EXECUTION BEFORE MEMORY: DIAGNOSING MEMORY FAILURES IN VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A memory test for a video world model should say whether the scene was set, whether the asked change was carried out, and whether the last frame shows the result. Current tests fold these into one number, so a low score is easy to read as forgetting. When the change is not carried out, the hidden state transition cannot appear, so the test is not really a test of memory. We introduce MIST, an evaluation that separates the change from the memory it should leave behind. The model first sets the scene while the change stays hidden, and only later is asked to show it, with leakage before that point kept under control. This separation locates the bottleneck: the model sets the scene but does not carry out the instructed change. The same gap appears in general memory evaluations. On that basis we propose M3-Agent (Make the Missing Move), which recursively revises the details of the input instruction so that the model successfully carries out the requested instruction. Relative to the baseline, it more than triples the MIST reveal score. On the benchmarks, WRBench's score T rises by 26.1% and MBench-T's score U by 52.7%. The method meets the problem from the execution side: it shows what the frozen model can do once the move is stated, and it sets a stricter baseline.

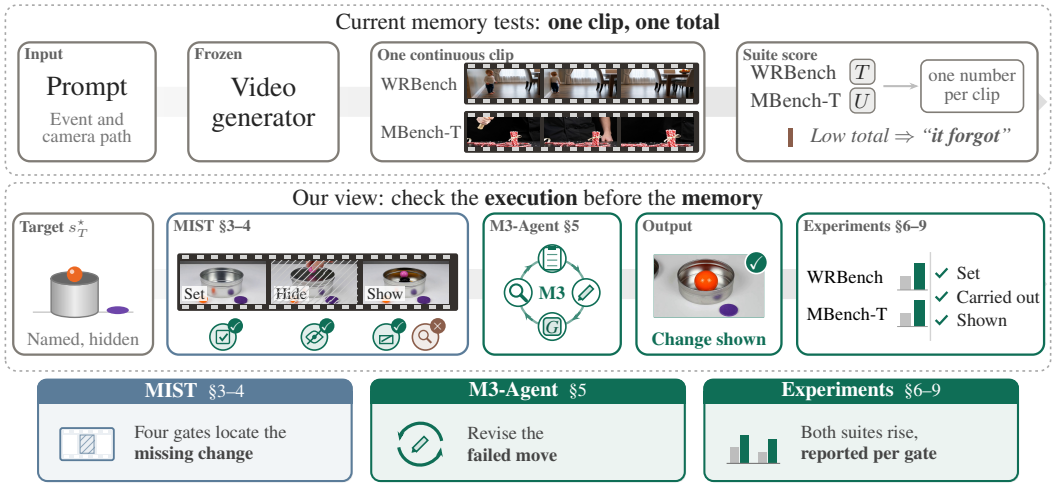

Figure 1: **A low memory score often means the change was never carried out.** Top: current tests turn one clip into one total, which is read as forgetting. Middle: MIST scores each gate, and M3-Agent revises the failed move until the change is shown. Bottom: the finding of each section. Tin frames are Internal MiniMax-H3 `R0-03`; film strips are WRBench and MBench-T clips.

1 INTRODUCTION

Memory tests for video world models ask about a scene the camera no longer shows: who sat down, what went into the box, whether the candy broke (Brooks et al., 2024; Bruce et al., 2024; Valevski et al., 2024). Object relations have to survive when the camera looks away, and actions have to update them (Shin et al., 2026; Ma et al., 2026; Chen et al., 2026a). Current tests fold these checks into one number: a person and a chair, an object left in a room, or a cube still in a cup (Lu et al., 2026; Zhang

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2026). A low number is easy to read as forgetting, yet it cannot say whether the scene was set, whether the asked change was carried out, or whether the last frame shows the result.

When the change is not carried out, the hidden transition cannot appear, so the test is not really a test of memory. Figure 2 shows the usual clip. The child never sits before the camera leaves. The camera comes back to candy that was never broken. A box stays shut. The task names the target, but the last frame never shows it. The score drops before memory is tested. Until that change happens, asking the model to remember targets the wrong stage. Tools that rewrite a video for quality, physics, or preference do not ask whether the asked change was shown (Lee et al., 2024; Xue et al., 2024; Liu et al., 2025a). What is needed first is an evaluation that separates the change from its result.

We introduce MIST to separate the two. The model first sets the scene while the change stays hidden, and only later is asked to show it, with leakage before that point kept under control. On Wan TI2V-5B, three 5-second shots and 43 cases, the scene is usually set and the hidden change does not leak, yet 41 of 43 baseline clips never show it. Asking only to open leaves most lids down, and naming the answer helps less than revising the failed move (§4). Looking at the last frame and revising the instruction makes the change appear, though the shown state can still be wrong. The bottleneck is therefore execution: the scene is set, but the instructed change is not carried out.

The same gap appears in general memory evaluations. Failed clips on WRBench skip the departure or the return, and failed clips on MBench-T return to an object that is still intact (§7). On that basis we propose M3-Agent (Make the Missing Move). It recursively revises the details of the input instruction so that the same model carries out the requested instruction. No memory module is added or trained. Relative to the baseline, MIST reveal rises from 0.147 to 0.492. On the benchmarks, WRBench's score T rises by 26.1% (0.452→0.570, n=40) and MBench-T's score U by 52.7% (0.292→0.446, n=30; 23 of 30 keep the baseline).

These results say where the score is lost. They do not show that the model stored the target state. A judged last frame is not stored memory (§8). They do show that the model can already carry out many missing changes once the instruction names them. The method meets the problem from the execution side: it shows what the frozen model can already do once the move is stated.

This paper makes three contributions (Fig. 1). **(1)** MIST, an evaluation that scores whether the scene is set, whether the hidden change leaks, and whether it is finally shown, so a missing change is not reported as forgetting (§3–4). **(2)** Evidence that failed official clips also miss the asked change: a skipped return on WRBench, an intact object on MBench-T (§7). **(3)** M3-Agent, which recursively revises the input instruction on the same model. It more than triples MIST reveal and raises T by 26.1% and U by 52.7% (§5, §7).

2 Related Work

A low memory score can come from failed persistence, a wrong state update, or a missing operation that never makes the state observable. The three lines of work below each report one total.

Video generators as world models. Large video generators are treated as simulators or interactive environments (Brooks et al., 2024; Bruce et al., 2024; Alonso et al., 2024; Wu et al., 2024b; Valevski et al., 2024; Xiang et al., 2024; Zhu et al., 2024; Huang et al., 2025a). Long-horizon systems add packing, retrieval, or persistent state to reduce drift (Henschel et al., 2024; Zhang et al., 2025; Yu et al., 2025; Hong et al., 2025; Wu et al., 2026; Garcin et al., 2026; Chen et al., 2026c; Wang et al., 2026). Neither line shows that a hidden state was kept and updated.

Hidden state and stage-resolved evaluation. A hidden state cannot be scored if the requested update never happens. The Video Shell Game tests rule-based updates across unseen swaps; StEvo-Bench and MemoBench study evolution during occlusion (Shin et al., 2026; Ma et al., 2026; Chen et al., 2026a). Understanding suites separate endpoint answers from continuous tracking (Yu et al., 2026; Chen et al., 2026b; Han et al., 2026b), while WRBench and MBench add a camera leave and a later check (Lu et al., 2026; Zhang et al., 2026). Temporal-composition and physical evaluations likewise distinguish whether a transition occurs from whether its endpoint is correct (Feng et al., 2024; Sun et al., 2024; Han et al., 2026a; Xie et al., 2026; Bansal et al., 2024; Meng et al., 2024; Li et al., 2025; Bansal et al., 2025). MIST applies that principle to delayed reveal. It does not replace Official WRBench, Official MBench, or the Video Shell Game hidden-swap test.

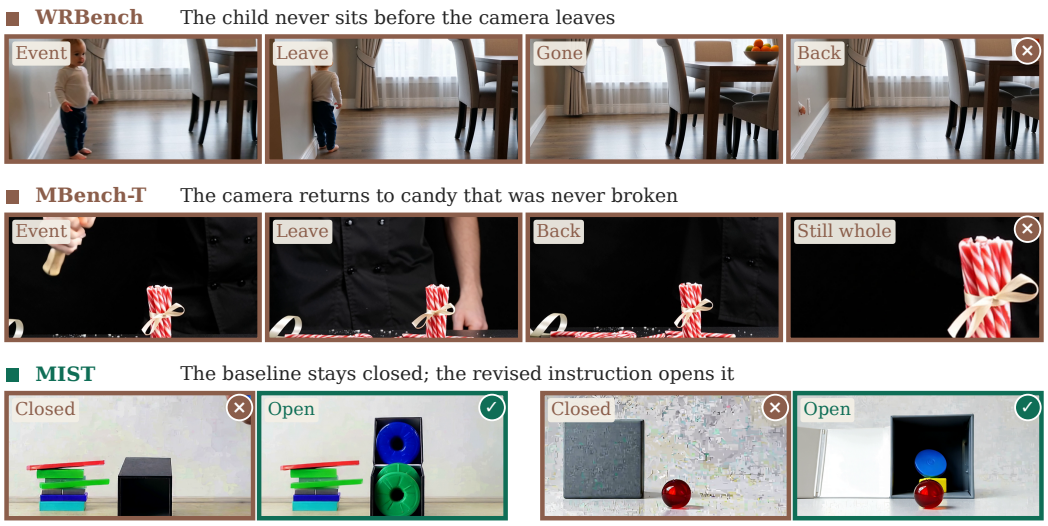

Figure 2: **The last frame does not show the asked-for result.** Three settings, not one run. (Top) WR-Bench: the child walks off before sitting, so the chair is empty when the camera comes back. (Middle) MBench-T: the camera returns to candy that is still whole. (Bottom) MIST: two baseline clips that stay closed, each next to the same case after the instruction is revised, where the box opens onto the hidden objects.

Judges and test-time refinement. Structured VLM judges help on appearance, but remain weakest on temporal and motion errors (Tong et al., 2024; Yang et al., 2025b; Liu et al., 2025c; Li et al., 2024). VideoRepair, PhyT2V, Video-T1, and VQQA already diagnose, rewrite, or search generated videos for quality, physics, or preference (Lee et al., 2024; Xue et al., 2024; Liu et al., 2025a; Song et al., 2026). Those tools target quality, physics, or preference; none asks whether the asked change was shown. M3-Agent instead revises the details of the input instruction so the model carries that change out. We claim no numbered comparison with the tools above. The broader map is in Section L.

3 MIST: SEPARATING SCENE SETUP FROM DELAYED STATE REVEAL

The cases in Fig. 2 show a missing final state, but not why it is missing. A closed lid and a change that was never carried out can produce the same total. MIST separates the two (Fig. 3a). The model first sets the scene while the change stays hidden, and only later is asked to show it.

