

HALLELUAI: A Hallucination-Aware AI System for Ultra-Realistic Image-to-Video Generation at Scale

Aniket Sakpal* Yang Jiang† Rouzbeh Davoudi‡ Shayan Hassantabar§
 Mani Najmabadi¶

July 28, 2026

Abstract

AI-generated video is increasingly used across marketing, product storytelling, and creative workflows, yet automated; high-precision quality control remains a major constraint to scaling production. We present HALLELUAI, an end-to-end system that moderates and regenerates image-to-video outputs to meet expert-level creative standards and deliver ultra-realistic videos with consistent end-user quality of experience (QoE) at scale. The system integrates a video moderation module that evaluates frame-level aesthetics, temporal motion fidelity, and fine-grained hallucination risks relative to the source image, with an agentic regeneration module that iteratively fixes failures through prompt refinement, controlled camera adjustments, targeted model or image switching, and structured retry strategies. The moderation logic is aligned with domain-specific creative guidelines and produces granular, machine-actionable feedback that directly drives regeneration. In human-in-the-loop evaluations with creative experts, HALLELUAI shows strong alignment and reliably outputs ultra-realistic, production-grade videos suitable for product and marketing placements at scale. This framework advances trustworthy AI generated video content by enforcing visual realism, brand safety, and strict input-image fidelity while enabling image-to-video generation at scale.

Keywords: image-to-video generation, hallucination detection, generative AI, diffusion models

1 Introduction

1.1 Motivation and Context

AI-generated video (AIGV) has rapidly become central to content production across visually intensive platforms such as digital advertising, social media, and product detail pages. Advances in diffusion-based image-to-video

models now enable large-scale video synthesis at negligible marginal cost, supporting rapid experimentation, personalization, and localization. For domains such as travel, real estate, and e-commerce—where visual realism directly influences trust and conversion, AIGV offers a scalable alternative to traditional, resource-intensive production pipelines.

The economic incentives are substantial. Generative video systems compress production timelines from weeks to minutes and reduce per-asset costs by orders of magnitude relative to conventional workflows. These advantages position AIGV as a foundational technology for high-throughput visual storytelling. However, realizing this potential in production environments requires robust mechanisms to guarantee quality, realism, and strict fidelity to the source image.

1.2 Problem Statement – Quality Challenges in Image-to-Video Generation

Despite rapid progress, image-to-video (I2V) models still produce artifacts that degrade quality of experience (QoE). Common failures include frame-level degradations (e.g., blur, noise, and brightness drift), temporal inconsistencies such as jitter and unnatural motion, and content hallucinations that introduce or alter scene elements relative to the source image [19, 26]. Hallucinations are particularly problematic because the input image serves as the ground truth. Deviations from it can misrepresent reality, reduce user trust, and create legal risks in domains such as travel and real estate [26]. Current evaluation methods do not adequately address these challenges. Distributional metrics including Inception Score (IS) [2], Fréchet Inception Distance (FID) [6], and Fréchet Video Distance (FVD) [20] assess dataset-level realism but not fidelity to a specific input image. Likewise, existing video quality assessment methods often overlook AIGV-specific artifacts such as cross-frame inconsistencies and subtle hallucinations [4, 5, 12, 24]. As a result, there is no reliable automated mechanism for detecting and mitigating these failures at scale. Recent experience with systems such as OpenAI’s Sora [18] further highlights the need for robust quality assessment and moderation frameworks for

*Expedia Group, Email: aniket17sakpal@gmail.com

†Expedia Group, Email: jiang.yang.james@gmail.com

‡Expedia Group, Email: rouzbeh.davoudi@gmail.com

§Expedia Group, Email: s_hassantabar@yahoo.com

¶Expedia Group, Email: maninajmabadi@gmail.com

reliable I2V deployment [9, 15, 16].

1.3 Gaps in Prior Work

Despite recent advances in generative video evaluation, existing approaches remain insufficient for production-grade, image-conditioned AI-generated video.

Per-asset, conditional evaluation: Metrics such as IS [2], FID [6], and FVD [20] are designed for model-level comparison rather than asset-level acceptance. They do not condition on a specific input image and therefore fail to assess source-image fidelity, temporal stability, camera-motion quality, and hallucinations. FVD can also be unstable and may mis-rank models under limited sample sizes [20]. Recent work further highlights the challenge of evaluating semantic consistency and multi-level structure in AI-generated video [10].

Actionability: Frameworks such as FVMD [12], EvalCrafter [14], VBench [7], and Video-Bench [17] improve interpretability and alignment with human judgment, while VQAScore, GenAI-Bench, and VideoScore extend VQA-based evaluation to generative content [5, 11, 13, 23]. However, these methods remain primarily evaluative: they do not enforce source-image fidelity, gate individual assets, or provide diagnostics that directly support remediation.

Hallucination detection: Existing approaches focus on prompt consistency in text-to-video generation [3], diffusion-model hallucinations [1], or broader multimodal settings using uncertainty- and self-consistency-based techniques [8, 22]. Recent multimodal research likewise emphasizes the need for stronger grounding and semantic consistency [21, 25]. However, these methods do not condition on a source image or detect fine-grained temporal and structural deviations.

Closed-loop deployment: Current benchmarks assess model capability but do not support per-asset gating, domain-specific compliance, or automated remediation. Consequently, the literature lacks end-to-end frameworks that combine expert-aligned evaluation, source-image fidelity, and iterative regeneration for deployment in high-trust domains such as travel and real estate [21, 25].

1.4 Contributions

This work introduces a production-oriented moderation and regeneration framework for image-to-video generation that addresses the above gaps.

