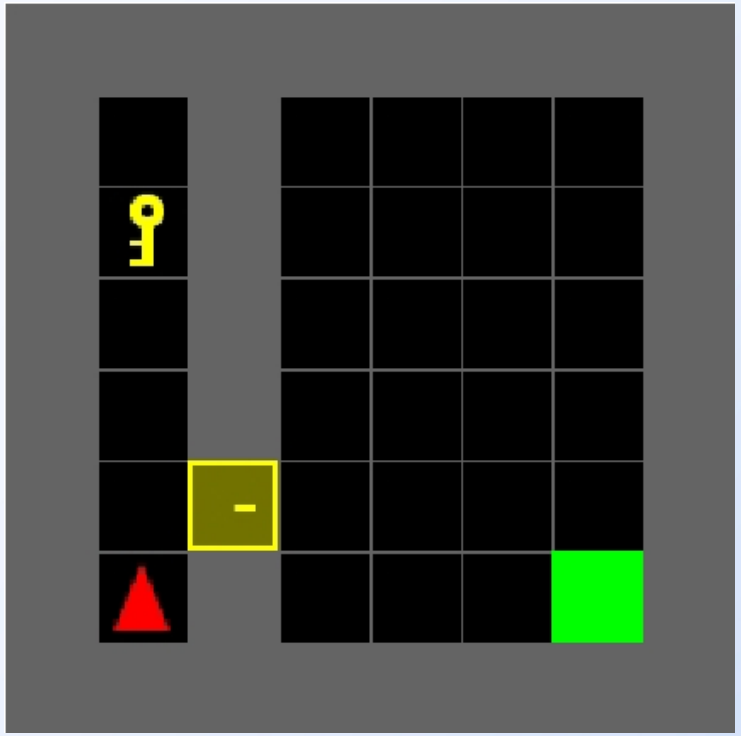
NeurIPS 2026 · Evaluations & Datasets Track

Towards Reliable VLM Judges:

State-Conditional Invariance and Presentation-Aware Diagnostics

Qinan Zhang*, Xinyu Wang*, Qihang Jin*,
Tianyuan Wang, Hongbing Luo, Pengfei Li, Wenjun Lv, Yu Kang, Jian Di†

University of Science and Technology of China



Simulator-grounded diagnostics · MiniGrid DoorKey

* Equal Contribution · † Corresponding Author

Why Do We Need Reliable VLM Judges?

VLMs increasingly decide whether an agent succeeded — but a verdict alone does not show why it was given



1 VLMs are becoming automated evaluators

Scalable assessment of multimodal and embodied agents, where human labels do not scale.

2 Correct predictions may rely on the wrong evidence

Accuracy and consistency cannot fully reveal why a judge reached its verdict.

3 How can we diagnose whether a judge is truly reliable?

Apply controlled changes to evidence presentation and to the task outcome, then audit the response.

Accuracy ≠ Consistency ≠ Reliability

Three different dimensions of judge behavior — not a strict mathematical inequality.

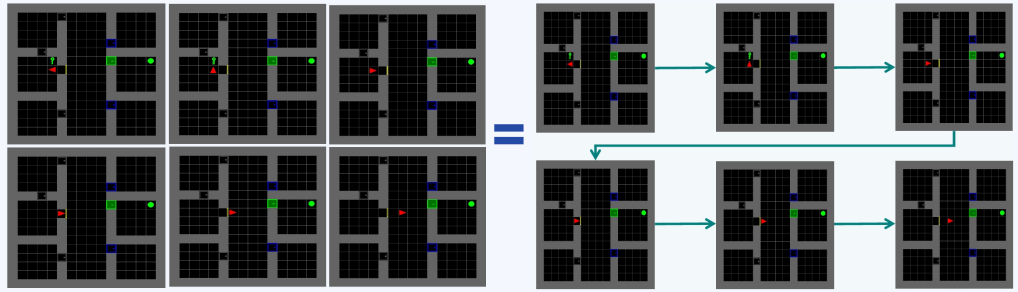
When Should a Judge Change Its Mind?