What the evaluation does. A delayed query asks who sits where, what is inside, or how many objects remain. In text, this is a question about a target state $s_T^\star$. For a video generator, the clip must also produce that state on screen. MIST names $s_T^\star$, hides it behind an opaque gap, and only then asks the generator to show it. Each case is three 5-second shots: the first establishes the scene, the second passes through the opaque gap, and the third requests the reveal. All reveal comparisons reuse the same locked first two shots. The baseline prompt and the gap never contain the answer; the ablations that do name it are marked as such (Section B). R0–R2 cover presence and count, binding, and history-update operations (Table 1; Section B).

Table 1: **MIST varies both the target state and the operation needed to reveal it.**

	Target state	**Required visible operation**
R0	Presence, absence, count, final state	Open, reveal, or hold the right objects
R1	Object–container or left–right binding	Place the object in the named slot or box
R2	History update: swap, remove, transform	Execute and reveal the *updated* state

Three observable gates. Let $x_{1:T} = G(\pi, c)$ denote a clip from a frozen generator G (Team Wan, 2025), prompt π, and image or caption context c, and let J be a fixed VLM judge (Liu et al., 2023a; Tong et al., 2024). We distinguish whether the required move occurs (I_{act}), whether the last frame

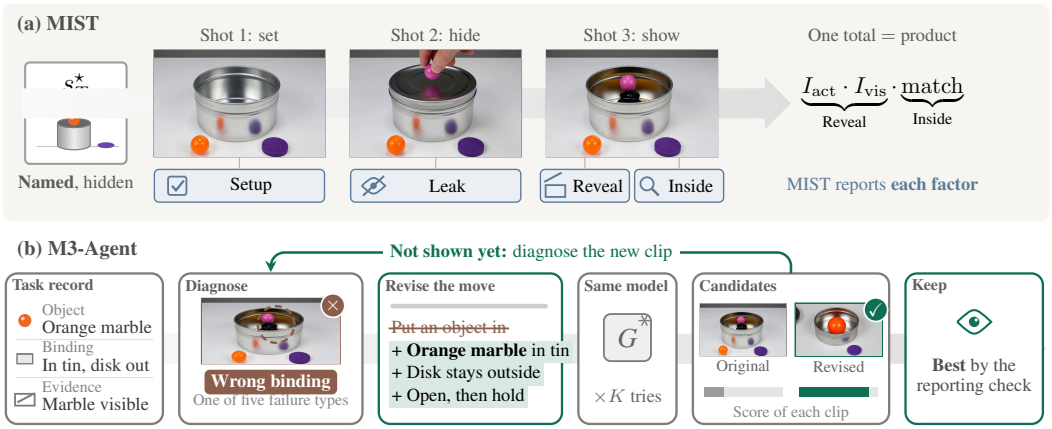

Figure 3: **MIST scores each gate of the reveal; M3-Agent revises the failed move until the change is shown.** Example: Internal MiniMax-H3 packed take of `R0-03`, not the Wan stitch and not Official T/U. (a) Setup and leak check the first two shots; reveal and inside split the last shot into the factors of Eq. (1). The raw take opens the tin onto another object. (b) The workflow on the same case: diagnose the last frame, revise the move rather than the answer, and generate again with the same model. The rewrite puts in the orange marble and opens to it; the original stays a candidate.

is readable (I_{vis}), and whether that frame matches the target ($\text{match}(x, s_T^\star) = J(x_T, s_T^\star)$). A score that requires all three is

$$\text{score}(x, s_T^\star) \;=\; I_{\text{act}}(x) \cdot I_{\text{vis}}(x) \cdot \text{match}(x, s_T^\star). \tag{1}$$

Eq. (1) says that a total collapses when any factor is missing, and the total alone cannot tell a missing action from an incorrect state. We therefore report the factors separately rather than their product (Fig. 3a): setup checks the first shot, leak the gap, and reveal and inside the last shot. Never-open is the thresholded action class, not the mean reveal: a low mean is not a never-open count (Section B). A valid setup followed by a closed lid places the failure on the reveal path, not on forgetting.

WRBench and MBench-T test the same gates in one continuous clip. Each publishes a *unique* score, T and U, that gives no credit when a step of this path is missing:

$$a \;\rightarrow\; \text{leave until unjudgeable} \;\rightarrow\; \text{return} \;\rightarrow\; \text{hold}(x_T), \tag{2}$$

where a is the requested event, such as a sit or a break. The score thus requires the executed move, not only a final match (§7).

4 Pre-experiments: the target is named, but the reveal is missing

We apply the MIST measures before designing an agent or evaluating an official suite. Two questions come first. At which gate does the baseline fail, and can naming the target fix it?

Does the scene fail before memory is tested? Setup is high (Table 2): in most first shots the objects and the container are in place. In 38 of 43 cases setup is at least 0.8 while the baseline reveal stays below 0.3. What is missing is the later move that shows the asked-for state.

Where does the baseline fail? Mean reveal stays low. When each clip receives one label, 41 of 43 baseline clips never open; one opens to an empty or blocked view, and one is correct. In the closed cases the final state cannot be evaluated at all, because the required action never exposes it. On those 41 clips, a hand comes in and does not open in 38, nobody comes in in 3, and the wrong object is opened in none. The shot length is not why the lid stays down (Section B), but the three-shot stitch shapes how the failure looks: on a packed 15s MiniMax-H3 take the same judge finds 3 of 43 never-open and 24 of 43 opened onto the wrong state (Section M). The failure moves from not opening to opening wrong; in both cases the asked change is not carried out. A low total here says nothing about whether the model forgot $s_T^\star$: persistence has not been tested yet.

Did the hidden gap leak the answer? No target is visible in the middle chunk: leak is clean for both the paired baseline and the later rewrite (Section B). So the reveal failure does not follow an early exposure of the answer; it occurs at the action gate in Eq. (1).

What changes after a targeted rewrite? We inspect the last shot, diagnose its failed move, and rewrite only that shot. With one kept clip per case, rather than best-of-five, reveal rises and inside match rises with it (Table 2). Setup remains flat. Per case, reveal rises in 35 of 43, stays the same in 6, and falls in 2; the mean paired change is 0.34 (95% CI 0.24 to 0.45). The failure class also changes. Of the baseline never-opens, rewrite leaves many closed, opens some onto a wrong state, and makes some correct (Fig. 4a). The table's open-but-wrong clips are those newly opened wrong states plus the initially correct case `R2-10`, which the rewrite can break. The rewrite exposes more states, but it does not guarantee that the exposed state is correct.

Table 2: **The rewrite changes the reveal gate, not the setup.** One kept clip per case, not best-of-five. Decimal rows are mean judge scores; count rows give each clip one ordered label (closed $\succ$ empty $\succ$ wrong $\succ$ correct). Inside match does not require a reveal; Correct does.

	Base	Rewrite	Δ
Setup	0.852	0.851	−0.001
Reveal	0.147	**0.492**	+0.344
Inside match	0.047	0.442	+0.395
Never opens	41	20	−21
Empty or blocked	1	0	−1
Opens, wrong inside	0	10	+10
Correct	1	13	+12

All three operation families move the same way, and setup stays flat in each: reveal goes from 0.132 to 0.549 on presence and count, from 0.147 to 0.464 on binding, and from 0.163 to 0.454 on a history update (Fig. 4c; Section B). Closed lids that remain are not piled in one family: 6 on presence and count, 6 on binding, and 8 on a history update. Openings onto the wrong state are failed updates, not noise in one family (Section B). The first pass revises 33 of 43 prompts. It most often names a wrong final state (32), a wrong count or binding (31), or an opening that cannot be seen (19); one prompt can carry several labels. The rewrite has to spell out both the move and the last-frame evidence, not only a verb.

> ***Finding 1.*** A hidden state cannot be judged until the model makes it visible; a visible opening can still expose the wrong state.

Is naming the target enough? The kept-clip diagnosis above locates the failure at the reveal gate; it does not yet explain what kind of instruction opens that gate. If the missing piece were only the answer, inserting $s_T^\star$ should suffice. If it were only a verb, "open and hold" should suffice. We allocate the same five tries in five ways (Table 3, Fig. 4): the answer sentence is present or absent, and the failed move is either rewritten or left unchanged. A fifth arm rewrites the move from the same last frame but forbids every target noun, color, and count (audited; 0/43 leaks).

The ordering in Table 3 separates these explanations. The bare "open and hold" instruction performs worst. Supplying the answer helps, but resampling the same sentence does not match a diagnosis-guided rewrite. Looking at the clip and changing the failed move gives the strongest best-of-five reveal and inside scores (0.763 against 0.684 for answer-only). With the target withheld the same rewrite reaches 0.174 and leaves 28 of 43 lids closed: on this probe the rewrite opens the lid only when the prompt also says what is inside. Each arm keeps, per case, the try whose last frame the judge scores highest, so no cell here is the one-clip mean of Table 2. Later tries are not better on average, so the best-of-five cell is a running best (Fig. 4d). Where the same cases have an unchanged-prompt control (Section B), extra draws of the same sentence do not explain the gain. The running best moves most between the second and third try, from 0.546 to 0.700, and the last two tries add little (Section B). On six cases with a matched resample, the same sentence ends at 0.580 and the rewritten move reaches 0.884. That subset is too small to replace the 43-case table.

> ***Finding 2.*** Rewriting the failed move beats naming the target alone, but only when the target is also named: with it withheld, the same rewrite opens almost nothing.

The same missing operation appears on Internal MiniMax-H3 (`R0-03`, Fig. 3): the raw take puts a different object in the tin and opens to it; the rewritten instruction puts the orange marble and opens to that. This is one qualitative packed take: not the n=43 stitch, not the equal-budget comparison in Table 3, and not Official T/U.

Asking only to open fails, naming the target helps, and revising the failed move helps most, but only with the target named. The next section turns look, diagnose, and change into a test-time workflow.

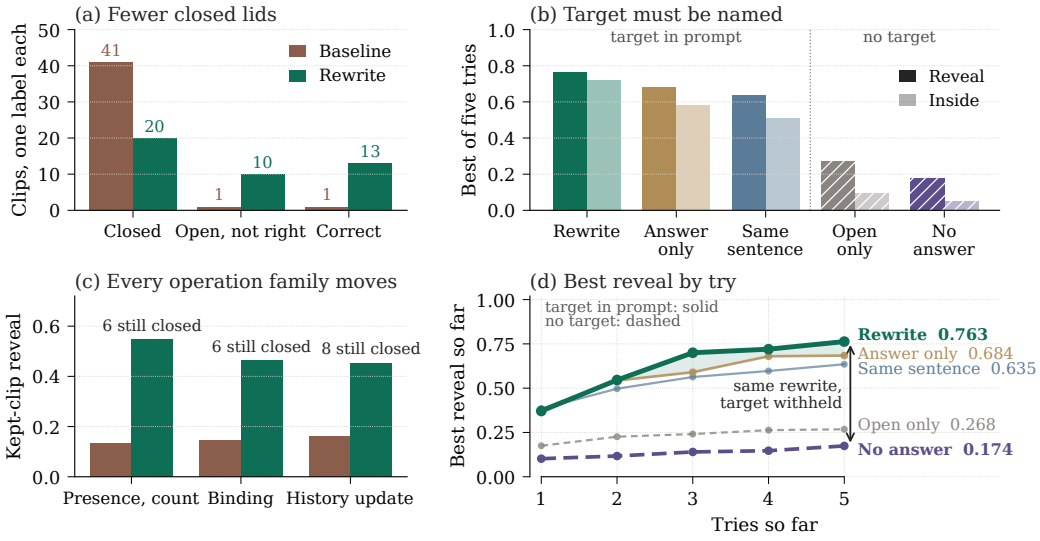

Figure 4: **Rewriting the failed move opens more lids, but only when the target is named.** (a) Failure classes for one kept clip per case (Table 2). (b) Best-of-five reveal (dark) and inside (light) for the five arms of Table 3; hatched arms have no target in the prompt. (c) Kept-clip reveal by operation family, with the lids that stay closed after the rewrite. (d) Best reveal so far as the five-try budget grows; dashed lines have no target; the arrow marks the same rewrite without the target.

