TEMPORAL-GEOMETRY CONSISTENCY FOR LOCAL STRUCTURAL DRIFT VERIFICATION IN GENERATED VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study local structural drift in generated videos using predicted geometry. We introduce Temporal-Geometry Consistency (TGC), a cohort-relative score that combines four structural and directional-temporal readouts of predicted normals and depth through within-generator ranks: directional coherence loss of the combined normal/depth channels (ϕ_{dir}), structural variation of a semantic–geometry agreement field (ϕ_{grad}), directional-residual coherence loss of the relation field (ϕ_{rel}), and an independent depth-derived adjacent-frame normal coherence loss (ϕ_{dep}). On a 407-pair GeneVA evaluation, TGC improves over the frozen Relation baseline by $+0.0367$ AUROC under equal-generator, prompt-cluster paired inference (95% CI $[+0.0040, +0.0693]$). Direct agreement-based normal/depth rerankers do not improve over Relation; a matched-encoding diagnostic shows that encoding-induced score changes concentrate in the amplitude-agreement readout, while matched lossless RGB shows no resolved ranking advantage ($+0.0023$, $[-0.0536, +0.0556]$). On the sealed 400-clip GeneVA fresh-headroom replication block, TGC attains 0.6868 pooled AUROC versus 0.5016 (GeCo) and 0.4913 (VIGOR, n=398 pairwise). In a pre-registered five-generator VideoPhy-2 evaluation, the pooled gain over Relation is modest ($+0.0036$ $[-0.0565, +0.0643]$) and positive in three of the five unseen generators, negative in the remaining two; matched appearance perturbations meet the frozen non-inferiority margin ($\mathrm{U}_{95} = +0.0075 < \delta = +0.05$). The paper contributes a frozen temporal-geometry score for post-hoc local-drift ranking with reproducible held-out gains, a controlled diagnostic of the failed agreement rerankers, and a pre-registered, fixture-reproducible evaluation protocol, while generalization to unseen generators remains open.

1 INTRODUCTION

Generated video has reached a quality where failures are rarely gross; they are local, brief and structurally subtle. A common class is *local structural drift*: a hand gains a finger, a face warps as it turns, an object surface slides against its support while the surrounding scene remains coherent. Such errors are easy for a human to find once pointed to but hard to catch in bulk review, and they are largely invisible to global quality scores (Huang et al., 2024; Liu et al., 2024; Wu et al., 2024): perceptual distances and text-assignment metrics (Zhang et al., 2018; Fu et al., 2023; Oquab et al., 2023; Unterthiner et al., 2018) can stay nearly unchanged across the defect, and monocular geometry estimators can be so smooth that a one-second local distortion leaves no aggregate trace. Fine-grained artifact-localization benchmarks make the same point from the annotation side: the defects concentrate in small spatiotemporal regions that clip-level scores aggregate away (Lin et al., 2025). Our setting is therefore not real-time detection but *post-hoc triage*: given a batch of generated clips, rank them so that a human reviewer sees the most likely local-drift cases first. The unit of interest is a relative ordering within a cohort, not a calibrated probability on a single video.

A natural cue for this task is predicted geometry. A locally drifting region contradicts itself over time—the surface normal or depth at a tracked point cannot be simultaneously coherent and deformed—so one would expect temporal consistency of predicted normals and depth to separate

