding in Vision-Language Models

Hongbo Liu, Jingwen He, Yi Jin, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Yangguang Li, Weichao Chen, Yu Qiao, Wanli Ouyang, Shengjie Zhao, Ziwei Liu

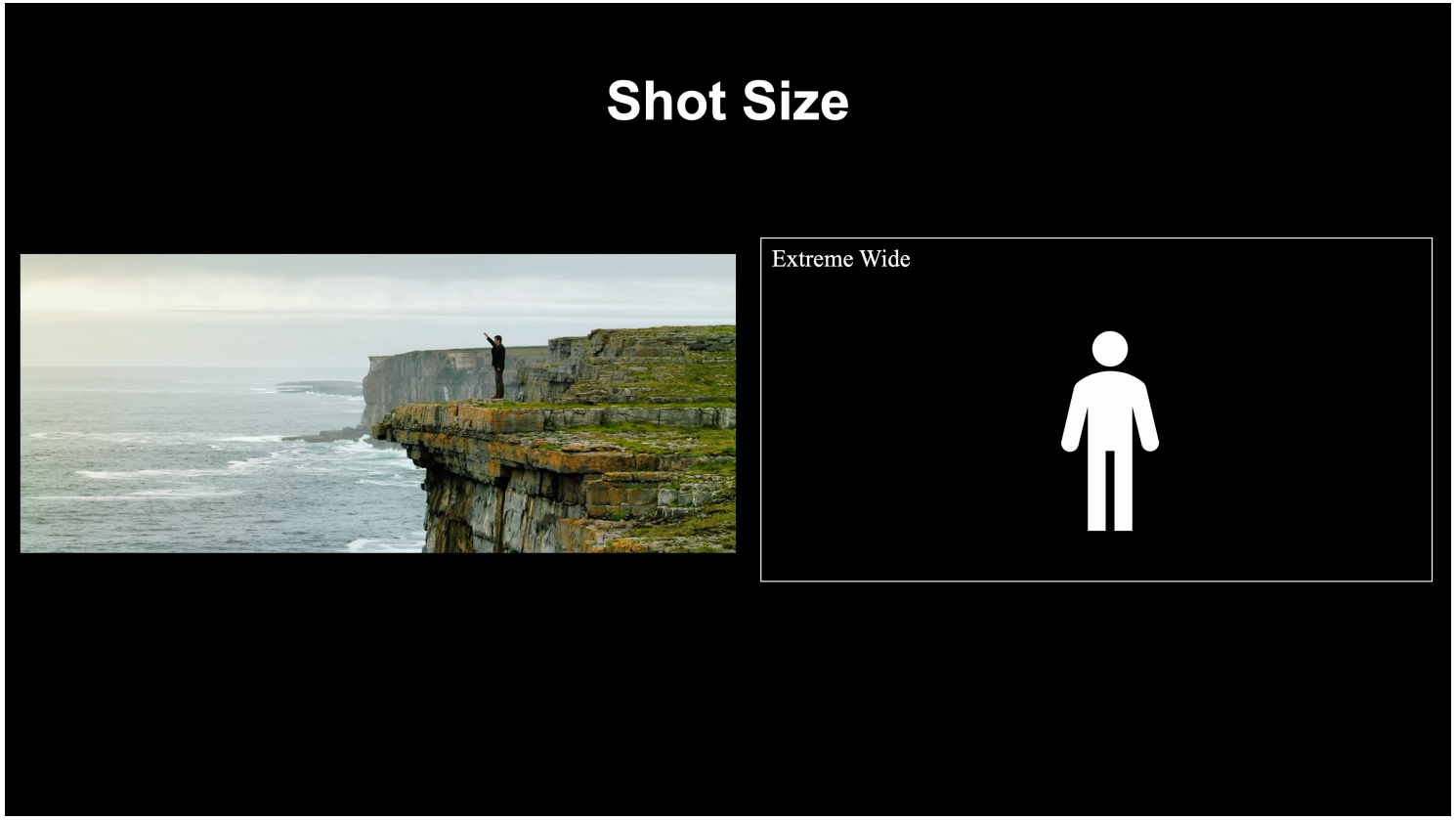
Challenges

- Cinematic Language Understanding
  - **Large-scale dataset and benchmark**
    - High quality movie shots datasets construction
    - Expert-level annotation
  - **Comprehensive evaluation and analysis**
    - Benchmarking cinematic understanding capabilities of leading MLLMs
    - Qualitative and quantitative analysis
  - **Strong MLLMs on Cinematography**
    - Fine-grained visual-terminology alignment
    - Enhanced spatial perception of camera position and orientation
    - Visual reasoning in cinematography