5 M3-AGENT: REVISE THE INSTRUCTION FOR THE MISSING CHANGE

The pre-experiments show that the workflow must revise the failed instruction, not append the target or repeat the same prompt. M3-Agent (Make the Missing Move) does that at test time (Fig. 3b): it reads the last frame, revises the details of the instruction for that change, and generates again with the same model. It does not add or train a memory module.

The workflow has three steps. First, a short task record lists the actor or object, its required binding, the operation, and the evidence that must appear in x_T; these fields come from the task card. Second, a VLM inspects the generated last frame against that record and assigns a visible failure mode, such as no action, wrong binding, missing departure, return to an intact object, or an occluded view. Third, the workflow rewrites the failed operation rather than the answer, and the frozen generator produces another candidate. This third step is the one that works in Table 3. The rewrite preserves the actor, object, target state, and temporal order; only the failed operation and its final-frame evidence clause may change. Diagnosis reads the rendered x_T once, never the generator's hidden states.

A worked case. Figure 3b traces the workflow on `R0-03`. The task record asks for the orange marble in the tin, the purple disk outside, and a visible marble at the end. The raw take opens the tin onto another object, which the record labels a wrong binding rather than a missing action. The revised instruction keeps the tin, the disk, and the order of events, and names the object that must go

Table 3: **With five tries, the rewrite helps only when the answer is also in the prompt.** Same judge; each arm keeps, per case, the try whose last frame the judge scores highest. Closed and Ok count kept clips that never open and that are fully correct. Look-then-rewrite exceeds answer-only, but the same rewrite with the target withheld falls below "open and hold."

What we do	Answer in prompt?	Reveal	Inside	Closed	Ok
Rewrite: look, then change the move	yes	**0.763**	0.721	5	29
Answer only: write the target, same move	yes	0.684	0.581	3	24
Same sentence: new seed only	yes	0.635	0.512	4	21
Open only: "open and hold"	no	0.268	0.093	6	3
Rewrite, no answer: look, change the move	no	0.174	0.047	28	2

in; the next take opens onto the orange marble. The raw take stays in the candidate set, so a revision that makes things worse is not kept.

Given K candidates, M3-Agent keeps the clip that best carries out the instruction,

$$\hat{x} \;=\; \arg\max_{k \le K} \text{score}\Big(x^{(k)}, s_T^{\star}\Big)\,, \tag{3}$$

and the original candidate remains in the set. The score here is the same check that reports the result: the MIST judge on MIST, and the official gates on WRBench and MBench-T. A reader that sees only the last frame, without that score, does not reliably beat the baseline (§8). A single revision is not always better: `R2-10`, for example, turns a correct case into an incorrect one, so the original clip stays eligible. Searching several instructions at test time is a known strategy (Snell et al., 2024; Oshima et al., 2025; He et al., 2025; Khalaf et al., 2025; Landesberg, 2026).

On MIST. The agent works with frozen TI2V-5B. It preserves the first two 5-second shots and rewrites only the final reveal shot, and the reveal prompt may include $s_T^{\star}$ (Table 3). The kept-clip analysis in Table 2 locks one outcome per case; Table 3 instead generates $K{=}5$ candidates per arm and applies Eq. (3).

On WRBench and MBench-T. The agent works on 10-second clips from frozen T2V-A14B and keeps each suite's official camera clause; on WRBench the caption stays `task-context none`. Candidate instructions establish the requested relation before departure and hold it after return. On MBench-T the candidate set must include the original clip: forcing another leave–return can preserve the camera motion while worsening the returned state. Selection is made per sample over a small instruction set. The task-record schema, diagnosis labels, candidate rules, and the selection record are in Section F.

6 Evaluation setup

We never merge numbers from different scorers. Every number is **Official** (the suite's own scorer), **Adapted** (a changed protocol, reported only in Section A), or **Internal** (our fixed judge). Table 4 lists the three settings; generation settings and budgets are in Sections A and B. The H3 row is Internal and qualitative (Fig. 3; Section K).

Table 4: **Three settings, three scorers.** T averages four official WRBench checks, $D3$, $D4$, $D5'$, and $D6'$, where the last two require the departure and the return; U averages MBench-T state and correctness. A missing return or an untriggered clip scores zero.

Setting	Generator	Clip	n	Scorer	Reported
MIST	Wan TI2V-5B	three 5s shots	43	Internal judge	setup, leak, reveal, inside
H3 pilot	MiniMax-H3	packed 15s	43	Internal judge	failure classes, examples
WRBench	Wan T2V-A14B	10s, `yaw_LR`	40	Official	$T = (D3{+}D4{+}D5'{+}D6')/4$
MBench-T	Wan T2V-A14B	10s, five captions	30	Official	$U = (\text{state}{+}\text{corr})/2$

7 Main results: the same diagnosis on official scores

Under the scorers fixed in §6, we ask whether official memory protocols fail at the same execution gates isolated by MIST. Inspecting failed clips on official WRBench and MBench-T shows the same missing action. The analysis transfers a diagnosis, not a number: Internal MIST, Official unique scores, and packed H3 pilots remain separate cells (Section M).

What do the official scores contain? Both suites already require the path in Eq. (2) (Lu et al., 2026; Zhang et al., 2026), so we report which gates change and which clip is kept per item, rather than correlating a total with its own parts (Section G).

What happens when we search that path? The two official protocols fail at different points, and those points map onto the two remaining MIST failure types. On WRBench, the camera often does not complete the required departure and return: a shut $D6$ gate is the same class as a never-open, so persistence is not yet measured. On MBench-T the camera returns, but the object is often still intact (Fig. 2): the path completed, yet the aftermath is the wrong world state, as in the open-but-wrong

Clip	U	State	Corr.	Trig.
Baseline (captions only)	0.292	0.350	0.233	0.93 / 0.83
Force leave–return	0.251	0.242	0.260	0.93 / 0.87
Force last-frame lock	0.228	0.208	0.247	0.97 / 0.97
Keep the better clip	**0.446**	0.492	0.400	1.00 / 1.00

Table 5: **Official MBench-T: selection can refuse a harmful rewrite.** Keeping the better clip for each check raises U by 52.7%; one kept clip per item gives 0.443. Trig. is how often the state and correctness checks fire. State here is not MIST reveal.

Clip	Vis.	D5	D6	T
Baseline (camera only)	0.648	16/40	18/40	0.452
One leave–return	0.690	23/40	14/40	0.496
Best-looking clip	**0.755**	20/40	13/40	0.514
Search, first set	0.681	29/40	23/40	0.550
+ second set	0.662	28/40	27/40	0.547
+ last step (kept)	0.654	**31/40**	**31/40**	**0.570**

Table 6: **Official WRBench: the search opens more return gates than visual quality.** The kept clip raises T by 26.1%; T zeros a missing return. Details in Section C.

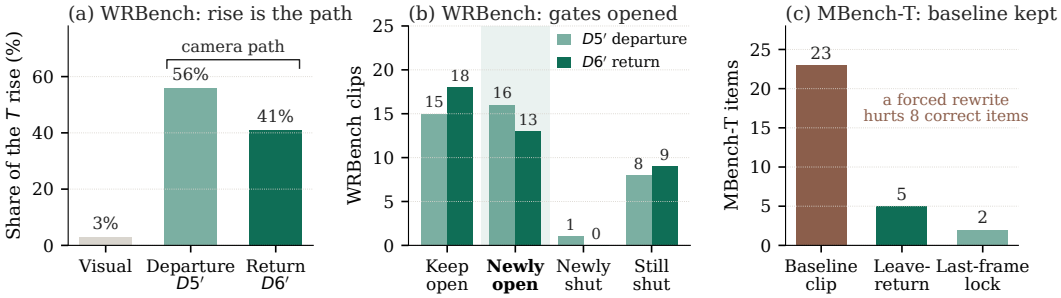

Figure 5: **The same diagnosis has different consequences on the two official suites.** (a) Share of the WRBench T rise from visual quality, the departure check $D5'$, and the return check $D6'$. (b) WRBench departure and return gates that stay open, newly open, newly shut, or stay shut between the baseline and the kept clip. (c) MBench-T items whose kept clip is the baseline, the leave–return instruction, or the last-frame lock. Panels keep separate denominators (Sections C and G).

MIST clips. So the search has to find a missing path in one suite, and keep the original clip when a fixed rewrite harms the other.

On WRBench, score T rises (Table 6) because the camera finishes the return, not because the frame looks better. The paired gain has a 95% bootstrap interval of $+0.070$ to $+0.166$ (Section F). About 3% of the rise is visual quality. A little over half of the rise is the departure, and about two fifths is the return (Fig. 5a). Thirteen returns newly open, none close, and nine stay shut (Fig. 5b). A second instruction set can lower the total; a later step restores it (Table 6; Section C).

On MBench-T the camera already comes back, so U follows the state after return, not whether the camera left. Fourteen of 30 cases stay hard: the object comes back unchanged or wrong, and a prompt that only says to hold the update fixes none of them. A forced rewrite also hurts 8 cases that were already right, so most items keep the original clip (Fig. 5c; Section G).

What do the ablations constrain? Table 3 rules out answer text or unchanged resampling as the sole explanation for the MIST rewrite. The official ablations add two boundaries. On MBench-T, a prompt that only asks to hold the update does not fix the hard cases, and a last-frame lock alone is worse than the baseline (Table 5). On WRBench, the best-looking clip and the clip kept for T are different clips (Table 6), so a readable frame is not the executed path (Section C).

