

Illusion or Integrity? Geometrical Consistency Metric for AIGC Video Quality Evaluation

YIFEI XUE*, Hunan University, China

YUANCHEN FEI†, Hunan University, China

HAO ZHANG, Hunan University, China

CHENZHI NIE, Hunan University, China

TIE JI, Hunan University, China

YIZHEN LAO, Hunan University, China

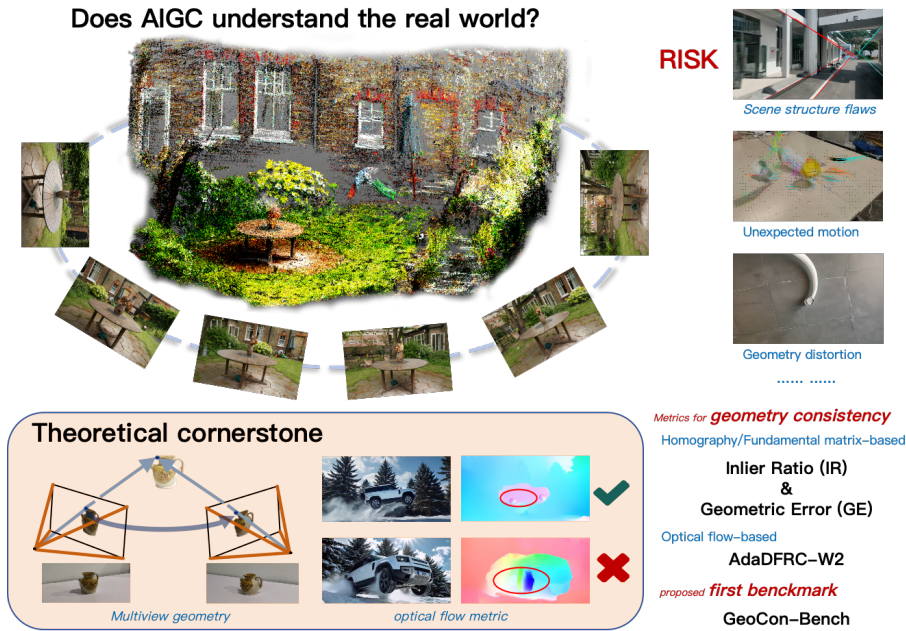


Fig. 1. To quantify how well AIGC-generated videos preserve physical-world geometry, we introduce GeoCon-Bench, a model-agnostic evaluation framework grounded in multiview geometry and optical-flow analysis. The figure shows (top) a generated camera sweep; (right) representative failure modes; and (bottom) the theoretical cornerstones we use. GeoCon-Bench scores clips via homography/fundamental-matrix metrics for rigid backgrounds and flow-based tests for dynamic regions, enabling reproducible auditing of geometric consistency.

*Both authors contributed equally to this research.

†Both authors contributed equally to this research.

Authors' Contact Information: Yifei Xue, Hunan University, China; Yuanchen Fei, Hunan University, China; Hao Zhang, Hunan University, China; Chenzhi Nie, Hunan University, China; Tie Ji, Hunan University, China; Yizhen Lao, Hunan University, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Recently, AI-driven video generation has attracted considerable attention. This surge increases demand for reliable video quality assessment (VQA) metrics to assess AI-generated content (AIGC) videos and guide model optimization. Existing work examines VQA via visual harmony, video-text consistency, and domain-specific alignment, yet lacks quantitative metrics for fidelity to physical laws. To address this limitation, we present a novel benchmark that evaluates the quality of AIGC videos through their compliance with physical principles, achieved by quantitatively measuring the geometrical consistency across frames extracted from the generated sequences, thus as a proxy to gauge the extent to which the generated videos conform to real-world physics rules. Specifically, **GeoCon-Bench** gates global motion via translation estimates, fits a homography or fundamental matrix to background correspondences, and reports complementary metrics (inlier ratio, geometric error). We also release a dataset of 20 scenes across six motion categories. Experiments on state-of-the-art AIGC models demonstrate GeoCon-Bench’s reliability as a video quality metric.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: T2V, AIGC, multi-view geometry, 3D reconstruction

ACM Reference Format:

Yifei Xue, Yuanchen Fei, Hao Zhang, Chenzhi Nie, Tie Ji, and Yizhen Lao. 2018. Illusion or Integrity? Geometrical Consistency Metric for AIGC Video Quality Evaluation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 35 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Generative video offers a tantalizing pathway to *lower the cost and latency* of 3D asset creation [20, 28]. Rather than capturing calibrated footage and running bespoke reconstruction pipelines, practitioners could in principle *feed an idealized video* to a reconstruction system and obtain geometry ready for downstream use. Yet this promise is ultimately gated by *photogrammetry and multi-view geometry* [11]: usable parallax, consistency with either homography or epipolar structure, and coherent occlusion handling are *prerequisites*, not niceties. When those prerequisites fail, error cascades are immediate and practical—point clouds fragment or rubber-sheet, camera poses drift, novel views shear, and AR overlays refuse to lock—even though the sequence may still look plausible to the eye [17, 19]. Recent progress in diffusion/transformer generators [1, 5, 14, 21, 43] increases perceptual quality but does not guarantee photogrammetric validity [2, 13, 32, 41, 49]. Accordingly, the *value proposition* of video-to-3D is determined not by aesthetics or prompt adherence, but by *geometric reliability* [37].

Perceptual plausibility ≠ photogrammetric correctness. [15, 16, 47]. Contemporary systems can score well on perceptual or distributional metrics while *systematically violating* multi-view constraints. Typical failure modes that matter for reconstruction include: (i) background slide where planar regions move inconsistently with the camera; (ii) rubber-sheet warps that break the assumption of a single rigid background model; and (iii) inconsistent scale-depth relations across time that a bundle adjuster cannot reconcile. These violations have been documented both qualitatively and quantitatively in generative pipelines [36], and they are *fatal* for multi-view reconstruction, pose estimation, and novel-view synthesis (NVS).

Mainstream evaluation emphasizes pixel fidelity, learned perceptual similarity, distributional distances, semantics, or composite leaderboards [6, 8, 15, 16, 24, 25, 33, 50]. While valuable for breadth, such axes are *insensitive to multi-view geometric validity*, especially in real-data settings where ground truth (GT) is unavailable. In practical platform

comparisons, GT is *usually absent*: real camera trajectories and scene geometry are proprietary or intractable to share; generated videos often have no physical camera at all; and synthetic GT, while useful scientifically, can *misalign with deployment distributions and sensors*, confounding cross-platform conclusions. Therefore, the community lacks a *GT-free, photogrammetry-aligned, auditable* protocol that can be reproduced across generators and data regimes.

Hence, we derive four design requirements for a practical geometry-consistency protocol: **R1 Regime awareness**: background modeling must switch between homography and fundamental descriptions based on *translation evidence*; **R2 Background–dynamic disentanglement**: rigid backgrounds and moving/non-rigid regions obey different assumptions and must be *scored against different objectives*; **R3 GT-free feasibility**: every stage must operate without GT, so that results are reproducible across platforms and datasets; **R4 Auditability and comparability**: decisions and scores must be *traceable* (with gate rationales, inlier diagnostics, and threshold sensitivity) so platform-level comparisons are defensible.

We propose a unified, GT-free protocol that (i) extracts translation evidence from pose logs to *gate* the background model, (ii) fits the model with *calibrated dual thresholds* tailored to each residual family, (iii) scores dynamic regions separately via an optical-flow-guided measure, and (iv) logs diagnostics for auditability. Concretely, a gate variable $\gamma_t \in \{H, F\}$ selects a homography when translation is negligible and a fundamental model otherwise. For background matches, we compute IR_H, GE_H under a homography residual with threshold τ_H and IR_F, GE_F under an epipolar/Sampson residual with threshold τ_F , explicitly allowing $\tau_H \neq \tau_F$ because the two residual families have *different statistical scales*. We accompany indicators with *threshold–score sensitivity scans* and *percentile alignment* to ensure cross-platform interpretability. Dynamic/non-rigid pixels are evaluated with a flow-guided measure (e.g., AdaDFRC-W2): forcing them into H/F would *pollute* inlier sets, destabilize robust fitting, and render IR/GE less diagnostic of background validity. The protocol logs gate rationales, inlier maps, threshold–score curves, and pose-log vs. fit sanity checks to support reproducible, inspector-friendly audits.

To our knowledge, this is the first *GT-free, regime-aware, photogrammetry-aligned* evaluation protocol for AIGC videos that is auditable across platforms. To summarize, our contributions are threefold:

- We introduce the first GT-free, regime-aware, and photogrammetry-aligned evaluation framework GeoCon-Bench for AIGC videos that supports auditable, cross-platform assessments.
- We design a translation-gated framework that adaptively selects between homography and fundamental modeling, proposes calibrated dual-threshold indicators for background consistency, and introduces AdaDFRC-W2, a Wasserstein-2 regional flow-consistency metric for robust evaluation of dynamic content.
- We release auditing assets including lightweight logs, coverage statistics, score curves, and a dedicated benchmark dataset that spans diverse scenes and motion regimes, enabling reproducible and inspector-friendly audits. *GeoCon-Bench will be publicly available after acceptance.*

2 Related Works

Mainstream evaluation families—pixel fidelity (PSNR/SSIM) [47], learned perceptual similarity (LPIPS) [49], distributional distances (FID/FVD) [42], semantic/prompt alignment (CLIP) [12, 34], and composite leaderboards—provide breadth but are largely *insensitive to multi-view geometric validity*, particularly under GT-free conditions [15, 16]. Their objectives emphasize local textures, global feature statistics, or text–image agreement, not whether homography/epipolar relations hold across viewpoints. As diffusion/transformer generators improve perceptual quality [2, 27, 32, 46], such blind spots become consequential: videos may look convincing yet remain unusable for 3D reconstruction, degrading pose

estimation, bundle adjustment, and novel-view synthesis. Recent multi-objective video benchmarks broaden axes (aesthetics, prompt adherence, temporal smoothness), but typically *do not enforce* photogrammetric constraints on real data without ground truth [15, 16, 18]. Our work complements this breadth with explicit, photogrammetry-aligned checks that are feasible and auditable in GT-free settings.

Photogrammetry and multi-view geometry formalize how scene structure and camera motion are recovered from image correspondences; practical pipelines rely on correspondence quality (inlier ratios), geometric residuals (reprojection or epipolar/Sampson) [9, 11], robust model fitting, and bundle adjustment [3, 4, 10, 22, 40]. However, thresholds inside these systems are typically tuned for *optimization stability* on particular datasets rather than for *cross-platform comparability* and *auditability*. Crucially, the appropriate background model depends on camera motion: negligible translation [30] supports a homography on planar/parallax-free backgrounds, whereas observable translation calls for epipolar geometry on backgrounds with depth variation. Modern monocular multi-view methods (e.g., DUS3R) expose *pose logs* that can be repurposed as *translation evidence* to inform such regime choices [45]. We build on these signals but reframe them as *comparable indicators* by (i) tying regime selection to translation evidence and (ii) calibrating *dual thresholds* (τ_H, τ_F) for model-appropriate residual families with sensitivity analyses, so reported IR/GE are interpretable across generators and datasets.

Moving or deforming content violates rigid-background assumptions; if dynamic pixels are forced into homography/fundamental fitting, they *pollute* inlier sets, destabilize robust estimation, and reduce the diagnostic power of IR/GE for the background. Optical flow and occlusion reasoning directly probe *motion plausibility*, *deformation continuity*, and *disocclusion behavior* [38, 39, 48], properties that matter for structure-from-motion [31] and view synthesis. In the generative setting, dynamics can be perceptually plausible while geometrically inconsistent across frames (e.g., flow fields that ignore parallax or violate occlusion ordering) [36]. We therefore *separate* background rigidity from dynamic coherence: H/F-based indicators summarize whether a *single rigid* model explains the background, while a flow-guided score targets non-rigid and occlusion dynamics. This separation is methodological rather than cosmetic, preserving interpretability and avoiding mixed-objective confounds.

GT-free or weakly supervised evaluations are standard in perceptual quality assessment, but GT-free *geometry-aware* protocols remain scarce for platform-level comparisons. Two challenges recur: *defensibility* of decisions (e.g., why a particular regime/threshold) and *traceability* of results (e.g., whether scores are stable across thresholds/unseen scenes). In practice, GT is usually unavailable because real camera trajectories and scene geometry are proprietary, intractable to share, or nonexistent for generated videos. Synthetic GT, while scientifically useful, often misaligns with deployment distributions and sensor data. We address these challenges by (i) grounding regime selection in pose-log translation evidence with sanity checks against fitted geometry, (ii) releasing *threshold-score sensitivity scans* and *percentile-aligned* summaries to ensure comparability, and (iii) logging inlier coverage and robust clip-level aggregation to make audits reproducible across systems.

Recent studies document how generative pipelines can violate physical/geometric constraints despite high perceptual scores [36], while multi-objective benchmarks emphasize user-perceived axes over explicit multi-view validity [15, 16]. We bridge these trajectories: retaining GT-free feasibility and breadth, but injecting *photogrammetry-aligned* checks that are auditable and reproducible. This reframes evaluation around practitioner questions: *Will this video support reconstruction? If not, where does geometry fail, and are failures systematic across content and generators?*

Prior work provides broad perceptual/semantic coverage and mature geometric estimators, but stops short of an integrated, GT-free standard that *enforces* multi-view validity while preserving auditability. We next formalize the latent construct and observable indicators in a concept layer, then instantiate the protocol and diagnostics in §4.

3 Concept

3.1 Construct, scope, and contextual factors

We treat *geometry consistency* as a latent construct with two facets: **background rigidity** and **dynamic-region coherence**. Perceived salience of a specific residual depends on contextual factors, including scene depth span, planarity, edge density, motion spectrum, and occlusion complexity [38, 39, 48]. Consequently, indicators must be interpretable under diverse conditions and aggregated robustly over time. The construct is operationalized without ground truth (GT-free) [15, 16] by checking self-consistency with photogrammetric constraints and by leveraging pose logs [45] as auxiliary evidence.

3.2 Regime-aware prior

Let I_t, I_{t+1} be consecutive frames with background matches $\{(x_i, y_i)\}_{i=1}^{N_b}$ and a background mask $\mathcal{M}_b$. From pose logs (e.g., DUST3R) we obtain a translation proxy τ_t [45] and define a gate

$$\gamma_t \in \{H, F\}, \quad \gamma_t = \begin{cases} H & \text{if } \tau_t < \tau_{\text{gate}}, \\ F & \text{otherwise.} \end{cases} \quad (1)$$

When $\gamma_t = H$ we fit a homography matrix $\mathbf{H}_t$ to background correspondences; when $\gamma_t = F$ we fit a fundamental matrix [11] $\mathbf{F}_t$. This prior encodes standard photogrammetric practice: homographies capture planar or parallax-free backgrounds under negligible translation, whereas epipolar geometry governs backgrounds with depth variation under translation.

3.3 Dual-threshold indicators for background rigidity

For a candidate model $g_t \in \{\mathbf{H}_t, \mathbf{F}_t\}$, let $e_i^{(g)}$ denote a model-appropriate residual. In our research, we utilize symmetric transfer error for $\mathbf{H}_t$, and Sampson/epipolar approximation for $\mathbf{F}_t$. Because these residual families have *different statistical scales and tails*, we calibrate *dual thresholds* (τ_H, τ_F) rather than forcing a single cutoff. Define the inlier set

$$\mathcal{S}_\tau^{(g)} = \{i \in \mathcal{M}_b \mid e_i^{(g)} \leq \tau\}, \quad (2)$$

and report two indicators [3, 40]:

$$\text{IR}_H = \frac{|\mathcal{S}_{\tau_H}^{(H)}|}{N_b}, \quad \text{GE}_H = \text{median}\{e_i^{(H)} : i \in \mathcal{S}_{\tau_H}^{(H)}\}; \quad \text{IR}_F = \frac{|\mathcal{S}_{\tau_F}^{(F)}|}{N_b}, \quad \text{GE}_F = \text{median}\{e_i^{(F)} : i \in \mathcal{S}_{\tau_F}^{(F)}\}. \quad (3)$$

IR answers whether a single rigid model explains the background at step t ; GE answers how tightly the consistent matches align when it does. To ensure cross-platform interpretability, we accompany each report with **threshold-score sensitivity scans** over a grid of (τ_H, τ_F) and with **percentile-aligned** summaries, mitigating absolute-scale differences across residual families and datasets. See §6.2 for threshold-score sensitivity scans, stability bands, and coverage definitions.

