Workshop on Who Verifies the Agents? Toward Reliable Agent Development, NeurIPS 2026

Red-Teaming the Grader:

Mutation-Validated Constraint Checking for Tool-Agent Benchmarks

Junha Jang | Someple Inc.

Incomplete final-state checks can accept wrong outcomes. Fixing the grader moved stored verdicts in both directions.

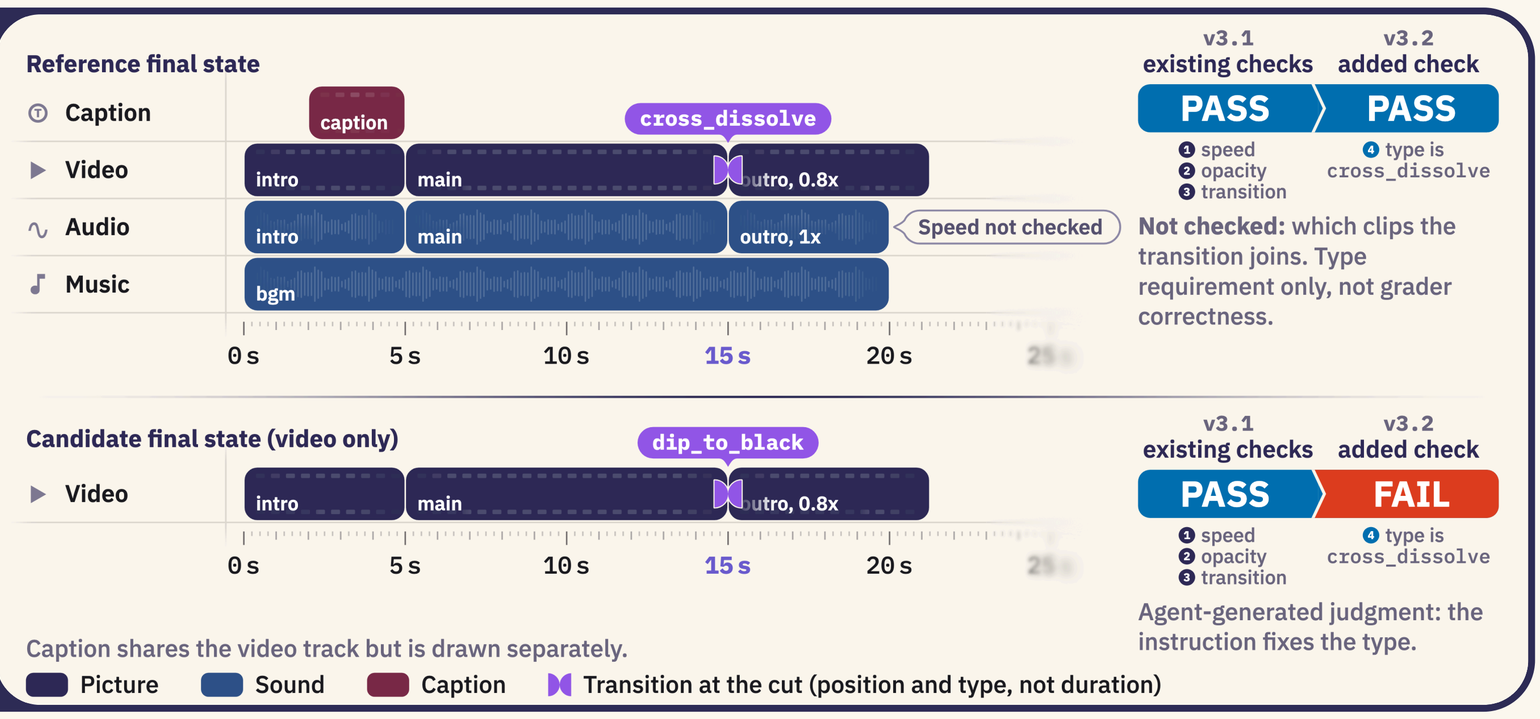
One case: the missing check

INSTRUCTION

The outro needs a slow-mo feel — drop it to 0.8x, lower its opacity to 0.9, and add a `cross-dissolve` coming in from the middle clip.

All calls succeed; **only the transition type differs**.

NLE-Bench is a benchmark the author built and is still developing: 800 scenarios (640 feasible) for tool-using agents in non-linear video editing, scored by deterministic final-state constraints with no model acting as judge. Audited before release.



Mutation probes

1. **Stored reference trajectory**: runs without error and passes; not verified correct.
2. **Probe**: drop a step, alter an argument, or run one deletion.
   ↓ *calls succeed and grader* PASS
3. **Candidate**: a passing mutation, *not proof of a defect*.
4. **Classify** against the instruction:
   agent-generated labels, **no human validation**.
   Added-check rule: candidate FAIL, reference PASS.

527 candidate sites agent-generated cause labels

305 defect — **193** freedom

defect: the instruction fixes the value, but nothing checked it.

freedom: the value is left open.

29 other
25 reference errors
4 probe artifacts
0 unresolved

Sites count unique probe locations, not distinct defects.