Can the judge answer without the video? A text-only control separates a text shortcut from visual evidence (Table 7). A held-out 8B reader names gold $s_T^\star$ from text, then collapses when the gold sentence is removed or shuffled. On the locked last frame it tracks the 122B inside check: almost no baseline clip is accepted, and the rewrite rate matches Table 2. Naming the target is easy; seeing the executed state is not. On these last frames the outside reader agrees with the 122B inside check on 0.93 of the baselines

Table 7: **A judge can name the target from text before the video shows it.**

What the judge sees	122B inside	8B
Gold sentence, no frame	–	0.91
Setup caption, no frame	–	0.40
Shuffled gold, no frame	–	0.07
Last frame, baseline	0.023	0.09
Last frame, rewrite	0.442	0.56

and 0.88 of the rewritten clips. It does not track the official gates: agreement is 0.50 on one WRBench return check, 0.35 on the other, and 0.43 on MBench-T state (Section D). The reread does not replace either official scorer (Han et al., 2025; Liu et al., 2025c; Chen et al., 2024; Jia et al., 2026).

Finding 3. On both official suites the score is lost before persistence is tested: WRBench clips skip the departure or return, and MBench-T clips return to an object that is still intact.

Finding 4. Text can identify the target without establishing that the video shows it; the last-frame check improves only after the failed move is rewritten.

8 Discussion: what execution evidence can establish

MIST separates the scene, the change, and the last frame (Eqs. (1)–(2)); M3-Agent revises the instruction. Neither shows what the model stored.

Measurement. A high reveal is not a memory claim: kept-clip reveal exceeds independent inside match (Table 2), and if the action never occurs, persistence is untested, whether the clip stays closed or opens onto the wrong state (§4).

Selection and cost. M3-Agent revises the instruction and keeps the clip that carries out the change, scored by the same check that reports the result. It generates five tries per arm on MIST and searches at most five clips per case on WRBench and three on MBench-T, including the original. The MIST rewrite names the target, and Table 3 shows that the same rewrite fails when that name is withheld. A last-frame reader that does not see the score gives $T = 0.466$ and $U = 0.248$; the first is within noise of the baseline and the second is below it (Section F). A rewrite can damage a correct case, so the original clip stays eligible.

Scope. The MIST study uses one generator family and a three-shot stitch; packed-15s MiniMax-H3 cells are Internal (Fig. 3; Sections K and M), not official totals. The held-out 8B check is a class-imbalanced reread, not human validation (Section D).

What a positive result would need. We cannot claim the backbone stored $s_T^\star$. That claim would require a target-blind selector, a held-out case split, and human labels on the exposed state. Our blind selector did not beat the baseline, and we provide neither of the other two. Until then, an opened lid shows that the state can be judged, not that it was remembered.

What a memory benchmark should report. The split does not need MIST's three-shot stitch. Per clip, a benchmark can report whether the scene was set, whether the requested event and camera path were carried out, and whether the last frame shows the target state. It can score the state only on clips that pass the execution check, say whether any selector saw the scorer or the target, and add a target-withheld control like the last row of Table 3. WRBench and MBench-T already compute most of these checks inside T and U (§7), so publishing them apart needs no new generation.

9 Conclusion

MIST splits one memory score into three questions: was the scene set, was the asked change carried out, and does the last frame show the result. On the internal three-shot test the scene stays set and the gap does not leak, yet 41 of 43 baseline clips never show the change, so persistence is not yet tested. Revising the failed instruction opens more lids than naming the target alone, though the shown state can still be wrong.

M3-Agent recursively revises the details of the input instruction so the same model carries out the requested change. Relative to the baseline, MIST reveal more than triples. On WRBench, T goes from 0.452 to 0.570. On MBench-T, U goes from 0.292 to 0.446, and 23 of 30 items keep the baseline clip, because a bad rewrite can damage a correct return. The method meets the problem from the execution side: it shows what the frozen model can already do once the move is stated.

On these cells the score is usually lost before persistence is tested, and our results do not establish that the backbone stored $s_T^\star$. A memory evaluation should therefore publish three checks, not one total: whether the scene was set, whether the change was carried out, and whether the last frame shows it.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and Francois Fleuret. Diffusion for World Modeling: Visual Details Matter in Atari. *arXiv preprint arXiv:2405.12399*, 2024. URL https://arxiv.org/abs/2405.12399.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating Physical Commonsense for Video Generation. *arXiv preprint arXiv:2406.03520*, 2024. URL https://arxiv.org/abs/2406.03520.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation. *arXiv preprint arXiv:2503.06800*, 2025. URL https://arxiv.org/abs/2503.06800.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking Intuitive Physics Understanding In Complex Synthetic Environments. *arXiv preprint arXiv:2506.09849*, 2025. URL https://arxiv.org/abs/2506.09849.

Tim Brooks, Bill Peebles, et al. Video generation models as world simulators. OpenAI technical report, 2024.

Jake Bruce et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.15391.

Haoyu Chen, Kaichen Zhou, Hang Hua, Kaile Zhang, Jingwen Qian, Wufei Ma, Haonan Chen, Chunjiang Liu, Yizhou Zhao, Xiaoyuan Wang, Weiyue Li, Alan Yuille, Paul Pu Liang, and Yilun Du. MemoBench: Benchmarking World Modeling in Dynamically Changing Environments. *arXiv preprint arXiv:2606.27537*, 2026a. URL https://arxiv.org/abs/2606.27537.

Junzhe Chen, Siyuan Meng, Yuxi Chen, Man Zhao, Wenyao Gui, and Xiaojie Guo. TOC-Bench: A Temporal Object Consistency Benchmark for Video Large Language Models. *arXiv preprint arXiv:2605.09904*, 2026b. URL https://arxiv.org/abs/2605.09904.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, Canyu Chen, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. MJ-Bench: Is Your Multimodal Reward Model Really a Good Judge for Text-to-Image Generation? *arXiv preprint arXiv:2407.04842*, 2024. URL https://arxiv.org/abs/2407.04842.

Zhifei Chen, Luozhou Wang, Guibao Shen, Dongyu Yan, Shuai Yang, Tianshuo Xu, Yihua Du, Wei Wang, Tianyi Gui, Lianghua Huang, and Yingcong Chen. ReWorld: An Interactive World Model with Long-Horizon Memory. *arXiv preprint arXiv:2608.23565*, 2026c. URL https://arxiv.org/abs/2608.23565.

Kaixin Ding, Xi Chen, Minghong Cai, Zhiyuan Xu, Yiyang Wang, Yuxiang Lu, Junyi Li, Shuyang Chen, Yuan Gao, Xin Tao, Pengfei Wan, and Hengshuang Zhao. PlayWorld: Benchmarking World Models with Agent Players over Long-Horizon Objectives. *arXiv preprint arXiv:2608.13552*, 2026. URL https://arxiv.org/abs/2608.13552.

Zicheng Duan, Jiatong Xia, Zeyu Zhang, Wenbo Zhang, Gengze Zhou, Chenhui Gou, Yefei He, Feng Chen, Xinyu Zhang, and Lingqiao Liu. LiveWorld: Simulating Out-of-Sight Dynamics in Generative Video World Models. *arXiv preprint arXiv:2603.07145*, 2026. URL https://arxiv.org/abs/2603.07145.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu jui Fu, Wenhu Chen, and William Yang Wang. TC-Bench: Benchmarking Temporal Compositionality in Text-to-Video and Image-to-Video Generation. *arXiv preprint arXiv:2406.08656*, 2024. URL https://arxiv.org/abs/2406.08656.

Bingjie Gao, Qianli Ma, Xiaoxue Wu, Shuai Yang, Guanzhou Lan, Haonan Zhao, Jiaxuan Chen, Qingyang Liu, Yu Qiao, Xinyuan Chen, Yaohui Wang, and Li Niu. RAPO++: Cross-Stage Prompt Optimization for Text-to-Video Generation via Data Alignment and Test-Time Scaling. *arXiv preprint arXiv:2510.20206*, 2025. URL https://arxiv.org/abs/2510.20206.

Samuel Garcin, Thomas Walker, Steven McDonagh, Tim Pearce, Hakan Bilen, Tianyu He, Kaixin Wang, and Jiang Bian. Beyond Pixel Histories: World Models with Persistent 3D State. *arXiv preprint arXiv:2603.03482*, 2026. URL https://arxiv.org/abs/2603.03482.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2VPhysBench: A First-Principles Benchmark for Physical Consistency in Text-to-Video Generation. *arXiv preprint arXiv:2505.00337*, 2025. URL https://arxiv.org/abs/2505.00337.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li jia Li, and Yongxin Ni. Video-Bench: Human-Aligned Video Generation Benchmark. *arXiv preprint arXiv:2504.04907*, 2025. URL https://arxiv.org/abs/2504.04907.

Xianjing Han, Bin Zhu, Shiqi Hu, Franklin Mingzhe Li, Patrick Carrington, Roger Zimmermann, and Jingjing Chen. OSCBench: Benchmarking Object State Change in Text-to-Video Generation. *arXiv preprint arXiv:2603.11698*, 2026a. URL https://arxiv.org/abs/2603.11698.

Yu Han, Wenhao Li, Yichao Cao, Hongyan Xu, Shuo Yang, Shan You, and Xiu Su. StateTrace: An Object-Centric Framework for Hidden-State Spatiotemporal Reasoning in Long Videos. *arXiv preprint arXiv:2608.18532*, 2026b. URL https://arxiv.org/abs/2608.18532.

Hangzhou He, Lunhao Duan, Shanshan Zhao, Kaiwen Li, Qing-Guo Chen, Weihua Luo, and Yanye Lu. Test-Time Scaling for Video Diffusion Models via Diagnosis-Guided Candidate Recycling. *arXiv preprint arXiv:2608.29322*, 2026a. URL https://arxiv.org/abs/2608.29322.

Haoran He, Jiajun Liang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Ling Pan. Scaling Image and Video Generation via Test-Time Evolutionary Search. *arXiv preprint arXiv:2505.17618*, 2025. URL https://arxiv.org/abs/2505.17618.

Ruozhen He, Meng Wei, Ziyan Yang, and Vicente Ordonez. EntityBench: Towards Entity-Consistent Long-Range Multi-Shot Video Generation. *arXiv preprint arXiv:2605.15199*, 2026b. URL https://arxiv.org/abs/2605.15199.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation. *arXiv preprint arXiv:2406.15252*, 2024. URL https://arxiv.org/abs/2406.15252.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text. *arXiv preprint arXiv:2403.14773*, 2024. URL https://arxiv.org/abs/2403.14773.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. RELIC: Interactive Video World Model with Long-Horizon Memory. *arXiv preprint arXiv:2512.04040*, 2025. URL https://arxiv.org/abs/2512.04040.

Kaiyi Huang, Yukun Huang, Xuefei Ning, Zinan Lin, Yu Wang, and Xihui Liu. GenMAC: Compositional Text-to-Video Generation with Multi-Agent Collaboration. *arXiv preprint arXiv:2412.04440*, 2024. URL https://arxiv.org/abs/2412.04440.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting Video Diffusion Models to Interactive World Models. *arXiv preprint arXiv:2505.14357*, 2025a. URL https://arxiv.org/abs/2505.14357.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. *arXiv preprint arXiv:2506.08009*, 2025b. URL https://arxiv.org/abs/2506.08009.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. *arXiv preprint arXiv:2311.17982*, 2023. URL https://arxiv.org/abs/2311.17982.