3.4 Why dynamics require a separate metric

Dynamic or non-rigid pixels on a moving object always violate the rigid-scene assumptions. That’s why including them in H/F fitting (i) corrupts inlier sets, (ii) destabilizes robust estimation, and (iii) degrades the diagnostic value of IR/GE for the background. Conversely, judging dynamics by background residuals rewards trivial solutions (e.g., shrinking inlier sets) and unfairly penalizes scenes with legitimate motion. We therefore adopt a *flow-guided* dynamic score to

capture motion plausibility, deformation continuity, and occlusion ordering:

$$\text{Dyn}_t = \text{RobAgg}\left(w_i \|\hat{\mathbf{u}}_i - \mathbf{u}_i\|_2 ; i \in \mathcal{M}_d\right), \quad (4)$$

where $\mathcal{M}_d$ is the dynamic-region mask, $\mathbf{u}_i$ is measured optical flow, $\hat{\mathbf{u}}_i$ is expected motion under consistency priors, w_i down-weights unreliable/occluded pixels, and RobAgg denotes a robust aggregator. We instantiate this metric with an *adaptive Wasserstein-2* [7] regional formulation (AdaDFRC-W2) in §4, which aggregates local motion-pattern discrepancies with occlusion-aware weighting. This score is *complementary* to background indicators: H/F-based IR/GE certify rigid-scene adequacy, while Dyn certifies that moving/occluding content evolves plausibly.

3.5 Intended use, GT-free positioning, and boundaries

Our protocol is an *engineering QA baseline* for model comparison, regression testing, and content-side risk triage under GT-free conditions. It complements human-subject studies rather than replacing them by enforcing photogrammetric validity without GT [23, 29]. It provides a *reproducible floor* upon which subjective evaluation can build.

Beyond the practical unavailability of GT in platform settings, GT-free protocols allow cross-platform comparisons on real data without disclosing proprietary trajectories or geometry and avoid mismatches introduced by synthetic GT. By aligning to photogrammetric self-consistency and releasing audit artifacts, we retain a hard, reproducible constraint that generalizes across devices, content, and generators.

Typical uses include (i) cross-generator comparisons with regime-stratified breakdowns; (ii) pipeline regression tests where gate agreement and sensitivity curves act as sentinels; and (iii) dataset curation by filtering clips with unstable regimes or insufficient background support. Known limitations include specularities, rolling-shutter distortions, extreme motions with low-SNR correspondences, and non-pinhole optics.

4 Methodology

Our methodology is designed to quantitatively answer a fundamental question: whether a generated video adheres to the geometric principles of a rigid 3D world observed by a moving camera. The framework operates without ground truth and is built on a core principle we term **mask-first, regime-aware geometry**. This principle dictates that we must first isolate independently moving objects from the static background (mask-first), and then apply the geometrically appropriate motion model to the background based on the nature of the camera’s movement (regime-aware).

4.1 Overall Framework

Given a generated video sequence $\{I_t\}_{t=0}^T$, our evaluation proceeds in three main stages, as illustrated in Fig. 2:

- (1) **Global Motion Regime Gating:** We first determine the dominant camera motion of the input video stream. Specifically, we distinguish between motion that is purely rotational (or static) and motion that includes a translational component. This decision is critical as it dictates which geometric model is appropriate for the background.
- (2) **Foreground/Background Separation:** We segment any objects that are moving independently of the camera’s ego-motion. This creates a clean background mask, ensuring that our analysis of the static scene’s rigidity is not contaminated by foreground motion.

- (3) **Consistency Evaluation:** We apply distinct, specialized metrics to the separated background and foreground regions. The background is evaluated for rigid-body consistency using classical geometric models, while each moving foreground object is evaluated for its internal geometric integrity.

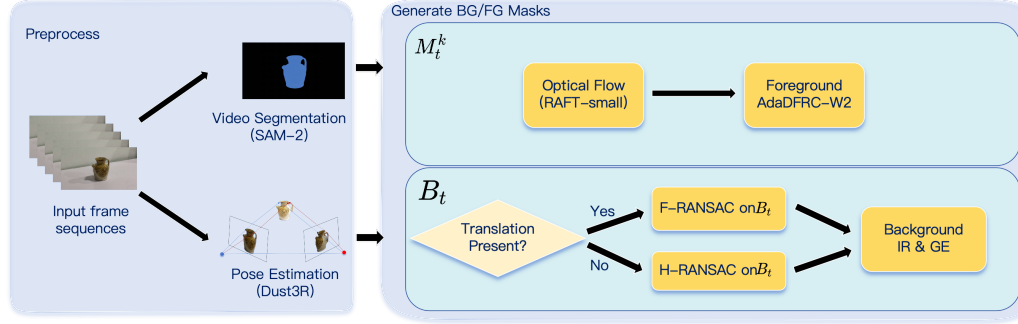


Fig. 2. Our evaluation workflow. A pose estimator (DUST3R) determines the camera motion regime to select between a Fundamental Matrix (F) or Homography (H) model. A segmentation model (SAM-2) [35] isolates moving objects. Background consistency is then evaluated on the static regions only, while foreground objects are scored for internal rigidity using AdaDFRC-W2.

4.2 Pose estimation and translation-aware gating

In photogrammetry, negligible translation justifies a single-plane/parallax-free background (homography), whereas observable translation requires epipolar geometry (fundamental). We therefore route by a translation-evidence proxy s_t from pose logs, with a hysteresis band $[\tau_\rho - \epsilon, \tau_\rho + \epsilon]$ to avoid flip-flop: $\gamma_t = H$ if $s_t < \tau_\rho - \epsilon$, $\gamma_t = F$ if $s_t > \tau_\rho + \epsilon$, and otherwise we keep γ_{t-1} .

To select the correct geometric model for the background, we must first characterize the camera’s ego-motion. We estimate the relative camera pose $(\mathbf{R}_t, \mathbf{t}_t) \in SE(3)$ between frames. Let I_t, I_{t+1} be adjacent frames. From deep matches $\mathcal{M}_{t,t+1} = \{(x_i, x'_i)\}_{i=1}^N$ (pixel coordinates in homogeneous form), we estimate the relative pose $(\mathbf{R}, \mathbf{t}) \in SE(3)$ using DUST3R or a PnP solver on the same correspondences [45]. Under DUST3R’s convention, the scene is normalized to unit scale along principal axes. We declare the pair *non-translation-dominant* (behaves like pure rotation) when

$$\|\mathbf{t}\|_\infty \triangleq \max\{|t_x|, |t_y|, |t_z|\} \leq \tau_t, \quad \tau_t = 0.04, \quad (5)$$

and *translation-dominant* otherwise. The choice of τ_t follows a first-order pinhole analysis: for a 3D point $\mathbf{P} = (X, Y, Z)^T$ the translation-induced image displacement satisfies

$$\|\delta \mathbf{x}\| \lesssim \frac{f}{Z} \|\mathbf{t}_\perp\|, \quad (6)$$

with focal length f (in pixels) and translation component $\mathbf{t}_\perp$ orthogonal to the view. Requiring $\|\delta \mathbf{x}\| \leq \tau_H$ (the inlier tolerance of homography fitting, typically ~ 1 px at our working resolution) yields $\|\mathbf{t}_\perp\| \leq (\tau_H/f) Z$. With median $Z \approx 1$ in DUST3R units and calibrated f , the practical bound $\|\mathbf{t}\|_\infty \leq 0.04$ keeps $\|\delta \mathbf{x}\|$ at the noise floor, so a single homography provides an adequate approximation. To stabilize decisions near the boundary, we adopt a narrow hysteresis in implementation (≤ 0.03 : homography branch; ≥ 0.05 : fundamental branch); in the gray zone, both models are evaluated, and the one with higher support is selected downstream. Section §6.2 empirically calibrates this gate by a threshold sweep.

4.3 Foreground/Background Separation

A core tenet of our method is that the geometric model for camera ego-motion applies only to the static parts of the scene. Therefore, we must first identify and mask out any independently moving objects. Using a video object segmentation model, we generate a set of masks $\{M_t^k\}$ for each moving object k at frame t . The static background mask B_t is then defined as the complement of the union of all foreground masks: $B_t = 1 - \bigcup_k M_t^k$. All subsequent analysis of the background’s geometric consistency is performed *exclusively* on feature correspondences found within the regions defined by B_t . This “mask-first” approach is crucial as it prevents moving objects from corrupting the estimation of the global camera motion model, which would otherwise lead to incorrect evaluations of background stability.

4.4 Background Consistency: IR and GE

We assess the behavior of the *static background* as a single rigid body across consecutive frames by fitting a global geometric model. Specifically, we utilize either a homography $\mathbf{H}$ for the planar/rotation regime or a fundamental matrix $\mathbf{F}$ for the translation/parallax regime. The analysis is conducted on *background-only* correspondences, employing a robust estimator. Models are estimated in normalized coordinates for numerical stability and *denormalized to pixels* for residual evaluation. All residuals are computed on the *content mask* (letterboxed padding ignored). Unless otherwise specified, frames are 1024×720 with letterboxing.

Inlier Ratio (IR) is the fraction of background correspondences whose geometric residual falls below the inlier test used by the robust estimator. A high IR indicates that a single rigid transformation explains the majority of the background motion, as expected for static scenes (cf. Fig. 3). Geometric Error (GE) summarizes the *accuracy* of alignment among the inliers only, reported as the median residual. Thus IR answers “does a single global model hold?”, while GE answers “how tightly do the consistent matches align?”. In our implementation, the *same inlier test used during robust estimation* is reused to define the inlier set for reporting.

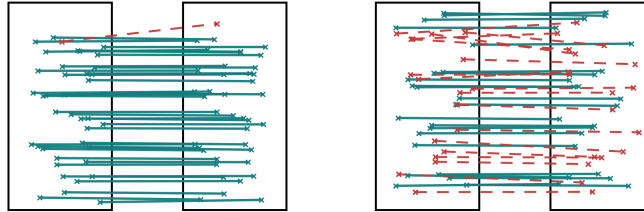


Fig. 3. Each panel shows two adjacent frames with matched feature correspondences. Left: high IR—most matches are inliers, indicating a rigid background. Right: low IR—many outliers, revealing background inconsistency.

Homography and epipolar residuals differ in scale and tail behavior; we therefore calibrate *dual thresholds* (τ_H, τ_F) rather than forcing a single cutoff. To make platform-level rankings defensible, we accompany per-clip IR/GE with (i) **threshold–score sensitivity scans** over (τ_H, τ_F) and (ii) **percentile-aligned** summaries (e.g., reporting GE_H at matched IR_H percentiles), which mitigate absolute-scale differences across datasets and generators.

Homography case (non-translation regime). In the non-translation regime, we model background motion with a homography $\mathbf{H}_t$. The error for each correspondence pair $(\mathbf{x}_1, \mathbf{x}_2)$ is evaluated using the *symmetric transfer error*, which

averages the forward and backward projection errors in the pixel domain:

$$\hat{\mathbf{x}}_2 = \pi(\mathbf{H}_t \mathbf{x}_1), \quad \hat{\mathbf{x}}_1 = \pi(\mathbf{H}_t^{-1} \mathbf{x}_2), \quad r_H^{\text{sym}} = \frac{1}{2} \|\mathbf{x}_2 - \hat{\mathbf{x}}_2\|_2 + \frac{1}{2} \|\mathbf{x}_1 - \hat{\mathbf{x}}_1\|_2, \quad (7)$$

where $\pi([a, b, c]^\top) = [a/c, b/c]^\top$ converts from homogeneous to Cartesian coordinates. Using this error metric, the two key consistency scores can be derived. We robustly estimate $\mathbf{H}_t$ (e.g., via PROSAC sampling) by finding the largest set of correspondences that satisfy the inlier condition

$$r_H^{\text{sym}} \leq \tau_H, \quad \text{with } \tau_H = 2.0, \text{ px.} \quad (8)$$

The **IR** is the fraction of correspondences classified as inliers, and the **GE** is the median symmetric transfer error among these inliers.

Fundamental case (translation/parallax regime). The estimation of the fundamental matrix $\mathbf{F}_t$ is highly sensitive to the scale and origin of pixel coordinates. To ensure numerical stability, we follow the standard practice of coordinate normalization. For a given set of N points $\{\mathbf{x}_i = [u_i, v_i, 1]^\top\}$, we first compute their centroid $\mathbf{c} = [\bar{u}, \bar{v}]^\top$ and a scaling factor $s = \sqrt{2/\bar{d}}$, where $\bar{d} = \frac{1}{N} \sum_i \|[u_i - \bar{u}, v_i - \bar{v}]^\top\|_2$ is the average distance from the centroid. The normalization matrix $\mathbf{T}$ is then constructed as:

$$\mathbf{T} = \begin{bmatrix} s & 0 & -s\bar{u} \\ 0 & s & -s\bar{v} \\ 0 & 0 & 1 \end{bmatrix}. \quad (9)$$

We compute $\mathbf{T}_1$ and $\mathbf{T}_2$ for the point sets in each image respectively, yielding normalized coordinates $\hat{\mathbf{x}}_1 = \mathbf{T}_1 \mathbf{x}_1$ and $\hat{\mathbf{x}}_2 = \mathbf{T}_2 \mathbf{x}_2$.

In this normalized space, we fit $\hat{\mathbf{F}}_t$ using a robust estimator (again, PROSAC can be used) exclusively in this normalized space. It finds an optimal normalized fundamental matrix $\hat{\mathbf{F}}_t$ by classifying a correspondence as an inlier if its Sampson distance, computed on normalized data, is below the unitless threshold $\tau_F = 0.1$. The **Inlier Ratio (IR)** is the fraction of correspondences in this final inlier set [9, 11].

Once the inlier set is identified, $\hat{\mathbf{F}}_t$ is transformed back to the original pixel domain via denormalization:

$$\mathbf{F}_t = \mathbf{T}_2^\top \hat{\mathbf{F}}_t \mathbf{T}_1. \quad (10)$$

Finally, the **Geometric Error (GE)** is defined as the *median* of the Sampson distances computed for all inlier pairs. This computation uses their original pixel coordinates $(\mathbf{x}_1, \mathbf{x}_2)$ and the denormalized matrix $\mathbf{F}_t$:

$$\text{GE} = \underset{(\mathbf{x}_1, \mathbf{x}_2) \in \text{Inliers}}{\text{median}} \left\{ \frac{|\mathbf{x}_2^\top \mathbf{F}_t \mathbf{x}_1|}{\sqrt{(\mathbf{a}_1^2 + (\mathbf{a}_2^2 + (\mathbf{b}_1^2 + (\mathbf{b}_2^2))} \right\}, \quad (11)$$

where $\mathbf{a} = \mathbf{F}_t \mathbf{x}_1$ and $\mathbf{b} = \mathbf{F}_t^\top \mathbf{x}_2$. This two-stage process ensures that the inlier selection is robust and independent of image resolution, while the final reported GE remains interpretable in the context of pixel-level deviations.

For **H** we use symmetric transfer error in *pixels* with threshold τ_H ; for **F** we use the *normalized* Sampson error with threshold τ_F (unitless). The dual-threshold design reflects fundamentally different residual scales and avoids ill-posed single-threshold comparisons. IR is the fraction of correspondences within the relevant threshold; GE is the median residual over inliers. Both metrics use the *same per-family threshold* to avoid moving standards. For brevity, in (12) we write a generic cutoff δ which denotes τ_H when $\gamma_t = \text{H}$ and τ_F when $\gamma_t = \text{F}$.