Current MLLMs not capable to understand cinematography

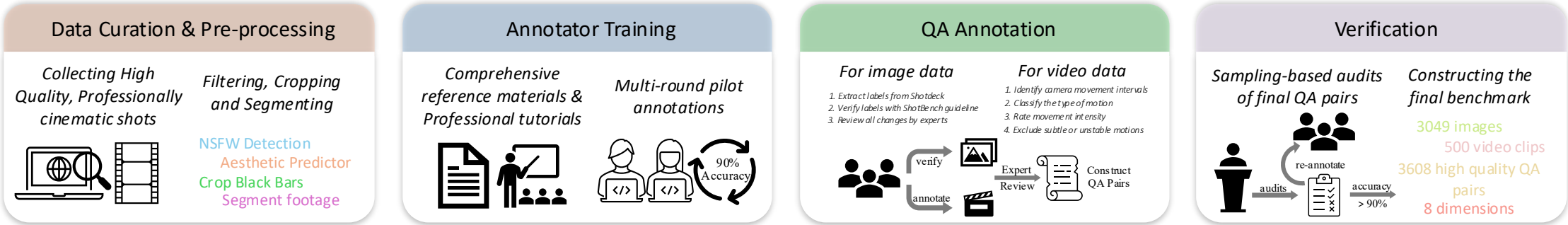


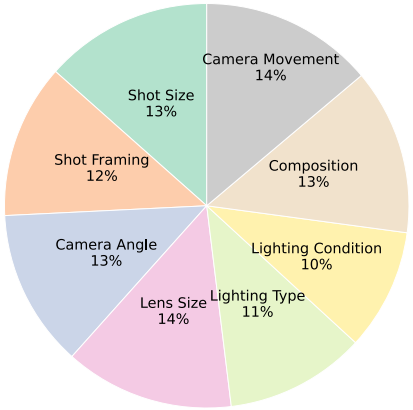
ShotBench: Expert-Level Cinematic Understanding in Vision-Language Models

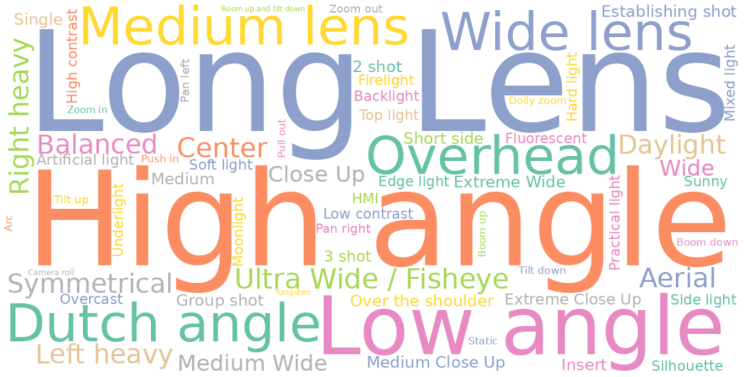


- Cover 8 core dimensions of cinematography language
- 3.5k images and videos curated from over two hundred Oscar-nominated films
- Expert annotated multiple-choice QA examples

