# VERIFY WHAT MATTERS: COST-AWARE LONG VIDEO GENERATION WITH GIVER

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Although recent agentic systems can generate long multi-shot videos from storyboards, physical inconsistencies can persist, such as fingers passing through a cup during lifting. Repairing these defects requires identifying failures and guiding regeneration without altering the requested content. To this end, we introduce GIVER, an agentic framework that embeds verification-and-repair loops into the generation process. Specifically, GIVER constructs a fixed Subject–Object Interaction Graph (SOIG), whose nodes encode storyboard-derived visual requirements and whose edges encode their logical prerequisites. Using this graph, GIVER verifies reusable assets and repairs them when needed, then runs two local loops: L1 checks the first frame's initial state, while L2 checks required video behavior and visible artifacts, retaining the selected first frame. Failed-node evidence guides targeted revisions to generation fields, and each new candidate is rechecked against the same requirements. On ViMax-Bench and ST-Bench-10, GIVER leads the compared pipelines in fidelity and physical plausibility under automated and blinded human evaluation, with automated severe-artifact rates of 0.70% and 2.80%, respectively. Additionally, our ablation study shows improved quality with approximately 24–25% fewer video generation and editing operations than two flat-verification configurations.

## 1 INTRODUCTION

A long video can be convincing for minutes and lose its credibility in a single shot. For example, when a character lifts a cup, his fingers may pass through cup's surface, undermining the physical plausibility of the whole video. While recent agentic systems (Huang et al., 2025a; 2026; Wu et al., 2025) can produce long multi-shot videos via coordinate planning and synthesis, such defects still persist. A human editor can address these defects by inspecting the clip, identifying the failed interaction, and revising the prompt before regeneration, often over multiple attempts. However, automating this diagnosis-and-repair process with agents remains challenging, especially when revisions must preserve the requested content. The practical challenge is therefore **to reduce artifact-induced rework without simplifying the requested content or lowering its visual quality**.

Automating this repair process requires a reliable link between visual diagnosis and the next generation request. A global score can indicate that one shot video is better than another, but does not necessarily explain what went wrong (Guan et al., 2025). By contrast, question-based evaluation, dependency-aware assessment, and localized visual feedback provide more specific diagnoses (Hu et al., 2023; Cho et al., 2024; Liang et al., 2024); for example, checking hand–cup contact requires first establishing that both the hand and cup are present. Yet identifying a failed check does not by itself specify how to revise the generation prompt. If the fingers pass through the cup, the revision should address the contact while retaining the requested lifting action, rather than remove the action to avoid the defect. Effective repair therefore requires keeping the original requirements fixed and using the observed failure to guide changes to the relevant parts of the prompt (Lee et al., 2026).

To address this limitation, we introduce GIVER (Graph-guided Iterative Verification and Evidence-linked Repair) around one principle: *keep the requirements fixed, and use visual evidence to guide repairs for failed requirements*. Specifically, we construct a Subject–Object Interaction Graph (SOIG) from the storyboard, where nodes encode atomic visual requirements and edges encode the logical prerequisites among them (Cho et al., 2024; Pothiraj et al., 2026). Keeping this graph fixed



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.

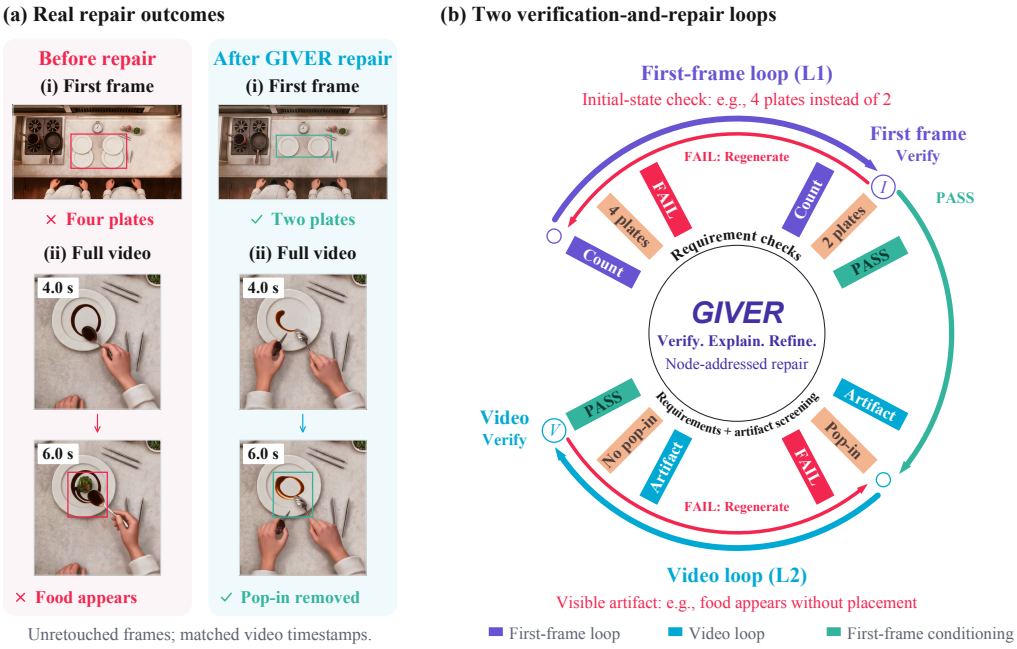


Figure 1: **Repair the initial state before repairing its evolution.** Real generated media show an initial-state mismatch (four plates instead of two) and a temporal artifact (unexplained food appearance). L1 and L2 have separate local regeneration paths; L2 retains the selected first frame. Asset verification precedes these loops. Source traces and evaluation protocols appear in Appendix C.

makes individual failures traceable across attempts, while its dependencies prevent a check from receiving positive credit when its prerequisites have not been established. Furthermore, to apply this representation during generation, GIVER draws on the filmmaking workflow of preparing a scene before filming it, embedding verification-and-repair loops into both first-frame and video generation. Accordingly, GIVER first verifies and repairs reusable assets, then runs two local verification-and-repair loops: L1 checks the first frame's initial state; L2 checks required behavior and visible video artifacts (Figure 1) In both loops, failed checks and their visual evidence guide revisions to the relevant generation fields. Each new candidate is then checked against all fixed requirements to determine whether the repair resolved the targeted failure without introducing new errors.

On ViMax-Bench Huang et al. (2026) and ST-Bench-10 Sun et al. (2025), GIVER achieves higher automated fidelity and physical-plausibility scores and lower severe-artifact rates than the compared pipeline adaptations. Blinded human ratings also favor GIVER on fidelity and physical plausibility. In a separate video-stage experiment with shared baseline-generated first frames, GIVER improves quality with approximately 24–25% fewer video generation and editing operations than two flat-verification configurations. **Our contributions are: (i)** a dependency-aware graph linking fixed requirements and node-level visual evidence to targeted repair; **(ii)** a staged pipeline combining asset verification and repair with separate first-frame and video loops; and **(iii)** automated and blinded human evaluations of video quality, plus a fixed-first-frame experiment on repair efficiency.

## 2 Related Work

Multi-shot video systems must coordinate a narrative and control the quality of its realization. We distinguish three control mechanisms by the decision they make: *what to generate*, *which candidate to retain*, and *what to change after observing an output*. A system can combine these mechanisms; the distinction concerns their control decisions, not agent names or a ranking of architectures.

**Planning and conditioning.** This family controls the information supplied to generation. Vlogger, VideoDirectorGPT, and MovieAgent decompose narratives into executable scene or shot plans (Zhuang et al., 2024; Lin et al., 2024; Wu et al., 2025). VideoStudio, VGoT, and StoryAgent carry entity references or visual context across shots (Long et al., 2024; Zheng et al., 2024;

Hu et al., 2024); Captain Cinema and STAGE use keyframes or structured storyboards to coordinate synthesis (Xiao et al., 2025; Zhang et al., 2025b). Related model-level approaches include hierarchical sequence generation in MovieDreamer and joint multi-shot modeling in HoloCine (Zhao et al., 2025; Meng et al., 2025). StoryMem and OneStory maintain visual memory for later shots (Zhang et al., 2025a; An et al., 2025). Their shared purpose is to make independently generated content consistent with a larger narrative. Such conditioning is complementary to checking whether the resulting images and clips correctly depict the requested objects, local interactions, and motion.

**Evaluation-guided selection.** Here, evaluation determines which candidate to retain or where to search next. Anim-Director evaluates visual candidates, and ViMax uses VLM-based best-of-$k$ selection (Li et al., 2024; Huang et al., 2026). AniMaker extends this mechanism to contextual clip evaluation with MCTS-driven search (Shi et al., 2025). Their common decision concerns *which realization to use*. Selection can improve quality without specifying a corresponding change to a generation field. ViMax also uses a graph to coordinate visual references across shots; GIVER instead represents prerequisites among checks within a shot. It retains candidates and links their failed checks to generation fields, using the diagnosis to direct the next repair request.

**Feedback-driven correction.** This mechanism uses generated outputs to revise later requests, intermediate content, or the workflow itself. DreamFactory iterates keyframes, while MAViS and PersonaVlog integrate staged review and refinement (Xie et al., 2024; Wang et al., 2025; Hou et al., 2025). AesopAgent and VideoWeaver improve workflows or reusable skills, and Hollywood Town allows bounded feedback between agents (Wang et al., 2024; Wei et al., 2026; 2025). FilMaster uses audience-oriented feedback in post-production (Huang et al., 2025a); Co-Director combines local refinement with final-video rewards for creative-direction search (Song et al., 2026). GIVER belongs to this family. Its contribution is not the existence of a feedback loop, but the fixed requirements and node-linked observations that connect a failed candidate to an eligible revision field. These requirements persist across attempts, so each revision is evaluated against the same target.

**Structured visual feedback.** Question-based evaluation decomposes alignment into atomic checks. TIFA evaluates image faithfulness, DSG introduces question dependencies, and ETVA and PQSG extend structured assessment to video alignment and physics (Hu et al., 2023; Cho et al., 2024; Guan et al., 2025; Pothiraj et al., 2026). NeuS-V uses temporal logic, while SG-PVR grounds video reward reasoning in scene graphs (Sharan et al., 2025; Kim et al., 2026). These are evaluation representations rather than complete long-video production pipelines. Related generation methods use physics-oriented prompt refinement, multi-agent critique, or localized correction (Xue et al., 2025; Long et al., 2026; Lee et al., 2026). Particularly close, Tyagi et al. (2026) combine alignment questions and artifact checks with prompt and parameter optimization for image-to-video generation. GIVER builds on these foundations through fixed graphs that connect the same requirements to fresh visual evidence across repair attempts, with checks for assets, first frames, and videos.

## 3 METHOD

GIVER converts storyboard requirements into a graph, checks each candidate against those requirements, and revises the generation fields associated with failed checks. Figure 2 shows the procedure, from reusable assets to clip assembly. We first define the task and the fixed requirement graph, then explain how node-level evidence determines acceptance and guides the controller's next repair.