- **Per-asset, conditional video moderation:** a domain-aligned module evaluating each video relative to its input image across visual quality, temporal motion, and source-image fidelity.
- **Machine-actionable, expert-aligned diagnostics:** structured failure taxonomies, severity assessments, and pass/fail decisions mapping failure modes to corrective actions.

- **Fine-grained, temporally aware hallucination detection:** a layer identifying object-, structure-, and identity-level deviations, temporal inconsistencies, and artifacts relative to the source image.
- **Closed-loop, agentic regeneration:** a coupled module translating moderation feedback into targeted actions—prompt refinement, camera-control adjustment, model switching, base-image selection—iterating until quality criteria are met.
- **Human-in-the-loop validation:** an evaluation protocol demonstrating alignment with expert judgment and real-world deployability.

2 System Overview

We propose an agentic closed-loop image-to-video generation system that unifies automated moderation with autonomous planning and targeted regeneration to enforce expert-defined creative quality and strict source-image fidelity at scale (Fig. 1). Given an input image and creative guidelines (1), the system generates an initial video (2), which is evaluated by the Video Moderation Module (3) acting as both gatekeeper and diagnostics engine. The module assesses outputs across frame-level visual quality (3a), temporal motion quality (3b), and hallucination relative to the source image (3c), producing a structured moderation report (4) with dimension-wise scores, rationales, risk indicators, and a PASS/FAIL decision enforced by a decision gate (5).

For failed outputs, control is passed to the Agentic Regeneration Module (7), which converts feedback into autonomous actions. A planning agent (7a) maps failure categories and severities to targeted interventions, including prompt refinement for motion constraints (7b), camera direction or pacing adjustments (7c), base-image substitution for artifact-prone regions (7d), and alternative model selection for model-specific failures (7e). These actions drive targeted regeneration (8), after which outputs are re-evaluated by the moderation module.

The loop iterates until acceptance criteria are satisfied and final approval is reached (6), or until iteration, latency, or compute limits are exceeded (Fig. 1). By coupling fine-grained diagnostics with autonomous planning and remediation, the system enables systematic quality control and scalable deployment of image-to-video generation.

3 Video Moderation Module

The Video Moderation Module is the system’s primary quality gate and diagnostic engine, assessing image-to-video outputs before approval or regeneration. For each candidate video, it evaluates domain-aligned creative and fidelity constraints across three dimensions: frame-level

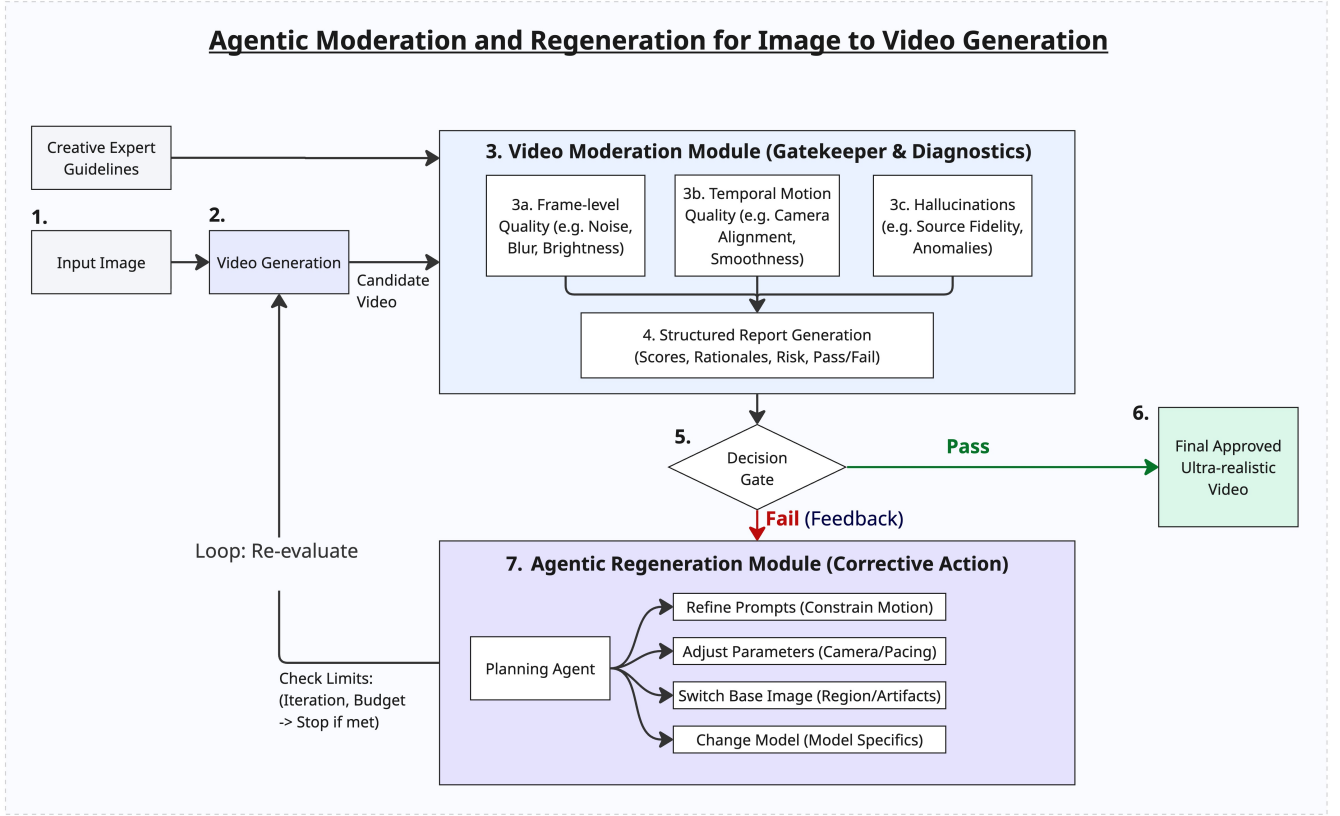


Figure 1: Architectural overview of the proposed closed-loop system for image-to-video generation. The framework integrates a diagnostic Video Moderation Module (3), calibrated by Creative Expert Guidelines, with an Agentic Regeneration Module (7). Candidate videos undergo iterative evaluation across frame-level, temporal, and fidelity dimensions (3a–c). Failures trigger a Planning Agent (7a) to execute targeted remediation strategies (7b–e) until the asset satisfies predefined quality thresholds or resource constraints are met.