Two kinds of controlled change call for opposite behavior

A · Representation change

Same trajectory, packaged two ways

Composite storyboard **Split frames**

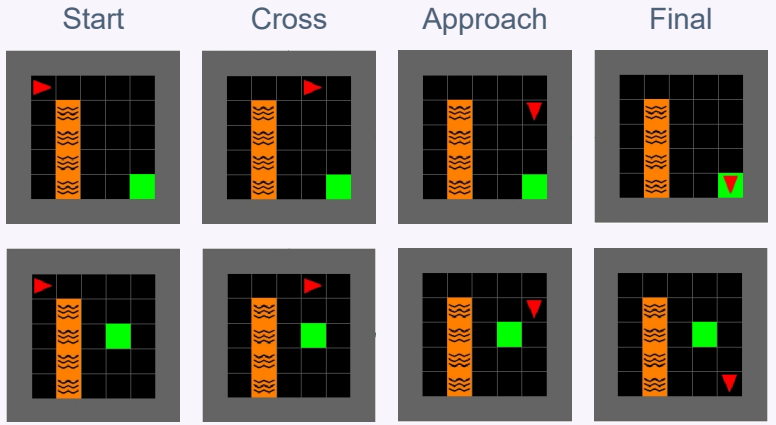


Success Success

Same states, actions and mission → verdict should remain stable

B · Causal change

LavaGap: simulator moves the goal, same actions replayed



Same action sequence + outcome-changing intervention → verdict should change

Invariant to how evidence is presented, but sensitive to what actually happens.

State-Conditional Invariance (SCI)

Conditional on trajectory semantics, nuisance presentation should not determine the verdict

$$\hat{Y} \perp \Phi \mid S$$

S = states, actions, mission ***Φ*** = nuisance presentation (layout, style, order, wording...)

Invariance to nuisance presentation is necessary but not sufficient: correct Full/CF discrimination is also required.

C1

Presentation Invariance

Stable across semantics-preserving changes?

Task C · temporal, visual, framing probes

C2

Action-conditioned Dynamics

Do next-state predictions depend on actions?

Task A · next state, shuffle, mask

C3

Representation Exchangeability

Same verdict across image packaging?

Task C · Composite vs Split

C4

Causal Sensitivity

Accept Full and reject the matched CF?

Task C · Full/CF pairs
Task D · reference-aided

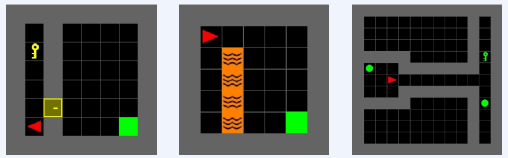
Tasks B and E are local perception and state-tracking controls.

Task C and reference-aided Task D use different input protocols — their Raw IDR is reported separately, never as one trend.

GridWM-Judge: Overall Pipeline

Generate auditable trajectories → probe them under SCI → assess what the judge sees and decides

1 Simulator-grounded generation



MiniGrid environments
↓
deterministic planner
↓
state · action · mission
+ verified outcome

Full

verified success

NoCue

early evidence removed, outcome preserved

CF

same actions + intervention, simulator-verified failure

2 SCI diagnostics

Presentation probes (C1)
Representation probes (C3)
Matched Full/CF pairs (C4)

A	action-conditioned transition
B / E	local perception controls
C	presentation / representation judging
D	matched causal discrimination (ref.-aided)

3 Judge assessment

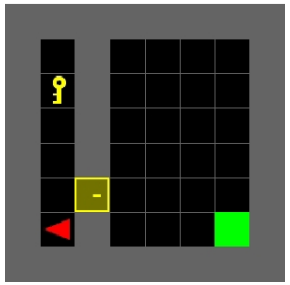
VLM sees model-facing evidence only
↓
predicts transitions, scene attributes, or Success/Failure
↓
diagnostic scoring
↓
correctness · invariance · causal discrimination
↓
failure taxonomy

Audit-only, not shown to the judge: gold labels · simulator state · reward · intervention & validation logs → scoring only

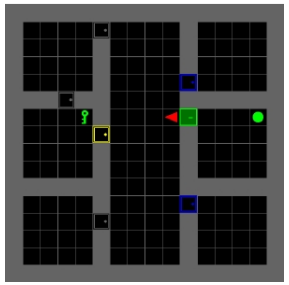
NoCue removes evidence — it is not a semantics-preserving presentation change. CF states may diverge after the intervention.