### 3.1 TASK AND OBJECTIVE

The input is a storyboard $X = (x_1, \ldots, x_N)$ with cast and scene specifications. Each shot $x_i$ describes an initial state, action, camera, and duration $d_i$. An upstream planner may produce this storyboard; our control procedure begins with it. Let $G_I$ and $G_V$ denote black-box image and image-conditioned video generators. Given selected references $\mathcal{A}_i$, image and video attempts follow

$$I_i^{(a)} \sim G_I(p_{i,I}^{(a)}, \mathcal{A}_i), \qquad v_i^{(b)} \sim G_V(p_{i,V}^{(b)}, I_i^*; d_i), \qquad y = \mathrm{Concat}(v_1^*, \ldots, v_N^*). \tag{1}$$

Here $a, b$ index attempts, $p$ is a generation request, and $I_i^*, v_i^*$ are the selected frame and clip. L2 holds $I_i^*$ fixed. The procedure uses only generator requests, without access to weights or gradients.

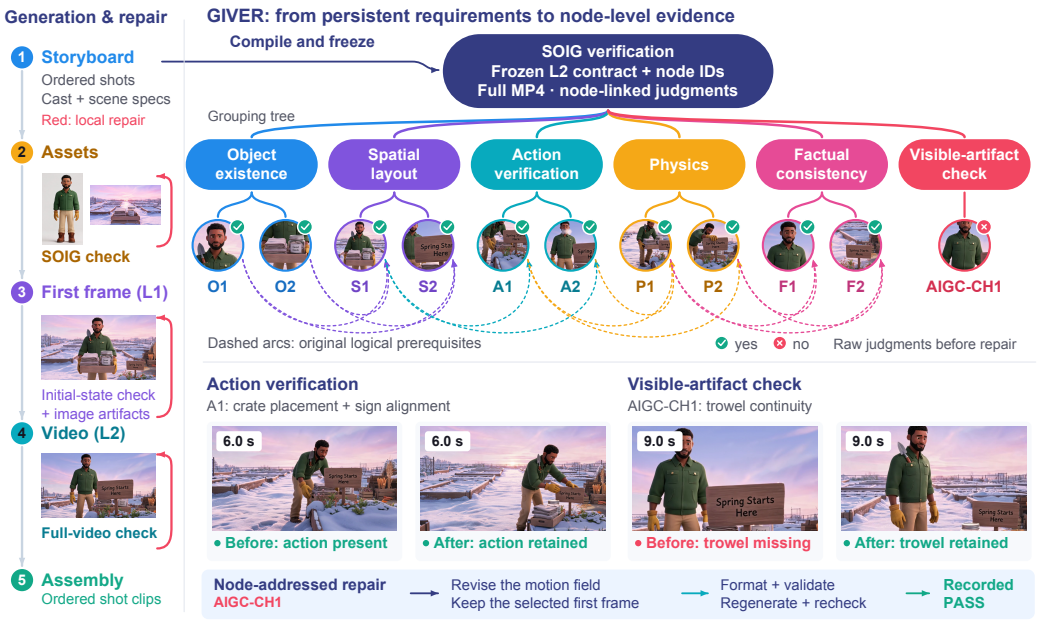


Figure 2: **From a storyboard to evidence-linked repair.** Assets, first frames, and videos undergo local review before ordered assembly. The expanded L2 tree groups atomic checks; dashed links represent logical prerequisites. Real frames show a satisfied action check (A1) and a failed trowel-continuity guard (AIGC-CH1), which guides motion-field revision while retaining the first frame. Badges denote recorded verifier judgments; Appendix C gives the recorded trace provenance.

The objective is fewer visible artifacts and fewer regeneration attempts while preserving fidelity to $X$. An artifact is a failure of visual integrity or physical plausibility, not simply a missing scripted action. We therefore assess fidelity, artifact severity, and effort separately. Stage budgets limit generation, verification, and refinement requests. A playable output selected at budget exhaustion is distinguished from an output that passes verification. This distinction separates the decision to stop spending requests from the evidence that a selected candidate meets the visual requirements.

### 3.2 A Fixed Graph of Visual Requirements

For shot $i$ and stage $\ell \in \{I, V\}$, the compiler builds a directed acyclic graph $\mathcal{C}_{i,\ell} = (V_{i,\ell}, E_{i,\ell})$. We call this set of requirements a *contract*. The image graph describes the initial state; the video graph adds actions and temporal requirements. Asset contracts are compiled separately. Each node contains an identifier, question, target condition, relevant entities, category, criticality, and evidence requirement. Categories cover object existence ($O$), spatial layout ($S$), action ($A$), physics ($P$), and factual consistency ($F$). An edge records a logical prerequisite: checking hand–cup contact requires identifying both the hand and cup. It does not describe a transition between shots. Nor does a category-grouping branch in Figure 2 imply a prerequisite: the dashed links encode those dependencies. The compiler validates node identities, edge endpoints, acyclicity, and stage compatibility, so that a still image is checked for its initial state, not for evidence of actions that unfold over time.

The graph remains fixed across attempts within its stage. Stages share the representation and storyboard, not an identical set of nodes. Runtime rules specify applicable checks, artifact guards, and any exclusions or waivers; these decisions are recorded separately from visual observations. This preserves three distinct objects: what the shot requires, what the verifier observes, and how the controller decides to proceed. A waiver can affect the controller's acceptance decision, but an observed failure must remain recorded as a failure rather than be presented as evidence of success.

### 3.3 Verification with Evidence and Dependencies

For a candidate, the verifier returns a record for each requested node $j$:

$$r_j = (a_j, o_j, z_j), \qquad a_j \in \{\text{yes}, \text{no}, \text{unknown}\}, \tag{2}$$

where $o_j$ describes the observation and $z_j$ specifies its supporting evidence. Each record also identifies the shot, stage, candidate, and node, so observations can be traced to the evaluated media. An unknown judgment denotes insufficient evidence, not an observed defect or a confirmed success.

L1 uses multi-turn requirement review with image reinspection and an artifact check. In the current L2 implementation, a separate conversation for each candidate receives the complete MP4 and queries the fixed nodes in dependency order, one node per turn. Earlier validated answers remain in context but cannot be overwritten. Video answers must refer to the clip: an optional first-frame attachment cannot establish later behavior. Artifact guards use the same interface. Checks on the response format and node coverage reject missing or mismatched identities. The resulting record identifies the evaluated requirement, links the answer to its supporting visual evidence, and preserves unresolved judgments for the controller instead of forcing a definite positive or negative verdict.

Let $V_s$ be the active scoring nodes, $b_j$ their policy-adjusted satisfaction, and $\mathrm{Anc}(j)$ the active ancestors of $j$. A failed prerequisite blocks positive credit below it:

$$\widetilde{b}_j = b_j \prod_{u \in \mathrm{Anc}(j)} b_u, \qquad s_{\mathrm{tree}} = \frac{1}{|V_s|} \sum_{j \in V_s} \widetilde{b}_j, \qquad s_{\mathrm{base}} = 1 + 9 s_{\mathrm{tree}}. \tag{3}$$

The empty-graph tree score is zero. Critical failures and artifact/physics vetoes additionally affect acceptance. Unknown evidence does not count as observed satisfaction; policy adjustments are logged separately. This diagnostic score controls generation and is not an external evaluation metric. Consider a contact node whose two object-presence prerequisites are satisfied: its own evidence must still establish contact. Conversely, if a prerequisite is unresolved, a downstream affirmative answer cannot independently establish the interaction. The raw answers distinguish an observed contact failure from insufficient evidence about an object. The graph determines how node answers support a decision while preserving the original observations for inspection and later repair.

### 3.4 Local Repair and Candidate Selection

**Algorithm 1** Local verification and repair in GIVER

**Require:** Fixed stage graph $\mathcal{C}$, initial request $p$, conditioning $c$, request limits $B$
$\mathcal{H} \leftarrow \emptyset$ ▷ Candidate records, including failed and unknown checks
**while** the stage request limits permit another attempt **do**
  Generate or edit a candidate $m$ using $(p, c)$; account for requests in $B$
  Validate media and obtain node observations against $\mathcal{C}$ within $B$
  **if** media or report validation fails **then**
    Record the failure; apply configured bounded error handling
    **continue** if permitted; otherwise **break**
  **end if**
  Compute the dependency-aware score and acceptance decision
  Append media, raw observations, policy decisions, and request to $\mathcal{H}$
  **if** accepted **then return** candidate and its recorded report
  Identify repair targets and eligible fields from failed or unresolved checks
  **if** no eligible repair or insufficient requests remain **then break**
  Revise and validate complete text for eligible fields; retain other fields
  Keep $\mathcal{C}$ and stage conditioning $c$ fixed for the next candidate
**end while**
Apply the configured fallback policy to $\mathcal{H}$, or report no selected output
**return** selected media, if any, with its original report and termination reason

The controller applies verification in production order. It first reviews reusable character and static scene/object assets with identity and image checks. Failed checks trigger bounded asset repair before the selected references are attached to downstream requests. L1 then generates a first frame and checks its initial state and image artifacts. For example, four plates instead of two calls for a change to the first-frame description. An accepted frame, or a recorded fallback at exhaustion, becomes the video condition. L2 checks the resulting clip for required behavior and visible artifacts, including temporal failures that a still image cannot reveal. **L2 does not return to L1**: video repair keeps the selected frame fixed. Selected clips are finally assembled in storyboard order for final delivery.

Within each local loop, failed or explicitly unresolved nodes identify eligible revision fields. The refiner receives their observations, candidate media, and current descriptions, then returns the *complete revised text* of each affected field. Other fields retain their values. L1 revises the image description; L2 revises video-generation fields without replacing its first frame. Before submission, the controller validates the update, reference attachments, and provider formatting. The graph is never rewritten to excuse a failure. A text revision can still affect unrelated pixels, so each new candidate is checked against all fixed requirements, including requirements that passed in earlier generation attempts.

Algorithm 1 summarizes this shared local procedure. Candidate records link the report, revised fields, parent candidate, and submitted request. Limits cover media generation, verification, refinement, response corrections, and retries, rather than only the outer loop count. An enabled fallback policy may select existing valid, scored media at exhaustion, but preserves its failed and unknown checks. Delivery is not relabeled as acceptance. Appendix A specifies execution details and the run-specific verification protocols; the video-stage experiment evaluates the graph-based repair configuration, not the isolated effect of the current one-node-per-turn interface used to query the video.

# 4 Experiments

We test three claims: GIVER reduces visible artifacts while preserving requested content; the quality gains are also apparent to human viewers; and structured video verification reduces repair operations relative to flat verification. Automated system comparisons, blinded human ratings, and a fixed-first-frame control experiment address these claims in turn. As in VBench, EvalCrafter, and T2V-CompBench (Huang et al., 2024b; Liu et al., 2024; Sun et al., 2025), we report separate scores for fidelity to requested content and for physical plausibility rather than combining both into one score.

## 4.1 Benchmarks and Evaluation Setup

**Tasks and generators.** ViMax-Bench (Huang et al., 2026) provides 35 themes: 12 single-character, 12 multi-character, and 11 multi-scene cases. We use its supplied scripts. Copyright restrictions limit our ST-Bench (Zhang et al., 2025a) evaluation to ten cases, which we denote *ST-Bench-10*. ViMax, MovieAgent, AniMaker, and GIVER are run on the same ten cases, and their generated outputs are scored by the MLLM evaluator; Table 1 reports the results. ST-Bench-10 uses five-second clips. The rendering stack uses Seedream 5.0 and Seedance 2.0, 720p, 16:9, with a shared stylized 3D setting. The ViMax system baseline uses 15-second clips, whereas the separate video-stage ablation uses ten-second clips and supplies the same baseline first frame to the repair configurations.