Let C_t^{bg} be the background correspondences for pair (I_t, I_{t+1}) . We dispatch to **F** (translation) or **H** (non-translation) using a translation surrogate $\tilde{\rho}_t$ with threshold τ_ρ (§4.2). Denote the residual of a match by r_i (either r_H^{sym} or r_F). With a *single global* inlier threshold δ (pixels), the inlier set and the per-pair metrics are

$$\mathcal{I}_t = \{ i \in C_t^{\text{bg}} \mid r_i \leq \delta \}, \quad \text{IR}_{\text{BG}}(t) = \frac{|\mathcal{I}_t|}{|C_t^{\text{bg}}|}, \quad \text{GE}_{\text{BG}}(t) = \text{median}\{ r_i \mid i \in \mathcal{I}_t \}. \quad (12)$$

Pairs with too few inliers (e.g., $|\mathcal{I}_t| < 30$ due to extreme blur/low texture) are marked invalid and excluded from aggregation.

For a clip, we summarize background consistency by robust medians over valid pairs:

$$\text{IR}_{\text{BG}}^{\text{clip}} = \text{median}_t (\text{IR}_{\text{BG}}(t)), \quad \text{GE}_{\text{BG}}^{\text{clip}} = \text{median}_t (\text{GE}_{\text{BG}}(t)). \quad (13)$$

Scene-balanced medians and 95% bootstrap CIs are reported across seeds within each category.

Within each category (C1–C6), we aggregate per-seed clip scores by a *scene-balanced* median, i.e., each seed (scene) contributes one score with equal weight: $\tilde{m}_{Ck} = \text{median}_s(m_s)$. To quantify uncertainty around the median, we report *95% nonparametric bootstrap confidence intervals* by resampling seeds with replacement ($B=10,000$ replicates) and taking the 2.5/97.5 percentiles of the bootstrap distribution of $\tilde{m}_{Ck}$. This avoids distributional assumptions and prevents categories or seeds with more regenerations from dominating the estimate. We use a family-specific inlier gate during robust fitting. Specifically, we use symmetric transfer error with $\tau_H=2$ px for a homography situation, and normalized Sampson with $\tau_F=0.1$ at the fundamental case. At report time, we apply a pixel-domain gate on GE fixed at $\delta=3.5$ px.

A geometrically consistent background exhibits *high* IR_{BG} and *low* GE_{BG} . Drops in IR indicate that a single global model cannot explain the pair (global inconsistency), while increases in GE at stable IR indicate imprecise alignment among otherwise consistent matches.

4.5 Foreground Consistency: AdaDFRC-W2

Dynamic/non-rigid pixels violate rigid-background assumptions. Including them in H/F fitting (i) pollutes inlier sets, (ii) destabilizes robust estimation, and (iii) degrades the diagnostic value of IR/GE for the background. Conversely, judging dynamics by background residuals rewards trivial solutions (shrinking inlier sets) and unfairly penalizes legitimate motion. A flow-guided score such as AdaDFRC-W2 directly targets motion plausibility, deformation continuity, and occlusion ordering, *complementing* IR/GE rather than conflating objectives.

To assess the geometric integrity of an individual moving object, which standard metrics fail to capture, we introduce the **Adaptive Dense Flow Regional Consistency (AdaDFRC-W2)** metric. This metric quantifies how well the internal motion of an object adheres to a rigid or near-rigid transformation, a key property of real-world objects.

Theoretical Foundation. The core principle of AdaDFRC-W2 is derived from the observation that the motion of a rigid object, when projected onto a 2D image plane under perspective projection, can be locally approximated by a 2D affine transformation. This approximation holds particularly well for objects that are relatively small in the field of view or distant from the camera. Consequently, for a geometrically consistent, rigid moving object, we expect its internal motion field—as captured by dense optical flow—to be highly structured and conform well to a single, dominant affine motion model. Any significant deviation from this model indicates non-rigid deformation, which in the context of AIGC, often manifests as generative artifacts like shearing, tearing, or texture "boiling." AdaDFRC-W2 is designed to measure the magnitude of this deviation precisely.

Derivation and Methodology. For each independently moving object k , identified by its mask M_t^k at time t , we compute its AdaDFRC-W2 score through a systematic, multi-step process detailed below. This process is designed to be robust and invariant to the object’s on-screen size and position.

- (1) **Adaptive Grid Normalization:** To standardize the analysis regardless of the object’s scale, we first define a bounding box around the mask M_t^k . This bounding box is then partitioned into a uniform $N \times N$ grid of cells. This adaptive grid ensures that our analysis has a consistent spatial resolution relative to the object itself.
- (2) **Robust Local Motion Summarization:** We compute a dense optical flow field from frame I_t to I_{t+1} within the object’s bounding box. For each grid cell (i, j) , we identify all the object pixels (as defined by M_t^k) that fall within it. A single, robust motion vector $\mathbf{v}_{i,j} \in \mathbb{R}^2$ is then computed for this cell by taking the median of the flow vectors of all its constituent pixels. The median is chosen for its robustness to outliers, effectively filtering out minor noise in the flow estimation. If a cell contains no object pixels, it is excluded from further analysis.
- (3) **Optimal Affine Model Fitting via Weighted Least-Squares:** The central step is to find the single affine transformation that best describes the collection of local motion vectors $\{\mathbf{v}_{i,j}\}$. An affine transformation models the motion of a point $\mathbf{p}$ as $\mathbf{p}' = \mathbf{A}\mathbf{p} + \mathbf{b}$, where $\mathbf{A}$ is a 2×2 matrix representing rotation, scaling, and shear, and $\mathbf{b}$ is a 2×1 translation vector. The motion vector is thus predicted as $\mathbf{v}_{\text{pred}}(\mathbf{p}) = \mathbf{p}' - \mathbf{p} = (\mathbf{A} - \mathbf{I})\mathbf{p} + \mathbf{b}$. We formulate this as a weighted least-squares problem. We seek the optimal affine parameters $(\mathbf{A}^*, \mathbf{b}^*)$ that minimize the sum of squared differences between the predicted motion and the observed local motion vectors $\{\mathbf{v}_{i,j}\}$. To give more importance to denser parts of the object, each cell (i, j) is assigned a weight $w_{i,j}$ equal to the number of object pixels it contains. The optimization problem is thus:

$$(\mathbf{A}^*, \mathbf{b}^*) = \arg \min_{\mathbf{A}, \mathbf{b}} \sum_{i,j} w_{i,j} \| (\mathbf{A}\mathbf{p}_{i,j} + \mathbf{b}) - \mathbf{v}_{i,j} \|_2^2 \quad (14)$$

where $\mathbf{p}_{i,j}$ is the geometric centroid of the grid cell (i, j) . This is a standard linear least-squares problem that can be solved efficiently in closed form.

- (4) **Score Calculation as Weighted Mean Squared Error (WMSE):** The AdaDFRC-W2 score for the object k at time t , denoted $\mathcal{D}_t^{(k)}$, is defined as the weighted mean squared error (WMSE) of this optimal fit. This value represents the residual, non-affine component of the motion field—the part that cannot be explained by a single rigid transformation.

$$\mathcal{D}_t^{(k)} = \frac{\sum_{i,j} w_{i,j} \| (\mathbf{A}^* \mathbf{p}_{i,j} + \mathbf{b}^*) - \mathbf{v}_{i,j} \|_2^2}{\sum_{i,j} w_{i,j}} \quad (15)$$

A low $\mathcal{D}_t^{(k)}$ score signifies high internal consistency, indicating that the object’s motion is well-described by a single affine model, which is characteristic of a rigid body. Conversely, a high score points to significant non-rigid deformation or other structural inconsistencies symptomatic of generative artifacts. The final consistency score for an object instance, $\mathcal{S}^{(k)}$, is computed by averaging $\mathcal{D}_t^{(k)}$ over its entire visible trajectory.

5 GeoCon-Bench Dataset: A Sophisticated Collection for Geometric Consistency

To transition geometric consistency from a qualitative afterthought to a first-class, quantitative evaluation target, we construct **GeoCon-Bench**, a standard, reusable, and expandable benchmark explicitly tailored to the challenges of geometry consistency among various prompt situations in AIGC videos. It is the first benchmark of its kind to systematically factorize geometric stressors, providing a controlled environment to probe model capabilities beyond semantic and aesthetic appeal. The dataset is designed to (i) rigorously test whether a single rigid motion model can

explain frame-to-frame changes, (ii) isolate independently moving objects for granular, per-object consistency analysis, and (iii) directly align with and validate our translation-aware gating mechanism.

5.1 Design Principles

Our benchmark is built on five core principles to ensure fairness, rigor, and utility:

- **Geometry-first, not aesthetics:** Prompts and scenes are constructed to emphasize camera motion, scene structure, and object rigidity over complex textures or artistic styles. This prevents models from hiding geometric flaws behind stylistic appeal and focuses the evaluation on structural integrity.
- **Orthogonal stressors:** Scenarios are carefully factored along controllable geometric axes (e.g., planarity of the background, presence of parallax, number of independent motion fields). This allows for targeted analysis of specific model weaknesses, such as a failure to model epipolar geometry while succeeding at projective transforms.
- **Model-agnostic generation:** To ensure a fair comparison, identical prompts and random seeds are used across all tested models (Sora [26], Runway, Wan [44]). Technical parameters such as resolution, frames per second (fps), and clip duration are also held constant.
- **Reproducibility:** Every generated sequence is accompanied by a detailed metadata record (prompt, seed, model version, resolution, fps, duration) and the scripts used for generation. This commitment to transparency allows for full verification and extension of our results.
- **Replaceable components:** The tools used in our evaluation pipeline (segmentation, feature matching, optical flow) are treated as default choices, not fixed requirements. The benchmark is structured to allow for drop-in alternatives (e.g., swapping RAFT for a future optical flow model), ensuring its long-term relevance as underlying technologies evolve.

5.2 Scenario Taxonomy

The dataset is partitioned into six distinct scenarios. We first include the **Static Scene** as a negative-control baseline. Its purpose is not to test complex 3D understanding but to evaluate temporal stability and the fidelity of executing a zero-motion command. It reveals texture flicker, unintended camera drift, and minor object instability. For a perfect static video, camera motion should be zero and appearance-consistency between frames should approach zero; any deviation directly quantifies inherent instability. Failing this simplest case indicates that errors in more complex scenarios likely stem from a lack of temporal coherence rather than motion understanding.

- 1) **Static Scene (Temporal Stability Baseline):** see above.
- 2) **Pure Camera Rotation (Homography-dominant):** A static 3D scene where the camera rotates around its optical center. This is a foundational test of projective consistency. *Expected Behavior:* The motion should be classified as *non-translation-dominant*. A single homography (H) should explain nearly all background feature matches, resulting in an Inlier Ratio (IR) approaching 100% and a very low Geometric Error (GE).
- 3) **Pure Camera Translation (Epipolar-dominant):** A static scene with significant depth variations (e.g., near and far objects) where the camera translates laterally or forward. This is a critical test of a model’s ability to synthesize motion parallax. *Expected Behavior:* The motion must be classified as *translation-dominant*. A single fundamental matrix (F) should model the epipolar geometry, yielding a high IR and low GE. Failure often manifests as scene "breathing" or warping, where distant objects do not remain stable relative to near ones.

- 4) **Static Background + Single Moving Object:** A static camera observes a scene where a single foreground object moves independently. This scenario tests the model’s ability to disentangle different motion fields. *Expected Behavior:* The background should remain perfectly rigid, passing the H/F test with high scores. The foreground object is evaluated for internal rigidity using AdaDFRC-W2 (Eq. (15)); a low score indicates a rigid object, while a high score reveals non-rigid deformations like stretching or shearing.
- 5) **Complex Camera Motion (Rotation+Translation):** Realistic, combined camera movements such as a dolly-zoom or arc shot. This probes the model’s ability to maintain a coherent 3D representation under more challenging, free-form motion. *Expected Behavior:* The motion should be classified as *translation-dominant*. The evaluation checks if a single fundamental matrix (F) can still robustly model the background motion, despite its complexity, and if foreground objects (if any) can be stably decomposed and tracked.
- 6) **Geometric Stress Tests:** Adversarial but physically-plausible setups designed to push models to their limits. This includes scenes with thin, detailed structures (prone to disappearing), reflective or transparent surfaces (challenging for feature matching), and occlusions. *Expected Behavior:* We anticipate a significant drop in performance across all metrics. These scenarios are diagnostic, designed to reveal failure modes like structural "breathing," object wobble, or non-rigid drift that might not be apparent in simpler scenes.

5.3 Prompt Templates and Lexicon

To ensure precise control over the generated geometry and minimize semantic ambiguity, we adopt a structured, modular prompt system. This forces the models to contend with the geometric request directly, rather than relying on semantic shortcuts. Each prompt uses four slots: Motion Command + Scene Composition + Object Specification + Qualitative Style. A comprehensive prompt lexicon, detailing the allowed vocabulary for each slot, is provided in an appendix (cf. Appendix A) to standardize wording.

Examples (abbrev.).

- **Pure Rotation:** “Single static shot. The camera rotates in place around the building facade. No translation. Background absolutely fixed.”
- **Pure Translation:** “Single static shot. Camera dollies right past trees and a building, revealing parallax (near vs far).”
- **Static BG + Object:** “Static camera. A red car drives a circular path in front of a fixed building. Background must stay rigid.”

5.4 Generation Protocol and Dataset Statistics

All clips in GeoCon-Bench are generated under a strict protocol. In addition to the released seed-image scenes, we include standardized videos from three representative AIGC models, including *Sora* (commercial), *Runway* (industry-leading), and *Wan* (open-source). Each clip is generated at a fixed resolution and 24 fps, with a duration of 5 seconds (120 frames). For each of the 20 unique prompts, we generate 3 clips per model using different seeds. We advocate for *scene-balanced* reporting, where metrics are first averaged within each scenario before being averaged across scenarios. This prevents a model’s high performance on an easy category from masking its failures on more challenging ones.

5.5 Metadata and Release Format

To maximize the benchmark’s utility and promote reproducible science, each clip is accompanied by a JSON metadata file containing: {prompt, model, model_version, seed, resolution, fps, duration, clip_id, masks_path}. We will publicly release (i) all prompts, (ii) all generated videos, (iii) all extracted foreground masks, (iv) evaluation configurations, and (v) scripts to reproduce all tables and figures in this paper.

5.6 Grounded Validation

To anchor our metrics and calibrate key thresholds, the benchmark includes a small subset of "grounded" scenarios where the intended geometry is unambiguous. These include textured planar surfaces for rotation checks (which should yield a perfect homography with $IR=100\%$) and scenes with clearly separated near/far layers for translation checks. These anchors provide a sanity check for the entire evaluation pipeline and are used to calibrate the gating threshold τ_p by observing the IR saturation across these known scenarios.

6 Experimental Validation

We conducted a comprehensive set of experiments to validate our proposed framework and evaluate the geometric consistency of state-of-the-art video generation models using a custom benchmark, GeoCon-Bench. The goal is not merely to rank models but to establish a standardized, interpretable procedure for judging geometric fidelity.

We investigate: **RQ1 (Gate validity)**: Does translation-evidence gating agree with photogrammetric fits across scenes and platforms? **RQ2 (Background rigidity)**: Are $IR_{H/F}$ and $GE_{H/F}$ stable under dual-threshold scanning and informative for cross-platform comparison? **RQ3 (Dynamic coherence)**: Does the flow-guided dynamic score (AdaDFRC-W2) capture complementary failures that background indicators cannot? **RQ4 (Comparability & auditing)**: Do percentile-aligned summaries and audit cards yield stable platform rankings and reproducible diagnostics?

6.1 Implementation Details

This section details the specific choices of algorithms and parameters used to implement our framework, clearly separating the practical implementation (cf. pseudocode Algorithm 1) from the theoretical methodology of Section 4.