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

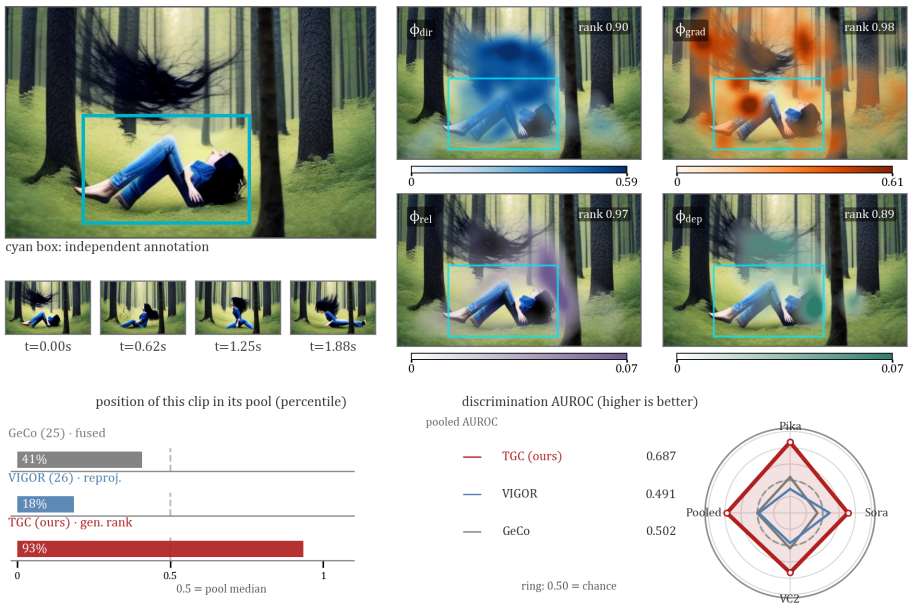


Figure 1: TGC flags a holdout local-drift clip that the two external geometry baselines read as benign: on the sealed fresh-headroom block, GeCo's fused score and VIGOR's reprojection error land at the 41st and 18th pool percentiles, while TGC's frozen four-factor rank places the clip in the top 6.5% of its generator cohort. The right panels overlay the four factor readouts on the independently annotated frame (cyan box); tile badges are within-generator rank fractions. The bottom strip shows pool positions, and a four-axis radar compares per-generator and pooled discrimination AUROC on the same block with the 0.50 chance ring (higher is better).

drift videos from clean ones. We previously built rerankers on this hypothesis and obtained a clean negative result. Beneath a frozen *Relation* baseline, three normal/depth/joint agreement rerankers failed to add reliable discrimination on 407 held-out pairs (0.6808 versus $0.6132/0.6294/0.6313$), and that negative conclusion survived early pilots and a corrected, frozen evaluation.

This negative result motivates our main contribution, Temporal-Geometry Consistency (`TGC`): instead of re-aggregating the same agreement amplitudes, it measures whether geometry is consistent over time in structure and direction. It combines four signals—directional coherence loss of the combined normal/depth channels, structural variation (gradient energy) of the semantic–geometry agreement field (ϕ_{grad}), the directional-residual coherence loss of the relation field, and an independent depth-derived adjacent-frame normal coherence loss—into a within-generator rank average. Evaluated with an equal-generator, prompt-cluster-aware paired bootstrap on the same held-out 407 pairs, it exceeds Relation by $\Delta = +0.0367$ (95% interval $[+0.0040, +0.0693]$), with the gain not attributable to rank transformation alone or pre-encoding access, nor obviously driven by a single generator.

A complementary controlled diagnostic examines the readout sensitivity behind the original verifier: using bounded, pre-registered generator runs with matched lossless RGB and lossy copies of the same clips, we find that encoding alone perturbs agreement-based scores substantially—the response across compression quality is non-monotonic and dominated by the agreement term—while preserving much of the underlying field structure, and that carrying lossless inputs shows no resolved ranking advantage ($+0.0023$, $[-0.0536, +0.0556]$ on the matched pool). Our third contribution is the evaluation apparatus itself: a decision statistic that separates descriptive pooled AUC from decision-grade uncertainty, a pre-specified appearance-control protocol with a non-inferiority margin for nuisance perturbations, and a small bundled numerical fixture that recomputes the reported main-table rows and pooled endpoints without models or downloads.

Appearance-control blocks show no degradation pattern, but their intervals are too wide to establish non-degradation, leaving the margin unresolved. The pre-registered larger appearance-control block

(200 bases, 400 base-clustered paired comparisons under two frozen nuisance families) completes and establishes control non-degradation at the frozen margin: $\mathrm{U}_{95}(\Delta_{\mathrm{ctrl}}) = +0.0075 < \delta = +0.05$.