**Baseline configurations.** We adapt ViMax, MovieAgent, and AniMaker to the same rendering backend. AniMaker retains its original full MCTS-based video search, including contextual candidate evaluation and search over video candidates. Table 1 compares the outputs from these configurations. The human study uses separately sampled, blinded four-system groups (Section 4.3). Appendix B records the benchmark cases and aggregation rules for each experiment.

**Post-generation assessment.** ShotEval-MLLM uses Doubao Seed 2.0 Pro with script-derived checks for fidelity, artifacts, and physics. Frame sampling combines 16 overview, 48 motion, and eight peak frames, with eight-fps local inspection. Artifact and global-commonsense judgments are prompt-blind; script and rule checks use the reference requirements. Assessment is run separately from generation but shares the internal verifier's model family. We therefore complement it with blinded human ratings rather than treating either model score as an independent perceptual ground truth (Liu & Zhang, 2025). Appendix B specifies metric denominators and exclusions.

## 4.2 Metrics and Automated Results

**SF** is script fidelity: passed atomic claims divided by all passed, failed, and uncertain claims. **Phys** is the 1–5 global physical-commonsense score; the ViMax-Bench aggregate conditions on physics-evaluable shots. **Severe%** is the fraction of evaluated shots with at least one confirmed artifact of severity 3, denoting an obvious generation failure. **Joint** counts physics-evaluable shots scoring at least 4, without an additional SF threshold. Its *denominator differs*: physics-evaluable shots on

ViMax-Bench, but all evaluated shots on ST-Bench-10. Joint values must therefore be compared within each benchmark block, since the two denominators measure different proportions.

Table 1: **Automated evaluation.** GIVER leads all four reported metrics on both benchmarks. Bold marks the best result. Results summarize available outputs from the adapted pipelines in Section 4.1; Joint has benchmark-specific denominators and is compared within each benchmark block.

| Benchmark | System | SF ↑ | Phys ↑ | Joint ↑ | Severe% ↓ |
|---|---|---|---|---|---|
| ViMax-Bench | ViMax | .769 | 3.760 | .512 | 3.50 |
| | MovieAgent | .753 | 3.790 | .531 | 4.20 |
| | AniMaker | .768 | 3.650 | .515 | 5.70 |
| | GIVER | **.780** | **4.010** | **.627** | **0.70** |
| ST-Bench-10 | ViMax | .687 | 3.944 | .278 | 9.70 |
| | MovieAgent | .700 | 4.278 | .361 | 5.60 |
| | AniMaker | .738 | 4.097 | .306 | 6.90 |
| | GIVER | **.791** | **4.361** | **.472** | **2.80** |

**Fewer artifacts without lower fidelity.** GIVER has the highest SF, Phys, and within-benchmark Joint, and the lowest Severe% in both blocks of Table 1. On ViMax-Bench, its severe-artifact rate is 0.70%, compared with 3.50–5.70% for the adapted baselines; SF is .780 versus .753–.769. ST-Bench-10 shows the same direction: 2.80% severe artifacts and .791 SF, versus 5.60–9.70% and .687–.738. Relative to the strongest baseline on Severe%, these are reductions of 80% and 50%, respectively. Higher fidelity accompanies the artifact reduction, so the improvement is not explained by a lower aggregate rate of satisfying the requested content. ViMax-Bench averages at case level, whereas ST-Bench-10 pools evaluated shots; comparisons are made within each benchmark.

**Qualitative comparison.** Figure 3 compares four methods at matched timestamps on two ViMax-Bench tasks. Woodworking exposes geometric distortion during contact; the wine-counter task exposes unexplained temporal changes. Labels identify observed defects, not a global judgment on every frame. GIVER preserves a flat workpiece and shows hand-mediated tool handling in the displayed sequences. Both panels retain all four methods to make the comparison explicit. The examples also explain the need for two kinds of evidence: object shape can be inspected in a frame, whereas unexplained appearance requires comparing what was visible before and after. These examples retain the requested objects and actions while illustrating their visual integrity over time.

### 4.3 Blinded Human Evaluation

Raters assessed script fidelity on a 0–1 scale and physical plausibility on a 1–5 scale. Systems were concealed behind neutral codenames, and shot order and within-shot system order were randomized. For ViMax-Bench, 16 raters each assessed 12 sampled shot groups with four systems, producing 768 valid ratings. The combined sample contains 37 shots and 148 system–shot videos. For ST-Bench-10, 25 raters assessed 24 shot groups, each containing one video per system, for 96 videos across the four systems. Of 2,400 submitted ratings, 78 had at least one out-of-range score and were excluded, leaving 2,322 valid ratings. Table 2 averages valid ratings within each system. The human study uses a separately sampled set, distinct from the automated evaluation set.

Table 2: **Blinded human evaluation.** GIVER receives the highest mean SF and Phys on both samples. Scales are 0–1 and 1–5, respectively. Counts are valid rating records, not independent videos; bold marks the best mean for each metric within each benchmark's human sample.

| | ViMax-Bench | | | ST-Bench-10 | | |
|---|---|---|---|---|---|---|
| System | SF ↑ | Phys ↑ | Ratings | SF ↑ | Phys ↑ | Ratings |
| ViMax | .577 | 3.266 | 192 | .526 | 3.625 | 579 |
| MovieAgent | .562 | 3.000 | 192 | .543 | 3.412 | 578 |
| AniMaker | .563 | 3.141 | 192 | .570 | 3.292 | 582 |
| GIVER | **.629** | **3.599** | 192 | **.657** | **3.731** | 583 |




Figure 3: **Qualitative comparison on two ViMax-Bench examples.** Each row shows one pipeline, with four matched-time frames per example. **(a)** ViMax develops a warped workpiece and AniMaker shows implausible block geometry. **(b)** AniMaker's background figure fades without leaving the view; MovieAgent's corkscrew appears without visible placement. GIVER shows a flat workpiece and hand-mediated tool handling in the illustrated sequences. Boxes locate the observations, and insets enlarge the same frames. These are selected examples, not estimates of failure frequency.

**Viewers also rate content and plausibility higher.** GIVER has the highest mean in both dimensions on both samples. Relative to the best baseline mean for each dimension, its SF margin is .052 on ViMax-Bench and .087 on ST-Bench-10; the corresponding physics margins are .333 and .106. These ratings support the quality advantage without exposing raters to the system identity or internal acceptance score. Agreement in the best-ranked system does not imply agreement on every individual video. Appendix B.4 reports video-level correlations and the handling of repeated ratings; our conclusions concern the observed means, without a statistical-significance claim.

### 4.4 Ablation: Video-Stage Verification

We study video-stage control in a ten-second generation experiment across all 35 ViMax-Bench themes. GIVER and both flat-verification variants use the baseline-generated first frame, with no new image-generation step. The comparison includes generation without SOIG verification, two event-list configurations (`abl1` and `abl2`), and GIVER's graph-based video verification and repair. Both flat variants retain the baseline frame; their run identifiers do not denote different first-frame strategies in this experiment. Table 3 includes 128 Type C GIVER clips from the separately implemented Type C pipeline. Appendix A.3 records the source of these Type C results, the retained evaluation samples, and the generation and verification protocols used by each configuration.

Table 3: **Video-stage verification with fixed first-frame conditioning.** GIVER improves all four quality metrics with fewer video operations than either flat configuration. Video Try sums generation and editing per selected shot. All settings cover 35 themes at ten seconds per clip; valid-report counts and Type C provenance are given in Appendix A.3, along with run-specific protocols.

| Metric | Generation only | Flat (`abl1`) | Flat (`abl2`) | GIVER |
|---|---|---|---|---|
| SF ↑ | .750 | .759 | .755 | **.766** |
| Physics ↑ | 3.790 | 3.757 | 3.811 | **3.932** |
| Joint ↑ | .566 | .552 | .560 | **.596** |
| Severe% ↓ | 5.5 | 4.8 | 5.5 | **3.6** |
| Video Gen / Shot | 1.000 | 1.745 | 1.736 | 1.574 |
| Video Edit / Shot | — | 1.111 | 1.174 | .609 |
| Video Try / Shot | 1.00 | 2.86 | 2.91 | 2.18 |

**Quality and effort measures.** SF and artifact assessment use the full script; the physics contract excludes camera motion. SF and Physics pool shot scores. Joint is the fraction of physics-evaluable shots scoring at least 4, and Severe% averages theme-level rates. Appendix A.3 reports retained evaluation counts and run-specific verification protocols for each video-stage configuration.

**Better quality with fewer video operations.** GIVER achieves .766 SF, 3.932 Physics, and .596 Joint, with a 3.6% severe-artifact rate. Both generation and editing decrease relative to flat verification, totaling 2.18 rather than 2.86 or 2.91 operations per selected shot: approximately 24–25% fewer operations. Generation only uses one operation but has lower quality. The shared first frame rules out better image synthesis as the source of these gains. This compares complete video-stage configurations, not individual graph edges. Appendix B.5 distinguishes video-operation counts from total costs including verification, prompt refinement, and other upstream model calls.

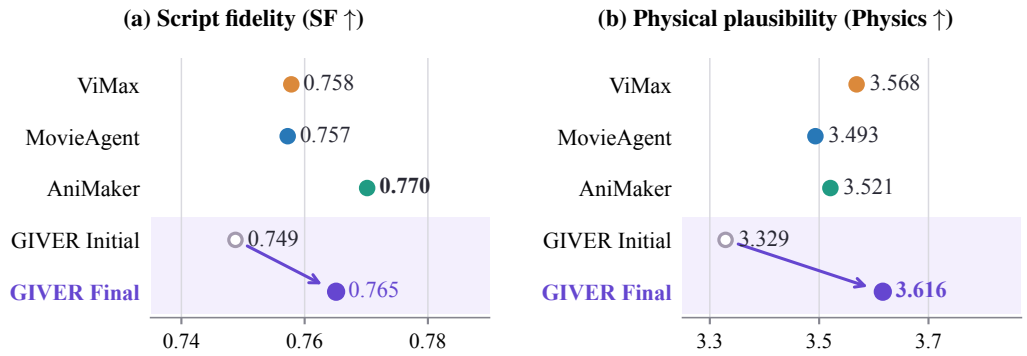


Figure 4: **Initial and final outputs on retried shots.** Means over 146 matched fifteen-second shots from 33 ViMax-Bench cases. Arrows connect the initial and final GIVER outputs. Baselines use the same shots. Historical runs were selected by case-level Physics (Appendix B.6).

**Initial-to-final quality on retried shots.** We separately examine fifteen-second shots with multiple video attempts. Across 177 paired shots from 34 cases, mean Physics rises from 3.373 to 3.678 and SF from .750 to .762. On the 146 shots also scored for all baselines (Figure 4), Physics rises from 3.329 to 3.616 and SF from .749 to .765. Final Physics is the highest mean on this subset; AniMaker has the highest SF. Each pair is evaluated against the original benchmark script, so both outputs are assessed for the same requested content. These gains describe selected historical runs, not uniformly beneficial retries; Appendix B.6 details run selection and paired score changes.