The abstract components of our framework were realized with the following state-of-the-art models:

- **Pose Estimation**: We used the official implementation of **DUST3R** for pairwise relative pose estimation. Its ability to handle uncalibrated images makes it robust for AIGC evaluation.
- **Video Segmentation**: Foreground object masks were generated using **SAM-2**, leveraging its streaming architecture for efficient video processing.
- **Optical Flow**: Dense flow fields for the AdaDFRC-W2 metric were computed with Recurrent All-Pairs Field Transforms (**RAFT-small**), chosen for its balance of accuracy and efficiency.

6.2 DUST3R Translation-Gate Calibration

Given the DUST3R relative pose between the first/last frames, we declare a pair *translation-dominant* (F) if $\|t\|_\infty > \tau_t$ and rotation/planar-like (H) otherwise, where $t_\infty \triangleq \|t\|_\infty = \max(|t_x|, |t_y|, |t_z|)$. Throughout the paper, background inlier gates are kept fixed to empirically chosen constants and are *not* ablated: $\tau_H=2$ px (homography reprojection) and $\tau_F=0.1$ (Sampson residual); when background consistency is required we use $IR \geq 0.97$ and $GE \leq 3.5$ px on the content mask (pixels, after denormalization).

Algorithm 1: Mask-first, translation-aware evaluation. Residuals are evaluated in *pixels* on the content mask; the background model (H or F) is chosen once from the DUS3R decision on (I_0, I_T) .

Input :Seed I_0 ; generated frames $I_{1:T}$; scales $S = \{1/4, 1/2, 1\}$;
DUS3R translation gate τ_t (default **0.04**); fixed background gates $\tau_H=2$ px, $\tau_F=0.1$.
Output: $\text{IR}_{\text{BG}}, \text{GE}_{\text{BG}}$; per-object $\mathcal{S}^{(O)}$; detected category.

A. DUS3R translation decision (first-last)
estimate pose $\mathbf{T} = (\mathbf{R}, \mathbf{t})$ between (I_0, I_T) ;
if *pose available* **then** set $t_\infty \leftarrow \|\mathbf{t}\|_\infty$, translation_flag $\leftarrow (t_\infty > \tau_t)$
else translation_flag $\leftarrow \text{F}$ (present-only policy)

B. SAM-2 segmentation (mask-first)
run SAM-2 $\rightarrow \{M_t^k\}$; define content mask (remove letterbox) and background $B_t = \text{content}(I_t) \setminus \cup_k M_t^k$;
build background-only correspondences $C_{t,s}^{\text{bg}}$ on $B_t \cap B_{t+1}$ for all t and $s \in S$.

C. Background geometry on masks-outside
for $t = 0$ **to** $T - 1$ **do**
 for $s \in S$ **do**
 if translation_flag **then**
 fit $\mathbf{F}_{t,s}$ on $C_{t,s}^{\text{bg}}$ (robust, gate τ_F); compute IR_F and median Sampson residual GE_F (pixels);
 else
 fit $\mathbf{H}_{t,s}$ on $C_{t,s}^{\text{bg}}$ (robust, gate τ_H); compute IR_H and median symmetric reprojection GE_H (pixels);
aggregate over time/scales by robust medians to obtain $\text{IR}_{\text{BG}}, \text{GE}_{\text{BG}}$;
(*optional*) also record best-of(H/F) per pair for analysis, without affecting the main scores.

D. Movers: AdaDFRC-W2
for each object O **and** $s \in S$ **do**
 estimate dense flow and occlusion inside O ; grid the bbox, take per-cell median flow $v_{i,j}$;
 fit affine $(\mathbf{A}, \mathbf{b})$ and compute weighted MSE per frame $\mathcal{D}_t^{(O)}$; average over the track $\rightarrow \mathcal{S}^{(O)}$.

E. Post-hoc categorization
combine translation_flag with $(\text{IR}_{\text{BG}}, \text{GE}_{\text{BG}})$ under fixed report gates ($\text{IR} \geq 0.97$, $\text{GE} \leq 3.5$ px) to assign one of
 $\{\text{Static}, \text{Rotation-dominant}, \text{Translation-dominant}, \text{StaticCam}+\text{MovingObjs}, \text{Complex}, \text{Stress}\}$.
return $\text{IR}_{\text{BG}}, \text{GE}_{\text{BG}}, \{\mathcal{S}^{(O)}\}$.

For the calibration of τ_t we use *present-only* pairs (i.e., DUS3R poses available; “fail” pairs have undefined t_∞), across all three platforms and all six prompt families. The sampling unit is a (*seed, platform, family*) triplet, yielding $N=300$ observations with valid t_∞ (**146** F-expected: C3/C5/C6; **154** H-expected: C1/C2/C4). This analysis is purely geometric and independent of semantic scores.

Fig. 4(a) shows smoothed densities of t_∞ for H-expected (C1/2/4) and F-expected (C3/5/6) families. Densities are reported as equal-width histograms over the central 99% range with a mild Gaussian smoothing (for visualization only; no decision is made on smoothed values). A clear valley emerges around $t_\infty \approx 0.04$, indicating a natural separation between the two regimes; choosing τ_t in that valley minimizes overlap of the class-conditional distributions.

To quantify the separation without assuming parametric forms, we estimate the empirical calibration curve $P(F | t_\infty)$ using equal-count binning and Wilson 95% confidence intervals (CIs): for each bin b with m_b samples and $\hat{p}_b$ the observed F fraction, we report

$$\left[\frac{\hat{p}_b + \frac{z^2}{4m_b}}{1 + \frac{z^2}{m_b}} \pm \frac{z}{1 + \frac{z^2}{m_b}} \sqrt{\frac{\hat{p}_b(1-\hat{p}_b)}{m_b} + \frac{z^2}{4m_b^2}} \right], \quad z=1.96. \quad (16)$$

Fig. 4(b) exhibits a smooth transition from H-like to F-like regimes. With $K=12$ equal-count bins (about 25 samples per bin), the 0.5 crossing lies in the interval $t_\infty \in [0.04, 0.07]$; notably, the bin centered at $t_\infty \approx 0.043$ has $\hat{p}=0.40$ with a

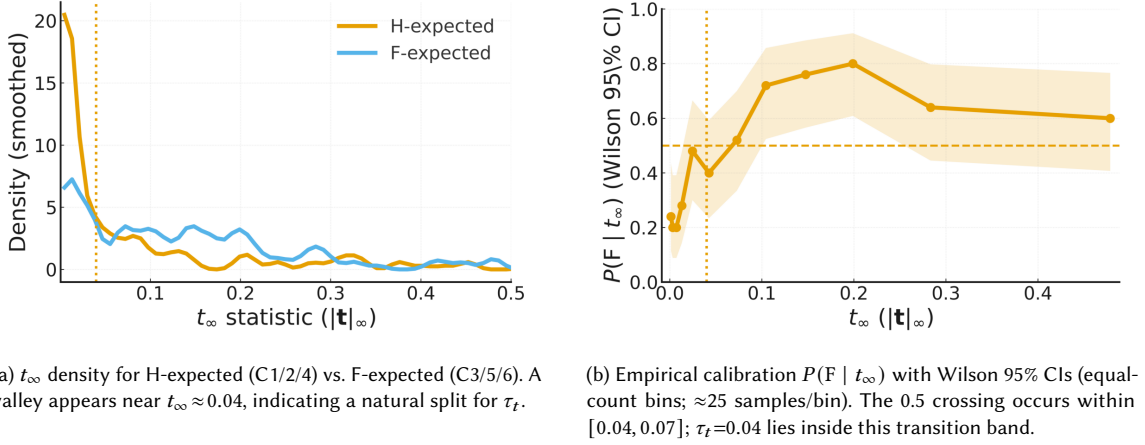


Fig. 4. Translation-gate calibration for pose estimation.

Wilson CI that *contains* 0.5, while the next bin centered at ≈ 0.073 has $\hat{p} = 0.52$ (CI also overlapping 0.5). Hence, $\tau_t = 0.04$ lies *inside* the statistically identified transition band, on the conservative side of the equiprobable boundary.

We therefore adopt $\tau_t = 0.04$ as the default translation gate: (i) it falls in the density valley separating the H/F families, thereby reducing class overlap; and (ii) it lies within the empirical calibration band where $P(F | t_\infty) \approx 0.5$, providing a principled knee-point that balances recall on F-expected cases (C3/5/6) against false positives on H-expected ones (C1/2/4). Complementary analyses in the Appendix B show that $\tau_t \in [0.04, 0.06]$ forms a broad performance plateau, and that $\tau_t = 0.04$ is either at or statistically indistinguishable from the maxima while being slightly more conservative on spurious F.

6.3 GeoCon Benchmark

We evaluate whether camera *translation* is dominant (“F”) or not (“H”) across three video-generation platforms (Runway, Sora, Wan) and six prompt categories. For each clip, we feed the *first* and *last* frames to DUST3R and extract the 4×4 relative pose; we compute Eq. 5 from the translation column and declare translation-dominant (“F”) iff $\|t\|_\infty > \tau$ with $\tau = 0.04$; otherwise we label “H”. Unless otherwise stated, all descriptive statistics below are computed over the *available* clip-pairs (i.e., no imputation), while some visualizations additionally show a conservative variant that treats missing pairs as F. For reporting raw coverage, we count a pair as “present” only if DUST3R returns a valid pose. Thus, the Sora split exhibits fewer “present” pairs not because videos were absent, but because DUST3R could not reconstruct a larger fraction of first/last-frame pairs. In all downstream analyses that depend on semantic consistency, these non-reconstructed pairs are treated as F by construction. Unless stated, statistics are computed over *present-only* pairs (DUST3R pose available). For robustness audits, we also report a conservative variant that treats missing pairs as F, reflecting that large endpoint displacement often prevents pose recovery.

We evaluate three platforms (Runway, Sora, Wan) over six categories and multiple seeds. DUST3R returns valid poses for 108 pairs on Runway, 92 on Sora, and 100 on Wan. The number of distinct seeds with at least one valid pair is 20 (Runway), 19 (Sora), and 20 (Wan), so the maximum possible pairs per platform are 120/114/120, leaving

12/22/20 missing (failure) pairs, respectively. Per-category coverage (present pairs divided by #seeds) is: Runway {100, 85, 95, 85, 95, 80}% for C1–C6; Sora {84.2, 84.2, 78.9, 78.9, 78.9, 78.9}%; Wan {90, 80, 80, 95, 85, 70}%.

On present pairs, the distributions are: Runway (count 108): mean 0.093, median 0.042, max 0.580; Sora (count 92): mean 0.129, median 0.050, max 0.691; Wan (count 100): mean 0.125, median 0.076, max 0.548. Overall, $21/300 = 7.0\%$ of decisions lie within ± 0.01 of the threshold $\tau=0.04$ (Runway 8.3%, Sora 5.4%, Wan 7.0%), indicating a non-negligible band of near-threshold cases. DUST3R failures indicate excessive translation and are treated as translation-present; such pairs are excluded from present-only aggregation but counted in all-pairs coverage.

As expected, semantics intended to contain translation (C3, C5, C6) yield higher F-rates, while C1/C2/C4 lean toward H. Per-platform F-rates (fraction of F among present pairs) are:

- **Runway:** C1 30.0%, C2 58.8%, C3 84.2%, C4 11.8%, C5 68.4%, C6 56.2%; overall 51.9%.
- **Sora:** C1 25.0%, C2 50.0%, C3 46.7%, C4 46.7%, C5 66.7%, C6 80.0%; overall 52.2%.
- **Wan:** C1 33.3%, C2 62.5%, C3 87.5%, C4 26.3%, C5 94.1%, C6 64.3%; overall 60.0%.

Notably, Sora’s C3 is closer to the decision boundary (only 46.7% F), whereas Wan is extremely translation-heavy on C5 (94.1% F).

By *design*, categories C1 (Static), C2 (Rotation-dominant), and C4 (Static camera + moving object) are expected to be **H-majority** (translation-free at the first/last endpoints), whereas C3/C5/C6 are expected to be **F-majority**. Empirically, we observe **C1 and C4 behave as expected (H-majority)**. F-rates for C1 are 30.0%/ 25.0%/ 33.3% (Runway/Sora/Wan); for C4 they are 11.8%/ 46.7%/ 26.3%. Thus, C1 is consistently H-majority; C4 is H-majority on Runway/Wan and borderline on Sora. **C2 deviates from expectation**. Despite the rotation-only intent, C2 shows elevated F: 58.8%/ 50.0%/ 62.5% (Runway/Sora/Wan), i.e., not H-majority on Runway/Wan and exactly balanced on Sora. This suggests either residual translation between endpoints, moving content that induces effective parallax, or reconstruction biases on low-parallax/rotation-heavy pairs. **C3 and C5 are strongly F-majority**, as intended: C3 has 84.2%/ 46.7%/ 87.5% F; C5 has 68.4%/ 66.7%/ 94.1% F (Runway/Sora/Wan). **C6 (stress) is also F-leaning**: 56.2%/ 80.0%/ 64.3% F.

Using the expected rule (C1/C2/C4 $\rightarrow$ H; C3/C5/C6 $\rightarrow$ F), an *anomaly* occurs when the DUST3R gate disagrees with the expectation. Present-only anomaly rates are:

- **Runway:** C1 30.0%, C2 58.8%, C3 15.8%, C4 11.8%, C5 31.6%, C6 43.8%; overall 31.5%.
- **Sora:** C1 25.0%, C2 50.0%, C3 53.3%, C4 46.7%, C5 33.3%, C6 20.0%; overall 38.0%.
- **Wan:** C1 33.3%, C2 62.5%, C3 12.5%, C4 26.3%, C5 5.9%, C6 35.7%; overall 29.0%.

Thus, C2 is systematically difficult (anomaly $\geq 50\%$ on Runway/Wan), while C1/C4 largely conform to expectations (lower anomalies, aside from Sora’s borderline C4).

When reconstruction failures are set to F prior to the check (our default protocol), overall anomaly rates become 33.3% (Runway), 39.5% (Sora), and 30.0% (Wan). By category: Runway {30.0, 65.0, 15.0, 25.0, 30.0, 35.0}%; Sora {36.8, 57.9, 42.1, 57.9, 26.3, 15.8}%; Wan {40.0, 70.0, 10.0, 30.0, 5.0, 25.0}% (C1–C6). These increases concentrate in H-expected categories (C1/C2/C4), reflecting that failures are interpreted as large-displacement F.

Fig. 5 merges cross-platform alignment and semantic consistency into a single panel. Each cell is color-coded by the DUST3R-derived $\|t\|_\infty$ and overlaid with the H/F gate; orange borders denote anomalies under the expected rule (C1/C2/C4 $\rightarrow$ H; C3/C5/C6 $\rightarrow$ F), while gray borders indicate agreement. Rows list seeds in the fixed order 1–20; columns enumerate category (C1–C6) $\times$ platform (R/S/W for Runway/Sora/Wan). Missing pairs follow our default convention (treated as F and colored with the global maximum $\|t\|_\infty$ to preserve a common scale). Under this convention, 46.7% of seed–category triplets are unanimous across platforms; of those, 76.8% are all-F, reflecting translation-dominant or

failure-inferred cases. The combined view highlights three trends: (i) C3 and C5 align with translation dominance across platforms; (ii) C1 and C4 are mostly H with scattered mismatches; (iii) C2 departs from its H expectation, showing noticeable F decisions and borderline behavior near the threshold.

