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
Expedia Group

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.

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Image/video synthesis**.

Keywords

image-to-video, generative AI, diffusion models, content moderation, hallucination

ACM Reference Format:
Aniket Sakpal Yang Jiang Rouzbeh Davoudi Shayan Hassantabar Mani Najmabadi, Expedia Group . 2026. HALLELUAI: A Hallucination-Aware AI System for Ultra-Realistic Image-to-Video Generation at Scale. In *Proceedings of ACM Multimedia Conference (ACMMM '26).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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, current image-to-video models frequently produce artifacts that undermine end-user quality of experience (QoE). These failures fall into three broad categories. First, frame-level degradations such as blur, noise, and contrast or brightness drift reduce visual clarity and aesthetic consistency. Second, temporal failures arise from unstable or misaligned camera motion, including jitter, incorrect motion direction, or unnatural pacing that breaks temporal coherence. Third, and most critically, models hallucinate content relative to the source image, introducing new structures, deforming objects, or generating implausible motion [17] and [26]

Hallucinations are particularly harmful in image-conditioned video generation because the input image establishes the ground truth. Deviations—such as inventing unseen scenery, altering property features, or modifying identity-bearing elements—can misrepresent reality, erode user trust, and create legal exposure in high-trust domains like travel and real estate [26].

Existing automated evaluation methods are poorly aligned with these requirements. Distributional metrics such as Inception Score(IS) [1], Fréchet Inception Distance (FID) [5], and Fréchet Video Distance (FVD) [18] provide coarse dataset-level realism signals but are unsuitable for per-asset acceptance and do not enforce fidelity to a specific input image. General-purpose video quality assessment methods emphasize natural video aesthetics and overlook fine-grained, AIGV-specific artifacts, particularly cross-frame object inconsistencies and subtle hallucinations [3, 4, 11, 23]

In addition, model-centric benchmarks evaluate model capability rather than output compliance and do not provide machine-actionable diagnostics for production gating or remediation. As a result, there is no reliable, automated mechanism to detect these failures or correct them at scale.

These limitations are not merely theoretical. Recent experience with state-of-the-art systems, such as OpenAI's Sora [15], shows

ACMMM '26, Rio de Janeiro, Brazil
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

that high perceptual realism does not ensure controllability or compliance. Early deployments highlight risks including hallucinated content, deepfakes, and copyright violations [17, 26], which—along with high computational cost have led to cautious rollout strategies. This underscores that robust quality assessment and moderation are prerequisites for reliable deployment of I2V systems [10, 14, 16].

1.3 Gaps in Prior Work

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

G1. Lack of per-asset, conditional evaluation: Widely used metrics such as IS [1], FID [5], and FVD [18] operate at the distribution level and target model comparison rather than asset-level acceptance. Such approaches do not condition on a specific input image, fail to capture QoE-critical properties such as temporal stability, camera-motion appropriateness, and hallucinations. FVD, in particular, can mis-rank models and require large sample sizes for stability [18]. Recent works on AI-generated video quality assessment further highlight the difficulty of capturing semantic consistency and multi-level structure in generated videos [9].

G2. Descriptive but non-actionable multi-aspect evaluation: Multi-aspect and unary frameworks such as FVMD [11], EvalCrafter [22], VBench [6], and Video-Bench [2] decompose video quality into interpretable dimensions and improve alignment with human judgment, often via multimodal large language models (LLMs). Classical VQA metrics (PSNR, SSIM, VMAF) are similarly misaligned with AIGV failure modes [21], motivating VQA-based scorers such as VQAScore and GenAI-Bench [12, 13]. Recent approaches such as MantisScore explicitly attempt to simulate fine-grained human feedback using large-scale annotated datasets, improving correlation with human judgments but still operating primarily as evaluative scorers rather than actionable gating systems [4]. However, these approaches remain descriptive: they do not gate individual generated assets, enforce strict source-image fidelity, or produce diagnostics that directly drive remediation.

G3. Insufficient hallucination detection for image-to-video generation: Existing hallucination detection methods either focus on prompt consistency in text-to-video generation (e.g., SoraDetector [8]), analyze diffusion-model hallucinations at a trajectory level [24], or address hallucinations in broader multimodal contexts using uncertainty- or self-consistency-based techniques [7, 20]. Recent multimodal advances further emphasize hallucination challenges and the need for stronger grounding and semantic consistency in video generation systems [19, 25]. None operate conditionally on a specific source image, reason temporally across frames, or detect fine-grained object- and structure-level deviations.