Karthik Inbasekar, Guy Rom, and Omer Shlomovits. WorldJen: An End-to-End Multi-Dimensional Benchmark for Generative Video Models. *arXiv preprint arXiv:2605.03475*, 2026. URL https://arxiv.org/abs/2605.03475.

Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian GE, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, Weilin Huang, and Ping Luo. Prompt-A-Video: Prompt Your Video Diffusion Model via Preference-Aligned LLM. *arXiv preprint arXiv:2412.15156*, 2024. URL https://arxiv.org/abs/2412.15156.

Ziheng Jia, Jiaying Qian, Zicheng Zhang, Xiaorong Zhu, Lancheng Gao, and Xiongkuo Min. RAVEN-Eval: Rubric-Guided Automatic Evaluation for AI Video Generation Models Based on LMM Preference Judgement. *arXiv preprint arXiv:2608.09111*, 2026. URL https://arxiv.org/abs/2608.09111.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio du Pin Calmon. Inference-Time Reward Hacking in Large Language Models. *arXiv preprint arXiv:2506.19248*, 2025. URL https://arxiv.org/abs/2506.19248.

Eddie Landesberg. When LLM Judge Scores Look Good but Best-of-N Decisions Fail. *arXiv preprint arXiv:2603.12520*, 2026. URL https://arxiv.org/abs/2603.12520.

Daeun Lee, Jaehong Yoon, Jaemin Cho, and Mohit Bansal. Self-Correcting Text-to-Video Generation with Misalignment Detection and Localized Refinement. *arXiv preprint arXiv:2411.15115*, 2024. URL https://arxiv.org/abs/2411.15115.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging Video Generation Models As World Models. *arXiv preprint arXiv:2502.20694*, 2025. URL https://arxiv.org/abs/2502.20694.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. VL-RewardBench: A Challenging Benchmark for Vision-Language Generative Reward Models. *arXiv preprint arXiv:2411.17451*, 2024. URL https://arxiv.org/abs/2411.17451.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-Time Scaling for Video Generation. *arXiv preprint arXiv:2503.18942*, 2025a. URL https://arxiv.org/abs/2503.18942.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving Video Generation with Human Feedback. *arXiv preprint arXiv:2501.13918*, 2025b. URL https://arxiv.org/abs/2501.13918.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and Evaluating Large Video Generation Models. *arXiv preprint arXiv:2310.11440*, 2023b. URL https://arxiv.org/abs/2310.11440.

Yuanxin Liu, Rui Zhu, Shuhuai Ren, Jiacong Wang, Haoyuan Guo, Xu Sun, and Lu Jiang. UVE: Are MLLMs Unified Evaluators for AI-Generated Videos? *arXiv preprint arXiv:2503.09949*, 2025c. URL https://arxiv.org/abs/2503.09949.

Zhe Liu, Jinghua Hou, Yuxiang Lu, Zhenya Yang, Xianzhe Fan, Junwei Luo, Junyi Li, Ruihua Han, Zhi Hou, and Hengshuang Zhao. StreamPI: Streaming Multimodal Temporal Modeling for Vision-Language-Action Models. *arXiv preprint arXiv:2608.26067*, 2026. URL https://arxiv.org/abs/2608.26067.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arık. VISTA: A Test-Time Self-Improving Video Generation Agent. *arXiv preprint arXiv:2510.15831*, 2025. URL https://arxiv.org/abs/2510.15831.

Jinpeng Lu, Dexu Zhu, Haoyuan Shi, Linghan Cai, Guo Tang, Yinda Chen, Jie Cao, Duyu Tang, Yi Zhang, Yong Dai, and Xiaozhu Ju. Current world models lack a persistent state core. *arXiv preprint arXiv:2606.20545*, 2026.

Ziqi Ma, Mengzhan Liufu, and Georgia Gkioxari. Out of Sight, Out of Mind? Evaluating State Evolution in Video World Models. *arXiv preprint arXiv:2603.13215*, 2026. URL https://arxiv.org/abs/2603.13215.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards World Simulator: Crafting Physical Commonsense-Based Benchmark for Video Generation. *arXiv preprint arXiv:2410.05363*, 2024. URL https://arxiv.org/abs/2410.05363.

NVIDIA et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. URL https://arxiv.org/abs/2501.03575.

Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-Time Text-to-Video Alignment with Diffusion Latent Beam Search. *arXiv preprint arXiv:2501.19252*, 2025. URL https://arxiv.org/abs/2501.19252.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video Diffusion Alignment via Reward Gradients. *arXiv preprint arXiv:2407.08737*, 2024. URL https://arxiv.org/abs/2407.08737.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Vignav Ramesh and Morteza Mardani. Test-Time Scaling of Diffusion Models via Noise Trajectory Search. *arXiv preprint arXiv:2506.03164*, 2025. URL https://arxiv.org/abs/2506.03164.

Joonghyuk Shin, Yicong Hong, Jaesik Park, and Xun Huang. Can video world models track unobserved world states? *arXiv preprint arXiv:2608.30692*, 2026.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. *arXiv preprint arXiv:2408.03314*, 2024. URL https://arxiv.org/abs/2408.03314.

Yiwen Song, Tomas Pfister, and Yale Song. VQQA: An Agentic Approach for Video Evaluation and Quality Improvement. *arXiv preprint arXiv:2603.12310*, 2026. URL https://arxiv.org/abs/2603.12310.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. *arXiv preprint arXiv:2407.14505*, 2024. URL https://arxiv.org/abs/2407.14505.

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. AlayaWorld: Interactive Long-Horizon World Modeling – Full Technical Report. *arXiv preprint arXiv:2607.18367*, 2026. URL https://arxiv.org/abs/2607.18367.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024.

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Yunhao Ba, Alex Wong, Celso M de Melo, and Achuta Kadambi. WorldBench: Benchmarking Physical Understanding of World Models by Isolating Physics Concepts. *arXiv preprint arXiv:2601.21282*, 2026. URL https://arxiv.org/abs/2601.21282.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion Models Are Real-Time Game Engines. *arXiv preprint arXiv:2408.14837*, 2024. URL https://arxiv.org/abs/2408.14837.

Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Qingyan Bai, Ka Leong Cheng, Yue Yu, Yixuan Li, Yihao Meng, Zichen Liu, Yanhong Zeng, Yujun Shen, and Qifeng Chen. WorldDirector: Building Controllable World Simulators with Persistent Dynamic Memory. *arXiv preprint arXiv:2607.02517*, 2026. URL https://arxiv.org/abs/2607.02517.

Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. AIGV-Assessor: Benchmarking and Evaluating the Perceptual Quality of Text-to-Video Generation with LMM. *arXiv preprint arXiv:2411.17221*, 2024. URL https://arxiv.org/abs/2411.17221.

Jiarui Wang, Huiyu Duan, Ziheng Jia, Yu Zhao, Woo Yi Yang, Zicheng Zhang, Zijian Chen, Juntong Wang, Yuke Xing, Guangtao Zhai, and Xiongkuo Min. LOVE: Benchmarking and Evaluating Text-to-Video Generation and Video-to-Text Interpretation. *arXiv preprint arXiv:2505.12098*, 2025. URL https://arxiv.org/abs/2505.12098.

Meiqi Wang and Shichao Li. Compiling and Benchmarking Task-State Horizons for Embodied Agents. *arXiv preprint arXiv:2608.08036*, 2026. URL https://arxiv.org/abs/2608.08036.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. *arXiv preprint arXiv:2407.15754*, 2024a. URL https://arxiv.org/abs/2407.15754.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are Scalable World Models. *arXiv preprint arXiv:2405.15223*, 2024b. URL https://arxiv.org/abs/2405.15223.

Meiqi Wu, Jiashu Zhu, Xiaokun Feng, Chubin Chen, Chen Zhu, Bingze Song, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Kaiqi Huang. ImagerySearch: Adaptive Test-Time Search for Video Generation Beyond Semantic Dependency Constraints. *arXiv preprint arXiv:2510.14847*, 2025. URL https://arxiv.org/abs/2510.14847.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-World: Scaling Interactive World Models to 1000-Frame Horizons via Pose-Free Hierarchical Memory. *arXiv preprint arXiv:2602.02393*, 2026. URL https://arxiv.org/abs/2602.02393.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Pandora: Towards General World Model with Natural Language Actions and Video States. *arXiv preprint arXiv:2406.09455*, 2024. URL https://arxiv.org/abs/2406.09455.

You-Zhe Xie, Yu-Hsuan Li, Jie-Ying Lee, Kaipeng Zhang, Yu-Lun Liu, and Zhixiang Wang. YoCausal: How Far is Video Generation from World Model? A Causality Perspective. *arXiv preprint arXiv:2605.30346*, 2026. URL https://arxiv.org/abs/2605.30346.

Ting-Bing Xu, Jiacheng Sui, Zhe Gao, Kewei Shi, Wenjin Yang, Zhicheng Liu, Zhaoxu Sun, Mingchao Sun, Hongyu Pan, Fan Jiang, Mu Xu, Qi Fan, Yang Gao, Yong Li, and Baoquan Chen. WorldRoamBench: An Open-World Benchmark for Long-Horizon Stability of Interactive World Models. *arXiv preprint arXiv:2606.31672*, 2026. URL https://arxiv.org/abs/2606.31672.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. PhyT2V: LLM-Guided Iterative Self-Refinement for Physics-Grounded Text-to-Video Generation. *arXiv preprint arXiv:2412.00596*, 2024. URL https://arxiv.org/abs/2412.00596.

Haolin Yang, Feilong Tang, Ming Hu, Qingyu Yin, Yulong Li, Yexin Liu, Zelin Peng, Peng Gao, Junjun He, Zongyuan Ge, and Imran Razzak. ScalingNoise: Scaling Inference-Time Search for Generating Infinite Videos. *arXiv preprint arXiv:2503.16400*, 2025a. URL https://arxiv.org/abs/2503.16400.

Sizhe Yang, Juncheng Mu, Tianming Wei, Chenhao Lu, Xiaofan Li, Linning Xu, Zhengrong Xue, Zhecheng Yuan, Dahua Lin, Jiangmiao Pang, and Huazhe Xu. MemoryWAM: Efficient World Action Modeling with Persistent Memory. *arXiv preprint arXiv:2606.20562*, 2026. URL https://arxiv.org/abs/2606.20562.

Yuhang Yang, Ke Fan, Shangkun Sun, Hongxiang Li, Ailing Zeng, FeiLin Han, Wei Zhai, Wei Liu, Yang Cao, and Zheng-Jun Zha. VideoGen-Eval: Agent-based System for Video Generation Evaluation. *arXiv preprint arXiv:2503.23452*, 2025b. URL https://arxiv.org/abs/2503.23452.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as Memory: Scene-Consistent Interactive Long Video Generation with Memory Retrieval. *arXiv preprint arXiv:2506.03141*, 2025. URL https://arxiv.org/abs/2506.03141.