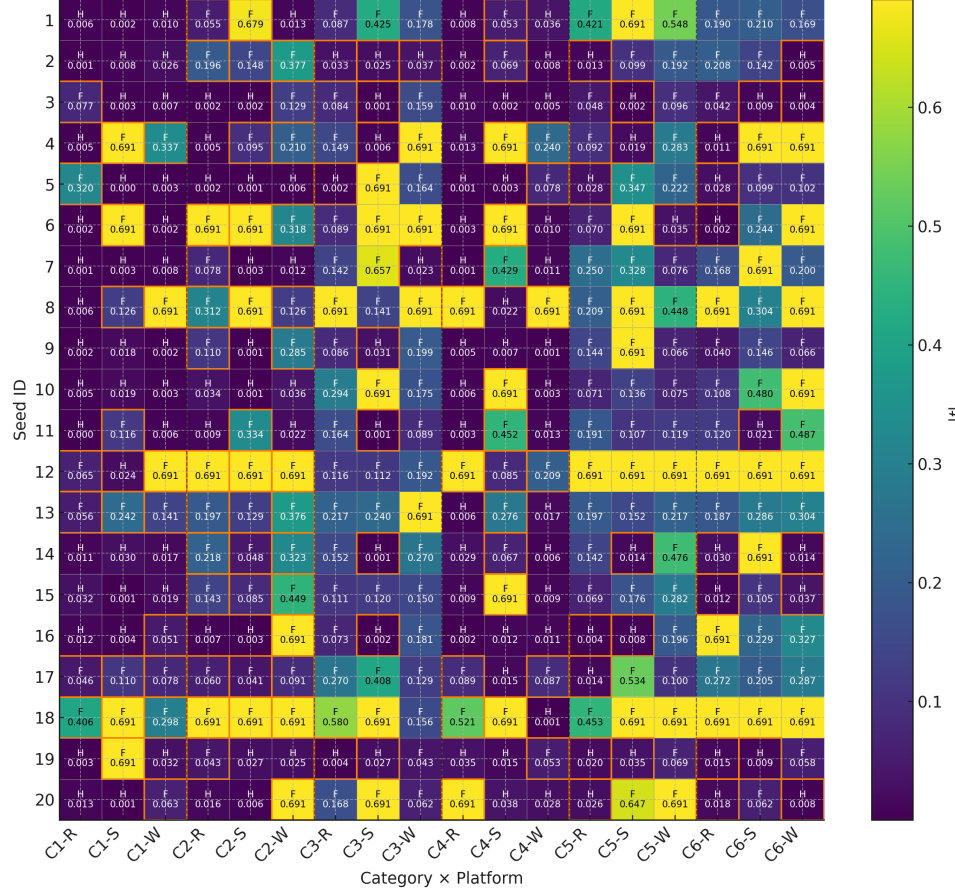


Fig. 5. Combined cross-platform alignment and semantic consistency. Color encodes $\|t\|_\infty$ from the DUST3R first–last pose; overlaid labels show the H/F decision. Orange borders mark anomalies under the expected rule (C1/C2/C4 $\rightarrow$ H; C3/C5/C6 $\rightarrow$ F), gray borders indicate matches. Missing pairs are treated as F and colored with the global maximum to keep a common scale. Rows follow the seed order 1–20; columns are category (C1–C6) $\times$ platform (R/S/W).

6.4 Sanity checks and gate–support agreement

For each clip, we read the DUST3R first–last pose $T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$, compute $t_\infty = \max(|t_x|, |t_y|, |t_z|)$, and apply a hysteresis gate: $t_\infty \leq 0.03 \Rightarrow H$, $t_\infty \geq 0.05 \Rightarrow F$, and $0.03 < t_\infty < 0.05$ is a tolerance band ("U"). Unless stated, Sankey plots drop U to avoid three-way clutter, whereas diagnostics report it explicitly.

Across all available tuples (platform $\times$ seed $\times$ prompt) we obtain $F = 156$ (52%), $H = 123$ (41%), and $U = 21$ (7%). Fig. 6a shows the global histogram of t_∞ with the two thresholds overlaid. The tolerance band $[0.03, 0.05]$ is sparsely occupied, supporting the defensibility of the gate.

We also visualize the full distributions with a violin plot in Fig. 6b. C1 (static scene) and C4 (static camera + moving objects) concentrate below 0.03 (H regime), C3 (translation-dominant) and C5 (complex camera motion) shift mass above 0.05 (F regime), while C2 (rotation-dominant) spans both sides, indicating that even small translations or scale drift can flip the decision. C6 (geometric stress tests) is intentionally unconstrained and shows an F-skewed but mixed distribution: 28/45 F, 14/45 H, 3/45 U (62.2% F, 31.1% H, 6.7% U), consistent with stress conditions that elicit both regimes.

We define gate margins as $m_H = 0.03 - t_\infty$ for H-gated tuples and $m_F = t_\infty - 0.05$ for F-gated tuples. Fig. 6c shows the margin histograms for H and F, respectively: both concentrate away from zero, indicating decisions are not brittle. Fig. 6d expose platform effects (left: m_H , right: m_F). Margins remain comfortably positive overall, but distributions differ across platforms, reflecting platform-specific motion priors under identical prompts.

On (seed, prompt) tuples where at least two platforms are non-U, 55/103 are unanimous (consensus) and 48/103 are split (F vs. H), yielding a 53.4% consensus rate. These differences are visible directly in the platform-conditioned margin violins in Fig. 6d. For H-gated tuples, the distributions of $m_H = 0.03 - t_\infty$ differ by platform, and for F-gated tuples, the distributions of $m_F = t_\infty - 0.05$ likewise separate across platforms. Margins remain largely positive, indicating decisions are not brittle, yet their platform-specific shifts corroborate the presence of distinct motion/rendering biases under identical prompts.

6.5 Background rigidity: distributional behavior, cross-model gaps, and effect sizes

We evaluate background rigidity using two complementary scalars computed per tuple: IR (image-space rigidity consistency; larger is better) and GE (background geometry error; smaller is better). We report both the **H** family (thresholded at $\tau_H = 2\text{px}$) and the **F** family (gate near $\tau_F \approx 0.10$), and we summarize per-prompt distributions (C1–C6) for each platform (Runway, Sora, Wan). Distributions are visualized with ridgelines (per prompt) and cross-family gaps are quantified with Gardner–Altman style effect-size panels (median differences with bootstrap 95% CIs).

Fig. 7 shows that **IR** concentrates in a high-rigidity regime with visibly broader right tails under the stress prompts **C4** (static camera + moving objects) and **C6** (geometric stress tests) across all three platforms. The corresponding **GE** ridgelines in Fig. 8 reveal rightward shifts (i.e., larger errors) for **H** relative to **F**, and again heavier upper tails in **C4/C6**. Together, the ridgelines indicate (i) background remains mostly rigid, but (ii) stress prompts inject a non-negligible fraction of difficult cases whose mass accumulates in the upper tails (higher GE / slightly lower effective IR).

To make platform-wise comparisons commensurate, we use median differences with bootstrap uncertainty. Fig. 9 reports $\text{Median}(\text{GE}_H) - \text{Median}(\text{GE}_F)$: the effect is *consistently positive* on all platforms, meaning **F** has lower background geometry error than **H**. The magnitude is **small on Runway** (on the order of a few 10^{-2}), **medium on Sora** (peaking in C5/C6), and **largest on Wan** (systematically the highest across C1–C6). This ranking matches the qualitative right-shift in the **GE** ridgelines for C4/C6 on Wan. Fig. 10 reports $\text{Median}(\text{IR}_H) - \text{Median}(\text{IR}_F)$: the effect is *consistently negative* (i.e., **F** yields slightly higher rigidity) with the strongest gaps on Sora/Wan and milder gaps on Runway. Across platforms, the signs are stable, the CIs are tight, and the ordering by prompt mirrors the stress semantics: **C4** and **C6** show the most significant gaps, while **C1–C3/C5** remain smaller and often overlap.

The **F** family, which enforces epipolar-consistent background structure, systematically reduces GE and *increases* IR (higher rigidity), with effects most pronounced under stress prompts (C4/C6)—see the rightward density mass in Fig. 8

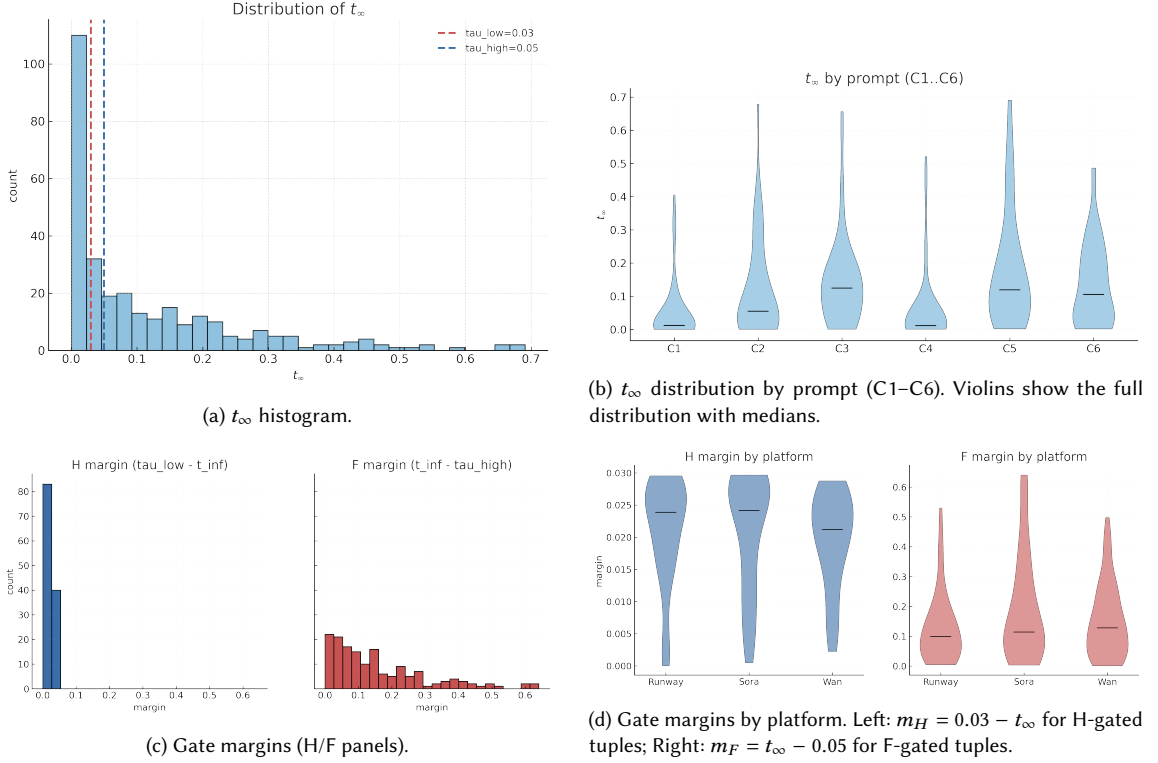


Fig. 6. Gate-support diagnostics with the hysteresis rule ($t_\infty \leq 0.03 \Rightarrow \text{H}$, $t_\infty \geq 0.05 \Rightarrow \text{F}$). (a) t_∞ histogram with thresholds, showing low occupancy of the tolerance band. (b) Prompt-wise t_∞ distributions (C1–C6; violins with medians): C1/C4 concentrate below 0.03, C3/C5 shift above 0.05, and C2 spans both regimes. (c) Gate margins $m_H = 0.03 - t_\infty$ and $m_F = t_\infty - 0.05$ (two panels) are largely positive, indicating stable decisions away from thresholds. (d) Platform-conditioned margin violins (left: m_H , right: m_F) reveal cross-platform differences while remaining comfortably above zero.

and the positive GE gaps in Fig. 9, together with the negative IR gaps in Fig. 10. Wan exhibits the largest GE gaps (H worse than F) across nearly all prompts; Sora shows medium gaps concentrated in C5/C6; Runway shows the smallest gaps. The same ranking is echoed in **IR**, supporting a consistent story across both metrics. (3) *Stress prompts magnify tails rather than medians*. Ridgelines show that medians stay near the high-rigidity regime, while the upper tail mass expands in C4/C6 (Fig. 7, Fig. 8). Consequently, the effect-size panels capture meaningful, prompt-aligned differences even when central tendency moves only slightly.

The Gardner–Altman representation in Fig. 9–10 (half-violin density + CI + point estimate) provides (i) robust ranking immune to outliers from stress cases; (ii) cross-platform comparability via a common horizontal scale; and (iii) visually verifiable uncertainty. These properties are crucial because stress prompts primarily affect the tails, which can mislead mean-based summaries but are faithfully captured by medians and their sampling distribution.

Across three datasets and six prompts, the evidence is consistent and convergent: *the F family dominates the H family for background rigidity*, reducing GE and improving IR, with the largest wins under C4/C6 and the strongest overall gaps on Wan, followed by Sora, then Runway. This agrees with the qualitative mass shifts in the ridgelines and the quantitative, uncertainty-aware effect sizes.

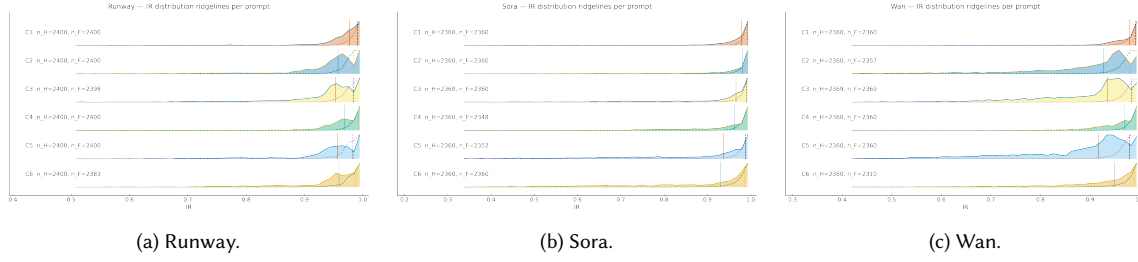


Fig. 7. IR distribution ridgelines per prompt (C1–C6). Each panel overlays H (solid, filled) vs. F (dashed) for a given platform. Stress prompts C4/C6 widen the upper tails across all platforms.

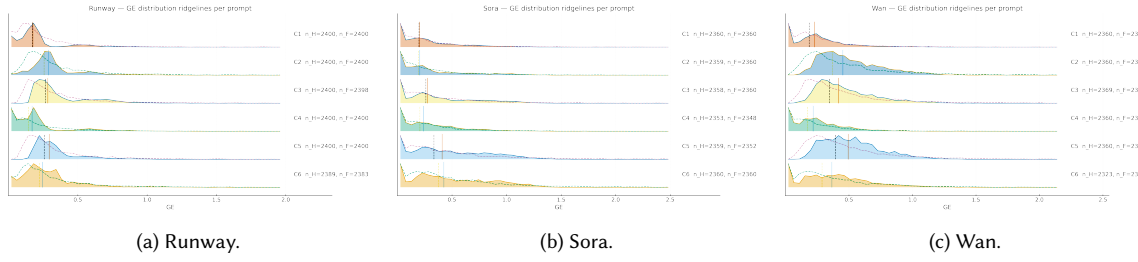


Fig. 8. GE distribution ridgelines per prompt (C1–C6). H shows right-shifted densities relative to F, with the largest gaps on Wan and pronounced tail mass in C4/C6.

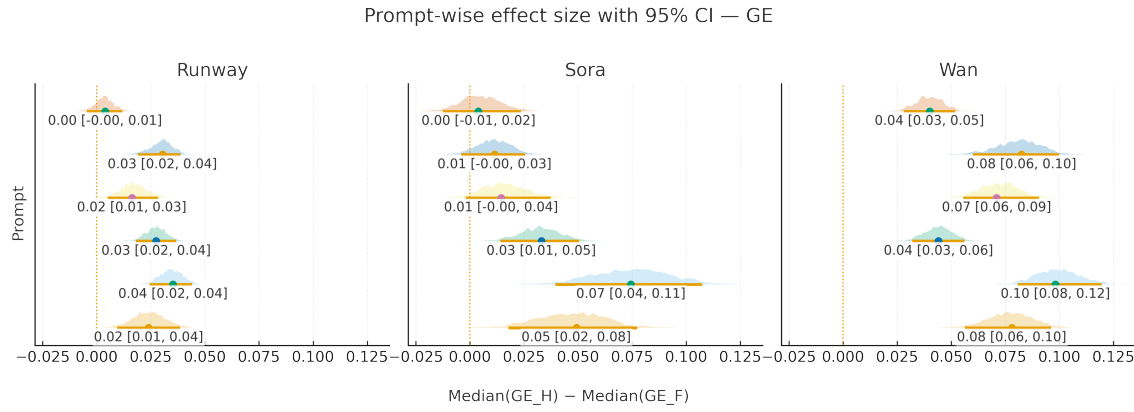


Fig. 9. Prompt-wise effect size with 95% CI — GE. Median difference $\text{Median}(\text{GE}_H) - \text{Median}(\text{GE}_F)$ per prompt and platform. Effects are positive across the board ($F < H$), smallest on Runway, medium on Sora, largest on Wan, and amplified for C4/C6.