G4. Absence of closed-loop, production-oriented systems: The literature lacks end-to-end frameworks that integrate expert-aligned evaluation, strict source-image fidelity, and automated regeneration into a closed loop. Existing benchmarks assess model capability, but do not support per-asset gating, domain-specific compliance, or iterative remediation required for scalable deployment in high-trust domains such as travel and real estate. Emerging research directions on unified multimodal evaluation and generation pipelines further highlight this gap but stop short of proposing full closed-loop production systems [19, 25]

1.4 Contributions

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

- Per-asset, conditional video moderation We introduce a domain-aligned moderation module that evaluates each generated video at the asset level, explicitly conditioned on its input image, across frame-level visual quality, temporal motion quality, and source-image fidelity.
- Machine-actionable, expert-aligned diagnostics The system emits structured failure taxonomies, severity assessments, and pass/fail decisions that move beyond descriptive scoring and map detected failure modes directly to corrective actions.
- Fine-grained, temporally aware hallucination detection We propose a hallucination detection layer designed specifically for image-to-video generation, capable of detecting object- and structure-level deviations, temporal inconsistencies, and identity-bearing visual artifacts relative to the source image.
- Closed-loop, agentic regeneration framework A tightly coupled regeneration module translates moderation feedback into targeted actions—such as prompt refinement, camera control adjustment, model switching, or base-image selection—and iterates until quality criteria are met or operational constraints are reached.
- Human-in-the-loop validation for production readiness. We present an evaluation protocol that demonstrates strong alignment with expert creative judgment and validates the system's suitability for real-world deployment.

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 (Figure 1). Starting from an input image and creative guidelines (1), the system produces an initial candidate video (2), which is evaluated by the Video Moderation Module (3) acting as both gatekeeper and diagnostics engine. This module assesses candidates along three complementary dimensions—frame-level visual quality (3a), temporal motion quality (3b), and hallucination detection relative to the source image (3c)—and emits a structured moderation report (4) with dimension-wise scores, rationales, standardized risk indicators, and a unified PASS/FAIL decision enforced by a decision gate (5).

When a candidate fails moderation, control is transferred to the Agentic Regeneration Module (7), which operationalizes feedback through autonomous decision-making. A planning agent (7a) interprets moderation signals and maps specific failure categories and severities to targeted corrective actions, including prompt refinement to constrain motion (7b), adjustment of camera direction or pacing parameters (7c), substitution of the base image in artifact-prone regions (7d), or selection of an alternative video generation model for model-specific failure modes (7e). These decisions drive

targeted video regeneration (8), after which regenerated candidates are re-submitted to moderation.

This agentic loop iterates until the video satisfies all acceptance criteria and receives final approval (6), or until predefined limits on iteration count, latency, or computational budget are reached (Figure 1). By coupling fine-grained, machine-actionable diagnostics with autonomous planning and remediation, the proposed architecture moves beyond unguided retries, enabling systematic improvement, scalable quality control, and reliable deployment of image-to-video generation systems.

3 Video Moderation Module

The Video Moderation Module functions as the system's primary quality gate and diagnostic engine, responsible for rigorously assessing image-to-video outputs before approval or regeneration. For each candidate video, the module evaluates compliance with domain-aligned creative and fidelity constraints across three complementary dimensions: frame-level visual quality, temporal motion quality, and hallucination detection relative to the input image. The evaluation produces a structured, machine-actionable report containing dimension-wise scores, calibrated risk levels, localized rationales, and a unified PASS/FAIL decision. Crucially, these outputs are designed to directly parameterize downstream regeneration actions, enabling targeted correction rather than unguided retries

To support fine-grained diagnosis and systematic remediation, the module is grounded in a hierarchical problem taxonomy (Figure X) that categorizes common failure modes observed in ultra-realistic image-to-video generation. This taxonomy provides a shared abstraction layer between moderation and regeneration, allowing detected failure types to be systematically mapped 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 $V = \{f_1, \ldots, f_T\}$ and a reference input image I. All frames and I are converted to grayscale when computing statistics. Each signal is summarized via normalized ratios (or percentages) and compared against calibrated thresholds to produce a unified PASS/FAIL decision with a concise reason.