High-quality Data Curation Pipeline



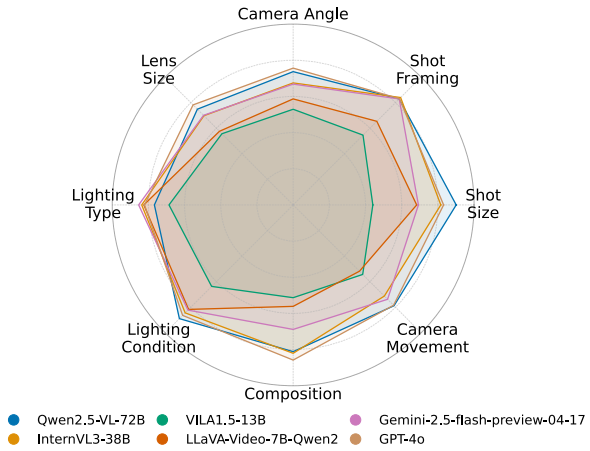


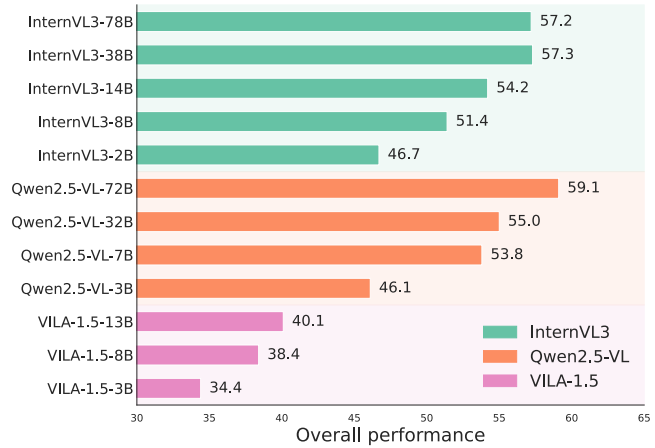


- Four-stage data construction pipeline
- Expert-in-the-loop for annotator training and verification

Analysis and Insights

Models	SS	SF	CA	LS	LT	LC	SC	CM	Avg
Open-Sourced VLMs									
Qwen2.5-VL-3B-Instruct	54.6	56.6	43.1	36.6	59.3	45.1	41.5	31.9	46.1
Qwen2.5-VL-7B-Instruct	69.1	73.5	53.2	47.0	60.5	47.4	49.9	30.2	53.8
LLaVA-NeXT-Video-7B	35.9	37.1	32.5	27.8	50.9	31.7	28.0	31.3	34.4
LLaVA-Video-7B-Qwen2	56.9	65.4	45.1	36.0	63.5	45.4	37.4	35.3	48.1
LLaVA-Onevision-Qwen2-7B-Ov-Chat	58.4	71.0	52.3	38.7	59.5	44.9	50.9	39.7	51.9
InternVL2.5-8B	56.3	70.3	50.8	41.1	60.2	45.1	50.1	33.6	50.9
InternVL3-2B	56.3	56.0	44.4	34.6	56.8	44.6	43.0	38.1	46.7
InternVL3-8B	62.1	65.8	46.8	42.9	58.0	44.3	46.8	44.2	51.4
InternVL3-14B	59.6	82.2	55.4	40.7	61.7	44.6	51.1	38.2	54.2
Internlm-xcomposer2d5-7B	51.1	71.0	39.8	32.7	59.3	35.7	35.7	38.8	45.5
Ovis2-8B	35.9	37.1	32.5	27.8	50.9	31.7	28.0	35.3	34.9
VILA1.5-3B	33.4	44.9	32.1	28.6	50.6	35.7	28.4	21.5	34.4
VILA1.5-8B	40.6	44.5	39.1	29.7	48.9	32.9	34.4	36.9	38.4
VILA1.5-13B	36.7	54.6	40.7	34.8	52.8	35.4	34.2	31.3	40.1
Instructblip-vicuna-7B	27.0	27.9	34.5	29.4	44.4	29.7	27.1	25.0	30.6
Instructblip-vicuna-13B	26.8	29.2	27.9	28.0	39.0	24.0	27.1	22.0	28.0
InternVL2.5-38B	67.8	85.4	55.4	41.7	61.7	48.9	52.4	44.0	57.2
InternVL3-38B	68.0	84.0	51.9	43.6	64.4	46.9	54.7	44.6	57.3
Qwen2.5-VL-32B-Instruct	62.3	76.6	51.0	48.3	61.7	44.0	52.2	43.8	55.0
Qwen2.5-VL-72B-Instruct	75.1	82.9	56.7	46.8	59.0	49.4	54.1	48.9	59.1
InternVL3-78B	69.7	80.0	54.5	44.0	65.5	47.4	51.8	44.4	57.2
Proprietary VLMs									
Gemini-2.0-flash	48.9	75.5	44.6	31.9	62.2	48.9	52.4	47.4	51.5
Gemini-2.5-flash-preview-04-17	57.7	82.9	51.4	43.8	65.2	45.7	45.9	43.5	54.5
GPT-4o	69.3	83.1	58.2	48.9	63.2	48.0	55.2	48.3	59.3
Ours									
ShotVL-3B Model Huggingface	77.9	85.6	68.8	59.3	65.7	53.1	57.4	51.7	65.1
ShotVL-7B Model Huggingface	81.2	90.1	78.0	68.5	70.1	64.3	45.7	62.9	70.1