Sihyun Yu, Nanye Ma, Pinzhi Huang, Hyunseok Lee, Shusheng Yang, June Suk Choi, Ellis Brown, Oscar Michel, Boyang Zheng, Jinwoo Shin, and Saining Xie. Benchmarking visual state tracking in multimodal video understanding. *arXiv preprint arXiv:2606.03920*, 2026.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Text-to-Time-lapse Video Generation. *arXiv preprint arXiv:2406.18522*, 2024. URL https://arxiv.org/abs/2406.18522.

Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. EWMBench: Evaluating Scene, Motion, and Semantic Quality in Embodied World Models. *arXiv preprint arXiv:2505.09694*, 2025. URL https://arxiv.org/abs/2505.09694.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models. *arXiv preprint arXiv:2504.12626*, 2025. URL https://arxiv.org/abs/2504.12626.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, Hao Liu, Chen Li, Jing Lyu, and Yueqi Duan. MBench: A comprehensive benchmark on memory capability for video world models. *arXiv preprint arXiv:2606.00793*, 2026.

Zengqun Zhao, Ziquan Liu, Yu Cao, Shaogang Gong, Zhensong Zhang, Jifei Song, Jiankang Deng, and Ioannis Patras. LatSearch: Latent Reward-Guided Search for Faster Inference-Time Scaling in Video Diffusion. *arXiv preprint arXiv:2603.14526*, 2026. URL https://arxiv.org/abs/2603.14526.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL https://arxiv.org/abs/2503.21755.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking Multi-task Long Video Understanding. *arXiv preprint arXiv:2406.04264*, 2024. URL https://arxiv.org/abs/2406.04264.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A Fine-Grained World Model for Robot Manipulation. *arXiv preprint arXiv:2406.14540*, 2024. URL https://arxiv.org/abs/2406.14540.

AI USE STATEMENT

We used generative AI tools to organize experiment reports, edit prose, and assist with analysis and figure-generation code. We reviewed all AI-assisted work and verified the reported claims and numbers against primary artifacts. We take responsibility for the final content of this submission.

ETHICS STATEMENT

This work evaluates generated videos and involves no human subjects or personal data. The VLM-based diagnostic may inherit biases or perceptual failures from its underlying models and should not be treated as a safety-critical evaluator.

REPRODUCIBILITY STATEMENT

Section 6 specifies populations, metric tracks, and selection boundaries. The appendix provides models, generation settings, protocol cards, MIST accounting, complete search budgets, statistical details, and judge-agreement checks, while separating Official, Adapted, Internal, and unsuccessful runs.

A PROTOCOL CARDS

Table 8: **Three tracks. Do not add across rows.**

	MIST (Internal)	**WRBench unique**	**MBench-T unique**
Backbone	TI2V-5B, 3×5s	T2V-A14B, 161f/10s	T2V-A14B, 161f/10s
Prompt	hide $s_T^\star$ in gap	official `yaw_LR`	native 5 captions
Gold in caption?	baseline no; rewrite arm yes	`task-context none`	no; lock is an intervention
Judge	Qwen3.5-122B	official scorer	official `mbench`
Score	reveal / inside	$T = (D3{+}D4{+}D5'{+}D6')/4$	$U = (\text{state}{+}\text{corr})/2$
N/A	never-open ≠ forget	missing return → 0 in T	untriggered → 0 in U
Seed class	9042	9042-class	8042 (official leftover split)
Execution head	last-shot rewrite, $K{=}5$	pack search, two selectors	pack search, refusal allowed

H3 packed-15s pilots are a fourth Internal cell, not a fourth official track (Table 16).

WRBench uses A14B, 161 frames / 10s, `yaw_LR`, its official scorer, and `task-context none`; MBench-T uses five native segments and official U. MIST is Internal. Adapted three-chunk MBench cond. 0 is omitted.

B MIST DIAGNOSTIC DETAILS

Kept-clip accounting. MIST is an Internal R0–R2 diagnostic. Kept labels are closed ≻ empty ≻ wrong ≻ correct. Of 41 baseline never-opens, rewrite leaves 19 closed, opens 9 to a wrong state, and makes 13 correct; `R2-10` closes. Conditional inside match is reported separately from this exclusive class: the independent check passes 19/43, while "correct" additionally requires a successful reveal and passes 13/43. Kept-clip reveal 0.492 is not best-of-five 0.763.

Decimal gate rows in Table 2 are means of judge scores in $[0, 1]$; exclusive counts apply an ordered thresholded rubric to the same clips. Hence mean reveal 0.147 is not $2/43$. Likewise, baseline mean inside is 0.047, while the strict 122B YES rate in Table 7 is $1/43 = 0.023$. For the rewrite, the mean and strict rate both round to 0.442 $(19/43)$. We keep these quantities named rather than treating them as interchangeable.

Gap leak. Paired gap leak is $0/43$ in both arms. A broader 51-row baseline export also contains no leak, but is not the paired denominator. Setup is not a leak; official captions remain `task-context none`, and a last-frame lock is an intervention.

Family and pass budget. Kept-clip reveal moves in every family; setup does not: R0 $0.132 \rightarrow 0.549$, R1 $0.147 \rightarrow 0.464$, R2 $0.163 \rightarrow 0.454$.

Best-of-five running-best reveal, distinct from kept-clip analysis, is 0.371, 0.546, 0.700, 0.720, and 0.763 for k=1 . . . 5; inside moves 0.256 → 0.721. Rewrite pulls away after k=3. A six-ID same-prompt resample ends at 0.580 and is not directly comparable.

C OFFICIAL SEARCH BUDGET

Official MBench-T is Table 5; adapted three-chunk cond. 0 is omitted. 0.446 is mostly keep-baseline, whereas fixed slogan / lock gives 0.251 / 0.228. Internal MBench-T 0.230 → 0.426, Internal WRBench 0.372 → 0.668, and prompt CLIP 31.528 → 32.303 are different scales. Selection keeps 23 baseline, five leave–return, and two last-frame clips.

Table 9: **Official vs. Internal, do not add across rows.** WRBench T zeros N/A. Vis. 0.755 uses a different clip on 24/40 IDs.

Suite	Cell	Base	M3	Δ	Track
WRBench/40	Official T	0.452	0.570	+0.118	Official
WRBench/40	Official vis. (other clip)	0.648	0.755	+0.107	Official
MBench-T/30	Official U	0.292	0.446	+0.154	Official
WRBench/40	Internal reveal	0.372	0.668	+0.296	Internal
MBench-T/12	Internal unique	0.230	0.426	+0.196	Internal

Table 10: **WRBench search budget, not pack versions.** Same clips as the main official table. Adding a second skill is not monotone. Peek-gate and max T are ceilings. Paper method is unique-head T=0.570. ΔT shares: vis. 3%, $D5'$ 56%, $D6'$ 41%. D6: keep 18 / gain 13 / lose 0 / closed 9.

What we do	Vis.	D5	D6	T
Camera only	0.648	16/40	18/40	0.452
Search over first skill	0.681	29/40	23/40	0.550
+ second skill	0.662	28/40	27/40	0.547
+ leftover repair	0.654	31/40	31/40	0.570
One skill, no search	0.690	23/40	14/40	0.496
Peek gate, then vis.	0.710	35/40	21/40	0.586
Max T (oracle)	0.707	36/40	24/40	0.601

D HELD-OUT LAST-FRAME JUDGE

Same locked last frames as Table 7. The main 122B judge is unchanged. Qwen3-VL-8B-Instruct rereads the frame and answers YES/NO to $s^{\star}_{T}$. Official T/U are not replaced. The main-text last-frame 0.023 is a strict YES rate; mean inside in Table 2 is 0.047.

Table 11: **Held-out 8B vs locked 122B / official gates.** MIST inside agreement is high. An 8B YES is not official D5/D6 or MBench state: those rubrics are not the same question.

Set	8B YES	Ref. rate	Agree
MIST last, base (vs 122B inside)	0.09	0.023	0.93
MIST last, M3 (vs 122B inside)	0.56	0.442	0.88
MIST last, base (vs 122B act)	0.09	0.070	0.84
MIST last, M3 (vs 122B act)	0.56	0.535	0.51
WRBench unique last (vs D5 gate)	0.33	0.78	0.50
WRBench unique last (vs D6 gate)	0.33	0.78	0.35
MBench winner last (vs state>0.5)	0.17	0.47	0.43

D.1 RUNS OUTSIDE THE MAIN CLAIM

Leftover I2V (not Official). Leftover-12 uses TI2V-5B aftermath stills, A14B I2V, and a 161-frame stitch (seed 10042), rather than frozen T2V-A14B. The I2V arm failed before producing clips (FSDP `no_sync`; 0/12), so it has no official score. Official U remains 0.446; scored v4 0.463 is a leftover note, not an abstract claim. The H3 12-case expansion was paused; completed six-case Internal pilots are reported separately in Section K. They show that the workflow can be instantiated on MiniMax-H3; they do not support a scale or official-protocol claim.

E EXTENDED MIST EVIDENCE

E.1 INFORMATION FLOW AND FAMILY CONSTRUCTION

MIST uses separately generated establish, opaque-gap, and reveal shots. The paired reveal analysis freezes the first two shots and regenerates only the last shot from the same pre-reveal context. The baseline reveal request and middle gap omit $s_T^\star$. Answer-only and diagnosis-guided arms may place the target in the reveal prompt; Table 3 marks that access explicitly. The prompt-channel audit confirms target access in the M3 and same-sentence templates and confirms its removal from the generic open-and-hold arm. Thus the latter arm is answer-free, whereas the same-sentence resample is an equal-compute control, not a gold-free control.

E.2 KEPT CLIPS, CANDIDATE PACKS, AND PASS BUDGET

Main-text Figure 4 reports the kept-clip classes, equal-budget arms, and running-best curve. The table below records the complete per-pass accounting used to draw panel (c).

Table 12: **Complete MIST running-best budget.** Per-pass means are non-monotone. "Inside of chosen" evaluates the reveal-argmax clip and is not the kept-clip inside 0.442.

k	Pass mean	Reveal best	Inside of chosen	Closed	Wrong	Correct	≥ 0.6
1	0.371	0.371	0.256	19	14	10	10/43
2	0.399	0.546	0.442	10	17	16	16/43
3	0.484	0.700	0.628	8	10	25	25/43
4	0.444	0.720	0.651	7	10	26	26/43
5	0.560	0.763	0.721	5	9	29	28/43

On the six IDs with a fully matched same-prompt compute control, running-best reveal is $0.410, 0.557, 0.561, 0.570, 0.580$, versus 0.884 for diagnosis-guided rewrite on those same IDs. This small matched subset shows that resampling explains part, but not all, of the observed lift; it is not a substitute for an all-43 factorial control. Constraints-only and ledger-only reach 0.189 and 0.215 on the same subset.