Constructing Auditable Trajectory Groups

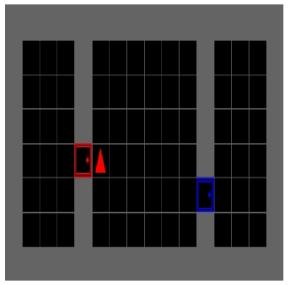
6 MiniGrid environment families · 5 tasks · 1,202 original (Tier-0) model-facing instances



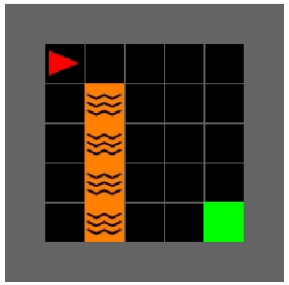
DoorKey



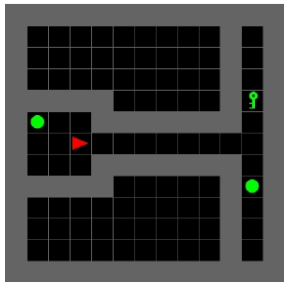
KeyCorridor



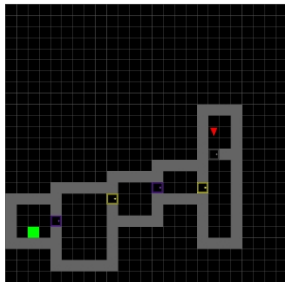
RedBlueDoor



LavaGap



Memory



MultiRoom

Variant	What changes	Simulator outcome	Diagnostic role
Full	Original planner rollout	Verified success	Correctness anchor
NoCue	Selected early evidence removed	Success (preserved)	Evidence dependence
CF	Localized intervention, same action sequence	Simulator-verified failure	Causal discrimination

- Matched Full/CF share the same action sequence; states may diverge after the intervention.
- NoCue removes information: an evidence-removal condition, not a C1 semantics-preserving perturbation.
- 1,202 counts the original benchmark only; Scale-2 expansions are reported separately.

Consistency ≠ Reliability

A stability trap in the original Task-C Composite/Split evaluation

100%

Composite/Split agreement

6%

Full acceptance

6%

Raw IDR

Perfect agreement does not mean reliable judgment.

GPT-5.4 gives the same verdict under both representations because it almost always outputs Fail.

$$IDR_{raw} = P(\text{accept Full AND reject matched CF})$$

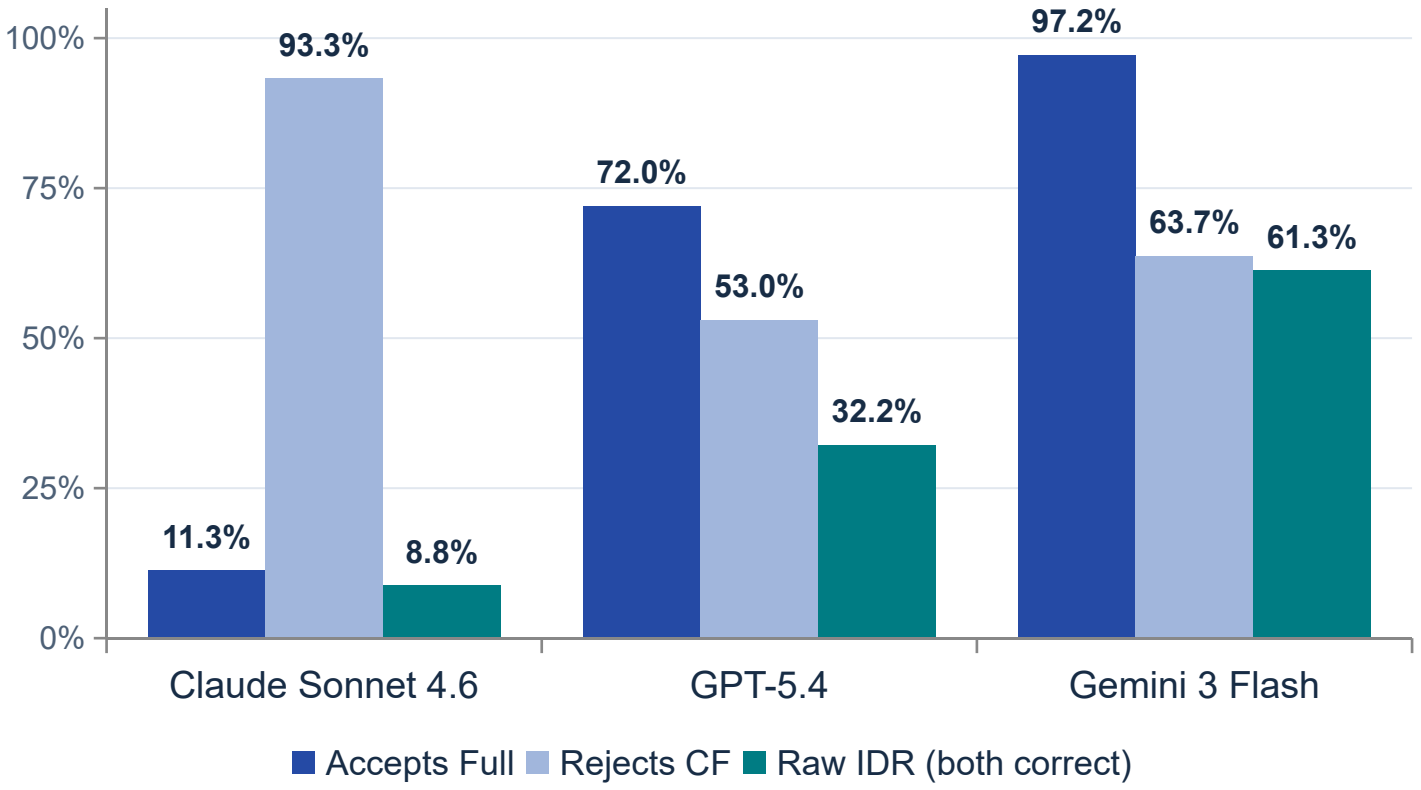
$$IDR_{raw} = p_{active} \times IDR_{cond}$$