- Half of the evaluated models attain an overall accuracy below 50%. Even the leading models like GPT-4o, fail to reach 60% accuracy.
- Within each series, larger models generally achieve higher accuracy.
- Trained by SFT and GRPO, our ShotVL model series achieves new SOTA performance in cinematography understanding.

ShotVL: SOTA MLLM in Cinematography Understanding

Question :What's the camera angle of this shot?

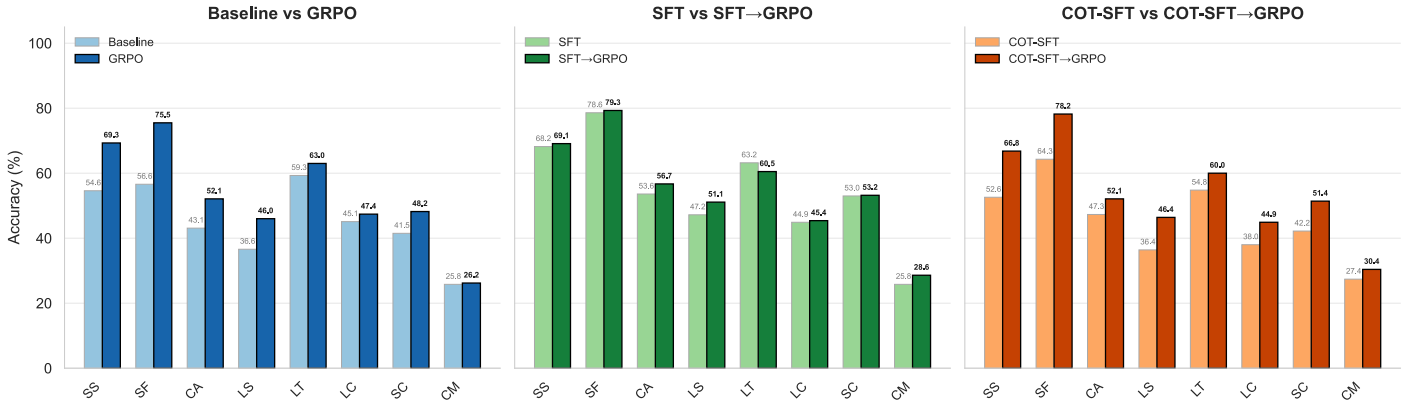
Options:

A. Low angle **B. Overhead** **C. Aerial** **D. High angle**

The most likely answer is:
D. High angle✖
The camera is positioned above the storefronts, looking slightly downward, which characterizes a high-angle shot.

ShotVL

<think>The image shows a street-level view of a snowy city scene, focusing on a storefront named "Twelvetrees Bookseller." This suggests that the camera is positioned at ground level, capturing the storefront and its surroundings without any elevated perspective. Therefore, the correct answer is A.</think><answer>A</answer>



- ShotVL outperforms GPT-4o on cinematography understanding.
- GRPO training after SFT can maximally enhance model performance.

Multimodal Reasoning

Within-Shot *Event Reasoning*

→

Cross-Shot *Cinematic Reasoning*

VChain: models event plausibility

VBench-2.0: evaluates event reasoning

Cut2Next: models coherent shot transitions

ShotBench: evaluates cinematic reasoning

Multimodal Reasoning

Within-Shot *Event Reasoning*

How humans understand causal logic

VChain: models event plausibility

VBench-2.0: evaluates event reasoning

→

Cross-Shot *Cinematic Reasoning*

How filmmakers compose stories

Cut2Next: models coherent shot transitions

ShotBench: evaluates cinematic reasoning

→ Toward **story-level human-like reasoning** in generative models

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Thank You

Ziwei Liu 刘子纬
Nanyang Technological University