E.3 RESIDUAL FAILURES

A second 122B pass over the last third of the 41 baseline never-open clips assigns one reason each: in 38 a hand or agent enters and does not open; in 3 no agent enters before the shot ends; in 0 the wrong object is opened. The 5-second shot length is not the main reason the lid stays closed. Of the 41 baseline never-open cases, rewrite leaves 19 never-open, opens nine to a wrong state, and makes 13 correct. The nine open-but-wrong outcomes include empty containers, an object outside its box, count multiplication, and inverted stacks (`R0-06`, `R0-05`, `R0-11`, `R2-14`). Diagnosis revises 33/43 first-pass prompts; its most frequent visible tags are wrong final state (32), wrong count or binding (31), opening not visible (19), contradictory extras (20), and an unstable hold (9). Rewrites are not monotone: `R2-10` moves from a correct baseline to never-open. After rewrite the exclusive leftover is $6/4/6$ never-open / wrong / correct on R0, $6/2/3$ on R1, and $8/4/4$ on R2. On the paired 43, 38 clips have setup ≥ 0.8 while baseline reveal stays below 0.3.

F M3-AGENT IMPLEMENTATION AND ORACLE ACCOUNTING

The intent ledger contains actor/object, binding, target state, temporal order, required operation, and final-frame evidence. Diagnosis reads rendered frames and selects among visible labels: no action, wrong actor or binding, missing departure, missing return, return to the pre-event state, occlusion, and unstable final hold. The rewrite may change only the failed operation and evidence clause; it preserves the case identity and temporal order.

Head A preserves the first two 5-second MIST shots and generates $K = 5$ last-shot candidates. Head B keeps each official temporal protocol: WRBench uses continuous `yaw_LR` with `task-context none`, while MBench-T uses five native caption segments. The original candidate is always eligible. Visual-quality and unique-score selectors remain separate, and score ties retain the earlier candidate.

All reported best-of-K and official keep-best cells are selected after generation with the target and relevant scorer. They are oracle coverage estimates: they ask whether the candidate pack contains a better executable path. They are not evidence that a learned or target-blind critic can identify that path at deployment. No selection score is back-propagated into the frozen generator.

Selectors that cannot see the scorer. Table 13 re-selects one clip per ID from the same official candidate packs without the official scorer. The blind selector is a frozen Qwen3-VL-8B that sees only the last frame and the setup sentence, asks whether the subject has returned (WRBench) or whether the event result is still visible (MBench-T), and never sees $s_T^\star$, `expected_reveal`, or any score. Paired bootstrap intervals use 4,000 resamples over IDs, seed 0. None of the scorer-free selectors separates from the baseline; keep-best does. The MBench-T single-clip keep-best is 0.443; the published 0.446 averages a state-oracle and a correctness-oracle pick.

Table 13: **Official T/U under selectors that do not see the scorer.** Same candidate clips as the main tables. 95% intervals are paired differences from the baseline. Do not add across rows.

Selector	WRBench T (n=40)	95% CI Δ	MBench-T U (n=30)	95% CI Δ
First clip (baseline)	0.452	—	0.292	—
First rewrite only	0.496	$[-0.027, +0.114]$	0.251	$[-0.172, +0.094]$
Random, 20 draws	0.487 ± 0.019	$[-0.008, +0.076]$	0.265 ± 0.030	$[-0.101, +0.027]$
Blind 8B, last frame	0.466	$[-0.044, +0.070]$	0.248	$[-0.107, +0.007]$
Keep-best (sees scorer)	0.570	$[+0.070, +0.166]$	0.443	$[+0.074, +0.243]$

G COMPLETE OFFICIAL-PROTOCOL EVIDENCE

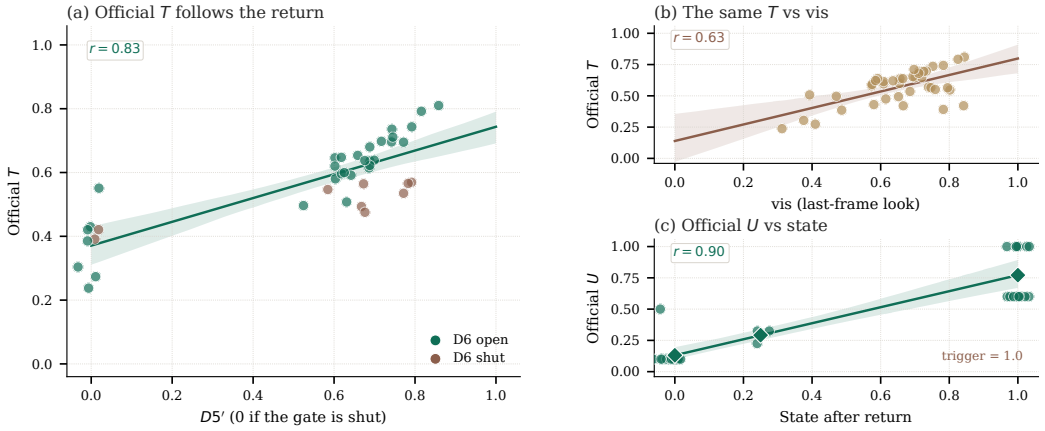

Figure 6: **Official-score component plots.** (a–b) WRBench T against $D5'$ and visual quality. (c) MBench U against returned state. Because these totals contain the plotted components, the panels describe score composition; they do not establish an independent causal correlation.

WRBench failure anatomy. For the official 40-case unique head, the T gain is concentrated in the leave–return path rather than visual quality. The component audit attributes approximately 3–4% of ΔT to visual quality, 55–56% to $D5'$, and 41% to $D6'$, depending on rounding. Thirteen $D6$ gates newly open, none close, and nine remain closed in the final $T = 0.570$ pack. Earlier search stages are non-monotone (Table 10), and a stricter departure instruction can recover return gates while lowering visual quality or $D5$. This is why pack size and visual quality are not treated as sufficient explanations.

MBench-T refusal and hard cases. The official five-segment harness improves the conditional state and correctness scores while retaining the original clip on most items. Fixed leave–return hurts eight cases whose baseline state score is 1.0. Fourteen of 30 cases remain hard after the first two arms: the leave–return trigger usually fires, but the object returns intact or in the wrong post-event state. A prompt-only hold-update attempt fixes 0/14. A last-frame lock alone gives $U = 0.228$; adding it to the candidate pack helps only because selection may refuse it. A later destroy-and-hold slogan is near

floor as a fixed arm and flips one additional case under oracle selection. These negative cells support the refusal interpretation; they are not additional headline methods.

Table 14: **Official MBench-T five-segment details.** Trigger and conditional scores are shown separately; they are not MIST metrics.

Metric	Baseline	Fixed leave–return	Keep-best
Prompt score	31.528	31.382	32.303
State trigger	0.933	0.933	0.967
State conditional	0.375	0.259	0.474
Correctness trigger	0.833	0.867	0.967
Correctness conditional	0.280	0.300	0.352

H EVIDENCE BOUNDARIES

Table 15: **Evidence boundaries used throughout the paper.**

Observation	Supported reading	Not established
Valid setup, clean gap, closed reveal	The visible reveal path is missing	The hidden state was forgotten
$41/43$ never-open, TI2V-5B 3×5s	Reveal path missing in this Internal stitch	Every generator or packed 15s take fails to open
Bo5 rewrite 0.763 vs answer-only 0.684; target withheld 0.174	The rewrite helps only with the answer in the prompt; keep-best is an oracle	Diagnosis separable from gold-in-prompt
Blind 8B selector T 0.466, U 0.248	Scorer-free selection does not beat the baseline	A deployable selector for either suite
Official keep-best T/U rises	Oracle coverage improves for that protocol	A deployable critic, or one transferable intervention
H3 n=6 Internal 8B reveal 1.000	The same cards can be packed as 15s	Official T/U, stored $s_T^\star$, or a scale claim
H3-43 MAE pass $0.605 \rightarrow 0.721$	Fewer hard cuts on packed 15s	Reveal, inside, or Official T/U
H3-43 122B reveal $0.387 \rightarrow 0.729$, never-open $3/43$	On a packed take the raw failure is wrong state, not a closed lid	That the stitch result transfers, or Official T/U

I TEXT-ONLY AND HELD-OUT JUDGE CONTROLS

The high baseline agreement in Table 11 occurs under strong class imbalance: almost every baseline frame is negative. Action agreement is much lower on rewritten MIST frames, and agreement with official WRBench/MBench gates is lower still. We therefore use the held-out model only as a robustness reread and do not replace official metrics or claim evaluator independence.

J SETTINGS, STATISTICS, AND REPRODUCIBILITY

MIST uses Wan2.2-TI2V-5B, three 5-second shots, and seed class 9042. The 122B judge reads six reveal-shot positions $\{0.90, 0.93, 0.95, 0.97, 0.99, 0.995\}$. Its prompt and parsing are frozen across arms. Paired confidence intervals use 4,000 bootstrap resamples over 43 cases with seed 0; the paired Wilcoxon test uses the same case pairs.

Official WRBench uses Wan2.2-T2V-A14B, 161 frames over 10 seconds, `yaw_LR`, `task-context none`, and the released scorer. Official MBench-T uses the same generation length and model, five native caption segments, and the released causal metrics. These settings differ jointly from MIST; TI2V-5B versus T2V-A14B is therefore not a controlled model-scale comparison. Internal, Adapted, and Official exports retain separate manifests and are never pooled.

K CROSS-GENERATOR PILOTS AND UNSUCCESSFUL RUNS

The MIST cards are the same objects and operations. The experiments are not. Table 16 lists what each reported cell actually is: generator, duration, judge, and claim. Headline MIST is Wan TI2V-5B, three 5-second shots, and the locked 122B judge. Official WRBench/MBench-T are Wan A14B, 10s, and the released scorers. MiniMax-H3 pilots are packed 15s Internal takes: the six-case table uses an 8B memory judge; the 43-case tables use an adjacent-frame continuity harness and, separately, the locked 122B reveal judge (Section M). Leftover demo stills, the six-case scored files, and the seed-61001 43-case files are three packs. Do not add a number from one row into another.

Table 16: **Same cards, different experiments.** Do not add across rows. Continuity is not reveal.

Cell	Generator	n	Judge	What it is
MIST kept reveal	TI2V-5B 3×5s	43	122B	$0.147 \to 0.492$; Internal diagnosis
MIST best-of-five	same	43	122B	0.763; oracle, answer in prompt
Official T	A14B, 10s	40	official	$0.452 \to 0.570$; leave–return
Official U	A14B, 10s	30	official	$0.292 \to 0.446$; mostly keep
H3 n=6 reveal	H3, 15s packed	6	Internal 8B	$0.435 \to 1.000$; tiny pilot
H3-43 MAE pass	H3, 15s packed	43	MAE ≥ 22	$0.605 \to 0.721$; not memory
H3-43 122B reveal	H3, 15s packed	43	122B	$0.387 \to 0.729$; never-open $3 \to 6$
Figs. 3, 7	H3, 15s packed	6 stills	eye	Leftover $\neq$ seed-61001

The six-case Internal 8B numbers stay in Table 17. The run labels H3 and Lingbot identify separate generation pipelines; neither pilot uses an official memory protocol. Prompt expansion is a front-end rewrite. We do not invent the cells that were not run: a VideoRepair or Video-T1 row under K=5 (the released Video-T1 code drives Pyramid-Flow, not Wan, and was not ported), Lingbot M3, the paused H3-12 expansion, or human labels on the 43 last frames.