Blur (Sharpness Degradation). Blur is quantified using the variance of the Laplacian operator. For each frame f_t, we compute a sharpness score $\mathcal{L}_t = \mathrm{Var}(\nabla^2 f_t)$. We then measure the relative sharpness drop Δ_{blur} with respect to the first frame f_1 (anchor), by comparing $\mathcal{L}_1$ to the minimum sharpness observed across the video.

Contrast (Low/High Contrast). Contrast is quantified using the standard deviation of grayscale intensities. Let $C_{\mathrm{ref}} = \sigma(I)$ be the reference contrast and $C_t = \sigma(f_t)$ be the frame contrast. We compute normalized contrast decrease $\Delta_\downarrow$ (relative drop vs. reference) and normalized contrast increase $\Delta_\uparrow$ (relative rise vs. reference). Optionally, we also track the within-video spread to capture excessive variability.

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

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

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

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

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

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

Brightness/Exposure (Low/High Exposure). Brightness is quantified using mean grayscale intensity. Let $B_{\mathrm{ref}} = \mu(I)$ and $B_t = \mu(f_t)$. We compute normalized exposure decrease $\Delta^B_\downarrow$ and increase $\Delta^B_\uparrow$ based on the minimum and maximum brightness observed in the video relative to the reference image.

Noise (Stochastic Artifacts). Noise is estimated using a Laplacian-variance-based score per frame (denoted N_t). Frames are marked as noisy when N_t exceeds a noise threshold τ_{noise}. The video is flagged if the percentage of noisy frames exceeds a small tolerance (e.g., 5%), capturing persistent stochastic artifacts.