On the pre-registered fresh blocks the frozen decision statistic crosses zero on VideoPhy-2 ($+0.0036$), on the Physion multi-entity block (-0.0208) and on the sealed GeneVA fresh-headroom block ($+0.0151$), so fresh superiority over Relation stays unresolved (Section 4.7 reports every pre-registered endpoint). Generalization to unseen generators therefore stays open, alongside the frozen candidate, the diagnosis and the protocol.

Contributions.

- **A frozen method (C1).** Temporal-geometry consistency (`TGC`): a support-weighted, within-generator rank ensemble of structural and directional temporal consistency factors, developed and frozen before labels of the held-out set were used.
- **A controlled readout diagnostic (C2).** Matched lossless/encoded evaluation isolates the score readout from the content; the non-monotonic quality response concentrates encoding-induced score variation in the agreement term (57–71% across pools), and lossless input shows no resolved ranking advantage.
- **A reproducibility-first evaluation (C3).** Equal-generator, prompt-cluster-aware paired inference with shared bootstrap views, a pre-specified appearance-control non-inferiority protocol, two external geometry baselines (GeCo, VIGOR), and a CPU-only numerical fixture that recomputes the main tables and pooled endpoints.

Throughout, *pre-registered* refers only to the frozen rules of Section 3.4; numeric statements never imply deployment readiness or a calibrated detector, and never reverse the earlier negative conclusion for the original rerankers.

2 Related Work

Generated-video evaluation. Automated evaluation of generative video is dominated by reference- and text-alignment metrics. Mean and distributional frame-level distances (Unterthiner et al., 2018; Zhang et al., 2018), deep perceptual similarity such as DreamSim over DINOv2 features (Fu et al., 2023; Oquab et al., 2023), and learned quality assessors for aesthetics and technical distortions (Wu et al., 2023; 2022; Lu et al., 2024; Ke et al., 2021) remain the standard automatic summaries, while comprehensive suites such as VBench and EvalCrafter and fine-grained aspect benchmarks such as FETV and T2VScore document how weakly a single global score tracks the fine-grained failures that matter for artifact triage (Huang et al., 2024; Liu et al., 2024; 2023; Wu et al., 2024). Physical-plausibility suites—VideoPhy and its action-centric successor VideoPhy-2, the earlier action-centric Physion, and the geometry-oriented PDI-Bench—target event-level physics rather than spatially local geometric drift (Bansal et al., 2025; 2026; Bear et al., 2021; Wu et al., 2026). Recent artifact-localization benchmarks make the complementary point explicitly: defects concentrate in small spatiotemporal regions, so clip-level scores conflate prompt failures, content errors and local structural artifacts (Lin et al., 2025). These benchmarks establish that global clip-level scores can underweight spatially localized, short-duration artifacts, but they mostly report aggregate model rankings and do not provide a working signal for clip-level triage inside one generator's output.

Generator landscape. The released-model landscape spans cascaded latent diffusion (Blattmann et al., 2023; Zhang et al., 2023; Chen et al., 2024), image-to-video animation priors (Xing et al., 2024; Ren et al., 2024), realtime latent transformers (HaCohen et al., 2024), and large transformer-based video models (Yang et al., 2025; Kong et al., 2024; Wan et al., 2025), with camera-controlled generation adding viewpoint conditioning (Zheng et al., 2024). Our cohorts are drawn from several of these families, and their between-generator scale differences motivate the within-generator rank design of our method.

Geometry-aware video analysis. The closest prior work evaluates geometry directly. GeCo uses residual motion and depth priors to flag deformation and occlusion inconsistency in generated video and releases single-video detection code (Gu et al., 2026); VIGOR constructs a geometry reward from pretrained geometry models and cross-frame reprojection error (Yin et al., 2026). Both are included as external geometry baselines in our protocol; their frozen adapters ran on the completed