visual quality, temporal motion quality, and hallucination detection relative to the input image. The evaluation yields a structured, machine-actionable report with dimension-wise scores, calibrated risk levels, localized rationales, and a unified PASS/FAIL decision—designed to directly parameterize downstream regeneration for targeted correction rather than unguided retries. The module is grounded in a hierarchical problem taxonomy (Fig. 1) categorizing common failure modes in ultra-realistic image-to-video generation, providing a shared abstraction layer that maps detected failures to corrective strategies such as prompt constraints, parameter adjustments, base-image substitution, or model switching.

3.1 Frame-Level Quality Signals: Blur, Contrast, Brightness, and Noise

We compute four complementary frame-level signals over a generated video $V = \{f_1, \dots, f_T\}$ and a reference input image I , with all frames and I converted to grayscale for statistics. Each signal is summarized via normalized ratios and compared against calibrated thresholds to produce a unified PASS/FAIL decision with a concise reason.

Blur (Sharpness Degradation). Blur is quantified via the variance of the Laplacian: for each frame f_t , we compute a sharpness score $\mathcal{L}_t = \text{Var}(\nabla^2 f_t)$, then measure the relative sharpness drop Δ_{blur} by comparing the anchor $\mathcal{L}_1$ to the minimum sharpness across the video.

Contrast (Low/High Contrast). Contrast is quantified via the standard deviation of grayscale intensities. With reference $C_{\text{ref}} = \sigma(I)$ and frame contrast $C_t = \sigma(f_t)$, we compute normalized decrease $\Delta_{\downarrow}$ and increase $\Delta_{\uparrow}$ relative to reference, optionally tracking within-video spread to capture excessive variability.

Brightness/Exposure (Low/High Exposure). Brightness is quantified via mean grayscale intensity. With $B_{\text{ref}} = \mu(I)$ and $B_t = \mu(f_t)$, we compute normalized exposure decrease $\Delta_{\downarrow}^B$ and increase $\Delta_{\uparrow}^B$ from the minimum and maximum brightness relative to the reference.

Noise (Stochastic Artifacts). Noise is estimated via a per-frame Laplacian-variance score N_t ; frames are marked noisy when N_t exceeds a threshold τ_{noise} , and

the video is flagged if the percentage of noisy frames exceeds a small tolerance (e.g., 5%), capturing persistent stochastic artifacts.

Algorithm 1 Unified Frame-Level Quality Detection (Blur, Contrast, Brightness, Noise)

Require: Reference image I , video $V = \{f_1, \dots, f_T\}$
Require: Thresholds $\tau_{\text{blur}}, \tau_{\downarrow}, \tau_{\uparrow}, \tau_{\downarrow}^B, \tau_{\uparrow}^B, \tau_{\text{noise}}, \tau_p$
Ensure: Decision $d \in \{\text{PASS}, \text{FAIL}\}$, reason r
 $d \leftarrow \text{PASS}, r \leftarrow \text{Acceptable Quality}$
 Convert I to grayscale once and compute $C_{\text{ref}} = \sigma(I)$,
 $B_{\text{ref}} = \mu(I)$

(1) Blur / Sharpness

Compute $\mathcal{L}_t = \text{Var}(\nabla^2 f_t)$ for all t
 $\Delta_{\text{blur}} \leftarrow (\mathcal{L}_1 - \min_t \mathcal{L}_t) / \mathcal{L}_1$
if $\Delta_{\text{blur}} > \tau_{\text{blur}}$ **then**
 $d \leftarrow \text{FAIL}; r \leftarrow \text{Sharpness Degradation}$
end if

(2) Contrast

Compute $C_t = \sigma(f_t)$ for all t
 $\Delta_{\downarrow} \leftarrow (C_{\text{ref}} - \min_t C_t) / C_{\text{ref}}$
 $\Delta_{\uparrow} \leftarrow (\max_t C_t - C_{\text{ref}}) / C_{\text{ref}}$
if $d = \text{PASS}$ **and** $\Delta_{\downarrow} > \tau_{\downarrow}$ **then**
 $d \leftarrow \text{FAIL}; r \leftarrow \text{Low Contrast}$
else if $d = \text{PASS}$ **and** $\Delta_{\uparrow} > \tau_{\uparrow}$ **then**
 $d \leftarrow \text{FAIL}; r \leftarrow \text{High Contrast}$
end if

(3) Brightness / Exposure

Compute $B_t = \mu(f_t)$ for all t
 $\Delta_{\downarrow}^B \leftarrow (B_{\text{ref}} - \min_t B_t) / B_{\text{ref}}$
 $\Delta_{\uparrow}^B \leftarrow (\max_t B_t - B_{\text{ref}}) / B_{\text{ref}}$
if $d = \text{PASS}$ **and** $\Delta_{\downarrow}^B > \tau_{\downarrow}^B$ **then**
 $d \leftarrow \text{FAIL}; r \leftarrow \text{Low Exposure}$
else if $d = \text{PASS}$ **and** $\Delta_{\uparrow}^B > \tau_{\uparrow}^B$ **then**
 $d \leftarrow \text{FAIL}; r \leftarrow \text{High Exposure}$
end if