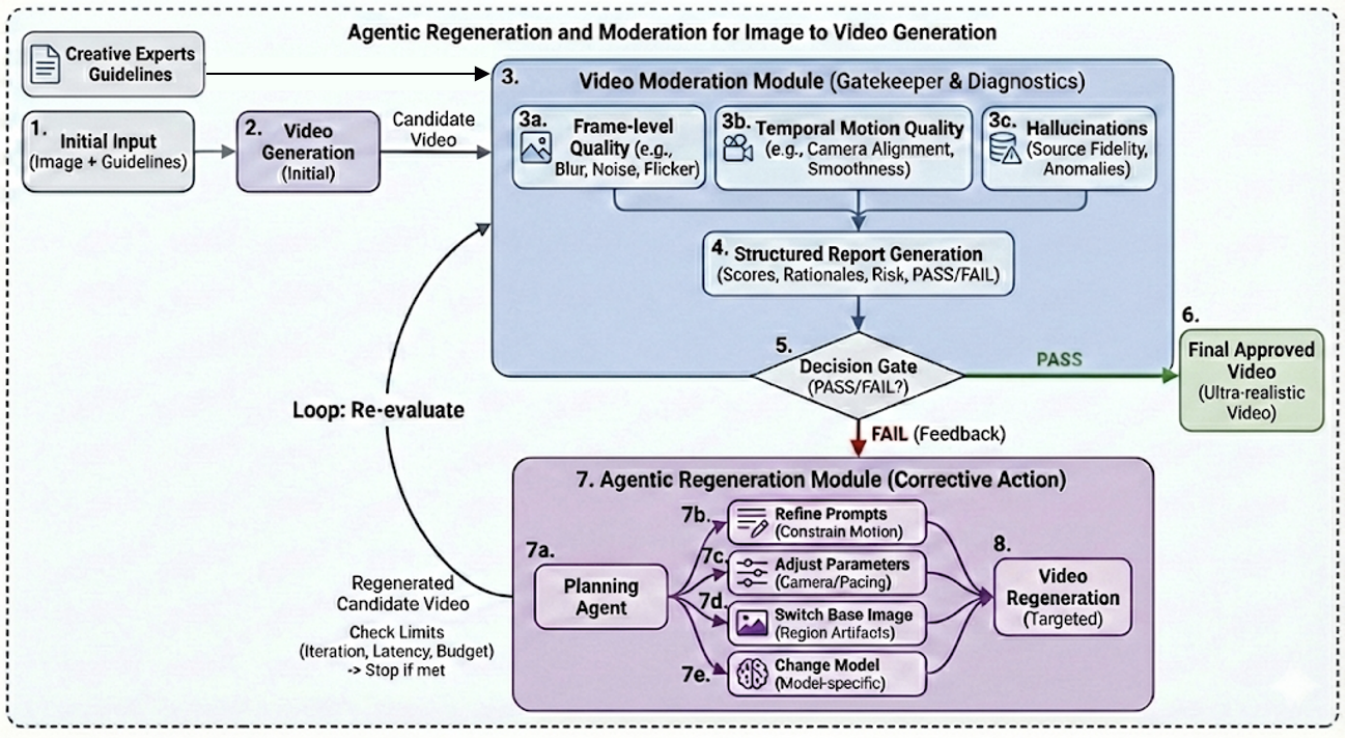

Figure 1: Architectural overview of the proposed closed-loop system for image-to-video generation. The framework integrates a diagnostic Video Moderation Module (3), calibrated by Creative Expert Guidelines, with an Agentic Regeneration Module (7). Candidate videos undergo iterative evaluation across 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.

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 motion pace is neither stagnant nor aggressive and remains free of jitter. We compute dense optical flow using RAFT to obtain a per-pixel displacement field between consecutive frames; from this we derive robust, frame-level motion statistics and 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. For each time step t, RAFT produces a dense displacement field $\mathbf{u}_t(x) \in \mathbb{R}^2$ mapping pixels from frame f_t to f_{t+1}. We compute the per-pixel motion magnitude $m_t(x) = \|\mathbf{u}_t(x)\|_2$ and summarize motion using a robust statistic: the mean magnitude of the top-q fraction of pixels (e.g., $q = 15\%$), denoted $M_t = \text{Mean}(\text{Top}_q(\{m_t(x)\}))$. This focuses on salient moving regions and reduces sensitivity to background noise. Video-level intensity is summarized by $\mu_M = \text{Mean}_t(M_t)$ and stability by $\sigma_M = \text{Std}_t(M_t)$. 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/decelerations). These outcomes are directly mapped to actionable remediation (reduce motion strength, tighten camera range, or stabilize).

3.2.2 Prompt Alignment Module. This module verifies that the observed camera motion matches the motion intent specified by the prompt (e.g., pan left-to-right, tilt up, dolly-in/zoom-in). We estimate camera motion using TAPIR point tracking: a set of salient points is initialized on the first frame and 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. Let $p_i^t \in \mathbb{R}^2$ denote the location of tracked point i at time t. We compute the net displacement $\Delta p_i = p_i^T - p_i^1$ and summarize motion by the dominant direction vector $\hat{\mathbf{v}} = \text{Normalize}(\text{Median}_i(\Delta p_i))$. We additionally use structured point groups (e.g., left-edge set $\mathcal{P}_L$ and top-edge set $\mathcal{P}_U$) to detect zoom/dolly: if $\mathcal{P}_L$ moves left while $\mathcal{P}_U$ moves up (points diverge away from the image center), the motion is consistent with a zoom-in/dolly-in; conversely, convergence toward the center indicates zoom-out/dolly-out. The observed motion label $\hat{y}$ is then compared against the prompt-specified intent y_{prompt} to yield an alignment score s_{align}, which triggers a FAIL when below threshold.

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

Require: Video $V = \{f_1, \ldots, f_T\}$, 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}
$d_{\text{motion}} \leftarrow$ PASS, $r_{\text{motion}} \leftarrow$ 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
$M_t \leftarrow \text{Mean}\big(\text{Top}_q(\{m_t(x)\})\big)$
end for
$\mu_M \leftarrow \text{Mean}_t(M_t), \quad \sigma_M \leftarrow \text{Std}_t(M_t)$
if $\mu_M < \tau_{\text{low}}$ **then**
$d_{\text{motion}} \leftarrow$ FAIL; $r_{\text{motion}} \leftarrow$ Low Motion (Stagnant)
else if $\mu_M > \tau_{\text{high}}$ **then**
$d_{\text{motion}} \leftarrow$ FAIL; $r_{\text{motion}} \leftarrow$ High Motion (Aggressive)
else if $\sigma_M > \tau_{\text{jit}}$ **then**
$d_{\text{motion}} \leftarrow$ FAIL; $r_{\text{motion}} \leftarrow$ 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 $d_{\text{motion}} =$ PASS **and** $s_{\text{align}} < \tau_{\text{align}}$ **then**
$d_{\text{motion}} \leftarrow$ FAIL; $r_{\text{motion}} \leftarrow$ 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 visual artifacts. We further define *new-structure hallucination* as the synthesis of objects or scene elements that are absent from the source image. This distinction is particularly important in high-trust domains (e.g., travel and real estate), where introducing nonexistent structures may misrepresent reality and increase legal and reputational risk. The following section details the annotation categories for each hallucination type.