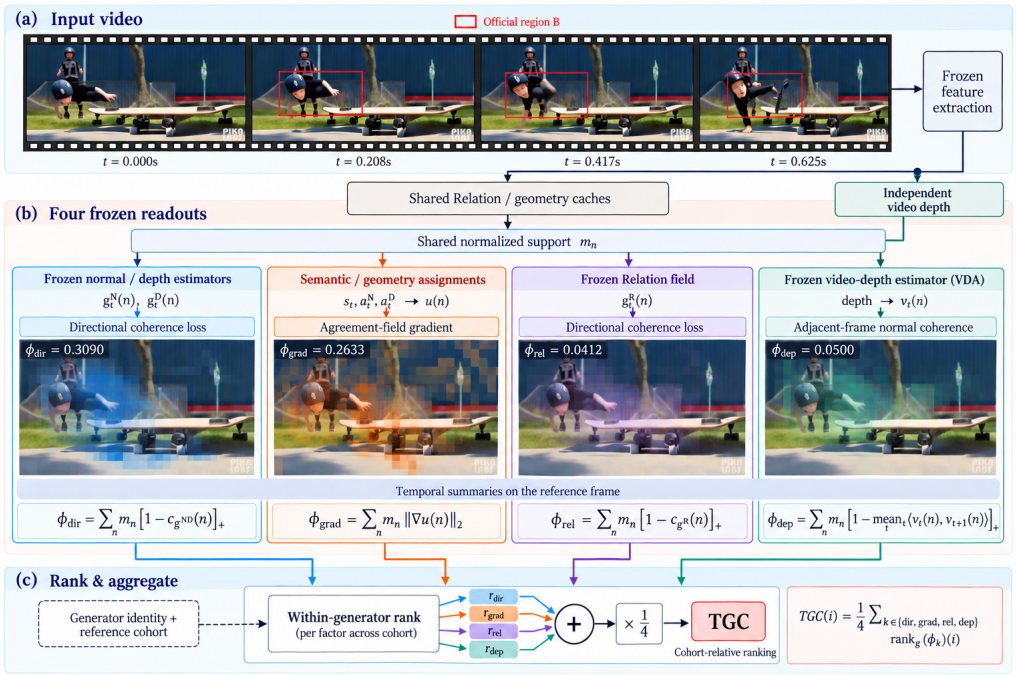


Figure 2: The frozen TGC pipeline on one held-out GeneVA local-drift clip, with the official drift region: (a) input video and frozen feature extraction; (b) the four factor readouts over the shared support with their measurement sources and loss formulas; (c) the within-generator rank average that yields the final TGC score.

public-label block and are reported in Section 4.7 (Table 6 in Appendix D). Unlike those metrics, our method operates on the same intermediate relation and geometry fields already computed by the production verifier, ranks clip-within-cohort, and is validated under a paired, generator-aware decision statistic.

Geometry estimation and temporal consistency. Our factors are built on monocular normal/depth prediction pipelines and inherit their biases. Depth foundation models scale monocular estimation to open-domain video (Ranftl et al., 2020; Bhat et al., 2023; Yin et al., 2023; Piccinelli et al., 2024; Yang et al., 2024a;b; Ke et al., 2024), and surface-normal estimation follows the same data-driven shift (Bae & Davison, 2024; Eftekhar et al., 2021); temporally coherent video depth supplies the independent depth grid of our pipeline (Chen et al., 2025). Dense tracking underlies the node supports, from RAFT-style flow to modern point trackers (Teed & Deng, 2020; Karaev et al., 2024; 2025; Doersch et al., 2022), and segmentation- and foundation-feature models represent the lineage of the semantic assignment on which the relation field is read (Kirillov et al., 2023; Oquab et al., 2023). We treat these as measurement components rather than competitors. Several of the automatic-quality works above note codec sensitivity of pixel-level metrics (Wu et al., 2023; 2022; Lu et al., 2024; Wu et al., 2024); Our diagnostic complements this literature by comparing matched lossless and encoded copies of identical generations and decomposing the resulting score variation at the readout level.