(4) Noise

Compute per-frame noise score N_t (Laplacian variance)
 and mark noisy frames where $N_t > \tau_{\text{noise}}$
 $p_{\text{noise}} \leftarrow \frac{1}{T} \sum_{t=1}^T \mathbb{I}[N_t > \tau_{\text{noise}}]$
if $d = \text{PASS}$ **and** $p_{\text{noise}} > \tau_p$ **then**
 $d \leftarrow \text{FAIL}; r \leftarrow \text{Noise Artifacts}$
end if

3.2 Temporal Motion Quality

This component ensures camera motion is cinematic, intentional, and domain-appropriate. It evaluates (i) *prompt alignment*—whether the dominant motion direction/style matches the prompt intent—and (ii) *motion intensity and smoothness*—whether the pace is neither stagnant nor aggressive and free of jitter. We compute

dense optical flow using RAFT to obtain a per-pixel displacement field between consecutive frames, deriving robust frame-level motion statistics that classify the video as *Low Motion*, *High Motion*, or *Acceptable*, with additional jitter flags for high-frequency instability.

3.2.1 Unpleasant Camera Motion (Intensity & Smoothness) Module

We estimate motion intensity from the RAFT flow magnitude. At each step t , RAFT produces a dense displacement field $\mathbf{u}_t(x) \in \mathbb{R}^2$ mapping pixels from f_t to f_{t+1} . We compute the per-pixel magnitude $m_t(x) = \|\mathbf{u}_t(x)\|_2$ and summarize motion via a robust statistic: the mean magnitude of the top- q fraction of pixels (e.g., $q = 15\%$), $M_t = \text{Mean}(\text{Top}_q(\{m_t(x)\}))$, focusing on salient moving regions and reducing background-noise sensitivity. Video-level intensity is $\mu_M = \text{Mean}_t(M_t)$ and stability $\sigma_M = \text{Std}_t(M_t)$. The module flags *Low Motion* when μ_M falls below a threshold (stagnant/slow pans), *High Motion* when μ_M exceeds a threshold (whip-pan/aggressive moves), and *Jitter* when σ_M or the high-frequency variation of M_t exceeds a threshold (abrupt accelerations). These outcomes map directly to actionable remediation (reduce motion strength, tighten camera range, or stabilize).

3.2.2 Prompt Alignment Module

This module verifies that observed camera motion matches the prompt-specified intent (e.g., pan left-to-right, tilt up, dolly-in/zoom-in). We estimate camera motion using TAPIR point tracking: salient points initialized on the first frame are tracked through the video to obtain 2D trajectories. To reduce object-motion confounds, we preferentially sample edge points (high-gradient regions) and aggregate trajectories into group-level displacement statistics. Letting $p_i^t \in \mathbb{R}^2$ denote tracked point i at time t , we compute net displacement $\Delta p_i = p_i^T - p_i^1$ and summarize motion by the dominant direction $\hat{\mathbf{v}} = \text{Normalize}(\text{Median}_i(\Delta p_i))$. We further use structured point groups (e.g., left-edge set $\mathcal{P}_L$, top-edge set $\mathcal{P}_U$) to detect zoom/dolly: if $\mathcal{P}_L$ moves left while $\mathcal{P}_U$ moves up (divergence from the image center), the motion is consistent with zoom-in/dolly-in, whereas convergence indicates zoom-out/dolly-out. The observed label $\hat{y}$ is compared against the prompt intent y_{prompt} to yield an alignment score s_{align} , triggering a FAIL when below threshold.

Algorithm 2 Temporal Motion Quality: Intensity & Smoothness + Prompt Intent Alignment (TAPIR)

Require: Video $V = \{f_1, \dots, f_T\}$, prompt motion intent

y_{prompt}

Require: Top-pixel fraction q (e.g., 0.15)

Require: Thresholds $\tau_{\text{low}}, \tau_{\text{high}}, \tau_{\text{jit}}$, alignment threshold τ_{align}

Ensure: Decision $d_{\text{motion}} \in \{\text{PASS}, \text{FAIL}\}$ and reason

r_{motion}

$d_{\text{motion}} \leftarrow \text{PASS}, r_{\text{motion}} \leftarrow \text{Acceptable Motion}$

(A) Unpleasant Camera Motion: Intensity & Smoothness (RAFT)

for $t = 1$ to $T - 1$ **do**

$\mathbf{u}_t \leftarrow \text{RAFT}(f_t, f_{t+1})$

$m_t(x) \leftarrow \|\mathbf{u}_t(x)\|_2$ for all pixels x

$M_t \leftarrow \text{Mean}(\text{Top}_q(\{m_t(x)\}))$

end for

$\mu_M \leftarrow \text{Mean}_t(M_t), \sigma_M \leftarrow \text{Std}_t(M_t)$

if $\mu_M < \tau_{\text{low}}$ **then**

$d_{\text{motion}} \leftarrow \text{FAIL}; r_{\text{motion}} \leftarrow \text{Low Motion (Stagnant)}$

else if $\mu_M > \tau_{\text{high}}$ **then**

$d_{\text{motion}} \leftarrow \text{FAIL}; r_{\text{motion}} \leftarrow \text{High Motion (Aggressive)}$

else if $\sigma_M > \tau_{\text{jit}}$ **then**

$d_{\text{motion}} \leftarrow \text{FAIL}; r_{\text{motion}} \leftarrow \text{Jitter / Instability}$

end if

(B) Prompt Intent Alignment (TAPIR)