6.6 Dynamic coherence: flow-guided plausibility and occlusion ordering

We assess dynamic plausibility with the per-frame AdaDFRC-W2 score, designed to increase when occlusion ordering or flow consistency becomes implausible. For each object track within a clip we summarize its temporal severity by the robust upper-tail statistic

$$\text{AdaPninety} = \text{quantile}_{0.9} \left(\{ \text{AdaDFRC-W2}_t \}_{t=1}^T \right), \quad (17)$$

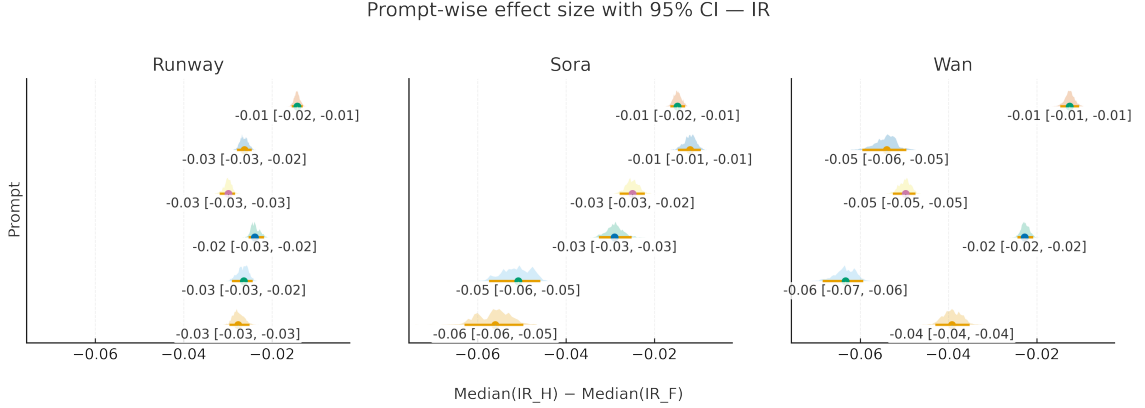


Fig. 10. Prompt-wise effect size with 95% CI — IR. Median difference $\text{Median}(\text{IR}_H) - \text{Median}(\text{IR}_F)$ per prompt and platform. Effects are negative ($F > H$) with the strongest gaps on Sora/Wan and the mildest on Runway; the ordering by prompt mirrors C4/C6 stress.

which captures the magnitude of the worst 10% frames while being insensitive to isolated spikes.

We rely on four complementary views. (i) *Beeswarm panels* group every object by prompt (C1–C6) and model, and overlay the per-prompt, per-model P80 thresholds (dashed ticks) and median markers (short bars), see Figs. 11a–11c. (ii) *Empirical Cumulative Distribution Function (ECDF) panels* show full distributions with the same P80 thresholds, exposing tail heaviness (right-shifts). (iii) *Anomaly-rate bars* report the fraction of objects above the *per-prompt* P80; these rates are primarily descriptive (they are near 20% by construction and not used for cross-prompt ranking). (iv) *Top-N sparkline sheets* plot the per-frame trajectories for the hardest objects in each model with summary statistics ($P50/P90/n$), revealing whether failures are bursty or persistent (Figs. 12a–12c).

Dynamic coherence is only defined when detectable objects exist. Across our data, the object counts per model are comparable, providing sufficient support for model- and prompt-conditioned analyses. Prompts C4 (static camera + moving objects) and C6 (geometric stress tests) contain the richest dynamic evidence, consistent with our protocol design.

We report consistent, model-agnostic evidence that dynamic coherence is most fragile under prompts C4 (static camera + moving objects) and C6 (geometric stress tests). Across all three models, the beeswarm panels reveal visibly taller stacks for C4/C6 while the medians remain comparatively stable, and the ECDFs shift to the right with fatter upper tails. This pattern indicates that failures concentrate in the worst part of the distribution rather than uniformly raising central tendency—exactly the behavior one would expect when motion- or geometry-induced occlusion ordering becomes implausible.

Although all models exhibit heavier tails for C4/C6, the magnitude of the extremes differs across models. The largest AdaPninety objects—and thus the most severe dynamic inconsistencies—are most pronounced on Runway and Sora, with Wan displaying comparatively lower maxima. This cross-model ordering is independently corroborated by the Top-N sparkline sheets: the largest-amplitude trajectories cluster in (Runway, C6) and (Sora, C6/C4), whereas Wan’s top traces remain lower in amplitude, albeit still clearly abnormal.

The temporal shapes of these failures are mechanistically distinct. The Top-N trajectories expose (i) brief burst spikes aligned with object enter/exit events or edge flicker; (ii) quasi-periodic rise–fall patterns consistent with recurrent

depth/flow flips; and (iii) long segments of persistent elevation indicative of sustained layer mis-ordering or non-rigid flow leakage. Notably, these dynamic signatures often arise without commensurate increases in background photogrammetric residuals, underscoring that they are genuinely dynamic (flow/ordering) failures rather than artifacts of the static fit.

We also clarify how prevalence is interpreted in our diagnostics. Because thresholds are defined *per prompt and per model* at the P80 level, the within-prompt anomaly fractions are expected to be near 20% by construction and are therefore not used to rank prompts. Instead, our conclusions rest on *severity* (how far the upper tail extends) and *tail shape* (ECDF right-shift), which together yield a clear and stable ordering in upper-tail mass: $C4 > C6 > (C1, C2, C3, C5)$. Joint checks with gate-support agreement (Sec. 6.4) further show a non-trivial subset of clips in which dynamic anomalies coexist with strong background fits ("dynamic-anomaly & background-good"), amounting to roughly a tenth of cases. This complementarity motivates including an explicit, object-aware dynamic score such as AdaDFRC-W2 alongside background rigidity fits in any holistic geo-consistency audit.

Two robustness choices make the conclusions stable: (i) summarizing per-object with AdaPninety instead of maxima prevents overreaction to isolated spikes, yet keeps genuine tail mass visible; (ii) auditing with model- and prompt-specific P80 thresholds prevents cross-model scale confounding while preserving comparability of *shapes* and *extremes*. Empirically, Top- N membership is stable under small changes in N (8–12) and under ± 2 percentile changes in the per-prompt thresholds.

Dynamic coherence is the dominant failure axis under motion- and geometry-stress prompts. It manifests as heavy-tailed AdaPninety distributions with clear bursty or persistent temporal signatures, and it is *not* predictable from background photogrammetric residuals alone. Therefore, any holistic geo-consistency audit must include an object-aware, flow-guided dynamic term like AdaDFRC-W2 in addition to background rigidity fits.

6.7 Quantitative Results

We present the scene-balanced average scores for each generative model across all categories of GeoCon-Bench in Table 1. Higher IR is better, while lower GE and AdaDFRC-W2 are better. We report scene-balanced averages over all seeds and categories of GeoCon-Bench. For each seed–category pair we first average frame-wise scores within a clip and then average equally across scenes. Background scores use H/F-based metrics (IR and GE) on the static layer, while foreground scores report AdaDFRC-W2 on moving objects. Higher IR is better; lower GE and AdaDFRC-W2 are better.

Table 1. Scene-balanced average results on GeoCon-Bench

Model	Background		Foreground
	IR (%) $\uparrow$	GE (norm) $\downarrow$	AdaDFRC-W2 $\downarrow$
Runway	97.0	0.274	18.54
Sora	96.6	0.398	21.79
Wan	96.2	0.394	12.76

Background geometry is consistently strong across models ($IR \approx 96\text{--}97\%$), with sub-pixel GE magnitudes on average. Foreground coherence (AdaDFRC-W2) varies more: Wan attains the lowest deformation on average (12.76), Runway is moderate (18.54), and Sora is the highest (21.79). A category-wise breakdown reveals model-specific weaknesses: Runway exhibits elevated foreground deformation in C4 (Static camera + moving object, 31.69) and C6 (Stress, 23.91);

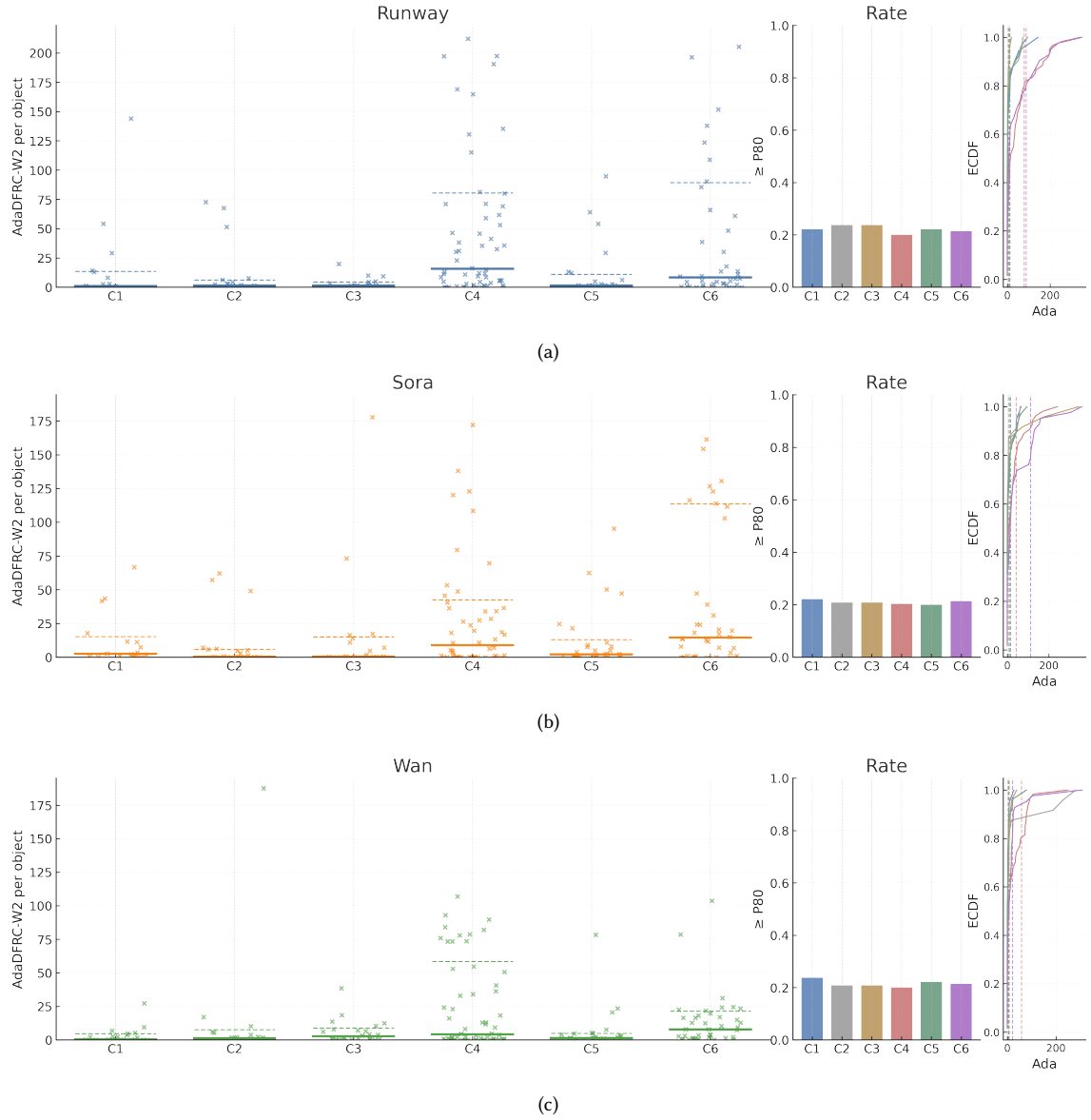


Fig. 11. Foreground rigidity via AdaDFRC-W2 across prompt families and platforms. Each row shows one platform—**(a)** Runway, **(b)** Sora, **(c)** Wan. Left: per-object AdaDFRC-W2 scores grouped by prompt family (C1–C6; lower is better). Each marker corresponds to one tracked object; points are horizontally jittered for visibility. Grey horizontal guidelines indicate the platform’s *R80* robustness threshold, used to summarize the heavy-tail portion of the distribution. Right: *Rate* bars report, per family, the fraction of objects whose score exceeds *R80* (higher bars imply more non-rigid/unstable motion). The ECDF inset shows the empirical cumulative distribution of AdaDFRC-W2 for that platform, with the vertical dashed line marking *R80*.

Sora peaks in C6 (36.84) and is also high in C3/C4 (22.83/22.36); Wan is most challenged in C2 (Rotation-dominant, 24.42) and C4 (18.94), while remaining low in C1/C5 (2.30/4.86). These patterns are consistent with the combined

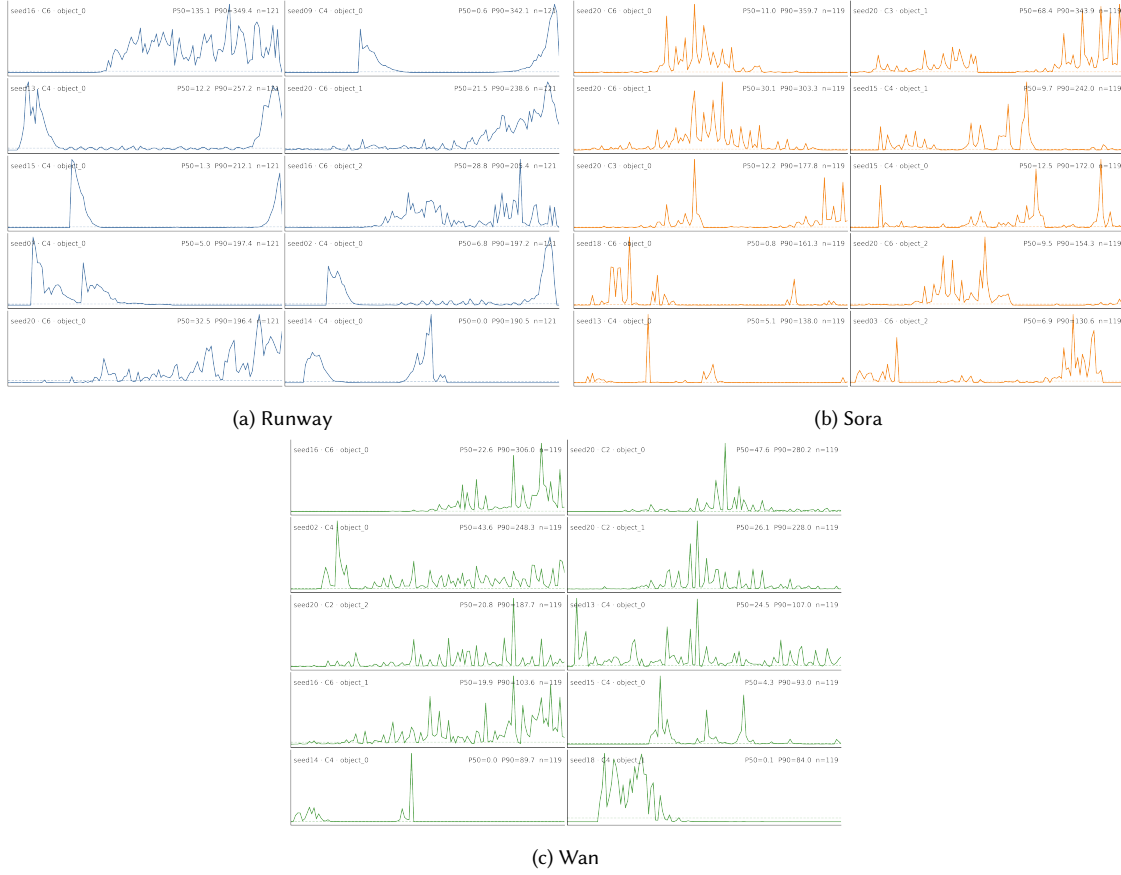


Fig. 12. Top- N hardest objects per model. Each sparkline shows per-frame AdaDFRC-W2 for one object; the page-level x/y labels indicate frame index (temporal order) and AdaDFRC-W2, respectively. Dashed lines mark the model-wise median of per-prompt P80. Right-side annotations report P50, P90, and the number of frames (n).

alignment/consistency map in Fig. 5, where translation-dominant or failure-inferred cells tend to coincide with larger foreground deformation.