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 that are not explained 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 the field of 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 of new objects**: Objects absent from the source image appear by entering the field of view with temporally continuous, physically plausible motion.
- **Camera-induced entry of new objects**: Previously unseen objects become visible as a consequence of global camera motion (e.g., panning, tilting, or zooming) rather than object motion.
- **Abrupt emergence of new objects**: Objects not present in the source image appear instantaneously without any reasonable physical interaction or explanation.

3.3.2 Hallucination Detection Module. As illustrated in Fig. 2, we propose a hallucination-detection module for identifying hallucinations in videos generated by an Image2Video model. Given a generated video, we uniformly sample a set of frames $F = \{f_1, \ldots, f_n\}$. To detect object hallucinations, we first extract a set of objects $O = \{o_1, \ldots, o_m\}$ from the input image using a multimodal large language model (MLLM). We then provide O, the input image, and the sampled frames F to an MLLM (e.g., GPT-4o) to identify object-level inconsistencies and anomalies across time, which are mapped to the predefined hallucination categories $C = \{c_1, \ldots, c_m\}$. For each category c_i, the MLLM outputs a severity score $S_o^i \in [0, 10]$ and a brief textual rationale when hallucination evidence is present. The object-hallucination decision d_{object} is obtained by aggregating the category-wise scores and thresholding each category 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, according to three cases: (1) natural entry of new objects, (2)camera-induced entry of new objects and (3)abrupt emergence of new objects. An MLLM produces a per-frame binary decision d_{new}^i indicating whether any new-structure hallucination is present. The video-level decision is computed as $d_{\text{new}} = \bigcup_{i=1}^n d_{\text{new}}^i$,