(B1) Point Initialization & Tracking:

Initialize an edge-biased point set $\mathcal{P}$ on the first frame f_1

Track points $\{p_i^t\}_{t=1}^T$ using TAPIR and compute net displacements

$\Delta p_i \leftarrow p_i^T - p_i^1$

(B2) Dominant Motion Estimation:

Estimate global camera translation

$\hat{\mathbf{v}} \leftarrow \text{Normalize}(\text{Median}_{i \in \mathcal{P}}(\Delta p_i))$

(B3) Zoom / Dolly Detection:

Partition points into spatial edge sets (e.g., left $\mathcal{P}_L$, top $\mathcal{P}_U$)

$\delta_L \leftarrow \text{Median}_{i \in \mathcal{P}_L}(\Delta p_i), \delta_U \leftarrow \text{Median}_{i \in \mathcal{P}_U}(\Delta p_i)$

(B4) Motion Intent Classification:

Infer observed motion label

$\hat{y} \leftarrow \text{ClassifyMotion}(\hat{\mathbf{v}}, \delta_L, \delta_U)$

(B5) Prompt Alignment Check:

Compute alignment score $s_{\text{align}} \leftarrow \mathbb{I}[\hat{y} = y_{\text{prompt}}]$

if $d_{\text{motion}} = \text{PASS}$ and $s_{\text{align}} < \tau_{\text{align}}$ **then**

$d_{\text{motion}} \leftarrow \text{FAIL}; r_{\text{motion}} \leftarrow \text{Prompt Misalignment}$

end if

3.3 Hallucinations

We evaluate two types of hallucination in image-to-video generation: *object hallucination* and *new-structure hallucination*. Object hallucination refers to failures in preserving temporal consistency, physically plausible motion,

and coherent object interactions across frames relative to the input image, leading to object drift, deformation, and other artifacts. *New-structure hallucination* is the synthesis of objects or scene elements absent from the source image—particularly important in high-trust domains (e.g., travel, real estate), where introducing non-existent structures may misrepresent reality and increase legal and reputational risk. The following section details the annotation categories for each type.

Hallucination Detection Prompt Template to VLLM:

The following are several frames in chronological order from an AI generated video, which may contain zooming, panning camera motion. First I will provide the definitions of hallucinations:

1. **Object Fusion:** Two or more objects unnaturally merge, exhibit boundary corruption, or interpenetrate.
2. **Implausible Transformation:** An object undergoes non-causal changes in shape, color, or identity that are not explained by interactions or scene dynamics.
3. **Object-Background Merge:** An object blends into the background despite being expected to remain visually distinct.
4. **Unnatural Object Movement:** Object motion violates physical constraints (e.g., unsupported motion, inconsistent trajectories) or lacks plausible causation.
5. **Implausible Disappearance:** An object abruptly vanishes without occlusion, exit from the field of view, or other reasonable explanation.
6. **Object Splitting:** A single object unnaturally divides into multiple distinct objects.
7. **Text Hallucination:** Textual elements (e.g., numbers, signs) change content, deform, or become illegible

Using the definitions from above, perform the following actions to identify hallucinations:

- **First**, identify Objects and Changes: Identify objects in the first frame and output a list. Remember to identify small objects, e.g. wine glass, vases etc.. Track the changes of the identified objects across frames and see how they change from the previous frame. Consider object identity across frames despite camera motion.
- **Second**, Categorize Changes as Hallucinations: If a change is detected, determine if it fits into any of the hallucination categories. Distinguish between camera motion effects and actual hallucinations. For example, if an object seems to disappear because the camera panned away, that's not 'implausible disappearance'.
- **Third**, Assign Intensity Scores: For each identified hallucination type, assign a score from 1 to 10 based on severity. The scoring guidelines:
 - 1-3: Minor, subtle, barely noticeable, could be ambiguous.
 - 4-6: Moderate, noticeable, clearly unnatural but not extremely jarring.
 - 7-9: Severe, very obvious, highly unnatural, breaks immersion significantly.
 - 10: Extreme, completely illogical, breaks the scene entirely.
- **Fourth**, Format Output as JSON: output the scores in the specified JSON format: {"Object Fusion": "integer", "Implausible transform": "integer", "Object Background Merge": "integer", "Object Movement": "integer", "Implausible Disappearance": "integer", "Object Split": "integer", "Unnatural Element Motion": "integer", "Text Hallucination": "integer"}. Even if no hallucinations are detected for a certain type, include it in the JSON with a score of 0.

Figure 2: Prompt template provided to the VLLM for hallucination detection. It defines seven hallucination categories and instructs the model to identify and track objects across frames, categorize changes as hallucinations (distinguishing them from camera motion), assign a 1–10 severity score per category, and output results as JSON (0 if none detected).

3.3.1 Hallucination Categories

1. Object Hallucination

- **Object Fusion:** Two or more objects unnaturally merge, exhibit boundary corruption, or

interpenetrate.

- **Implausible Transformation:** An object undergoes non-causal changes in shape, color, or identity unexplained by interactions or scene dynamics.
- **Object-Background Merge:** An object blends into the background despite being expected to remain visually distinct.
- **Unnatural Object Movement:** Object motion violates physical constraints (e.g., unsupported motion, inconsistent trajectories) or lacks plausible causation.
- **Implausible Disappearance:** An object abruptly vanishes without occlusion, exit from view, or other reasonable explanation.
- **Object Splitting:** A single object unnaturally divides into multiple distinct objects.
- **Text Hallucination:** Textual elements (e.g., numbers, signs) change content, deform, or become illegible.

