Efficient Multimodal Reasoning and Inference Optimization

Video Generation Evaluation: Survey and Empirical Audit

Yiming Liao

Faculty Advisor: Prof. Bo Yuan

WinLab Summer Internship

PROJECT FOCUS

Can benchmark scores explain the quality we actually see?

Background: Efficient generation should preserve quality

Examples for good visual video and bad visual video

VISIBLE SUCCESS

Smooth motion; the person and scene stay coherent.

VISIBLE FAILURE

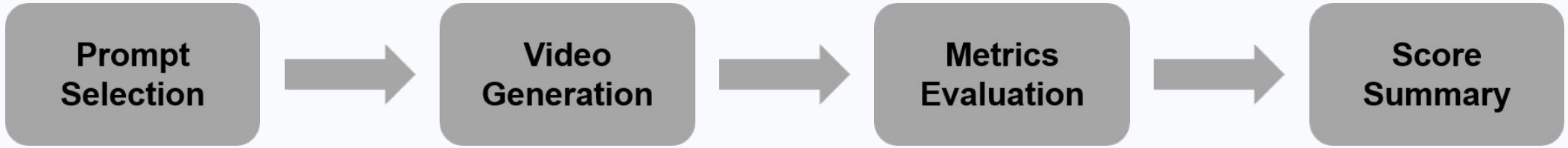
The chair duplicates, bends, and changes shape.

Efficient generation → **Preserve quality** → **Evaluate it reliably**

SURVEY OVERVIEW

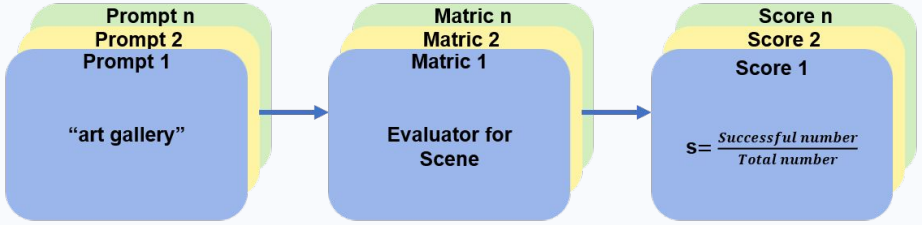
Benchmark survey: how video quality becomes a score

Basic Pipeline

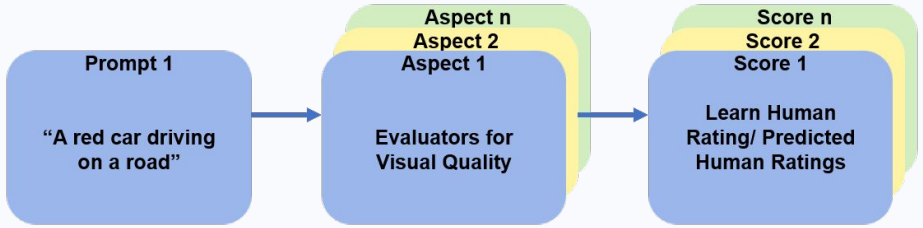


Four benchmark designs studied

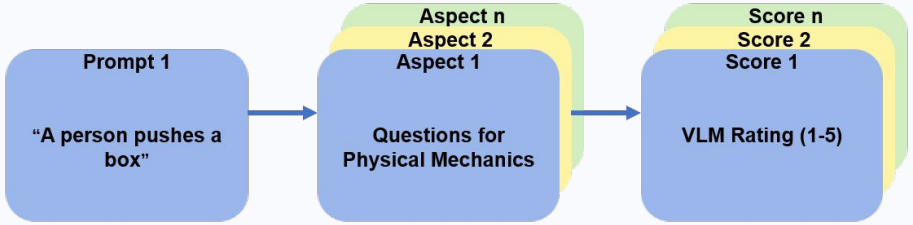
- **VBench series** controlled capability tests



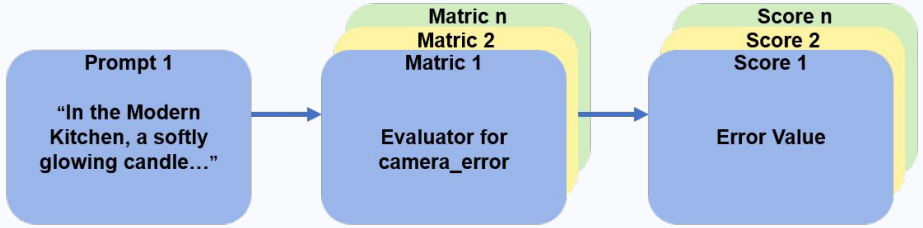
- **EvalCrafter** many metrics + human calibration



- **WorldJen** AI questions + VLM ratings



- **WorldScore** static + dynamic video metrics



Main question: does the score agree with what a person can see?

BENCHMARK LIMITATIONS

Limitations observed in video generation benchmarks

1. **Separate tests can leave coverage gaps**
   Background stability and prompt alignment may be tested on different videos.

2. **Hard thresholds create abrupt decisions**
   Similar motion can receive opposite 0/1 decisions.

3. **Evaluators can miss visible content**
   Tag2Text missed a visible art gallery.

4. **Unrelated failures can affect metric scores**
   Collapse or deformation can distort motion scores.

5. **Averages hide generation reliability**
   The same mean can represent stable outputs or mixed successes and failures.

LIMITATION 3 EXAMPLE · VBENCH SCENE

Example: An evaluator can miss content that viewers clearly see

Metric: SCENE

Prompt: “art gallery”

SCENE SCORE: 0.000

The art gallery is visible, but Tag2Text omitted the required scene label.

SCENE SCORE: 1.000

Tag2Text included the required scene label.

Tag2Text is the captioning model used by this VBench metric.

Method: Tag2Text captions each frame; the scene score checks whether the required scene words appear.

LIMITATION 4 EXAMPLE · WORLDSCORE

Example: Generation collapse can dominate a motion score

Metric: Motion accuracy

Prompt:
“The jellyfish gently pulsates its bell, propelling itself upward through the water”

VC1 · collapsed · score −96.39

VC2 · mostly static · score −5.08

Both videos miss the requested action, but collapse receives a much larger penalty than static output.

Q&A

Questions?

Efficient Multimodal Reasoning and Inference Optimization