## 5 Discussion and Limitations

The results support a practical design choice: retain the task requirements and use specific evidence to guide repair. The graph makes decisions inspectable, but perception still depends on the verifier's temporal coverage, an issue also observed in video-understanding evaluation (Nagrani et al., 2025; Fu et al., 2025). Keeping the first frame fixed saves repeated image generation but prevents L2 from revisiting an error missed by L1. Cross-shot requirements are a natural extension beyond the present shot-level contracts. Our experiments establish gains in measured clip quality and video-operation counts. Assessing full-story continuity, total API expenditure including verification, and human inspection time would require measurements beyond the reported clip scores and request counts.

## 6 Conclusion

GIVER connects fixed, dependency-linked requirements to visual evidence and targeted repair for multi-shot video generation. Asset verification and repair precede separate first-frame and video loops; video repair retains the selected first frame. Automated and blinded human evaluations rank GIVER first among the compared pipeline adaptations in fidelity and physical plausibility. The fixed-first-frame experiment shows lower severe-artifact rates with about 24–25% fewer video generation and editing operations than two flat-verification configurations. Actionable verification identifies failed checks, guides repair with evidence, and rechecks candidates against fixed requirements.

### Ethics Statement

Long-video generation can enable deceptive synthetic media, unauthorized likeness use, and biased portrayals. Our checks address visible defects and requested content, not consent, authenticity, or comprehensive content safety. The evaluation uses generated media and reports human ratings in aggregate; individual rater identities are not included in the manuscript. Automated judges and human ratings can reflect cultural and perceptual biases. Generated examples should be identified as synthetic media and should not be presented as authentic recordings or evidence of real events.

### Reproducibility Statement

Appendices A to C document the control interface, model identifiers, run-specific protocols, metric denominators, human-study aggregation, and figure provenance. The video-stage experiment specifies its fixed conditioning, Type C source, and media-operation accounting separately from the main system comparison. These details support reproduction of the reported evaluation and interpretation of each experimental setting; exact media regeneration remains subject to proprietary API behavior.

### AI Use Statement

AI assistants supported literature retrieval and synthesis, methodological discussion, code assistance, manuscript drafting and editing, and figure-production assistance. Generative models also produce and evaluate the experimental media as described in the method and evaluation sections. The authors are responsible for the scientific claims, references, data interpretation, and final manuscript.

## References

Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, Chenyang Zhang, Tao Xiang, Fanny Yang, Serge Belongie, and Tian Xie. OneStory: Coherent multi-shot video generation with adaptive memory, 2025. URL https://arxiv.org/abs/2512.07802.

Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. VPO: Aligning text-to-video generation models with prompt optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15636–15645, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cheng_VPO_Aligning_Text-to-Video_Generation_Models_with_Prompt_Optimization_ICCV_2025_paper.html.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/446905abfec9db8856f7f7465bbd5552-Abstract-Conference.html.

Meng Chu, Senqiao Yang, Haoxuan Che, Suiyun Zhang, Xichen Zhang, Shaozuo Yu, Haokun Gui, Zhefan Rao, Dandan Tu, Rui Liu, and Jiaya Jia. VisionDirector: Vision-language guided closed-loop refinement for generative image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.19243.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Jing Gu, Ashwin Nagarajan, Tejas Polu, Kaizhi Zheng, Ruijian Zha, Jie Yang, and Xin Eric Wang. CCC: Enhancing video generation via structured MLLM feedback. In *2nd Workshop on Test-Time Adaptation: Putting Updates to the Test at ICML*, 2025. URL https://openreview.net/forum?id=B4eaJfAbCP. Workshop poster.

Kaisi Guan, Zhengfeng Lai, Yuchong Sun, Peng Zhang, Wei Liu, Kieran Liu, Meng Cao, and Ruihua Song. ETVA: Evaluation of text-to-video alignment via fine-grained question generation and answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2503.16867.

Ruozhen He, Meng Wei, Ziyan Yang, and Vicente Ordonez. EntityBench: Towards entity-consistent long-range multi-shot video generation, 2026. URL https://arxiv.org/abs/2605.15199. Preprint.

Xiaolu Hou, Bing Ma, Jiaxiang Cheng, Xuhua Ren, Kai Yu, Wenyue Li, Tianxiang Zheng, and Qinglin Lu. PersonaVlog: Personalized multimodal vlog generation with multi-agent collaboration and iterative self-correction, 2025. URL https://arxiv.org/abs/2508.13602. Preprint.

Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. StoryAgent: Customized storytelling video generation via multi-agent collaboration, 2024. URL https://arxiv.org/abs/2411.04925.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Hu_TIFA_Accurate_and_Interpretable_Text-to-Image_Faithfulness_Evaluation_with_Question_Answering_ICCV_2023_paper.html.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. FilMaster: Bridging cinematic principles and generative AI for automated film generation, 2025a. URL https://arxiv.org/abs/2506.18899.

Lingxuan Huang, Sizhe He, Hengji Zhou, Liqiang Nie, Lianghao Xia, and Chao Huang. ViMax: Agentic video generation, 2026. URL https://arxiv.org/abs/2606.07649.

Yidong Huang, Zun Wang, Han Lin, Dong-Ki Kim, Shayegan Omidshafiei, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Planning with sketch-guided verification for physics-aware video generation, 2025b. URL https://arxiv.org/abs/2511.17450. Preprint.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024. doi: 10.1162/tacl_a_00713. URL https://aclanthology.org/2024.tacl-1.78/.

Hyomin Kim, Junghye Kim, Joanie Hayoun Chung, Yoonjin Oh, Kyungjae Lee, Sungbin Lim, and Sungwoong Kim. Plan-and-verify video reward reasoning with spatio-temporal scene graph grounding, 2026. URL https://arxiv.org/abs/2606.11838. Preprint.

Daeun Lee, Jaehong Yoon, Jaemin Cho, and Mohit Bansal. Self-correcting text-to-video generation with misalignment detection and localized refinement. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 36464–36489, 2026. doi: 10.18653/v1/2026.findings-acl.1817. URL https://aclanthology.org/2026.findings-acl.1817/.

Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, and Min Zhang. Anim-Director: A large multimodal model powered agent for controllable animation video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024. URL https://arxiv.org/abs/2408.09787.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katherine M. Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liang_Rich_Human_Feedback_for_Text-to-Image_Generation_CVPR_2024_paper.html.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent multi-scene video generation via LLM-guided planning. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2309.15091.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-time scaling for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18671–18681, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_Video-T1_Test-time_Scaling_for_Video_Generation_ICCV_2025_paper.html.

Ming Liu and Wensheng Zhang. Is your video language model a reliable judge? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/dc4f891373d19087d1ddda33e81e00e4-Abstract-Conference.html.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_EvalCrafter_Benchmarking_and_Evaluating_Large_Video_Generation_Models_CVPR_2024_paper.html.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan O. Arik. VISTA: A test-time self-improving video generation agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6021–6032, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Long_VISTA_A_Test-Time_Self-Improving_Video_Generation_Agent_CVPR_2026_paper.html.

Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. VideoStudio: Generating consistent-content and multi-scene videos. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2401.01256.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Ryosuke Matsuda, Keito Kudo, Haruto Yoshida, Nobuyuki Shimizu, and Jun Suzuki. SLVMEval: Synthetic meta evaluation benchmark for text-to-long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Matsuda_SLVMEval_Synthetic_Meta_Evaluation_Benchmark_for_Text-to-Long_Video_Generation_CVPR_2026_paper.pdf.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, Yujun Shen, and Huamin Qu. HoloCine: Holistic generation of cinematic multi-shot long video narratives, 2025. URL https://arxiv.org/abs/2510.20822.

Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, Cordelia Schmid, and Tobias Weyand. MINERVA: Evaluating complex video reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23968–23978, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Nagrani_MINERVA_Evaluating_Complex_Video_Reasoning_ICCV_2025_paper.html.

Atin Pothiraj, Jaemin Cho, Yue Zhang, Elias Stengel-Eskin, and Mohit Bansal. Physics question scene graph: Fine-grained evaluation of physical plausibility in text-to-video generation. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.25306.

Candace Ross, Melissa Hall, Adriana Romero-Soriano, and Adina Williams. What makes a good metric? evaluating automatic metrics for text-to-image consistency. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=LFfktMPAci.

S. P. Sharan, Minkyu Choi, Sahil Shah, Harsh Goel, Mohammad Omama, and Sandeep Chinchali. Neuro-symbolic evaluation of text-to-video models using formal verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2411.16718.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. AniMaker: Multi-agent animated storytelling with MCTS-driven clip generation. In *SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025. doi: 10.1145/3757377.3764009. URL https://arxiv.org/abs/2506.10540.

Haoyuan Shi, Yunxin Li, Nanhao Deng, Zhenran Xu, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. MSVBench: Towards human-level evaluation of multi-shot video generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24034–24058, 2026. doi: 10.18653/v1/2026.findings-acl.1203. URL https://aclanthology.org/2026.findings-acl.1203/.

Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/dfd0bd56e8a6f82d1619f5d093d5f9ca-Abstract-Conference.html.

Yale Song, Yiwen Song, Nick Losier, Nathan Hodson, Ye Jin, Rhyard Zhu, Yan Xu, Daniel Vlasic, Carina Claassen, Jasmine Leon, Khanh G. LeViet, Zack Chomyn, Joe Timmons, Brett Slatkin, Scott Penberthy, and Tomas Pfister. Co-Director: Agentic generative video storytelling, 2026. URL https://arxiv.org/abs/2604.24842. Preprint.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_T2V-CompBench_A_Comprehensive_Benchmark_for_Compositional_Text-to-video_Generation_CVPR_2025_paper.html.

Aman Tyagi, Hemanth Boinpally, Jonathan Chen, Douglas Gebert, and Steven Hickson. Beyond trial-and-error: Agentic optimization for image-to-video adherence, 2026. URL https://arxiv.org/abs/2608.12290. Preprint.

Jiuniu Wang, Zehua Du, Yuyuan Zhao, Bo Yuan, Kexiang Wang, Jian Liang, Yaxi Zhao, Yihen Lu, Gengliang Li, Junlong Gao, Xin Tu, and Zhenyu Guo. AesopAgent: Agent-driven evolutionary system on story-to-video production, 2024. URL https://arxiv.org/abs/2403.07952.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul Debevec, and Ning Yu. MAViS: A multi-agent framework for long-sequence video storytelling, 2025. URL https://arxiv.org/abs/2508.08487.

Jianhui Wei, Jie Tan, Hengchuan Zhu, Xiaotian Zhang, Yan Zhang, Ziyi Chen, Daoan Zhang, Wei Xu, and Zuozhu Liu. VideoWeaver: Evaluating and evolving skills for agentic long video generation, 2026. URL https://arxiv.org/abs/2606.08091.

Zheng Wei, Mingchen Li, Zeqian Zhang, Ruibin Yuan, Pan Hui, Huamin Qu, James Evans, Maneesh Agrawala, and Anyi Rao. Hollywood town: Long-video generation via cross-modal multi-agent orchestration, 2025. URL https://arxiv.org/abs/2510.22431. Preprint.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent CoT planning, 2025. URL https://arxiv.org/abs/2503.07314.

Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain Cinema: Towards short movie generation, 2025. URL https://arxiv.org/abs/2507.18634.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawendé F. Bissyandé, and Saad Ezzini. DreamFactory: Pioneering multi-scene long video generation with a multi-agent framework, 2024. URL https://arxiv.org/abs/2408.11788.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. PhyT2V: LLM-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18826–18836, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xue_PhyT2V_LLM-Guided_Iterative_Self-Refinement_for_Physics-Grounded_Text-to-Video_Generation_CVPR_2025_paper.html.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. StoryMem: Multi-shot long video storytelling with memory, 2025a. URL https://arxiv.org/abs/2512.19539.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. STAGE: Storyboard-anchored generation for cinematic multi-shot narrative, 2025b. URL https://arxiv.org/abs/2512.12372.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. MovieDreamer: Hierarchical generation for coherent long visual sequences. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7d4647db960780c8b5a4e7d9e4a58d68-Abstract-Conference.html.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention, 2024. URL https://arxiv.org/abs/2412.02259.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, Xuanyang Zhang, Xianfang Zeng, Zhewei Huang, Gang Yu, and Chi Zhang. ViStoryBench: Comprehensive benchmark suite for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhuang_V

iStoryBench_Comprehensive_Benchmark_Suite_for_Story_Visualization_CVPR_2026_paper.html.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8806–8817, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhuang_Vlogger_Make_Your_Dream_A_Vlog_CVPR_2024_paper.html.

# A CONTROL AND IMPLEMENTATION DETAILS

## A.1 GENERATION PROCEDURE

The independently generated first-frame route proceeds as follows.

1. Validate the storyboard, generator interfaces, reference attachments, and stage budgets. Prepare reusable assets; apply identity-focused or static-image checks and bounded repair.
2. Compile separate first-frame and video contracts from the shot and asset metadata. Validate graph identities and logical prerequisites. Record which checks apply at runtime and keep policy adjustments separate from the fixed requirements and raw visual observations.
3. Generate a first-frame candidate, perform L1 requirement review and image-artifact screening, and retain the observations and the controller's recorded acceptance decision.
4. On a repairable non-pass, revise eligible image fields and generate or edit another candidate. Apply only configured, bounded same-image rechecks for unresolved evidence. Select a passing frame, or an explicitly permitted scored fallback at resource exhaustion.
5. Bind the frame; check formatting and repair application before video submission.
6. Verify the resulting clip, record node evidence and the acceptance decision, and revise video fields when a repair target and budget remain. The selected frame remains fixed.
7. Select the video under the recorded acceptance or exhaustion policy. Preserve any failed or unknown checks. Assemble selected clips in storyboard order and report which shots produced an output, separately from whether those outputs passed their quality checks.

## A.2 VERIFICATION, POLICY, AND CANDIDATE RECORDS

L1's second requirement-review turn reinspects the image with provisional observations; it is not merely an answer-formatting step. Image artifacts and portrait identity use their applicable checks. The current L2 conversation maintains a frozen node order and candidate-specific media context. Each answer identifies its node and references the video. Missing, duplicate, or mismatched identifiers invoke bounded response handling rather than automatic acceptance. Earlier validated answers remain in context and cannot be revised by a later response in the same candidate session.

Full-video attachment is not exhaustive observation of all frames. Perception depends on provider preprocessing and sampling, and prerequisite answers can bias later judgments. Artifact guards distinguish insufficient visibility and authored occlusion from confirmed unexplained disappearance. Node-level evidence is therefore a diagnostic record, not a formal proof of correctness.

The base graph, runtime applicability, raw answers, policy adjustments, and gate outcome are distinct objects. Critical failures or artifact/physics vetoes can prevent acceptance despite a high tree score. An unresolved or dependency-blocked node's zero score contribution is not itself a confirmed artifact. A policy waiver does not retrospectively change a raw observation to `yes`.

The refiner returns complete revised text for eligible fields, not an unconstrained replacement storyboard. Validation checks node references, field eligibility, usable text, and an effective revision. Formatting the revised request for a provider is a separate operation. Fields outside the update scope remain unchanged, but semantics inside a rewritten field and the resulting pixels may change.

Only valid media with a usable report can enter scored-candidate retention. A fallback records the original verdict, failed nodes, uncertainty, and termination reason. A parser failure, invalid media file, or missing repair target is not permission to relabel a candidate as passed. Accounting distinguishes physical generation, editing, verification, refinement, and failed requests; overlapping service calls must be accounted for separately rather than summed to estimate wall-clock latency.

## A.3 RUN CONFIGURATIONS AND PROTOCOL VERSIONS

The model identifiers are `doubao-seedream-5-0-260128` for images, `doubao-seedance-2-0-260128` for videos, and `doubao-seed-2-0-pro-260215` for generation/verification reasoning. Reported media settings are 720p, 16:9, with a shared stylized 3D ap-

pearance. The implementation's image-generation and video-submission limits are ten and five per shot, respectively; these limits are caps, not observed averages or evidence of equal budgets.

**Current interface and experimental protocols.** Section 3.3 describes the current one-node-per-turn L2 interface. The reported generation cohorts also include earlier multi-turn and consensus-review implementations with a video-edit backend. They share the node-linked verification-to-repair design; conversation and editing protocols are run-specific. The ten-image/five-video implementation caps are not substituted for historical run limits when reproducing those results.

**Fixed-first-frame video-stage experiment.** Table 3 uses the ten-second control experiment rather than the default full-pipeline end-to-end ablation. Its protocol supplies the baseline-generated first frame to `abl1`, `abl2`, and GIVER, with no additional first-frame synthesis. The baseline uses its own keyframe selector and no SOIG video verification. Both flat configurations use event-list video verification and refinement, while GIVER uses graph-based video verification and editing. The two flat run identifiers do not denote different image-verification policies in this experiment.

For context, in the *default E2E configuration*, `abl1` instead reuses a graph-verified GIVER frame, and `abl2` regenerates the frame using a flat event list. Those image-stage interventions are disabled in the control experiment reported here. Its results therefore examine video-stage verification and repair with a shared initial frame, rather than the effect of graph verification at the image stage.

**Type C source and evaluation coverage.** The final GIVER aggregate uses a separate Type C implementation for the 11 multi-scene cases, contributing 128 evaluated clips. Across all 35 themes, the valid report counts are 410, 419, 429, and 425 for baseline, `abl1`, `abl2`, and GIVER, respectively. Counts reflect generated outputs and report availability, including a small number of JSON failures. All conditions cover the same themes, but retained shot records are not a one-to-one matched set. The Type C source is therefore identified separately for reproducibility.

The control evaluator uses four-case, eight-shot concurrency and excludes camera-motion requirements from the physics contract. SF retains the full reference script; artifact assessment is prompt-blind and receives neither the generation prompt nor the reference script. These settings and the ten-second duration are specific to Table 3, not the system comparison or the separate verification-only tool, which re-scores existing media without running generation or repair.

# B Evaluation Protocols and Aggregation

## B.1 Benchmark Cases and Baseline Configurations

ViMax-Bench has 12 Type A, 12 Type B, and 11 Type C themes. The system comparison uses the reported case-level aggregates; the video-stage experiment spans all 35 themes with its own retained output set. The main-result summary does not enumerate a per-system case denominator, so the benchmark size is not substituted for that denominator. The system comparison specifies a 15-second ViMax baseline, whereas the video-stage experiment uses ten-second clips. The latter is a separate generation experiment, not a duration-normalized reaggregation of the main table.

Copyright restrictions limit our evaluation of ST-Bench (Zhang et al., 2025a) to ten cases: `artist`, `black`, `cat`, `dog`, `japan`, `phd`, `daiyu`, `jane`, `qin`, and `robinson`. We refer to this subset as ST-Bench-10. ViMax, MovieAgent, AniMaker, and GIVER are run on these same ten cases with five-second clips. The MLLM evaluator scores their generated outputs, with results reported in Table 1.

ViMax, MovieAgent, and AniMaker are instantiated with a common nominal image/video backend. AniMaker retains its original full MCTS-based video search, including contextual candidate evaluation and search over video candidates. The main comparison evaluates the resulting pipelines; sharing a rendering backend does not impose equal generation budgets.

## B.2 Post-Generation Evaluator

The evaluator is `doubao-seed-2-0-pro-260215` using the `prompts_aigc_gaffe` rubric. The script-fidelity route uses the reference script to compile atomic requirements, record observa-

tions, and verify them. Artifact assessment is prompt-blind: neither the coarse scan nor local detail verification receives the generation prompt or reference script. The physics route separates script-derived prerequisites and rules from global physical-commonsense judgments that receive only visual evidence and timestamps. Overview sampling merges 16 uniform frames with up to eight motion-peak frames selected from a 48-frame scan, after timestamp deduplication and chronological sorting. Detail verification samples local intervals at eight fps; the maximum image dimension is 1024 pixels.

Artifact severity is 0 for a non-artifact, 1 for a defect apparent only under close/framewise inspection, 2 for a defect noticeable during ordinary viewing, and 3 for an obvious generation failure. Confirmed events receive surface, structural, or temporal categories. Event counts depend on proposal and sampling policies; they do not exhaust every defect in a clip. The evaluator is invoked separately from the control judge, so the reported score is not copied from the generation loop. However, both use the same model family and may therefore share perceptual errors when interpreting the video.

Figure 5 summarizes the prompt-blind artifact route. Complete evaluation prompt templates appear in Appendix E.

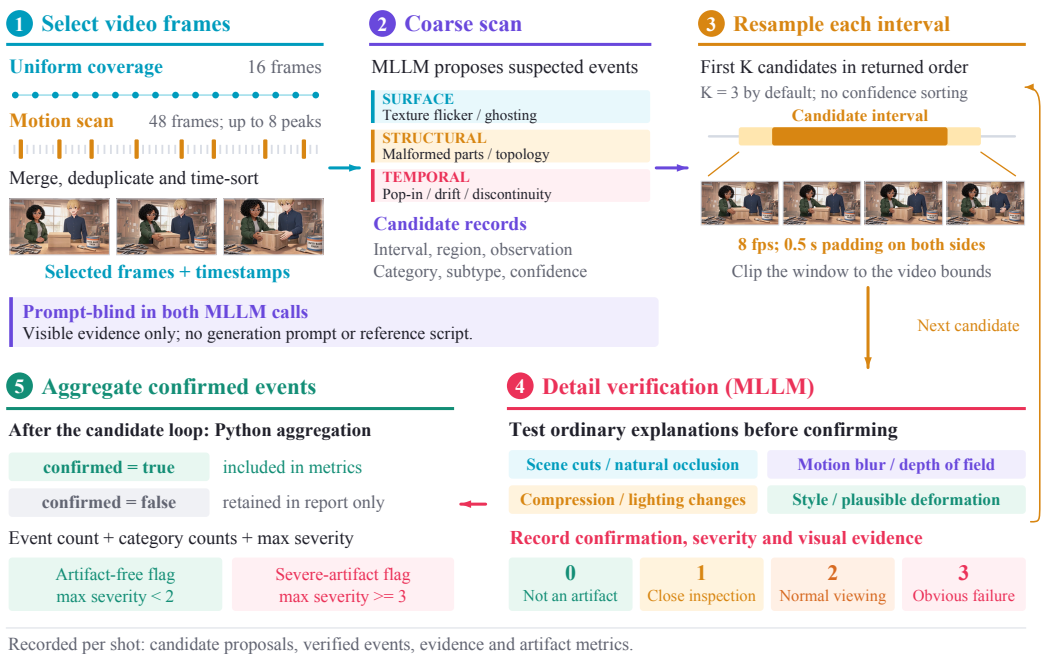


Figure 5: **Prompt-blind artifact assessment in ShotEval-MLLM.** Uniform and motion-peak frames are merged in temporal order for an MLLM coarse scan. The first $K$ proposed events are inspected using denser local frames with temporal padding. Neither MLLM call receives the generation prompt or reference script: the coarse scan uses video frames and timestamps, and detail verification additionally receives the suspected event. Detail verification tests ordinary explanations and returns confirmation, severity, and visual evidence. Python aggregates confirmed events into event counts, category counts, and severity-based shot flags; unconfirmed events remain in the report but do not contribute to these metrics. The displayed settings use 16 overview frames and up to eight motion peaks from a 48-frame scan, with the default detail budget $K = 3$, eight-fps local sampling, and 0.5-second padding on each side. Real frame strips illustrate sampling rather than a recorded evaluator verdict; only a subset of sampled frames is shown.

