Efficient Multimodal Reasoning and Inference Optimization

Yiming Liao | Faculty Advisor: Prof. Bo Yuan | WINLAB Summer Internship 2026

RUTGERS
WINLAB | Wireless Information Network Laboratory

Background and Objective

- **Reliable evaluation is required for efficient generation.**

  Faster generation is useful only when visible video quality is preserved.

  Objective: determine whether benchmark scores agree with what viewers see.

COHERENT DYNAMIC VIDEO

Natural motion; stable body, clothing, and background.

OBVIOUS GENERATION FAILURE

The chair duplicates, bends, and changes shape across frames.

Benchmark Survey

- **VBench series**
  Dimension-specific prompts and specialist evaluators produce capability profiles.
- **EvalCrafter**
  Real-world prompts and 17 proxy metrics produce aspect-level scores.
- **WorldJen**
  Complex prompts and prompt-specific VLM questions produce ratings and rankings.
- **WorldScore**
  Each video receives multiple diagnostic scores rather than one overall score.

Audit Methodology

- **Compare benchmark designs**
  Prompts, evaluators, human roles, and score outputs.
- **Inspect official scoring code**
  16 VBench 1.0 and 18 VBench 2.0 dimensions.
- **Validate scores against videos**
  Same-prompt samples with large score differences.
- **Run a cross-model case study**
  VideoCrafter1 and VideoCrafter2 across all 9 WorldScore metrics.

Displayed cases are representative video-level audits, not the complete metric set.

Empirical Results

Observed failure mechanisms

- **Metric construct mismatch**
  Frame similarity measures how little a video changes; it does not prove prompt correctness.
- **Threshold-induced decision**
  A fixed threshold converts continuous motion estimates into a binary outcome.
- **Evaluator recognition error**
  Tag2Text or a VLM can omit content that is clearly visible to viewers.

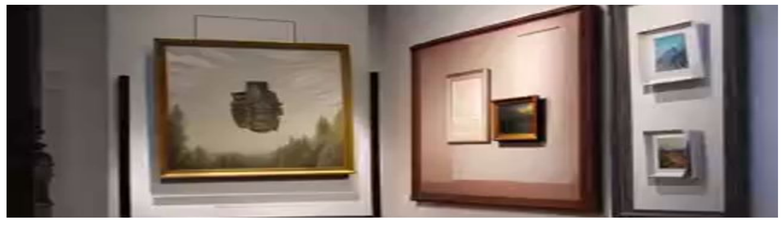
VBench case — Scene evaluator

Prompt: art gallery

scene score: 0.000

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

scene score: 1.000



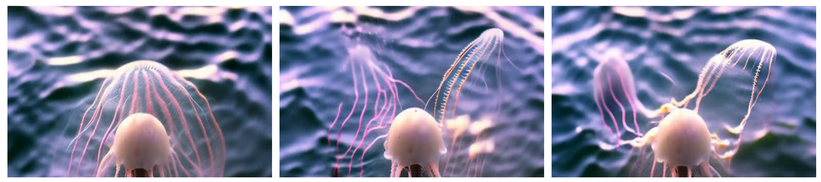
Tag2Text generated the required scene label for this gallery.

- **Scene pipeline**
  Video frame → Tag2Text caption → keyword match → scene score
- **Observed issue**
  The final score can change because the evaluator misses a word—not because the visible scene is absent.

WorldScore case — Action score

Prompt: same requested jellyfish action

motion_accuracy: -96.39



Generation collapse; the requested action fails.

motion_accuracy: -5.08

Coherent but static output; the requested action still fails.

- **Observed issue**
  The score magnitude separates collapse from static output, but both videos fail the requested action.

Limitations of Current Benchmarks

- **Proxy metrics measure only part of video quality**
  Frame similarity or pixel change does not establish prompt correctness.
- **Hard thresholds can distort continuous behavior**
  Small differences around a threshold may produce different final decisions.
- **Scores depend on evaluator reliability**
  Captioning models, classifiers, and VLMs may miss content visible to viewers.
- **Task success and generation quality are not evaluated separately**
  A visually coherent video may receive a higher score even when it still fails the requested action.
- **Aggregate results can hide video-level contradictions**
  Model averages may conceal individual failures and quality trade-offs.

Current benchmarks provide useful but incomplete evidence of generation quality and task success.

References: VBench (CVPR 2024); VBench++ (TPAMI 2025); VBench 2.0 (2025); EvalCrafter; WorldJen; WorldScore; official HunyuanVideo-1.5 and Sora demos. Dynamic examples use three frames; the scene case uses one representative frame per video.