3 Method

3.1 Setup and field sources

All methods share a fixed node grid, the supports, and the frozen relation/geometry extraction caches (the full readout-to-score flow and one holdout readout are visualized in Figures 2 and 1); the exact byte-hashes travel with the reproducibility package. At each node n we have, over sampled frames t: unit directional sequences of the normal and depth channels $g_t^N(n), g_t^D(n)$; the relation field direction sequence $g_t^R(n)$; the semantic, normal and depth assignment indicators $s(n), a^N(n), a^D(n)$ used by Relation (its time-pooled node-level split responses); a support map $m_n \geq 0$, $\sum_n m_n = 1$;

and, independently, a depth grid from a video depth estimator (Chen et al., 2025) from which per-frame surface normals can be derived. Define the temporal directional coherence of a unit-vector sequence $g_t(n)$ as

$$c_g(n) \;=\; \frac{1}{T}\sum_{t=1}^{T}\frac{\langle g_t(n),\,\hat{g}(n)\rangle}{\|\hat{g}(n)\|}, \qquad \hat{g}(n)=\frac{1}{T}\sum_{t=1}^{T} g_t(n), \tag{1}$$

with the convention $c_g(n)=0$ when $\hat{g}(n)=0$ (the frozen implementation adds a small denominator epsilon), and the smoothed agreement field

$$u(n) \;=\; \tfrac{1}{2}\,\mathrm{nbh}\big[a^{\mathrm{N}}(n)=s(n)\big] \;+\; \tfrac{1}{2}\,\mathrm{nbh}\big[a^{\mathrm{D}}(n)=s(n)\big], \tag{2}$$

where nbh is the four-neighbour mean over the node grid. The field is built from the node-level split responses shared with Relation and carries no separate temporal averaging; the spatial gradient in Eq. 4 is computed by four-neighbour finite differences on the same grid.

3.2 Four frozen factors

Directional coherence loss of combined normal/depth channels (ϕ_{dir}). With combined unit vectors $\bar{g}_t^{\mathrm{ND}}(n)=u_0(g_t^{\mathrm{N}}(n)+g_t^{\mathrm{D}}(n))$, where $u_0(x)=x/\|x\|_2$ with the convention $u_0(0)=0$, the raw factor is the support-weighted coherence loss over trajectories:

$$\phi_{\mathrm{dir}} \;=\; \sum_{n} m_n\,\big[1-c_{\bar{g}^{\mathrm{ND}}}(n)\big]_+, \quad \text{with } [x]_+=\max(x,0). \tag{3}$$

Structural variation of the agreement field (ϕ_{grad}). The agreement field's spatial L2-gradient magnitude is summarized under the same support:

$$\phi_{\mathrm{grad}} \;=\; \sum_{n} m_n\,\big\|\nabla u(n)\big\|_2, \tag{4}$$

which reads *where* agreement changes in space instead of how high it is on average. **Directional residual coherence loss of Relation (ϕ_{rel}).** Relation's own direction sequence enters only through its coherence loss,

$$\phi_{\mathrm{rel}} \;=\; \sum_{n} m_n\,\big[1-c_{g^{\mathrm{R}}}(n)\big]_+, \tag{5}$$

never through Relation's magnitude or final clip score. **(the depth-derived factor ϕ_{dep}) Depth-derived adjacent-frame normal coherence loss.** From the independent depth grid, compute per-frame normals $\nu_t(n)$—distinct from the frozen shared depth-direction channel $g_t^{\mathrm{D}}(n)$—and their cross-frame cosine consistency, summarized as

$$\phi_{\mathrm{dep}} \;=\; \sum_{n} m_n\,\Big[1-\tfrac{1}{T-1}\textstyle\sum_{t=1}^{T-1}\langle \nu_t(n),\nu_{t+1}(n)\rangle\Big]_+. \tag{6}$$

3.3 Aggregation and operational boundary

Each factor is converted to its within-generator rank fraction: within each generator g, clips are ordered by ascending factor score and assigned the ordinal positions $1,\dots,|g|$ of the frozen sort, and $\mathrm{rank}_g(x)(i)$ is that position divided by $|g|$; equal scores are absent from the packaged per-clip factor ledgers. Between-generator scale differences are thereby absorbed without per-generator tuning, and the final score is the frozen equal-weight average

$$\mathtt{TGC}(i) \;=\; \tfrac{1}{4}\Big(\mathrm{rank}_g(\phi_{\mathrm{dir}})+\mathrm{rank}_g(\phi_{\mathrm{grad}})+\mathrm{rank}_g(\phi_{\mathrm{rel}})+\mathrm{rank}_g(\phi_{\mathrm{dep}})\Big)_{(i)}. \tag{7}$$

When the independent depth field is unavailable, the three-factor variant omits the last term; where the depth-derived factor ϕ_{dep} exists, removing it is reported as an ablation and never substitutes for Eq. 7 as the primary score. All weights, the support map, and the factor formulas are frozen; Eq. 7 is the only primary candidate.