Five modification classes

What is run instead of the stored trajectory

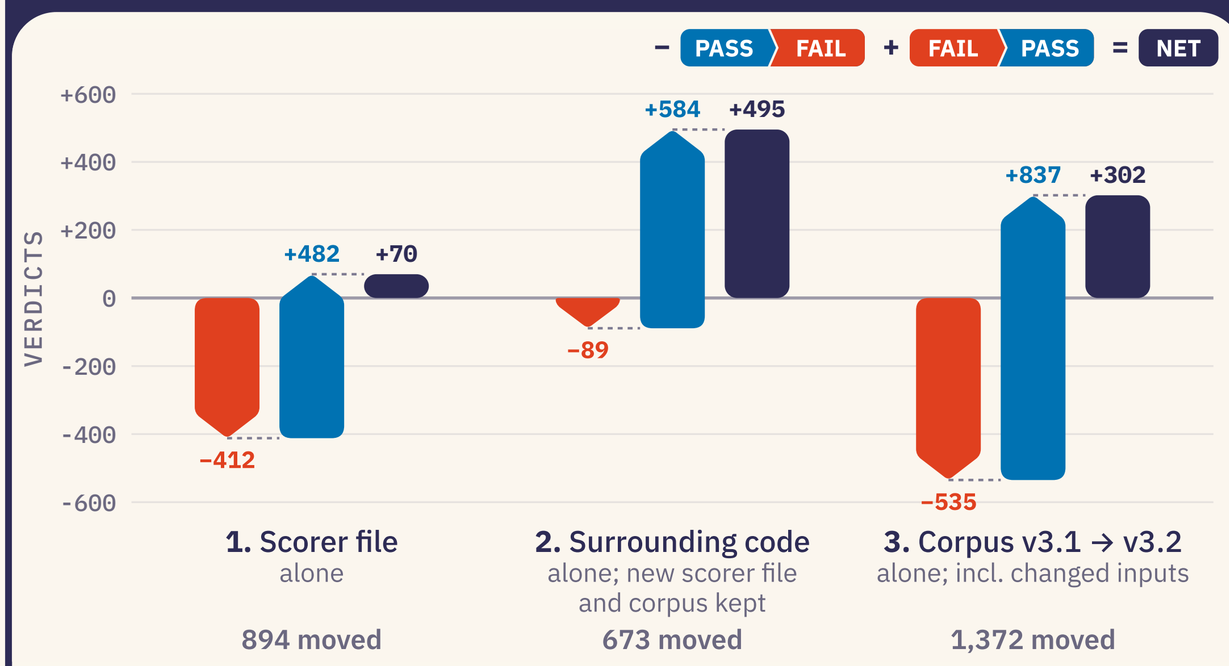
Checks on the final state:
1. `video_clip_3` speed = 0.8
2. `video_clip_3` opacity = 0.9
3. a new transition exists
4. its type is `cross_dissolve` added in v3.2

CLASS	MODIFICATION	STEP 1	STEP 2	V3.1 (1 2 3)	V3.2 (1 2 3 4)
Stored reference		`update_clip` clip_id="video_clip_3" speed=0.8 opacity=0.9	`add_transition` clip_id_1="video_clip_2" clip_id_2="video_clip_3" transition_type="cross_dissolve" duration=…	PASS	PASS
M1	One step left out	update_clip …	left out	FAIL	FAIL
M2	A single call deleting a step's target	remove_clip clip_id="video_clip_3"		FAIL	FAIL
M3	One clip argument pointed at another clip	update_clip clip_id="video_clip_1" … *call fails (clip overlap)*	add_transition … *not reached*	—	—
M4	One number increased by 7	update_clip speed=0.8+7 …	add_transition …	FAIL	FAIL
	M4-full: as M4, each eligible numeric argument of every step, one at a time	update_clip …	add_transition duration=…+7 …	PASS candidate	PASS freedom
M5	One string prefixed or set to another valid kind	update_clip …	add_transition transition_type="dip_to_black" …	PASS candidate	FAIL defect

Re-judging stored final states

One change per step, **in this order**; **a net total hides how many moved.**

894 verdicts changed by the scorer file alone (step 1); net **+70** successes. **No evaluated agent was rerun.**



Among 85,898 re-judged feasible records. For 74 scenarios with changed inputs the verdicts are **counterfactual**: agents were not rerun on them.

Why verdicts move both ways

PASS FAIL **False successes**: direction unenforced · deletion read as change · existence check ignores its filter · preservation blind spot

FAIL PASS **False failures**: non-existence check ignores its filter

Five representative defect classes from the benchmark's history; not frequencies, not a breakdown of the 894.

Three checks for other graders

1. Inspect **accepted outcomes**, not only failures. → case
2. Record **which scorer file and code** produced each verdict. → bar 2
3. Report **fail-to-pass and pass-to-fail** separately. → bars

Limits

- One deterministic grader, one editing domain; probes near stored trajectories, feasible scenarios only.
- Agent-generated judgments; no human validation.
- LLM-assisted scenarios; blind spots shared with agent labels not measured.
- A passing reference shows neither that an added check is required nor that other valid solutions pass.
- What the repositories allow, and do not: paper Sec. 9 (Paper QR).

Paper
OpenReview

LinkedIn
Junha Jang
linkedin.com/in/junha-jang

SOMEPLE
Someple Inc.
someple.com

Audit records: github.com/Junha-Jang/nle-bench-audit
Benchmark snapshot (dev split): github.com/Junha-Jang/nle-bench-artifact

CC BY-NC 4.0, plus permission to reproduce or check this paper's results (exclusions: see each LICENSE.md)

junha.jang@someple.com