Across scenes, the 95% confidence intervals (CI) for background IR are Runway [96.35, 97.55], Sora [95.91, 97.29], Wan [95.41, 97.06] (in %); for GE (px): Runway [1.90×10^{-4} , 2.48×10^{-4}], Sora [2.70×10^{-4} , 3.65×10^{-4}], Wan [2.78×10^{-4} , 3.52×10^{-4}]; for AdaDFRC-W2: Runway [12.32, 24.76] (63 scenes), Sora [12.74, 30.84] (64), Wan [6.57, 18.95] (66).

The results indicate clear differences in the geometric consistency of the evaluated models. A more detailed breakdown by scenario reveals specific strengths and weaknesses. For instance, most models perform well on pure rotation scenarios, achieving high IR with the homography model, but struggle with pure translation, where maintaining epipolar constraints for parallax is more challenging. The AdaDFRC-W2 scores are particularly revealing in categories with moving objects, quantifying the degree of unnatural deformation.

6.8 Analysis: Decoupling Geometric and Semantic Quality

We investigate whether semantic alignment, measured by CLIP, predicts geometric coherence of the generated videos. For each scene (seed–category) and platform, we robustly fit a single background motion model on background-only correspondences, using either a homography $\mathbf{H}$ (planar/rotation regime) or a fundamental matrix $\mathbf{F}$ (translation/parallax regime). Models are estimated in normalized coordinates for numerical stability, and residuals are *denormalized to pixels* before scoring on the content mask (letterboxing excluded). Under either family $g \in \{\mathbf{H}, \mathbf{F}\}$ we report two model-agnostic indicators: *Inlier Ratio (IR)*—the fraction of correspondences passing the family-specific inlier test—and *Geometric Error (GE)*—the median *pixel* residual among those inliers. Hence IR answers whether a single global model sufficiently explains the background motion, while GE quantifies how tightly the consistent matches align, both defined for either $\mathbf{H}$ or $\mathbf{F}$.

We analyze the coupling of CLIP with geometry from: (i) an **auto-selected** view, where each scene adopts the better-fitting family, $\text{IR}_{\text{best}} = \max(\text{IR}_H, \text{IR}_F)$ and $\text{GE}_{\text{best}} = \min(\text{GE}_H, \text{GE}_F)$; and (ii) a **per-family** view that keeps $\mathbf{H}$ and $\mathbf{F}$ separate to avoid bias toward either residual family.

Fig. 13-(a,b) plot overall CLIP against IR_{best} (higher is better) and GE_{best} in pixels (lower is better), respectively, pooling Sora, Runway, and Wan. Across platforms, the scatter shows no strong monotonic trend; the decile medians overlaid in each panel remain flat or erratic. Practically, we observe numerous clips with high CLIP but low IR_{best} or elevated GE_{best} (px), as well as the converse. This directly illustrates that semantic fidelity is not a reliable proxy for geometric consistency.

Fig. 13-(c) reports *reliability by CLIP deciles*: for each CLIP bin we plot the proportion of scenes meeting a pixel-domain geometry criterion, with 95% Wilson confidence bands. We consider two thresholds: (1) $\text{IR}_{\text{best}} \geq 0.97$; and (2) $\text{IR}_{\text{best}} \geq 0.97$ and $\text{GE}_{\text{best}} \leq 3.5$ px. If CLIP were predictive of geometry, the curves should increase monotonically with CLIP. Instead, both curves remain flat or mildly non-monotonic, indicating that higher CLIP bins do not consistently translate into better geometry. Fig. 13-(d) overlays the same decile analysis for the two residual families (solid: IR-only; dashed: IR+GE). The shapes of the $\mathbf{H}$ and $\mathbf{F}$ curves are similar and remain weakly dependent on CLIP. Thus the decoupling phenomenon is not an artifact of favoring one family: whether a scene is better explained by a homography or a fundamental matrix, CLIP still fails to predict geometric coherence.

The findings, which are consistent across various platforms and resilient to the selection of residual family, indicate a significant decoupling between semantic and geometric quality. Therefore, evaluating video generators as “world simulators” requires reporting geometry-oriented metrics (IR/GE in the pixel domain) in addition to semantic scores. In our supplementary material, we provide counterexample lists (high-CLIP with poor geometry and vice versa) and per-family rank correlations, further corroborating this conclusion.

Across three generators and all six categories, CLIP is a poor predictor of geometry: both scatter views (CLIP vs. IR_{best} , CLIP vs. GE_{best} in pixels) and reliability-by-CLIP deciles remain flat or mildly non-monotonic. All residuals are evaluated in the pixel domain on the content mask (letterboxing excluded) with a single gate fixed at $\delta = 3.5$ px and a stringent rigidity criterion $\text{IR}_{\text{best}} \geq 0.97$. Per-family overlays ($\mathbf{H}/\mathbf{F}$) behave similarly, ruling out residual-family bias; higher CLIP does not reliably imply better geometry.

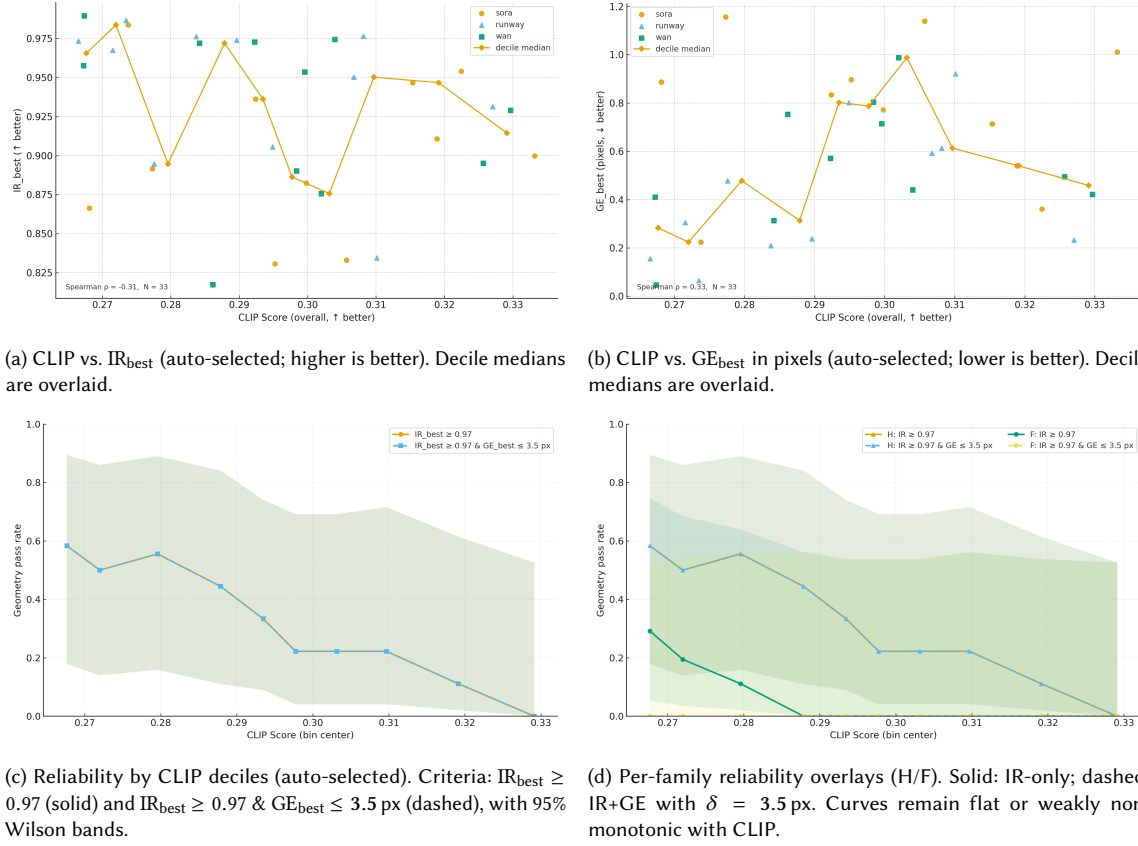


Fig. 13. Semantic vs. geometric decoupling. All residuals are evaluated in the pixel domain on the content mask (letterboxing excluded). Points aggregate Sora, Runway, and Wan over all seed–category scenes.

6.9 Perceptual Evaluation

Fig. 14 assembles row-wise exemplars. The leftmost cell of each row shows the seed image for context; the three cells to its right visualize the background geometry obtained from adjacent-frame matches on the content mask (letterboxing ignored) for three platforms. For each cell, we report the *Inlier Ratio* (IR) and the *Geometric Error* (GE, median residual in pixels after de-normalization). Two visual regularities are consistent across rows. Scenes that are visually rigid to a human—e.g., static rooms or static landscapes viewed with small camera motion—systematically present *high* IR (most background correspondences agree with one global model) and *low* GE (aligned inliers are tight). The reconstructed point clouds look compact with clean planes and straight edges; tiny reprojection jitter is perceptually negligible. When humans perceive background instability—e.g., subtle “breathing” of buildings, bending railings, or tearing near depth discontinuities—the geometry mirrors that: IR drops (more outliers against a single model) and/or GE rises (inlier alignment loosens). In rows where DUST3R fails to return a pose for the first–last pair (because the displacement is too large or the texture is degenerate), the reconstruction is sparse or absent. By our evaluation policy, those cases are labeled *F*, and they correlate with visible large parallax or scene changes. Overall, the qualitative impressions match the

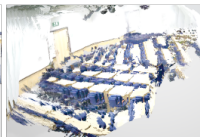
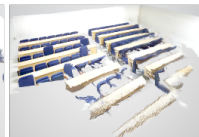
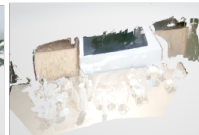
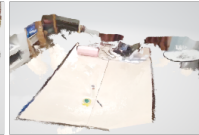
	cf. img	Runway		Sora		Wan	
Seed01 C5							
		Classroom	94.8% 0.787	95.4% 1.240	94.4% 0.834		
Seed06 C5							
		Stairs	99.7% 0.294	97.5% 0.527	85.1% 0.717		
Seed11 C3							
		Model Furniture	96.0% 0.716	66.1% 0.901	97.4% 0.548		
Seed14 C2							
		Office Table	97.8% 0.441	98.6% 0.577	95.4% 0.831		

Fig. 14. Background geometry in practice. For each row, the left cell shows the seed image; the three cells on the right visualize background reconstructions and report IR/GE per platform. Missing or extremely sparse reconstructions are treated as F under our DUS3R gate and also exhibit poor visual stability.

quantitative trend: visually cleaner reconstructions coincide with $IR \geq 0.97$ and $GE \leq 3.5$, whereas visually degraded ones violate one or both.

Fig. 15 presents object-centric optical-flow slices used by our foreground metric. Horizontally, we sample representative seeds; vertically, we juxtapose platforms. For each object mask and frame pair, hue encodes the *direction* of the per-cell *median* flow (magnitude is normalized for visualization only). When an object undergoes near-rigid motion (e.g., a pedestrian translating without limb swing emphasis, or a car with minimal rolling artifacts), the flow direction is coherent within the mask and stable over time. Our AdaDFRC-W2—computed by fitting an affine flow on the object support and measuring a Wasserstein-2 transport cost from that rigid prediction to the observed per-cell medians with adaptive reliability weights—stays low. Non-rigid behavior (articulated limbs, fluttering texture, rolling shutter-like distortions) and occlusion events create sharp direction changes within the mask. The hue then varies within a single frame or flips between frames; these spatio-temporal inconsistencies raise the transport needed from the rigid hypothesis to the observed distribution, yielding a higher AdaDFRC-W2. Cross-platform differences in these hue patterns align with our quantitative ranking of foreground consistency.

The two visual panels ground our metrics in human perception: high IR/low GE tracks "background feels rigid," while low/high signals "background looks wrong"; low AdaDFRC-W2 matches "object moves as a solid," while high values expose non-rigid or unstable motion. These qualitative correspondences validate that our thresholds produce behavior that is not merely numerical but perceptually meaningful.

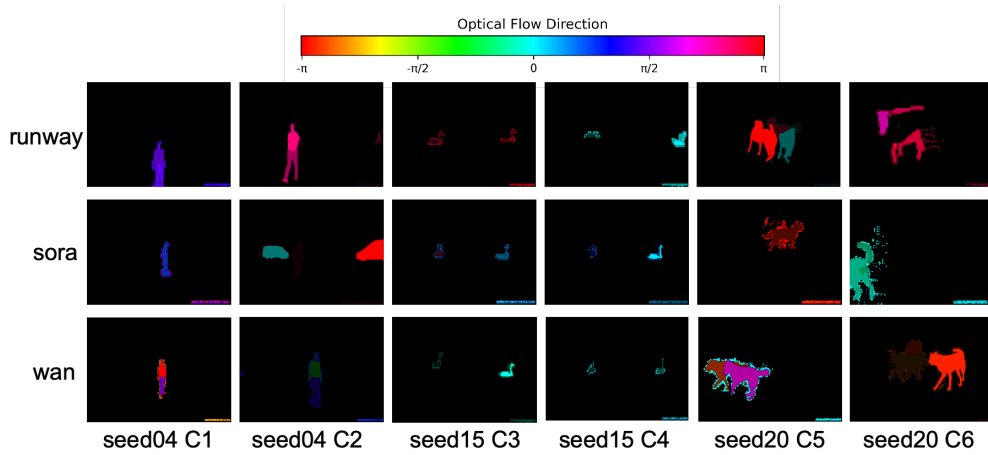


Fig. 15. Foreground motion slices used by AdaDFRC-W2. Columns pick representative seeds; rows juxtapose platforms. Within each object mask and time step, color encodes the direction of the per-cell *median* optical flow. Directional coherence across the mask indicates near-rigid motion and leads to low AdaDFRC-W2; intra-mask direction changes or flips expose non-rigid/unstable dynamics and increase AdaDFRC-W2.

7 Discussion

7.1 What GeoCon-Bench Measures—and What It Does Not

GeoCon-Bench operationalizes *geometric consistency* in AIGC video along two complementary, model-agnostic axes. For the *background*, we test whether a single global transformation explains adjacent-frame motion: either a homography (H) or a fundamental matrix (F), with the model routed by a DUST3R translation gate on the first/last frames. For *both* models, we report (i) the inlier ratio (IR), defined as the fraction of background correspondences that fit the chosen model under a family-specific inlier gate, and (ii) the geometric error (GE), defined as the median residual among those inliers. Inliers are defined with *family-specific* gates, while *all* GEs are evaluated and reported *in the pixel domain after denormalization on the content mask*. At the clip level, a background is called geometrically consistent when $IR \geq 0.97$ and $GE \leq 3.5$ px. For the *foreground*, AdaDFRC-W2 regresses the best affine motion inside each object mask and summarizes non-rigid residuals, exposing elastic artifacts and layer slippage that a rigid-background test cannot see. This decomposition yields interpretable diagnostics: low IR or high GE indicates camera–scene inconsistency (violations of projective/epipolar geometry), whereas high AdaDFRC-W2 flags dynamic incoherence of movers. Crucially, GeoCon-Bench *does not* measure semantic fidelity; it is designed to complement, not replace, semantic metrics. The dissociation we observe between semantics and geometry underscores why a dedicated geometric axis is necessary.