### B.3 Metric Definitions

The names below are those of the evaluation reports, not newly proposed metrics.

- **SF / Atomic Fidelity**: for a shot, PASS claims divided by PASS + FAIL + UNCERTAIN claims. Uncertain claims stay in the denominator and do not count as passes.

- **Phys / Physics**: a 1–5 global physical-commonsense score. ViMax-Bench's system comparison averages available scores on physics-evaluable shots. ST-Bench-10 and the video-stage experiment pool available shot scores. For the latter, the physics contract excludes camera-motion requirements while retaining the other physical requirements of the shot.
- **Joint, ViMax-Bench**: shots with global commonsense at least 4 divided by physics-evaluable shots; it does not additionally threshold SF. An evaluable shot without a commonsense score stays in the denominator but cannot contribute to the numerator.
- **Joint, ST-Bench-10**: the evaluator's `joint_semantic_physics_success` flags divided by all evaluated shots. The code sets this flag when the shot is physics-evaluable and its global commonsense is at least 4; no separate SF threshold is required. Nonevaluable or unscored shots contribute no success. Both remain in the all-shot denominator.
- **Joint, video-stage experiment**: the proportion of physics-evaluable shots with global commonsense at least 4, matching the ViMax-Bench main-table definition. This differs from some exports' all-shot success flag, whose denominator includes nonevaluable shots.
- **Severe%**: the percentage of evaluated shots with at least one confirmed severity-3 event. ViMax-Bench reports the mean of case-level percentages; ST-Bench-10 pools evaluated shots. In Table 3, case-level percentages are macro-averaged over the 35 themes.
- **Video Gen / Shot and Video Edit / Shot**: counts of video generation and editing operations divided by selected output shots, called *promoted shots* in the run logs. This denominator counts selected outputs, rather than only shots with valid post-generation evaluation reports.
- **Video Try / Shot**: the sum of generation and editing operations per selected shot in Table 3. The single-generation baseline is 1.00; editing also counts toward the total.

ViMax-Bench's main comparison uses case-level aggregation; ST-Bench-10 pools evaluated shots. In Table 3, SF and Physics pool valid shot scores, Joint conditions on physics evaluability, and Severe% averages theme-level percentages. Values are compared among configurations within the same experimental setting, since these aggregation rules weight cases and shots differently.

### B.4 Human Assessment

Each displayed shot group contains four anonymized system videos and the corresponding requested content. Raters assign script fidelity and physical-plausibility scores on their original scales. Video order within shots and shot-group order are randomized; codenames conceal system identities. Figure 6 shows the ViMax-Bench rating interface, with system names added only for illustration.

ViMax-Bench has 16 raters, 12 groups per rater, and four system videos per group: 768 valid system–shot ratings. The combined sample contains 148 unique videos over 37 shots. For ST-Bench-10, 25 raters assess 24 shot groups, each containing one video per system ($24 \times 4 = 96$ unique videos across four systems), producing 2,400 submitted ratings. Seventy-eight records contain at least one score outside SF's [0,1] or physics' [1,5] range and are excluded. The remaining counts are 579 for ViMax, 578 for MovieAgent, 582 for AniMaker, and 583 for GIVER. Table 2 preserves the original system means, without rescaling physics to [0,1].

The reporting script computes the system means from valid rating records; equivalently, these are per-video human means weighted by the number of valid ratings. For correlation analysis, repeated ratings are first averaged within each system–shot video and then paired with that video's automated score. Original-scale SF/physics Pearson correlations are .022/.483 for ViMax-Bench and .487/−.026 for ST-Bench-10. Thus aggregate system rankings and video-level agreement describe different aspects of the evaluation. We report the observed system means without a statistical-significance claim, since repeated ratings of the same video are not independent samples.

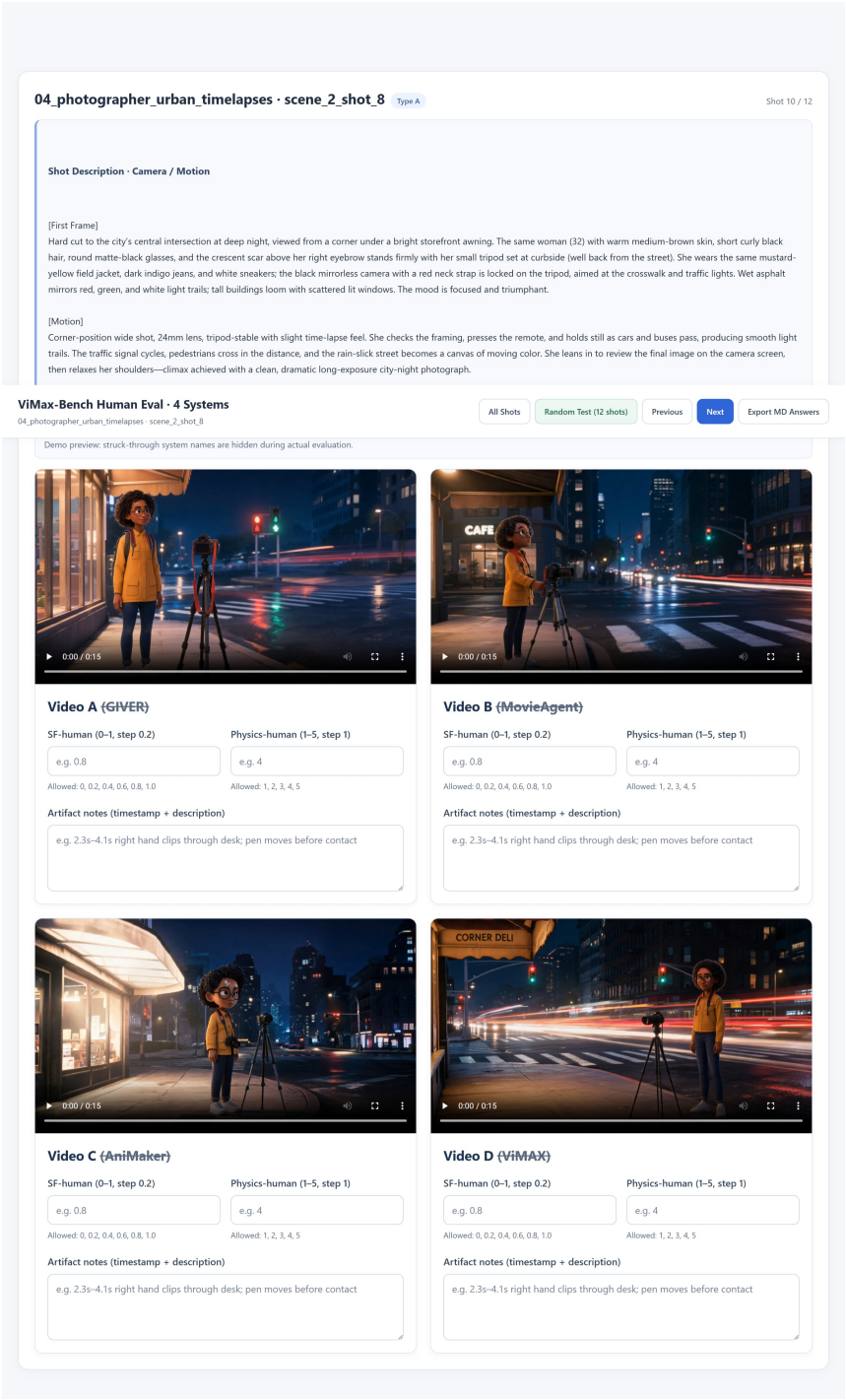


Figure 6: **Human-evaluation interface, shown with illustrative system labels.** Participants viewed the reference shot description and four anonymized videos, rated script fidelity (SF-human, 0–1 in increments of 0.2) and physical plausibility (Physics-human, 1–5 in increments of 1), and could record artifact timestamps and descriptions. The struck-through system names are annotations added solely for this illustration; these names were entirely hidden during the actual evaluation.

### B.5 Accounting Boundaries

System-comparison GIVER video-request means are 1.51 on ViMax-Bench and 1.88 on ST-Bench-10. ViMax and MovieAgent records generally stop after the first successful video request; configured transport retry limits do not mean three videos were generated.

The video-stage report separates Video Gen / Shot from Video Edit / Shot, then adds them as Video Try / Shot. Reported component sums are 1.000, 2.856, 2.910, and 2.183 for baseline, `abl1`, `abl2`, and GIVER. The last value is about 23.6% and 25.0% below the two flat configurations, respectively. Rounded totals in Table 3 are 1.00, 2.86, 2.91, and 2.18. The baseline edit entry follows the summary's dash notation; its total describes a single image-to-video operation.

Video generation and editing may have different prices, so Video Try measures media operations rather than a verification-inclusive monetary budget. Asset preparation, requirement compilation, judge calls, prompt refinement, failures, and caching contribute additional costs. Dollar expenditure, wall-clock latency, and human-inspection time are outside the reported measurements.

### B.6 Paired Analysis of Retried Shots

**Candidates and evaluation.** This analysis compares the first video candidate with the final selected output for shots with more than one recorded video attempt. The first candidate is `video_000/candidate.mp4`; the final output is the shot's selected `video.mp4`, not necessarily its last generated candidate. The original benchmark script is the reference for both outputs. The evaluation uses the AIGC artifact rubric and excludes camera-motion requirements from the physics contract. Only pairs with numeric SF and Physics scores for both outputs enter the aggregate. This yields 177 shots from 34 ViMax-Bench cases. The fifteen-second setting is separate from the ten-second experiment in Table 3; scores from these two experiments are not pooled.