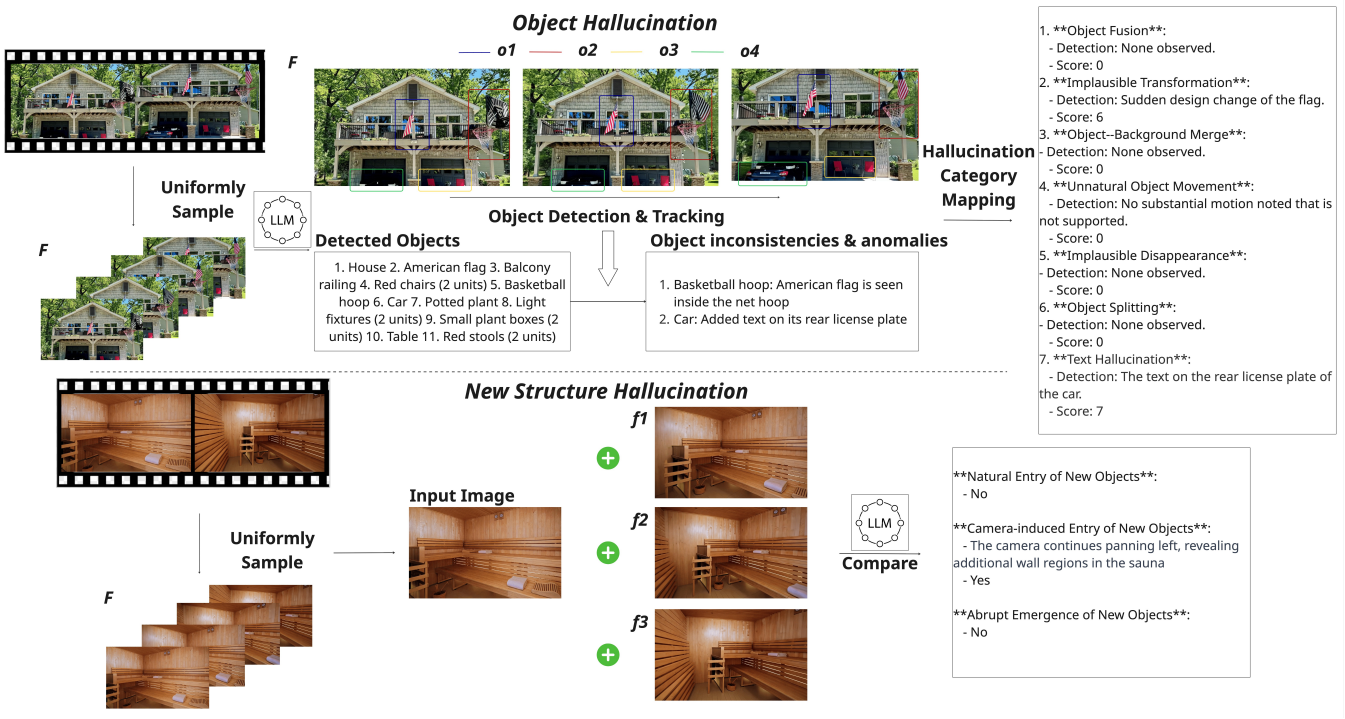

Figure 2: Hallucination detection module for Image2Video generation. Uniformly sampled frames are analyzed with an MLLM to (top) detect and track input-image objects, map temporal inconsistencies to predefined object-hallucination categories, and aggregate category scores for a final decision; and (bottom) compare each sampled frame with the input image to identify new-structure hallucinations (natural entry, camera-induced entry, or abrupt emergence)

3.4 Scoring and Risk Assessment

The moderation module aggregates signals across all evaluation dimensions into a unified, machine-actionable quality verdict.

Aggregation. Frame-level, temporal consistency, and hallucination detection scores are combined using domain-calibrated weights that reflect their relative impact on user trust, perceived realism, and overall aesthetics. In addition to numeric aggregation, the system produces a structured narrative summary that contextualizes each detected issue, providing concise definitions and illustrative examples to clarify its relevance to creative standards and Quality of Experience (QoE).

Decision and Risk Classification. An overall `PASS/FAIL` decision is produced based on thresholded criteria derived from expert moderation guidelines. The system further assigns a standardized risk level (*No*, *Low*, *Medium*, or *High*) linked to specific failure categories and their associated severities.

Machine-Actionable Feedback. Outputs are structured as a tuple of failure codes, severity levels, and recommended remediation levers (e.g., reduce motion intensity, avoid generating or modifying specific objects, adjust exposure, or switch generative models). A language model is used to normalize terminology, prioritize issues, and ensure consistent severity mapping across assets and distribution channels.

This structured report not only functions as a gating mechanism but also serves as input to the agentic regeneration module, enabling targeted and iterative corrections that efficiently converge toward videos satisfying brand, safety, and creative standards.

4 Agentic Regeneration Module

The Agentic Regeneration Module transforms moderation findings into targeted corrective actions through 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 by the moderation module, and the loop continues until quality thresholds or operational limits are reached.

Inputs and Outputs. Inputs include 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 consist of a next-generation plan (revised prompt, selected actions, updated parameters, chosen model or base image) and execution artifacts (regenerated video with a full action log for reproducibility).

Action Policy. The agent operates over a constrained library of interventions: (i) prompt refinements enforcing motion clarity, object preservation, fidelity to the source image, and aesthetic stability; (ii) camera and motion parameter adjustments to reduce jitter, overshoot, or blur; (iii) base image selection when artifacts cluster around occlusions or ambiguous regions; and (iv) model switching or parameter tuning for model-specific failure modes. Selection follows a cost- and efficacy-aware policy that prioritizes low-cost, minimally invasive actions, escalates for high-severity or repeated failures, and permits bundled actions when interactions are likely.

Iteration and Governance. The loop (Generate → Moderate → Regenerate) terminates upon PASS with acceptable risk, budget exhaustion, or escalation to human review for persistent high-severity issues. All decisions—prompts, parameters, seeds, model versions, base image identifiers, and rationales—are logged to ensure auditability and continuous policy improvement.

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

5 Evaluation

The system is optimized using 58 AI-generated video clips, with respect to the 3 moderation components. We then evaluate the proposed 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, 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, 2 and 3

Table 1: PASS/FAIL outcomes in Shadow Mode.