7.2 Cross-Platform Trends and Where Models Diverge

Because prompts are categorized into six families (C1–C6), the benchmark supports per-regime comparisons rather than a single pooled score. Three regularities emerge. *First*, families intended to be H-majority—C1 (static), C2 (rotation-dominant), and C4 (static camera + moving objects)—concentrate mass at high IR and low GE under the homography model; when discrepancies arise, they typically reflect (i) small end-to-end translational drift that pushes a minority of seeds into the F branch, or (ii) segmentation leakage admitting moving edges into the background set, both surfaced as mild but systematic penalties. *Second*, families expected to be F-majority—C3 (translation-dominant), C5 (complex

camera motion), and **C6** (geometric stress)—exhibit broader dispersion and a higher rate of F decisions; platforms diverge most in *how* parallax is realized (amount of depth-induced motion, stabilization priors), which manifests as different balances of IR versus GE. *Third*, failure handling matters: present-only results isolate pure geometry where DUST3R succeeds, whereas a conservative fail→F view additionally counts first/last configurations whose displacement prevents pose recovery—an informative signal under stress conditions. At the *seed-aligned* level (same seed across platforms), two divergence types are actionable. One is *routing disagreement*: the same seed–family pair routes to H on one platform but to F on another, indicating distinct camera priors or depth realism; such cases dominate off-diagonal cells in cross-platform alignment maps and co-occur with GE shifts. The other is *agreement with different margins*: all platforms agree on H (or F), yet one sits closer to the IR/GE boundary—these “near-fail” seeds are highly sensitive to training or decoding changes and thus serve as high-yield regression tests. Finally, the anomaly distribution separates families where dispersion is intrinsic (**C6**) from those where instability is a warning sign (**C2**, **C4**). Together, these trends motivate reporting *per-family*, *seed-aligned* summaries in addition to pooled scores: they are diagnostic, reproducible, and immediately actionable for model development.

7.3 Semantic–Geometric Dissociation and Why It Matters

High semantic alignment (e.g., CLIP) does not guarantee geometric correctness. We substantiate this with two independent lenses. (i) *Thresholded reliability*. Under a fixed, pixel-domain reporting gate ($IR \geq 0.97$, $GE \leq 3.5$ px), a substantial fraction of clips that score highly on CLIP fail the background rigidity check or exhibit elevated AdaDFRC-W2. This persists whether the background model is auto-selected by DUST3R or forced per-family (H for C1/2/4, F for C3/5/6), indicating that the dissociation is not a routing artifact (see Fig. 13). (ii) *Distributional stability*. Across CLIP deciles, both the pass rate for background consistency and the distribution of foreground residuals change weakly, if at all, despite large semantic swings—again pointing to orthogonality. Mechanistically, text–vision alignment systems reward object/texture plausibility and global style, whereas our geometry axis constrains camera–scene relations and rigidity. Video generators inherit stabilization or motion-style priors that visually improve coherence yet can suppress or mis-shape parallax, and they often synthesize movers with elastic shortcuts that evade semantic penalties but are penalized by AdaDFRC-W2. The implication is practical: leaderboards and ablations should report semantics and geometry as *two orthogonal axes*. GeoCon-Bench provides a drop-in geometric axis—background IR/GE in the pixel domain, plus object-level AdaDFRC-W2—so that model builders can diagnose failure modes that semantic metrics systematically miss. Releasing per-clip pose logs, masks, and fixed gates is essential for *reproducibility* and *actionable debugging*.

7.4 Limitations and Future Work

Our study is deliberately scoped to geometric behaviors that are most consequential for perception-centric use. We evaluate static or near-rigid backgrounds with a single global model and summarize foreground motion with an affine proxy and distributional residuals, routed by DUST3R into either planar/rotation or translation/parallax regimes. This design prioritizes interpretability and reproducibility, and within that operating regime, our metrics exhibit stable behavior and align with human judgments. Several methodological boundaries remain. First, the routing currently relies on a two-frame statistic; clip length, baseline, or near-degenerate configurations can bias this decision. Second, residuals are computed in the pixel domain and are mildly sensitive to sensor- or codec-induced artifacts such as rolling shutter and heavy compression. Third, background–foreground separation inherits any errors in segmentation, so boundary perturbations may affect marginal cases even when coverage is reported and failure cases are handled consistently. Finally, our prompt families emphasize common scene regimes rather than exhaustive coverage. These caveats point to

concrete extensions: multi-window or temporally smoothed routing that is invariant to clip length; degeneracy-aware estimation with explicit parallax and conditioning checks; residual models that account for sensor-domain effects; uncertainty- and occlusion-aware masks with consistency checks between pose and segmentation; and layered or depth-harmonized background models that retain the current interpretability while broadening applicability. Expanding the dataset with calibrated trajectories and controlled synthetic scenes will further enable stress testing and fine-grained calibration, moving the framework toward a more comprehensive standard for assessing generative video as a geometric process.

Acknowledgments

To Robert, for the bagels and explaining CMYK and color spaces.

References

- [1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. 2024. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.
- [2] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. 2020. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Trans. Graph.* 39, 4, Article 75 (Aug. 2020), 13 pages. doi:10.1145/3386569.3392457
- [3] Ondrej Chum and Jiri Matas. 2005. Matching with PROSAC-progressive sample consensus. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. IEEE, 220–226.
- [4] Ondrej Chum, Jiri Matas, and Josef Kittler. 2003. Locally optimized RANSAC. In *Joint pattern recognition symposium*. Springer, 236–243.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [6] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. 2025. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983* (2025).
- [7] Yonatan Dukler, Wuchen Li, Alex Lin, and Guido Montufar. 2019. Wasserstein of Wasserstein Loss for Learning Generative Models. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 1716–1725. <https://proceedings.mlr.press/v97/dukler19a.html>
- [8] Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. 2023. Aigebench: Comprehensive evaluation of image-to-video content generated by ai. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 3, 4 (2023), 100152.
- [9] Mohammed E Fathy, Ashraf S Hussein, and Mohammed F Tolba. 2011. Fundamental matrix estimation: A study of error criteria. *Pattern Recognition Letters* 32, 2 (2011), 383–391.
- [10] Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.
- [11] Richard Hartley and Andrew Zisserman. 2004. *Multiple View Geometry in Computer Vision*. Cambridge University Press.
- [12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).
- [13] 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).
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [15] Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.
- [16] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yanan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. 2024. Vbench+: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503* (2024).
- [17] Eunhae Jang, Hui Min Lee, Sangwook Lee, Yongnam Jung, and S. Shyam Sundar. 2025. Too Good to be False: How Photorealism Promotes Susceptibility to Misinformation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 531, 8 pages. doi:10.1145/3706599.3719796
- [18] Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhui Chen. 2024. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems* 37 (2024), 79889–79908.

- [19] Negar Kamali, Karyn Nakamura, Aakriti Kumar, Angelos Chatzimparmpas, Jessica Hullman, and Matthew Groh. 2025. Characterizing Photorealism and Artifacts in Diffusion Model-Generated Images. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.
- [20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.
- [21] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125* (2023).
- [22] Karel Lebeda, Jiri Matas, and Ondrej Chum. 2012. Fixing the locally optimized ransac—full experimental evaluation. In *British machine vision conference*, Vol. 2. Citeseer Princeton, NJ, USA.
- [23] Yan Li, Yingdan Wu, Yang Ming, Yong Zhang, and Zhesheng Cheng. 2025. GeoAT: Geometry-Aware Attention Feature Matching Network. *IEEE Access* (2025).
- [24] Xiao Liu, Xinhao Xiang, Zizhong Li, Yongheng Wang, Zhuoheng Li, Zhuosheng Liu, Weidi Zhang, Wei Qi Ye, and Jiawei Zhang. 2024. A survey of ai-generated video evaluation. *arXiv preprint arXiv:2410.19884* (2024).
- [25] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22139–22149.
- [26] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177* (2024).
- [27] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. 2024. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048* (2024).
- [28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.
- [29] Riku Murai, Eric Dexheimer, and Andrew J Davison. 2025. MAST3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 16695–16705.
- [30] David Nistér. 2004. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence* 26, 6 (2004), 756–770.
- [31] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. 2017. A survey of structure from motion*. *Acta Numerica* 26 (2017), 305–364.
- [32] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.
- [33] Bowen Qu, Xiaoyu Liang, Shangkun Sun, and Wei Gao. 2024. Exploring aigc video quality: A focus on visual harmony video-text consistency and domain distribution gap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6652–6660.
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.
- [35] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Juntong Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714* (2024). <https://arxiv.org/abs/2408.00714>
- [36] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. 2024. Shadows Don't Lie and Lines Can't Bend! Generative Models don't know Projective Geometry... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 28140–28149.
- [37] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4104–4113.
- [38] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8934–8943.
- [39] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*. Springer, 402–419.
- [40] Philip HS Torr and Andrew Zisserman. 2000. MLESAC: A new robust estimator with application to estimating image geometry. *Computer vision and image understanding* 78, 1 (2000), 138–156.
- [41] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).
- [42] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new metric for video generation. (2019).
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [44] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei

- Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wenten Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).
- [45] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20697–20709.
- [46] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. 2025. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision* 133, 5 (2025), 3059–3078.
- [47] 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.
- [48] Haoifei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. 2022. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8121–8130.
- [49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
- [50] Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. 2025. Q-Bench-Video: Benchmark the Video Quality Understanding of LMMs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 3229–3239.

A Prompt Component Lexicon

This appendix provides an expanded vocabulary for the prompt components defined in our framework.

Table 2. Expanded Prompt Component Lexicon

Component	Category	Vocabulary Examples
Motion Command	Foundational	“This is a single continuous shot”, “The camera is locked down”
	Camera Rotation	“pans left”, “tilts up”, “rolls clockwise”, “orbits around”
	Camera Translation	“dollies in”, “trucks left”, “cranes up”, “moves in a straight line”
	Figurative & Phenomenological	“as if on a cinematic dolly track”; “creating true motion parallax”; “as if mounted on a heavy concrete pillar”; “shadows stretch realistically”; “reflections move accurately across the surface”
Qualitative Style	Realism/Fidelity	“hyper-realistic”, “photorealistic”, “8K resolution”, “sharp focus”, “deep depth of field”, “shot on IMAX film”, “documentary style”
	Lighting	“cinematic lighting”, “golden hour”, “dramatic lighting”, “studio lighting”, “neon glow”, “volumetric lighting”
	Artistic Medium	“in the style of a vintage film”, “black and white film noir”, “impressionist painting”, “stop-motion animation”
Scene Composition	Environment	“a bustling city street”, “a serene forest path”, “a minimalist interior”, “a futuristic laboratory”, “a Victorian-era room”
	Geometric Structure	“a long corridor with repeating arches”, “a skyscraper with a grid of windows”, “a spiral staircase”, “a checkerboard floor”
Object Specification	Material/Texture	“reflective chrome”, “matte black finish”, “rough stone texture”, “smooth glass”
	Object State	“a pristine new car”, “a rusted, abandoned vehicle”, “a blooming flower”

B Additional Evidence for DUST3R Translation-Gate Calibration

We reassess the DUST3R translation gate using only pairs with available poses (present-only), across all platforms and prompt families. This analysis is geometric and independent of semantic scores; background inlier gates remain

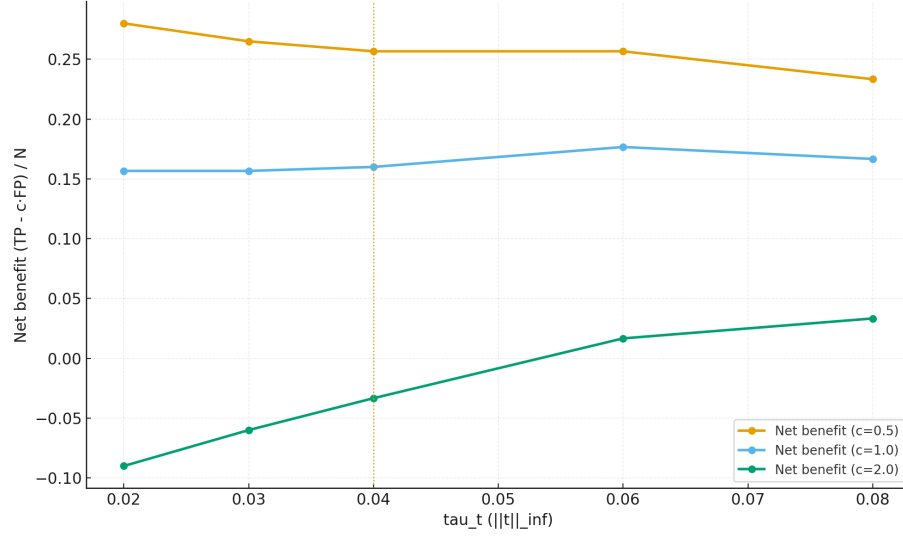


Fig. 16. Decision-curve analysis (present-only): net benefit $(TP - c \cdot FP) / N$ across $c \in \{0.5, 1, 2\}$. The choice $\tau_t=0.04$ lies near the maxima and remains robust to c .

fixed throughout the paper ($\tau_H=2$ px, $\tau_F=0.1$, with $IR \geq 0.97$ and $GE \leq 3.5$ px when background consistency is evaluated elsewhere).

We quantify the operating trade-off between recall on F-expected scenes (C3/5/6) and spurious F on H-expected ones (C1/2/4) via the net benefit

$$NB(\tau_t) = \frac{TP - c \cdot FP}{N}, \quad (18)$$

for false-positive cost ratios $c \in \{0.5, 1, 2\}$. Fig. 16 shows that $\tau_t=0.04$ sits close to the maxima simultaneously for all c and is thus robust to reasonable changes in cost.

We perform $B=400$ bootstrap resamples over the observation set (sampling units are *(seed, platform, family)*), compute balanced accuracy for each $\tau_t \in \{0.02, 0.03, 0.04, 0.06, 0.08\}$, and record both the mean $\pm$ 95% CI and the best threshold per resample. Fig. 17 (left) reveals a broad performance plateau over $[0.04, 0.06]$ where confidence bands strongly overlap; the histogram (right) concentrates the best-threshold selections in the same interval, with a visible mode at 0.04–0.06. Hence $\tau_t=0.04$ is either at or statistically indistinguishable from the maxima while being slightly more conservative on false positives.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

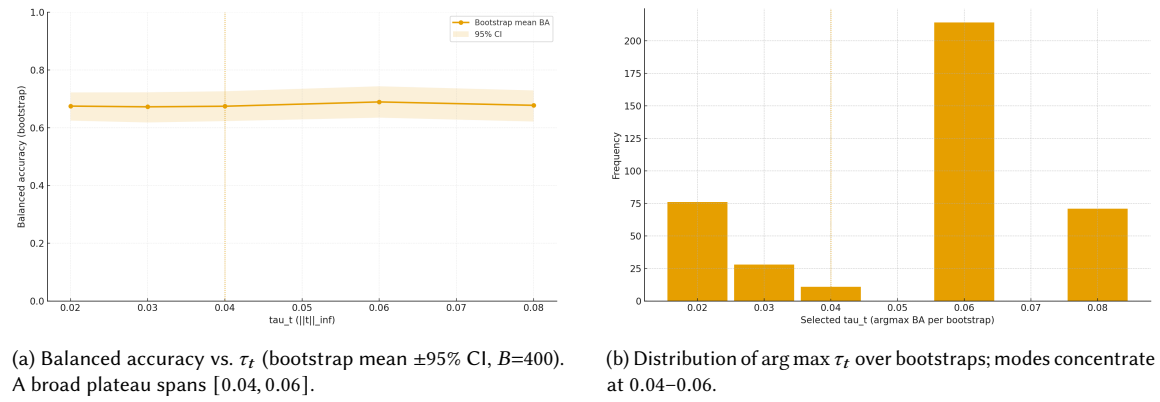


Fig. 17