a conditional-probability decomposition, not an independence assumption

Original Task-C Composite/Split · GPT-5.4 · values rounded as in the main table. Not comparable with Scale-2 / Task D.

High CF Rejection ≠ Causal Discrimination

Scale-2 · reference-aided Task D · 432 matched groups · 9 environment configurations (not 9 independent families)



Claude Sonnet 4.6

93.3% CF rejection

but only

8.8% correct pairs

It rarely accepts genuine successes (11.3%).

A judge must accept genuine successes AND reject matched counterfactual failures.

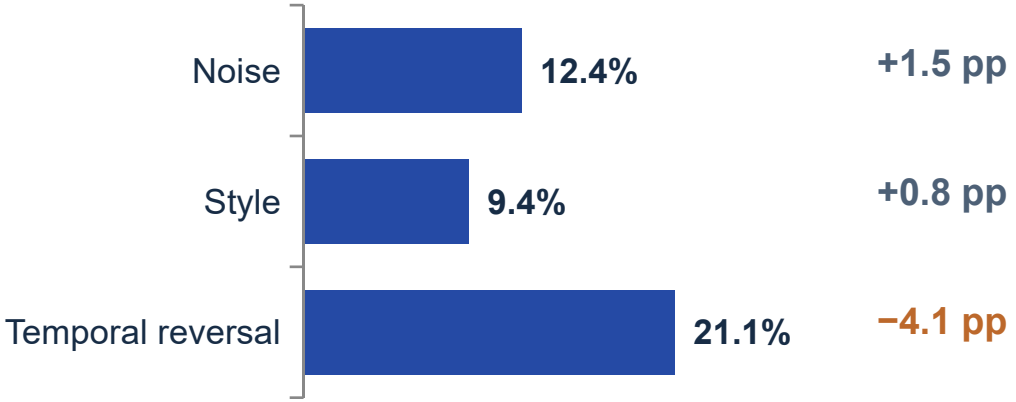
Raw IDR group-bootstrap 95% CI: Claude 6.2–11.6% · GPT-5.4 27.8–36.6% · Gemini 3 Flash 56.7–66.0%. Environment-level intervals are wider — no strong ranking claims.

Presentation and Representation Fragility

Verdicts can shift with evidence presentation and representation.