	Expert	HALLELUAI
Total Clips	146	146
PASS	39 (27%)	33 (23%)
FAIL	107 (73%)	113 (77%)

Overall Outcomes.

Table 2: Confusion matrix: HALLELUAI moderation vs. Expert decisions.

	Expert PASS	Expert FAIL
Our AI system PASS	29	4
Our AI system FAIL	10	103

Table 3: 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%

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 4.

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 currently provide comprehensive evaluation for image-to-video generation, we compared our system against user-generated content (UGC) video quality assessment (VQA) models, such as DOVER [23] and COVER [3]. Additionally, we compared our system to multimodal large language models (MLLMs) using zero-shot prompting. The following is an example prompt designed for MLLMs, which aims to evaluate the three components addressed by our system: frame-level quality, temporal motion quality, and hallucination. Frames are uniformly sampled from the video at 2 fps and provided as input to the MLLMs for evaluation.

As shown in Table 5 , UGC video quality assessment models, such as DOVER and COVER, were unable to differentiate between PASS and FAIL videos, as these models were not designed for image-to-video generation tasks. Furthermore, these models lack metrics for hallucination detection. The MLLM-based method exhibited low precision due to its inability to detect artifacts in AI-generated videos. This finding demonstrates that without detailed and comprehensive prompting, MLLMs cannot reliably identify artifacts in AI-generated video content.

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. Out of 1,158 clips approved by

Table 4: 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	(1/4 FP) 97%
	Brightness	100%
	Noise	100%
Temporal Motion Quality	Prompt Alignment	100%
	Motion Intensity	(1/4 FP) 97%
Hallucination Detection	Object Hallucination	(1/4 FP) 97%
	New Structure Hallucination	(1/4 FP) 97%
	Text Hallucination	100%

Algorithm 3 Example Prompts Designed for MLLMs to Evaluate Frame-level Quality, Temporal Motion Quality, & Hallucination in Image2Video Generation

Input: AI-generated video (created from the first frame as input image)
Task: Assess the following criteria (PASS/FAIL for each)
1. Frame Quality
- Blur: Is sharpness maintained?
- Contrast: Are levels appropriate?
- Brightness: Is exposure consistent?
- Noise: Is the frame free of artifacts?

2. Motion Quality
- Camera Motion: Is it smooth and pleasant?
- Temporal Smoothness: Are transitions fluid?

3. AI Hallucination
- Temporal Consistency: Do objects maintain their identity?
- Physical Plausibility: Is motion realistic?
- Object Coherence: Is there any drift or deformation?

Required Output Format:
FRAME QUALITY: Blur=[P/F], Contrast=[P/F], Brightness=[P/F], Noise=[P/F]
MOTION QUALITY: Camera=[P/F], Smoothness=[P/F]
HALLUCINATION: Consistency=[P/F], Physics=[P/F], Coherence=[P/F]
REASON: [One sentence explanation]
OVERALL: [PASS/FAIL]

Table 5: Performance Comparison of Different Evaluation Methods

Methods	Precision	Recall	F1-score	Accuracy
DOVER	0.20	0.21	0.20	0.56
COVER	0.27	0.33	0.30	0.58
Qwen3-VL-8B-Instruct	0.28	1.00	0.43	0.30
NOVA 2 lite	0.25	0.74	0.37	0.34
HALLELUAI moderation	**0.88**	**0.74**	**0.81**	**0.87**

the system, 1,126 were also accepted by experts, yielding a precision of 97%, 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. As a result, the system is not just filtering errors but actively improving clips, leading to 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 updates to thresholds or regeneration policies do not degrade previously validated performance.

In production, the system was deployed at scale, generating approximately 70,000+ videos, further validating its reliability and consistency in real-world conditions.

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

We presented HALLELUAI, an expert-aligned, production-oriented system that turns image-to-video quality assurance into a controllable, closed-loop process. HALLELUAI couples per-asset moderation—frame-level aesthetics, temporally grounded motion quality, and source-image–conditioned hallucination detection—with an agentic regeneration policy that translates failures into targeted fixes rather than unguided retries.

Human-in-the-loop validation shows strong agreement with creative experts and supports deployment as a high-precision gate for scalable AIGV. By enforcing strict input-image fidelity, emitting machine-actionable diagnostics, and maintaining auditable decision trails, HALLELUAI bridges the gap between benchmark-style evaluation and real-world creative governance. To our knowledge, this is the first integrated moderation-and-regeneration framework purpose-built for ultra-realistic, image-conditioned video generation at scale.