2. New Structure Hallucination:

- **Natural entry:** Objects absent from the source image appear by entering the field of view with temporally continuous, physically plausible motion.
- **Camera-induced entry:** Previously unseen objects become visible due to global camera motion (e.g., panning, tilting, zooming) rather than object motion.
- **Abrupt emergence:** Objects not in the source image appear instantaneously without reasonable physical explanation.

3.3.2 Hallucination Detection Module

As illustrated in Fig. 3, we propose a hallucination-detection module for videos generated by an Image2Video model. Given a generated video, we uniformly sample frames $F = \{f_1, \dots, f_n\}$. To detect object hallucinations, we first extract objects $O = \{o_1, \dots, o_m\}$ from the input image using a multimodal large language model (MLLM), then provide O , the input image, and F to an MLLM (e.g., GPT-4o) to identify object-level inconsistencies across time, mapped to the predefined categories $C = \{c_1, \dots, c_m\}$. For each category c_i , the MLLM outputs a severity score $S_o^i \in [0, 10]$ and a brief rationale when evidence is present, following the prompt template shown in Fig. 2. The object-hallucination decision d_{object} aggregates the category-wise scores, thresholding each against τ_{object}^i . To detect new-structure hallucinations, we compare each sampled frame f_i with the input image to determine whether novel objects appear under three cases: (1) natural entry, (2) camera-induced entry,

and (3) abrupt emergence. An MLLM produces a per-frame binary decision d_{new}^i , and the video-level decision is $d_{\text{new}} = \bigcup_{i=1}^n d_{\text{new}}^i$.

3.4 Scoring and Risk Assessment

The module aggregates frame-level, temporal-consistency, and hallucination scores via domain-calibrated weights into a unified verdict, paired with a narrative summary contextualizing each issue against creative standards and Quality of Experience (QoE). A PASS/FAIL decision follows from thresholded expert-derived criteria, with a standardized risk level (*No*, *Low*, *Medium*, *High*) tied to specific failures. Outputs form a tuple of failure codes, severities, and remediation levers (e.g., reduce motion, adjust exposure, switch models), serving both as a gating mechanism and as input to the agentic regeneration module for targeted, iterative corrections.

4 Agentic Regeneration Module

The Agentic Regeneration Module converts moderation findings into targeted corrective actions via a closed-loop planner-executor process. A planning agent interprets frame-level, temporal, and hallucination signals relative to the source image, selects a minimal yet effective intervention set, and issues an updated generation request; the regenerated asset is re-scored, looping until quality thresholds or operational limits are met.

Inputs and Outputs. Inputs comprise the source image and metadata (e.g., campaign, channel), prior generation configuration (prompt, model, seed, motion parameters), and the structured moderation report (failure codes, severities, rationales, PASS/FAIL). Outputs comprise a next-generation plan (revised prompt, actions, parameters, chosen model or base image) and execution artifacts (regenerated video with a full action log for reproducibility).

Action Policy. The agent draws on a constrained intervention library: (i) prompt refinements enforcing motion clarity, object preservation, source-image fidelity, and aesthetic stability; (ii) camera/motion adjustments to reduce jitter, overshoot, or blur; (iii) base-image selection when artifacts cluster around occlusions or ambiguous regions; and (iv) model switching or tuning for model-specific failures. A cost- and efficacy-aware policy prioritizes low-cost, minimally invasive actions, escalates for high-severity or repeated failures, and bundles actions when interactions are likely.

Iteration and Governance. The loop (Generate $\rightarrow$ Moderate $\rightarrow$ Regenerate) terminates on PASS, budget exhaustion, or escalation to human review for persistent issues, logging all decisions for auditability. Coupling struc-