C1 Presentation Invariance

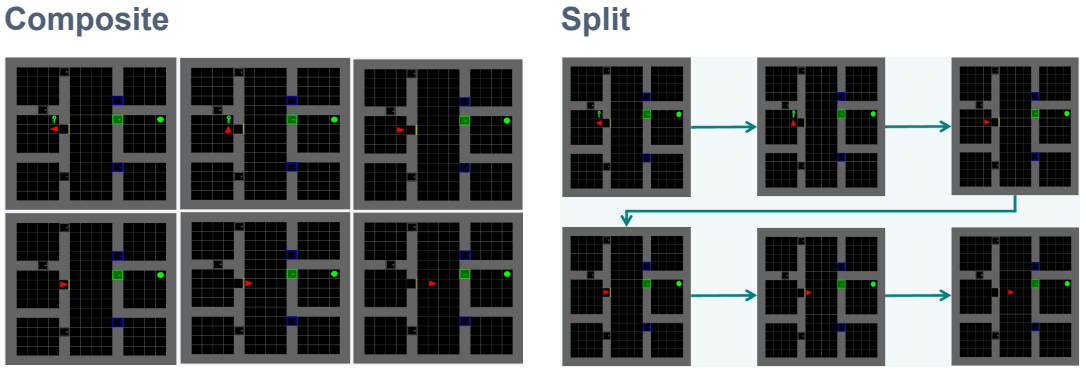
Matched probes: correctness flips per comparison



Small net changes hide large two-way movement: both degradation and rescue occur.

None of 180 probe-level comparisons survives the specified Holm correction — descriptive evidence only.

C3 Representation Exchangeability



Same underlying evidence, different image packaging

GPT-4o raw IDR **22%** Composite → **72%** Split

Across-model contrast p = 0.063: a trend, not significant at 0.05. Prompt wording and run conditions also differed historically, so the gap is not attributable to layout alone.

The evidence interface itself is a diagnostic dimension.