**Historical run selection.** The source export selects historical runs using case-level Physics scores. If the selected source is a batch run, the analysis uses that run's candidate history. If it is a packaged output bundle, the resolver selects the highest-Physics batch with multi-attempt shots; its fallback selects the batch containing the most multi-attempt shots. Selection therefore uses post-generation evaluation, rather than a score-independent assignment of runs. We report the resulting paired changes descriptively: they do not identify the average causal effect of retrying over all runs, or separate feedback-guided repair from an equal-budget random-resampling alternative. This selection is distinct from the controller's within-run choice of its final candidate during video generation.

**Paired scores and baseline intersection.** Means give equal weight to each retained shot, not to each case. Across all 177 pairs, SF changes from .7496 to .7619 (a mean difference of .0123), while Physics changes from 3.3729 to 3.6780 (a mean difference of .3051). Physics improves on 70 shots, is unchanged on 67, and decreases on 40. The corresponding SF counts are 103, 2, and 72. The median Physics change is zero, so a positive average does not imply that most individual shots improve. Intermediate candidates are not evaluated in this comparison; it does not establish a monotonic improvement with attempt count or determine how many attempts a shot needs.

The baseline comparison matches the case and shot identifiers and retains only shots with all five outputs' SF and Physics scores. Its intersection contains 146 shots from 33 cases; all methods in Figure 4 use this same subset. SF means are .7578, .7572, and .7701 for ViMax, MovieAgent, and AniMaker, versus .7488 and .7651 for GIVER Initial and Final. Physics means are 3.5685, 3.4932, and 3.5205, versus 3.3288 and 3.6164. These baseline values contextualize the selected subset, rather than replace the full-benchmark system comparison. We make no statistical-significance or verification-inclusive cost claim from these paired scores on the matched shot subset.

## C Illustrative Media and Figure Provenance

Figure 1 uses real images and frames from a historical cooking-scene repair trace. The initial pair displays a count mismatch and its corrected realization. Video candidates are shown at matched timestamps, exposing unexplained food appearance in the earlier clip. Rectangles are annotations of the observed defects, not learned localization outputs or pixel-level masks used for repair.

Figure 2 uses a historical ViMax-Bench gardening trace. The expanded contract has category-grouping branches and thirteen logical prerequisite links; the former visualize organization, whereas the latter define dependencies. The recorded A1 action remains satisfied at 6.0 seconds. At 9.0 seconds, AIGC-CH1 identifies a missing garden trowel, prompting revision of the motion field while retaining the first frame. The selected subsequent clip records a pass and matches the saved evaluation video. The matched close-ups use the same crop window and timestamp in the source frames.

Both traces predate the one-node-per-turn protocol. The gardening example used multi-turn requirement review and two full-video continuity reviews. Frames are unretouched except for documented crops and overlays; Table 3 reports the separate quantitative video-stage experiment.

Figure 3 uses two ViMax-Bench tasks from the blinded comparison bundle: a mentor–student woodworking interaction and a wine-counter interaction. All four systems are shown at the same timestamps within each task. The woodworking frames are at 0, 6, 10, and 12 seconds; the wine-counter frames are at 0, 1.5, 11, and 14 seconds. The 32 main frames retain their full field of view. Five detail insets enlarge marked regions from the same frames; they do not introduce additional observations from different timestamps. Red marks observed defects, gray marks an earlier comparison state, and green marks the illustrated GIVER behavior. Frames, insets, and timestamps are not altered to simulate a repair. The cases were selected to illustrate geometric and temporal failure modes, not sampled to estimate their frequency. Unmarked frames do not establish that a clip is artifact-free.

## D Additional Context for Verification and Evaluation

**Where additional computation is spent.** The three mechanisms in Section 2 distinguish control decisions rather than agent names. Related single-clip optimization methods further clarify the scope of our efficiency experiment. Video-T1 searches generation trajectories and prunes branches (Liu et al., 2025), whereas SketchVerify evaluates inexpensive rendered motion sketches before full synthesis (Huang et al., 2025b). CCC supplies structured multimodal critique (Gu et al., 2025). VPO instead learns a prompt optimizer using text- and video-level feedback (Cheng et al., 2025); its training procedure is distinct from a test-time local repair loop. These approaches allocate computation to different objects: trajectories, plans, critiques, or learned prompts. Our reported operation count measures the generation and editing performed by the video-stage controller with a fixed first frame. It does not compare internal compute or establish an optimum over all search strategies.

**From specific feedback to a repair target.** Decomposed feedback has precedents in image generation. Divide, Evaluate, and Refine uses assertion-level VQA feedback to improve alignment (Singh & Zheng, 2023); Rich Human Feedback identifies problematic regions and misrepresented prompt terms (Liang et al., 2024). VisionDirector checks structured visual goals and incorporates staged edits and rollback (Chu et al., 2026). These methods motivate separating a diagnosis from a single aggregate score. In GIVER, the additional requirement is to keep the shot-level check identifiable across media candidates and connect it to an allowed text-field update. The graph is a record of requirements and prerequisites, not a guarantee that the visual judge is correct. The marked regions explain observed defects; they are not spatial masks passed to the generator for editing.

**Correction requires evidence.** Self-Refine demonstrates the usefulness of iterative language feedback (Madaan et al., 2023), while studies of self-correction show that improvements depend on the feedback available and the task (Huang et al., 2024a; Kamoi et al., 2024). We do not assume that repeating a judgment will make it correct. A new candidate must provide new visual evidence for the same requirement, and uncertain evidence stays uncertain unless a later check resolves it. Similarly, work on reward-model overoptimization shows why maximizing a proxy alone is not sufficient evidence of improved task performance (Gao et al., 2023). The external quality tables and blinded ratings are therefore separate from the diagnostic score used to accept a candidate.

**What a benchmark measures.** Broad video-quality suites such as VBench and EvalCrafter cover multiple perceptual and alignment dimensions (Huang et al., 2024b; Liu et al., 2024); T2V-CompBench emphasizes compositional requirements (Sun et al., 2025). Story-oriented benchmarks target another level of assessment: ViStoryBench evaluates coherent story visualization, MSVBench studies multi-shot video evaluation, and EntityBench focuses on entity consistency across longer

shot sequences (Zhuang et al., 2026; Shi et al., 2026; He et al., 2026). SLVMEval uses controlled synthetic perturbations to examine long-video evaluation metrics (Matsuda et al., 2026). Our ViMax-Bench and ST-Bench-10 experiments use multi-shot tasks, but the reported artifact and physics metrics evaluate individual clips. Improvements in these clip-level metrics do not, by themselves, establish better pacing or cross-shot continuity across the complete assembled story.

**System rankings and individual judgments.** Automatic evaluation is useful for coverage, but its behavior depends on the judge and the questions it is asked. Studies of text–image metrics and video-language judges examine these reliability issues directly (Ross et al., 2024; Liu & Zhang, 2025). MINERVA identifies temporal localization and visual perception as important sources of video-reasoning errors, while Video-MME evaluates understanding across video lengths and modalities (Nagrani et al., 2025; Fu et al., 2025). This motivates reporting both automated and human assessment and preserving their distinct units of analysis. The system means in Table 2 summarize viewer ratings; the correlations in Appendix B.4 ask whether two evaluators agree on individual system–shot videos. Agreement on the best system does not make these statistics equivalent.

# E PROMPTS FOR MLLM-BASED EVALUATION

This section provides the prompt templates used by ShotEval-MLLM to assess script fidelity, visible artifacts, and physical commonsense after generation. They are separate from GIVER's repair prompts. The evaluator uses `doubao-seed-2-0-pro-260215`. The system and user message bodies below preserve the original instructions, scoring rubrics, and output schemas. Braced placeholders denote runtime inputs; role headings and line wrapping are presentation only.

**Inputs and prompt visibility.** Overview images combine 16 uniformly sampled frames with up to eight motion peaks selected from a 48-frame scan, after timestamp deduplication and chronological sorting. Images are attached separately from the text. The placeholder {`frame_manifest`} lists their image indices and timestamps. Neither artifact call nor the global physical-commonsense call receives the generation prompt or reference script. Script-fidelity observation and physics-contract construction do use the reference requirements. Thus *prompt-blind* refers to withholding the generation prompt or reference script, not to withholding the evaluation instructions.

| Template | Function | Visual input |
|---|---|---|
| SF-1 | Compile the reference script into atomic claims | None |
| SF-2 | Observe evidence for one claim | Overview frames |
| SF-3 | Judge the claim from observations and dependencies | None; text evidence |
| AR-1 | Propose visible artifact intervals | Overview frames |
| AR-2 | Confirm an event and assign severity | Dense local frames |
| PH-1 | Compile applicability and physical prerequisites | None |
| PH-2 | Determine whether prerequisites visibly occur | Overview frames |
| PH-3 | Score global physical commonsense | Overview frames |

The templates cover the model outputs used by the reported metrics. Joint and Severe% are computed from those outputs, not elicited through additional rating prompts. Separate key-phenomenon, event-order, rule-grounding, and pairwise interfaces are not reproduced here because they do not enter these reported metric definitions.

## E.1 SCRIPT FIDELITY

SF-1 constructs the atomic claims from the reference script. SF-2 receives that script, one claim, and sampled images; SF-3 receives the observation record and prerequisite results, without an additional image attachment. Claims are processed after their prerequisite results are available. The shot-level SF score is the number of PASS claims divided by PASS, FAIL, and UNCERTAIN claims together. An unresolved claim is not counted as a pass.

**SF-1 input.** {`script`} is the reference shot script. The call is text-only; its output supplies the claims used by SF-2 and SF-3.

**SF-1: Script contract construction**

```
[System message]
You convert a generated-video shot script into atomic, visually verifiable
claims.
Return JSON only. Do not judge a video and do not add requirements that are
not stated or visually implied.
Each claim must contain exactly one fact. Psychological states must be
translated only when the script gives a visible manifestation.

[User message template]
Compile the following shot script into a semantic contract.

SHOT SCRIPT:
{script}

Allowed claim types:
entity, attribute, number, action, interaction, spatial, temporal_order,
state_change, end_state, camera.
```

**SF-1: Script contract construction (continued)**

```
Rules:
1. Every claim must be independently visually decidable.
2. Mark core=true only for facts essential to the shot's main content.
3. Add dependencies using claim IDs when a claim cannot be evaluated unless
another claim passes.
4. Do not include physical plausibility judgments here. Only encode what
should visibly occur.
5. Do not infer exact colors, counts, directions, camera motion, or object
states unless specified.
6. Use concise declarative English claims.

Return this schema:
{
  "shot_summary": "...",
  "claims": [
    {
      "id": "E1",
      "type": "entity",
      "claim": "A woman is visible.",
      "core": true,
      "dependencies": []
    }
  ]
}
```

**SF-2 inputs.** {script}, {claim}, and {claim_id} specify the reference and target claim. {frame_manifest} corresponds to the attached overview images. This call describes evidence rather than returning the final verdict.