References

[1] Shane Barratt and Rishi Sharma. 2018. A Note on the Inception Score. *arXiv preprint arXiv:1801.01973* (2018).

[2] Fu et al. 2023. A Comprehensive Benchmark and Toolkit for Evaluating Video Generative Models. *arXiv preprint arXiv:2311.16103* (2023).

[3] Chenlong He, Qi Zheng, Ruoxi Zhu, Xiaoyang Zeng, Yibo Fan, and Zhengzhong Tu. 2024. Cover: A comprehensive video quality evaluator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5799–5809.

[4] et al. He. 2024. MantisScore: Fine-grained Video Quality Assessment with Human-Aligned Feedback. *arXiv preprint arXiv:2406.15252* (2024).

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems* 30 (2017).

[6] Huang et al. 2023. VBench: Comprehensive Benchmark Suite for Video Generative Models. *arXiv preprint arXiv:2311.16103* (2023).

[7] Saurabh Kadavath et al. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221* (2022).

[8] Li et al. 2024. A Unified Hallucination Detection for Large Text-to-Video Models. *arXiv preprint arXiv:2405.04180* (2024).

[9] et al. Li. 2025. MSA-VQA: Multi-Scale Assessment for AI-Generated Video Quality. *arXiv preprint arXiv:2501.02706* (2025).

[10] X. Li et al. 2025. Video Diffusion Generation: A Comprehensive Review and Open Problems. *Artificial Intelligence Review* 58, 4 (2025), 1234–1256.

[11] Liu et al. 2024. Fréchet Video Motion Distance: A Metric for Evaluating Motion Consistency in Video Generation. *arXiv preprint arXiv:2407.16124* (2024).

[12] Liu et al. 2024. GenAI-Bench: Evaluating and Improving Compositional Text-to-Visual Generation. *arXiv preprint arXiv:2406.13743* (2024).

[13] Liu et al. 2024. VQAScore: Evaluating Text-to-Visual Generation with Image-to-Text Models. *arXiv preprint* (2024).

[14] 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. 2025. Controllable Video Generation: A Survey. *arXiv preprint arXiv:2507.16869* (2025).

[15] OpenAI. 2026. Sora: OpenAI Text-to-Video Model. Product documentation and public reports, URL: https://openai.com/research/sora.

[16] Z. Pan et al. 2026. Reasoning Diffusion for Text-Image-to-Video Generation. In *International Conference on Learning Representations (ICLR)*. arXiv preprint arXiv:2601.01234.

[17] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. 2024. ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation. *arXiv preprint arXiv:2402.04324* (2024).

[18] Thomas Unterthiner et al. 2018. Towards Accurate Generative Models of Video: A New Metric & Challenges. In *International Conference on Learning Representations Workshop*.

[19] Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. 2024. AIGV-Assessor: Benchmarking and Evaluating the Perceptual Quality of Text-to-Video Generation with LMM. *arXiv preprint arXiv:2411.17221* (2024).

[20] Xuezhi Wang et al. 2023. Chain-of-Thought Is Not Needed! Self-Consistency Improves Chain of Thought Reasoning in Large Language Models. *arXiv preprint arXiv:2203.11171* (2023).

[21] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[22] Wu et al. 2024. EvalCrafter: Benchmarking and Evaluating Large Video Generation Models. In *CVPR*.

[23] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF international conference on computer vision*. 20144–20154.

[24] Yao et al. 2024. Understanding Hallucinations in Diffusion Models through Mode Interpolation. *arXiv preprint arXiv:2406.09358* (2024).

[25] Zhichao Zhang, Wei Sun, Xinyue Li, Jun Jia, Xiongkuo Min, Zicheng Zhang, Chunyi Li, Zijian Chen, Puyi Wang, Fengyu Sun, Shangling Jui, and Guangtao Zhai. 2024. Benchmarking Multi-dimensional AIGC Video Quality Assessment: A Dataset and Unified Model. *arXiv preprint arXiv:2407.21408* (2024).

[26] Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. 2024. Identifying and Solving Conditional Image Leakage in Image-to-Video Diffusion Model. *arXiv preprint arXiv:2406.15735* (2024).