Local Perception and Action Controls

Different bottlenecks cannot be explained by one top-level judge accuracy

Task B **Perception · GPT-5.4 · 144 rows**

61.6% mean scene-component score

vs **11.1%** items reaching strict Acc@0.90

Strict exact-match failure ≠ total perceptual blindness

Task A **Action use · GPT-5.4**

25.2% next-state accuracy

0.0 pp action-shuffle effect

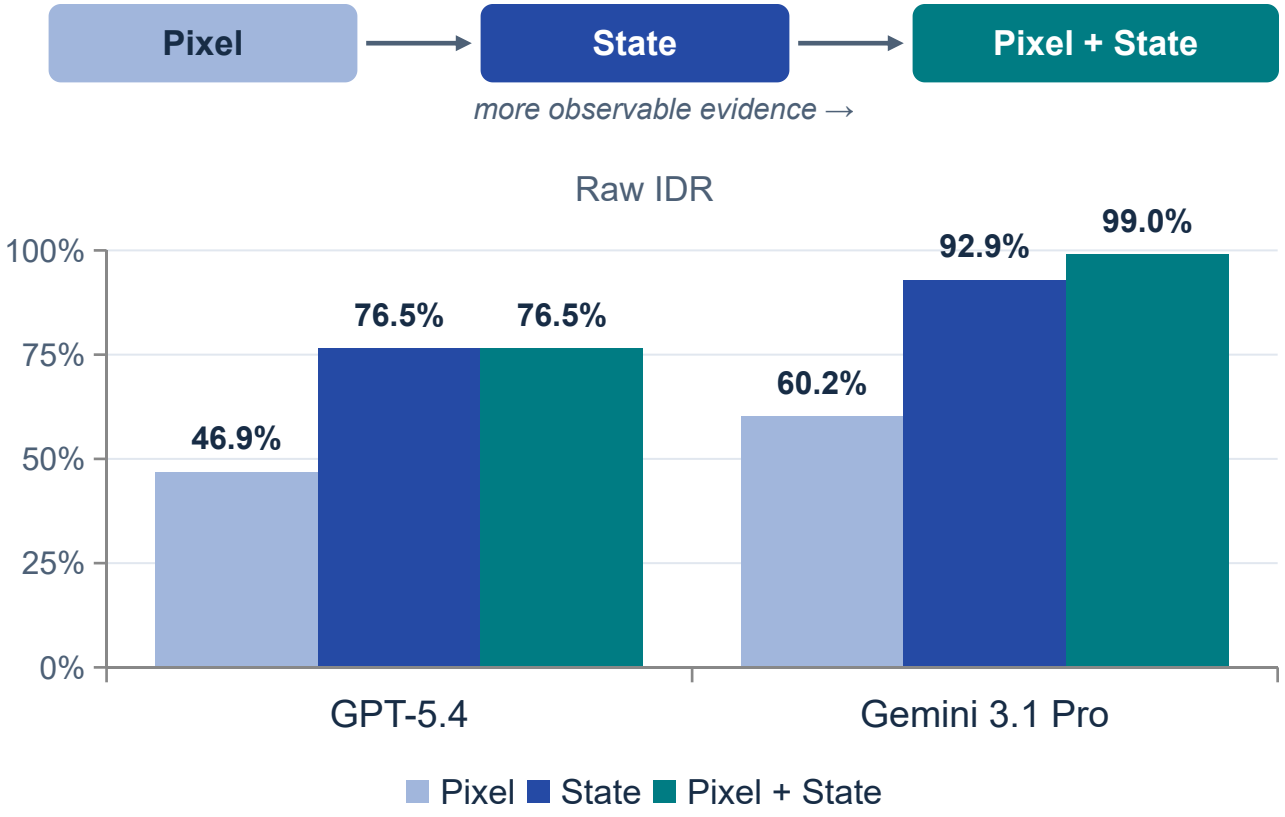
−1.9 pp action-mask effect

Small or mixed ablation effects: limited evidence of action-conditioned reasoning under this protocol (floor effects possible).

Output pipeline also matters — separate Gemini 3 Flash parsing-fix check: valid parsing 20.8% → 99.3%; Acc@0.90 9.0% → 45.1% (additional run, not a replacement).

Oracle-State: Where Failures Arise

Evidence-condition (perception-controlled) experiment · 98 shared groups · 9 environments — not a C1/C3 equivalence test



GPT-5.4 objective recovery

40.5% → **90.5%**

Pixel → State

State facts exclude the exact objective, gold verdict and reward.

State evidence suggests grounding and objective-recovery bottlenecks — it does not uniquely locate the internal failure mechanism.

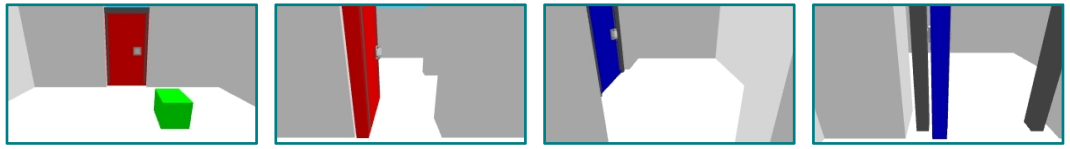
GPT-5.4's group-level interval for the gain excludes zero; its environment-level interval includes zero.

SCI Diagnostics Extend to Simulated 3D

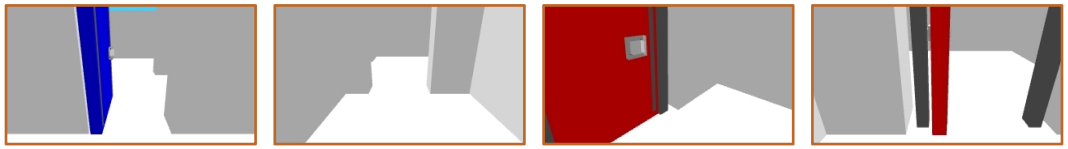
Habitat-Sim/HSSD and MiniWorld use separate protocols — reported separately, never pooled

MiniWorld · first-person RGB, matched pair

Full · goal reached



CF · goal not reached



Required order: red door → blue door. CF opens blue first.

MiniWorld	Trajectory acc.	Task-D-style acc.
Kimi K2.5	82%	54%
Gemini 3 Flash	81%	78%

Habitat-Sim / HSSD · reference-aided Task D

100 matched query groups · 18 synthetic indoor scenes

Judge	Accepts Full	Rejects CF	Raw IDR
GPT-5.4	58%	52%	**16%**
Gemini 3 Flash	97%	21%	**20%**

Gemini accepts almost every success, yet only one in five pairs is fully correct.

MiniWorld Task-D-style accuracy is an aggregate score, not the paired Raw IDR used for Habitat and MiniGrid.

Scope-limited simulated-3D evidence — not a claim about real-robot deployment.

Key Takeaways and Resources

1	**A reliability principle**	SCI separates nuisance invariance from causal sensitivity.
2	**An auditable benchmark**	GridWM-Judge enables simulator-grounded diagnostic testing.
3	**A multidimensional evaluation**	Correctness, stability and causal sensitivity expose failure modes.

Reliable judges should change their minds for the right reasons.

Paper: https://arxiv.org/abs/2611.13596

Code: https://github.com/xinyuw922-pixel/Towards-Reliable-VLM-Judge

Data: https://huggingface.co/datasets/nahco318/Towards-Reliable-VLM-Judges