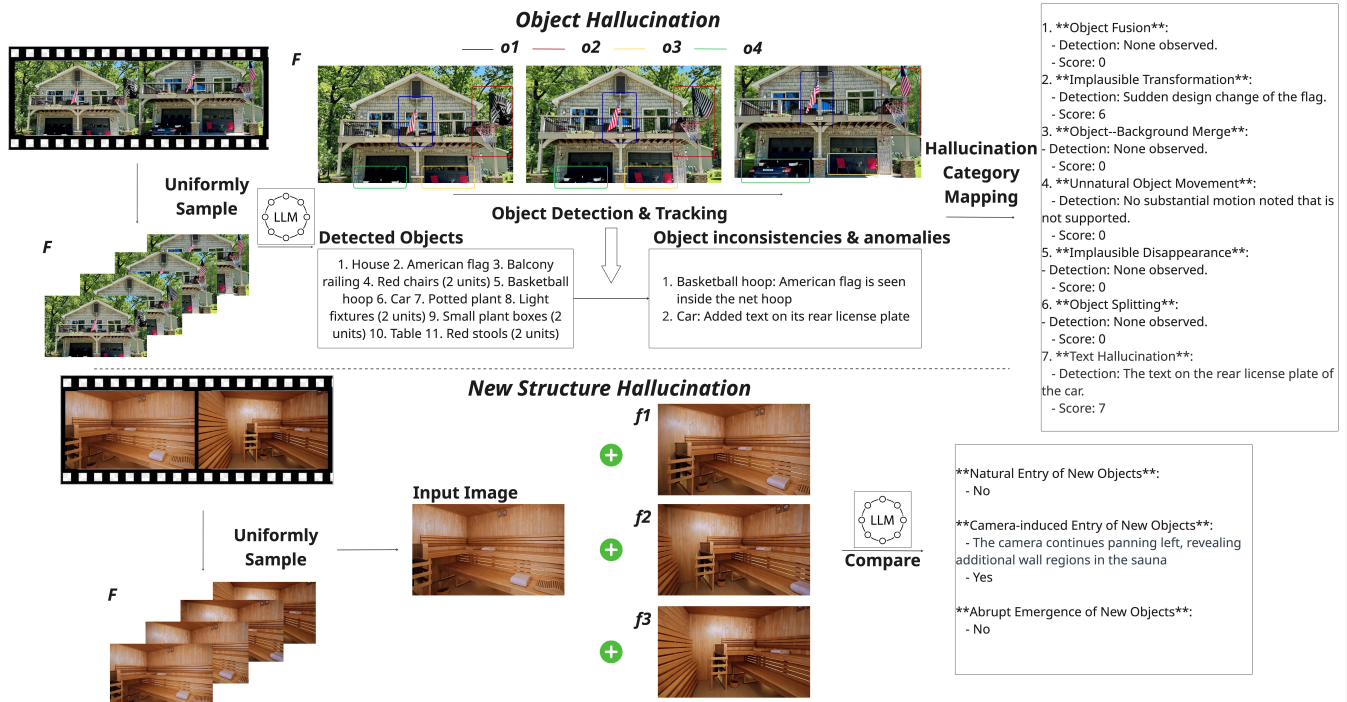


Figure 3: Hallucination detection module for Image2Video generation. Uniformly sampled frames are analyzed with an MLLM to (top) detect and track input-image objects, map temporal inconsistencies to predefined object-hallucination categories, and aggregate category scores for a final decision; and (bottom) compare each sampled frame with the input image to identify new-structure hallucinations (natural entry, camera-induced entry, or abrupt emergence)

tured diagnostics with a cost-aware strategy, the module converts failing candidates into compliant assets while preserving input-image fidelity and minimizing manual oversight.

Table 1: HALLELUAI moderation Performance Metrics Compared to Expert Decisions

Metric	Value
Overall PASS/FAIL Agreement	86.9%
Precision on AI-approved assets	88%
Recall on AI-approved assets	74%

5 Tuning and Evaluation

Threshold calibration and parameter tuning were performed using 58 AI-generated video clips across the three moderation components. For hallucination detection, we used GPT-4o as the underlying LLM. Prompts were iteratively refined to elicit accurate reasoning for each hallucination category. Recall, defined as the proportion of hallucination-containing videos correctly detected, was used as the primary metric for selecting the final prompts. Category-specific hallucination thresholds were then optimized to maximize sensitivity for detecting high-risk videos, as defined in Section 3.4. We then evaluate the proposed tuned system using a human-in-the-loop protocol designed to measure alignment with expert creative judgment and readiness for production deployment. The protocol consists of three stages: **Shadow Mode** (calibration), **Pseudo Production** (gating performance), and **Production QA** (longitudinal monitoring).

5.1 Stage 1: Shadow Mode

In Shadow Mode, unseen 146 AI-generated video clips were independently evaluated by creative expert reviewers, the HALLELUAI moderation system, and industry benchmark methods, without cross-visibility of decisions. Each clip was standardized through trimming and normalizing the aspect-ratio before evaluation.

5.1.1 Creative Expert Alignment

First, we evaluated the alignment of moderation systems by comparing their output to reviews from human creative experts on the video dataset. The HALLELUAI moderation system achieves 88% precision and an agreement rate of 87%. The results are summarized in Table 1

Agreement Analysis. Precision is the key metric, as it directly reflects the quality and reliability of videos approved by the system.

The system exhibits conservative gating behavior, filtering a portion of expert-approved clips while limiting false approvals.

Stage-Level Diagnostic Breakdown To analyze performance across moderation components, we decompose precision by category, as shown in Table 2.

The majority of precision-impacting errors arise from fine-grained object-level hallucinations. Motion intensity thresholds contribute primarily to false negatives, indicating conservative rejection behavior. Frame-level degradations contribute minimally to disagreement.

5.1.2 Benchmark Comparison

Since no open-source systems provide comprehensive evaluation for image-to-video generation, we compared our system against UGC video quality assessment (VQA) models such as DOVER [24] and COVER [4], as well as multimodal large language models (MLLMs) via zero-shot prompting (See Fig. 4 for prompt details). The MLLM prompt was designed to evaluate the three components our system addresses—frame-level quality, temporal motion quality, and hallucination—with frames uniformly sampled at 2 fps.

As shown in Table 3, UGC VQA models like DOVER and COVER could not differentiate PASS from FAIL videos, as they were not designed for image-to-video generation and lack hallucination-detection metrics. The MLLM-based method showed low precision, unable to detect artifacts in AI-generated videos—demonstrating that without detailed, comprehensive prompting, MLLMs cannot reliably identify such artifacts.

Prompt: Below is an image generated video using the first frame of the video. Please check the quality of the video based on the following criterias: 1. Frame-Level Quality Signals: Blur (Sharpness Degradation), Contrast (Low/High Contrast), Brightness (Low/High Exposure), and Noise (Stochastic Artifacts). 2. Temporal Motion Quality: Unpleasant Camera Motion (Intensity & Smoothness). 3. Hallucination: failure in preserving temporal consistency, physically plausible motion, and coherent object interactions across frames relative to the input image, leading to object drift, deformation, and other visual artifacts. Please return PASS/FAIL for each criteria. Please return an overall PASS/FAIL for the video

Figure 4: Evaluation Prompt Designed for MLLMs to Evaluate Frame-level Quality, Temporal Motion Quality and Hallucination in Image2Video Generation

5.2 Stage 2: Pseudo Production (Gating Performance)

In Pseudo Production mode, we evaluate the combined moderation + regeneration system end-to-end by sending only system-approved clips for human creative review. Of 1,158 system-approved clips, 1,126 were also accepted by experts, yielding 97% precision and indicating that the

vast majority of surfaced clips are production-ready. This precision is not directly comparable to Shadow Mode, since outputs here have undergone iterative regeneration and refinement before evaluation—the system is not merely filtering errors but actively improving clips, driving higher agreement with expert judgment.