**SF-2: Evidence observation**

```
[System message]
You are the observation stage of an evidence-grounded video-script evaluator.
You must describe only visible evidence relevant to one atomic claim. Do not
issue the final PASS/FAIL verdict.
The images are ordered video frames. Their timestamps are provided in the
prompt.
Never infer hidden events between sampled frames. Use UNCLEAR when evidence is
 insufficient.
Return JSON only.

[User message template]
SHOT SCRIPT:
{script}

ATOMIC CLAIM:
{claim}

FRAME TIMESTAMPS:
{frame_manifest}

Locate evidence relevant to the claim. Describe only what can be visibly
verified.
For actions and temporal claims, identify an evidence interval when possible.

Return:
{
  "claim_id": "{claim_id}",
  "observability": "CLEAR|PARTIAL|UNCLEAR",
  "observations": [
    {
      "start_time": 0.0,
      "end_time": 0.0,
      "description": "visible fact only"
```

**SF-2: Evidence observation (continued)**

```
    }
  ],
  "missing_evidence": "short explanation or null"
}
```

**SF-3 inputs.** The claim record supplies {claim}, {claim_type}, {core}, and {claim_id}. Prerequisite verdicts and evidence populate {dependency_results}; the SF-2 output supplies {observation}. This call receives text only, with no image attachment.

**SF-3: Claim verification**

```
[System message]
You are the verification stage of an evidence-grounded video-script evaluator.
Decide whether one atomic claim is satisfied using only the supplied
observations and claim dependencies.
No evidence means no PASS. Return JSON only.

[User message template]
ATOMIC CLAIM:
{claim}

CLAIM TYPE:
{claim_type}

CORE CLAIM:
{core}

DEPENDENCY RESULTS:
{dependency_results}

OBSERVATION STAGE OUTPUT:
{observation}

Decision policy:
- PASS: clear visible evidence directly satisfies the entire atomic claim.
- FAIL: clear evidence contradicts the claim, or a required event/entity is
absent despite adequate visibility.
- UNCERTAIN: sampling, occlusion, blur, ambiguity, or failed dependencies
prevent a reliable judgment.
- A temporal-order claim can PASS only when all required events are themselves
 observed and their order is clear.
- An attribute/action claim cannot PASS when its referenced entity dependency
failed.

Return:
{
  "claim_id": "{claim_id}",
  "verdict": "PASS|FAIL|UNCERTAIN",
  "confidence": 0.0,
  "evidence": [{"start_time": 0.0, "end_time": 0.0, "description": "..."}],
  "failure_reason": "wrong_entity|missing_entity|wrong_attribute|
missing_action|wrong_action|missing_interaction|wrong_spatial_relation|
wrong_event_order|wrong_state|camera_mismatch|insufficient_evidence|null"
}
```

### E.2 PROMPT-BLIND ARTIFACT ASSESSMENT

AR-1 proposes candidate intervals from the overview frames. AR-2 receives one candidate and dense images from its local interval, with timestamps but no reference script. The detail defaults inspect the first three returned candidates at eight fps, with 0.5-second padding on each side and the window clipped to the video bounds. Candidates are processed in returned order, not re-ranked by confidence. Only confirmed events enter artifact aggregation. A shot is severe when its maximum confirmed severity is at least 3; Severe% aggregates these shot flags according to the experiment's stated weighting. Both stages judge the visible clip without access to the reference script.

**AR-1 input.** The attached overview images and {frame_manifest} are the only shot-specific inputs. No script-derived requirements are supplied.

**AR-1: Artifact coarse scan**

```
[System message]
You detect visible generative artifacts in a video without access to its
prompt or script.
Judge only intrinsic visual/temporal defects. Do not penalize missing
requested content, unusual but possible events, intentional style, normal
motion blur, depth of field, compression, or camera cuts.
The images are ordered frames with timestamps. Return JSON only.

[User message template]
FRAME TIMESTAMPS:
{frame_manifest}

Scan for candidate artifact events in these categories:
- SURFACE: texture boiling, local flicker, ghosting, unexplained color/
exposure jumps, smeared or melting textures, repeated texture, contaminated
boundaries.
- STRUCTURAL: malformed face/hand/body/object, extra or missing parts, merging
/splitting, unstable topology, incorrect occlusion boundary, foreground-
background adhesion, reflection/shadow geometry mismatch.
- TEMPORAL: abrupt appearance/disappearance without cut or occlusion, identity
/shape drift, teleportation, discontinuous local motion, incompatible
foreground/background motion, non-rigid background warping during camera
motion.

Identify only candidate intervals worth detailed verification. Avoid
concluding from a single low-quality frame.

Return:
{
  "candidates": [
    {
      "category": "SURFACE|STRUCTURAL|TEMPORAL",
      "subtype": "short_snake_case_label",
      "start_time": 0.0,
      "end_time": 0.0,
      "region": "visible region or object",
      "observation": "what appears suspicious",
      "confidence": 0.0
    }
  ]
}
```

**AR-2 inputs.** One AR-1 event record supplies {candidate}. Dense local images are attached with their timestamps in {frame_manifest}. The output reports confirmation, severity, and the alternative explanations tested for that event.

**AR-2: Artifact detail verification**

```
[System message]
You verify one suspected generative artifact using densely sampled frames from
 a short interval.
You must actively test ordinary explanations before confirming an artifact.
Return JSON only.

[User message template]
SUSPECTED EVENT:
{candidate}

DENSE FRAME TIMESTAMPS:
{frame_manifest}

Determine whether the suspected event is a genuine visible artifact.
Explicitly test these alternatives when relevant:
scene cut, natural occlusion, motion blur, depth of field, compression,
lighting change, intentional stylization, physically plausible deformation.

Severity rubric:
0 = not an artifact.
1 = only visible under close frame-by-frame inspection.
2 = noticeable during normal viewing but does not dominate the shot.
3 = immediately obvious and strongly exposes the video as generated or broken.

Return:
{
  "confirmed": true,
  "category": "SURFACE|STRUCTURAL|TEMPORAL",
  "subtype": "...",
  "start_time": 0.0,
  "end_time": 0.0,
  "region": "...",
  "bbox_normalized": [0.0, 0.0, 1.0, 1.0],
  "severity": 0,
  "confidence": 0.0,
  "observation": "specific visible evidence across frames",
  "alternative_explanations": [
    {"name": "motion blur", "accepted": false, "reason": "..."}
  ]
}
```

### E.3 Physical Commonsense and Evaluability

PH-1 derives a physics contract from the reference script, including an applicability flag and semantic prerequisites. In experiments configured to exclude camera-motion requirements, the script is filtered before this call; the template itself is unchanged. PH-2 checks whether those prerequisites visibly occur. It does not score physical correctness. For an applicable contract, PH-3 independently scores visible physical commonsense from frames and timestamps only, even if PH-2 cannot establish evaluability. An inapplicable contract yields no physical-commonsense score.

Phys uses the numeric PH-3 score with the selection and aggregation rules specified for each experiment. Joint requires `physics_evaluable=true` and a numeric PH-3 score of at least 4; it has no additional SF threshold. Its denominator remains experiment-specific: physics-evaluable shots for the ViMax-Bench and fixed-first-frame reports, and all evaluated shots for ST-Bench-10. The auxiliary phenomenon, event-order, and rule fields in PH-1's original schema are preserved verbatim; they do not define an extra condition for these reported Joint scores.

**PH-1 input.** {script} is the reference script, or its camera-filtered form for the specified experiments. No generated frames are supplied to this text-only compilation call.

**PH-1: Physics contract construction**

```
[System message]
You convert a shot script into an observable physics contract for later video
evaluation.
Return JSON only. Focus on ordinary real-world physical behavior, not
mathematical derivations.
AIGC CONTEXT: The video that will be judged is AI-generated and may contain
gaffes (穿帮). Include ordinary body-integrity / impenetrability / object-
persistence rules so later judges can catch extra limbs, fusion, floating, and
 sudden replacements.

[User message template]
Build a physics contract for the following shot script.
The evaluated video is AIGC and may have 穿帮; keep body-integrity and
impenetrability rules even for simple actions.

SHOT SCRIPT:
{script}

Rules:
1. Set applicable=false when no meaningful motion, interaction, material
response, fluid behavior, optics, or gravity-related event is expected.
2. Semantic preconditions state what must visibly occur before physics can be
evaluated.
3. Expected key phenomena must be directly observable in video.
4. Expected event order must describe causal temporal order.
5. Applicable rules should be short neutral statements. Prefer categories:
gravity, inertia, momentum, collision, impenetrability, material_response,
fluid_dynamics, buoyancy, reflection, shadow, object_persistence,
kinematic_continuity.
6. Forbidden outcomes must be concrete visible failures.
7. Do not over-specify outcomes when multiple real-world outcomes would all be
 plausible.

Return this schema:
{
  "applicable": true,
  "semantic_preconditions": ["..."],
  "expected_key_phenomena": ["..."],
  "expected_event_order": ["event 1", "event 2"],
  "applicable_rules": [
    {"category": "impenetrability", "rule": "Objects should not pass through
each other during contact."}
  ],
  "forbidden_outcomes": ["..."]
}
```

**PH-2 inputs.** {preconditions} comes from PH-1. {frame_manifest} lists the attached overview images. This stage determines physics_evaluable, not the 1–5 quality score.

**PH-2: Semantic prerequisites**

```
[System message]
You evaluate whether the visible prerequisites for a physical event occurred
in a video.
This stage does not judge whether the physics is correct. Return JSON only.
AIGC CONTEXT: The clip is AI-generated video and may contain visible gaffes (
穿帮). Judge only pixels; do not invent off-screen body parts to satisfy a
precondition.

[User message template]
```

**PH-2: Semantic prerequisites (continued)**

```
SEMANTIC PRECONDITIONS:
{preconditions}

FRAME TIMESTAMPS:
{frame_manifest}

For each precondition, judge whether it is visibly satisfied.
Return:
{
  "items": [
    {
      "precondition": "...",
      "verdict": "PASS|FAIL|UNCERTAIN",
      "evidence_time": [0.0, 0.0],
      "observation": "..."
    }
  ],
  "physics_evaluable": true,
  "reason": "..."
}
```

**PH-3 input.** Only overview images and {frame_manifest} are supplied. Neither the script, the physics contract, nor the PH-2 verdict is included in this prompt.

**PH-3: Global physical commonsense**

```
[System message]
You are a prompt-blind evaluator of visible physical commonsense in a video.
Judge only the behavior shown in the frames. Do not assume intended events
from any script.
Return JSON only.
AIGC CONTEXT: The video is AI-generated and may have production gaffes (穿帮).
Extra limbs, interpenetration, floating objects, melting geometry, and
unexplained appearance/disappearance are physical violations, not style. Do
not give the benefit of the doubt for off-screen explanations.

[User message template]
FRAME TIMESTAMPS:
{frame_manifest}

This clip is AIGC and may contain 穿帮. Considering ordinary real-world
behavior, rate the physical plausibility and naturalness of the visible
process. Penalize visible AIGC gaffes.

Rubric:
1 = severe and obvious violations of everyday physics.
2 = multiple clear implausible behaviors.
3 = understandable process with one or more significant anomalies.
4 = largely plausible with only minor anomalies.
5 = natural and physically credible throughout the observable process.

Separately consider motion continuity, object persistence, contact/interaction
 response, gravity/material behavior when applicable, and causal timing.

Return:
{
  "score": 1,
  "confidence": 0.0,
  "observed_process": "neutral description",
  "violations": [
    {"type": "...", "time": [0.0, 0.0], "description": "..."}
  ]
}
```