Table 17: **Six-case Internal transfer pilots.** Small n, changed generators, and an Internal judge prevent a model-scale or generalization claim.

Pipeline / arm	Overall	Establish	Middle shot	Reveal
H3 raw	0.500	0.682	0.785	0.435
H3 + prompt expansion	0.937	0.842	0.503	1.000
H3 + M3	0.937	0.747	0.685	1.000
H3 + M3 + expansion	0.855	0.829	0.665	0.879
Lingbot raw	0.035	0.008	0.025	0.042
Lingbot + expansion	0.311	0.790	0.569	0.190

The pilots are mixed rather than monotone: H3 prompt expansion lowers the middle-shot score while raising reveal, stacking M3 and expansion is worse than either arm alone, and Lingbot remains near floor. The diagnose–rewrite workflow can be instantiated on MiniMax-H3 as a packed 15s prompt; we do not infer scale or official-protocol transfer. The planned 12-case H3 expansion was paused before scoring, and Lingbot M3 arms were not run.

The featured leftover `R0-03` packed 15s strip is Figure 3 in the main text, not a cell in Table 17. Additional leftover and 43-case pairs follow.

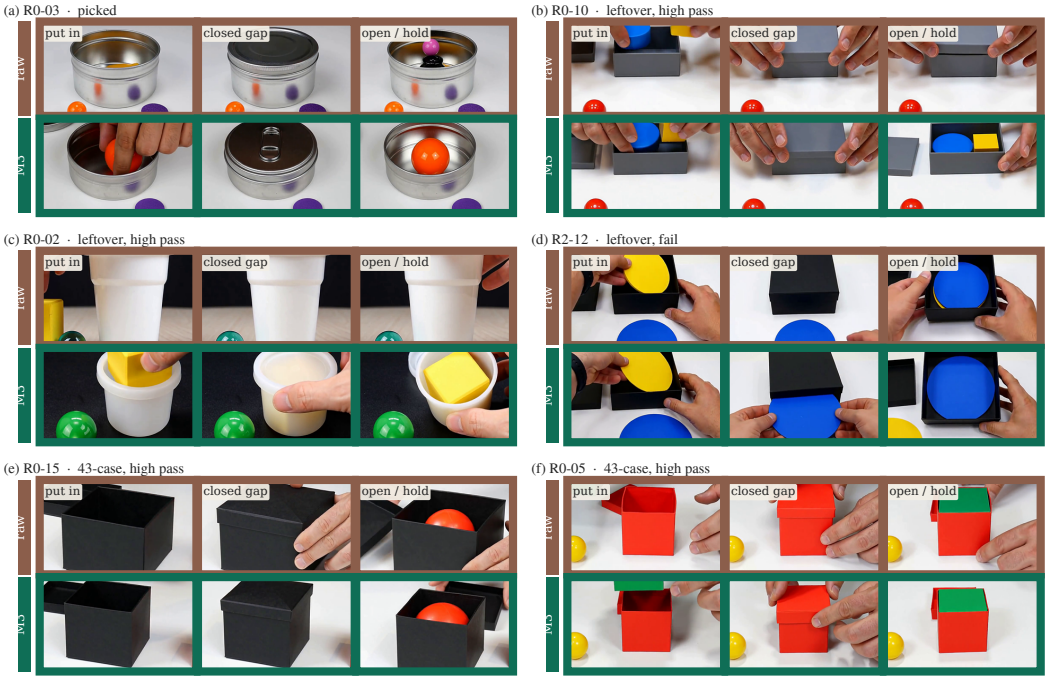

Figure 7: **H3 raw versus M3, a diagnostic plate rather than a brochure.** Each block is put-in, closed gap, then open/hold. (a) featured leftover pick. (b–c) leftover high-pass cases that were not featured. (d) leftover fail (`R2-12`): both arms fail the continuity harness, and the last frame exposes the wrong object. (e–f) 43-case Internal pairs from the seed-61001 run, not the leftover demo files. Teal frames are M3. These stills are Internal qualitative evidence, not Official T/U, not Table 17, and not reveal scores.

The MBench-T leftover I2V arm also changed the protocol: it combined TI2V-5B aftermath stills with an A14B I2V stitch. It produced 0/12 scoreable clips because of an FSDP `no_sync` failure and has no official score. Adapted MemoBench and SANA-WM smoke tests are likewise excluded from the main claim because their generation or metrics are not protocol-comparable. Reporting these exclusions prevents failed or non-native cells from silently becoming evidence for the main story.

L Extended literature map and positioning

This section records the wider 2024–2026 literature used to bound the paper's claims. Entries are grouped by the question they answer, rather than presented as evidence for one aggregate notion of "memory."

What enables long and persistent rollouts? Diffusion and latent-video foundations establish the rendering lineage (Ho et al., 2020; Blattmann et al., 2023; Team Wan, 2025), while physical-AI platforms broaden the world-model framing (NVIDIA et al., 2025). Recent systems address exposure bias, out-of-sight dynamics, bounded memory, and revisit stability through self-conditioned training or explicit state (Huang et al., 2025b; Duan et al., 2026; Team et al., 2026; Xu et al., 2026). Agent-driven and embodied evaluations further test whether action-conditioned worlds remain usable over long objectives (Ding et al., 2026; Wang & Li, 2026; Yang et al., 2026; Liu et al., 2026). EntityBench separately measures recurring entities across multi-shot gaps (He et al., 2026b). These works motivate long-horizon memory, but architectural success in one protocol does not resolve MIST's execution-versus-persistence ambiguity.

Which benchmark dimensions are adjacent to MIST? WorldBench and IntPhys 2 isolate physical concepts and object permanence, whereas EWMBench and VBench 2.0 separate scene, motion, semantics, and intrinsic faithfulness (Upadhyay et al., 2026; Bordes et al., 2025; Yue et al., 2025; Zheng et al., 2025). LongVideoBench and MLVU test long-context understanding rather than

generator-side state production (Wu et al., 2024a; Zhou et al., 2024). ChronoMagic-Bench separates amount of change from temporal coherence, and T2VPhysBench evaluates physical laws with manual judgments (Yuan et al., 2024; Guo et al., 2025). MIST instead fixes a target state and asks whether the generated operation makes that state judgeable after an opaque gap.

How has video generation been refined at test time? Reward gradients and human-feedback rewards improve video generation through learned objectives (Prabhudesai et al., 2024; Liu et al., 2025b). GenMAC and Prompt-A-Video use collaborative or preference-aligned prompt revision (Huang et al., 2024; Ji et al., 2024). ScalingNoise, VISTA, RAPO++, and ImagerySearch broaden search over noise, prompts, candidates, or evaluators (Yang et al., 2025a; Long et al., 2025; Gao et al., 2025; Wu et al., 2025). LatSearch and noise-trajectory search move selection inside diffusion inference, while GEARS diagnoses and recycles candidates (Zhao et al., 2026; Ramesh & Mardani, 2025; He et al., 2026a). Accordingly, M3-Agent's novelty is not generic test-time scaling. Its contribution is the operation-constrained response to a delayed-reveal diagnosis, and its score-aware selection remains an oracle coverage analysis.

What can automated video evaluators establish? VBench and EvalCrafter established multidimensional evaluation beyond single perceptual totals (Huang et al., 2023; Liu et al., 2023b). VideoScore, AIGV-Assessor, and LOVE learn fine-grained human-facing quality or task judgments (He et al., 2024; Wang et al., 2024; 2025). VideoGen-Eval decomposes prompts into checkable questions, while WorldJen and RAVEN-Eval study rubric design, score compression, ordering, and confidence (Yang et al., 2025b; Inbasekar et al., 2026; Jia et al., 2026). CLIP-style alignment and vision-centric VLMs provide useful readers (Radford et al., 2021; Liu et al., 2023a; Tong et al., 2024), but no one reader is treated here as ground truth for latent state. This is why MIST reports gate-specific judgments, a held-out reread, and explicit disagreement with official scorers.

M FORTY-THREE-CASE INTERNAL H3: REVEAL AND CONTINUITY

The headline $41/43$ never-open rate is a Wan TI2V-5B 3×5s stitch. The same 43 cards can be packed into one MiniMax-H3 15s take. Generation is MiniMax-H3 T2VA, 15s, short-edge 768, seed 61001. Raw uses the packed prompt without the last-frame state sentence; M3 adds that sentence. Prompt expansion is not used.

Same judge, different failure. Table 18 runs the locked 122B reveal judge with the Table 2 prompt and positions on these 43 takes. The raw take opens in 40/43 cases; its failure is a wrong exposed state (24/43), not a closed lid. The never-open cell is therefore a property of the stitch protocol, not of generators. These are Internal packed cells, one clip per arm, no keep-best; they are not Official T/U and are not on the same population as Table 2.

Table 18: **Internal 43-case H3 reveal** (locked 122B, same positions as MIST). Inside is the match rate of the exposed state. Not the stitch population; not Official.

Arm	Reveal	Inside	Never-open	Empty	Wrong	Ok
H3 raw	0.387	0.256	3	6	24	10
H3 + M3	0.729	0.674	6	1	8	28

Continuity. Table 19 asks only whether the take hard-cuts. The six-case 8B memory judge stays in Table 17.

Table 19: **Internal 43-case H3 continuity** (adjacent-frame harness). Hard-cut if gray MAE ≥ 22. Not a memory score.

Arm	n	Harness pass	Mean max MAE
H3 raw	43	0.605	19.8
H3 + M3	43	0.721	17.93

Table 20: **Internal 43-case continuity by family.** Adjacent-frame harness only. M3 raises the pass rate but is not monotone: 12 cases fail raw and pass M3, while 7 pass raw and fail M3. Both arms pass on 19/43. Not a memory score and not Table 17.

Family	n	Raw pass	M3 pass	Both pass
R0	16	0.625	0.688	9
R1	11	0.545	0.818	5
R2	16	0.625	0.688	5
All	43	0.605	0.721	19

Figure 7(e–f) are two high-pass pairs from this 43-case seed-61001 run. Panels (a–c) are leftover high-pass files; (d) is leftover `R2-12`, where both arms fail the harness and the last frame shows the wrong object. Leftover files are not the same files as this table: `R0-02` is harness-PASS in the leftover pack and FAIL in the 43-case run. Continuity only checks that the 15s take does not hard-cut; do not read $0.605 \rightarrow 0.721$ as reveal or Official T/U. Reveal on these clips is Table 18.