5.3 Stage 3: Production QA (Longitudinal Monitoring)

Following calibration, ongoing monitoring employs random sampling (1–5%) of system-approved assets for expert review. Precision levels of 97% remain consistent with Pseudo Production Mode results, and regression testing ensures that threshold or regeneration-policy updates do not degrade previously validated performance. Deployed at scale, the system generated approximately 70,000+ videos, further validating its reliability.

5.4 Evaluation Summary

These results demonstrate that structured moderation combined with agentic regeneration can function as a reliable production-grade quality control layer for ultra-realistic image-conditioned video generation.

6 Conclusion

HALLELUAI is an expert-aligned, production-oriented system that makes image-to-video quality assurance a controllable, closed-loop process, coupling per-asset moderation—aesthetics, motion quality, and source-image-conditioned hallucination detection—with an agentic regeneration policy that converts failures into targeted fixes rather than blind retries. Human-in-the-loop validation shows strong agreement with creative experts, supporting deployment as a high-precision gate for scalable AIGV. By enforcing input-image fidelity, emitting machine-actionable diagnostics, and maintaining auditable decision trails, it bridges benchmark-style evaluation and real-world creative governance—to our knowledge, the first integrated moderation-and-regeneration framework purpose-built for ultra-realistic, image-conditioned video generation at scale.

References

- [1] Sumukh K Aithal, Pratyush Maini, Zachary C Lip-ton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *Advances in neural information processing systems*, 37:134614–134644, 2024.
- [2] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Table 2: Stage-level contribution to Precision and False Positives (FP). Percentages reflect the proportion of total observed errors attributable to each signal.

Module	Sub-Category	Precision
Frame-Level Quality	Blur	100%
	Contrast	97%
	Brightness	100%
	Noise	100%
Temporal Motion Quality	Prompt Alignment	100%
	Motion Intensity	97%
Hallucination Detection	Object Hallucination	97%
	New Structure Hallucination	97%
	Text Hallucination	100%

Table 3: Performance Comparison of Different Evaluation Methods

Method	Prec.	Rec.	F1	Acc.
DOVER	0.20	0.21	0.20	0.56
COVER	0.27	0.33	0.30	0.58
Qwen3-VL-8B	0.28	1.00	0.43	0.30
NOVA 2 Lite	0.25	0.74	0.37	0.34
HALLELUAI	0.88	0.74	0.81	0.87

- [3] Zhixuan Chu, Lei Zhang, Yichen Sun, Siqiao Xue, Zhibo Wang, Zhan Qin, and Kui Ren. Sora detector: A unified hallucination detection for large text-to-video models. *arXiv preprint arXiv:2405.04180*, 2024.
- [4] Chenlong He, Qi Zheng, Ruoxi Zhu, Xiaoyang Zeng, Yibo Fan, and Zhengzhong Tu. Cover: A comprehensive video quality evaluator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2024.
- [5] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyang Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, 2024.
- [6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.
- [7] Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.
- [8] Saurabh Kadavath et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- [9] Siqi Kou, Jiachun Jin, Zetong Zhou, Ye Ma, Yungang Wang, Quan Chen, Peng Jiang, Xiao Yang, Jun Zhu, Kai Yu, et al. Think-then-generate: Reasoning-aware text-to-image diffusion with llm encoders. *arXiv preprint arXiv:2601.10332*, 2026.
- [10] Jiaze Li, Haoran Xu, Shiding Zhu, Junwei He, and Haozhao Wang. Multilevel semantic-aware model for ai-generated video quality assessment. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.
- [11] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024.
- [12] Liu et al. Fréchet video motion distance: A metric for evaluating motion consistency in video generation. *arXiv preprint arXiv:2407.16124*, 2024.
- [13] Liu et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.
- [14] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejong 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*, pages 22139–22149, 2024.
- [15] Wenping Ma, Xiaoting Yang, Licheng Jiao, Lingling Li, Xu Liu, Fang Liu, Puhua Chen, Yuting Yang,

- Mengru Ma, Long Sun, et al. Video diffusion generation: comprehensive review and open problems. *Artificial Intelligence Review*, 58(11):338, 2025.
- [16] Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, Zeyu Wang, Zhifeng Li, Xiu Li, Wei Liu, Dan Xu, Linfeng Zhang, and Qifeng Chen. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.
- [17] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiayi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *Computational Visual Media*, 2025.
- [18] OpenAI. Sora: Openai text-to-video model, 2026. Product documentation and public reports, URL: <https://openai.com/research/sora>.
- [19] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhui Chen. Consist2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.
- [20] Thomas Unterthiner et al. Towards accurate generative models of video: A new metric & challenges. In *International Conference on Learning Representations Workshop*, 2018.
- [21] Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. Aigv-assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with lmm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18869–18880, 2025.
- [22] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- [23] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [24] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20144–20154, 2023.
- [25] Zhichao Zhang, Wei Sun, Li Xinyue, Jun Jia, Xiongkuo Min, Zicheng Zhang, Chunyi Li, Zijian Chen, Wang Puyi, Sun Fengyu, et al. Benchmarking multi-dimensional aigc video quality assessment: A dataset and unified model. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(9):1–24, 2025.
- [26] Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. *Advances in Neural Information Processing Systems*, 37:30300–30326, 2024.