One operational constraint follows from the rank transform: the score needs the generator identity of each clip and a reference cohort of clips of that generator to form ranks. The output is therefore *cohort-relative*: it ranks clips within one generator's batch and is not a calibrated standalone-video probability, and it inherits whatever the reference cohort contains. The scientific task (post-hoc local-drift ranking) is unchanged from the earlier pilots; the interface is narrowed. Fixed-reference and generator-agnostic normalization are predeclared *variants* to be tested separately, never adopted as the primary on the basis of fresh-evaluation performance.

Table 1: Main held-out evaluation (GeneVA 407 pairs; 143 Pika-1 / 75 Sora / 189 VC-2). Pooled AUROC is descriptive; the decision statistic is the equal-generator, prompt-cluster paired bootstrap of the gain over Relation with shared views. Historical normal/depth/joint rerankers add no reliable value; Q10 re-compresses the same clips.

Condition	Relation	Normal	Depth	Joint	TGC
Released (407 pairs)	0.6808	0.6132	0.6294	0.6313	0.7177
Q10 (same clips)	0.6657	0.5920	0.6148	0.6137	0.7175
Gain over Relation (pooled, released)	$+0.0369\,[+0.0090,\ +0.0648\,]$, $P=.995$				
Gain over Relation (equal-generator, released)	$+0.0367\,[+0.0040,\ +0.0693\,]$, $P=.986$				
Gain over Relation (equal-generator, Q10)	$+0.0429\,[+0.0073,\ +0.0781\,]$				

3.4 Evaluation protocol

Data and labels. The primary evaluation is the GeneVA set of 407 local-drift/No-Issue pairs (Kang et al., 2026) (143 Pika-1 / 75 Sora / 189 VC-2, released encoding plus a Q10 re-encode); the representation was developed on a separate generator (Wan (Wan et al., 2025)) and frozen before these labels were used. Wan/LTX (HaCohen et al., 2024) (Q27) and GeneVA-exact/Physion (Bear et al., 2021) are development-adjacent supporters, with Relation development sets marked. "Positive' means human-confirmed strict local drift and "negative' fully normal; other-issue and uncertain clips are excluded and reported, never folded into a class. The frozen fresh confirmation is composed of *Fresh A* (VideoPhy-2 public labels, 170 pairs, five unseen generators (Bansal et al., 2026)), *Fresh B* (Physion multi-entity strict, three fresh generators, 12 pairs retained after the frozen quota shortfall) and *Fresh C* (sealed GeneVA fresh-headroom, a same-engine replication of 200 pairs / 400 clips with hero quotas $67/67/66$ under the same strict-drift rule, all prior clips and prompts excluded and the manifest sealed before extraction). The pre-registered primary fresh union averages A and B with equal block weights (182 pairs); the three-block union A+B+C (382 pairs) is reported alongside, and C is supplementary evidence, never fresh-generator confirmation. **Baselines.** Relation is the frozen in-house ablation anchor; the historical normal/depth/joint rerankers remain negative references; the depth factor ϕ_{dep} alone is the predeclared internal comparator; GeCo (Gu et al., 2026) and VIGOR (Yin et al., 2026) (with VBench (Huang et al., 2024) dimensions as a recommended broad check) are the external comparators.

Decision statistic. Pooled AUROC is descriptive; decisions use the equal-generator, prompt/source-cluster paired bootstrap with shared views: a 95% percentile interval, per-generator rows and leave-one-generator-out sensitivity (Appendix A). On GeneVA the released cluster keys are all distinct, so cluster resampling reduces to clip-level resampling within generator. Cohort-relative ranks are computed once per full cohort and reused in every bootstrap view (Appendix A).

Appearance controls. For a clean clip and pre-specified appearance-perturbed copies (exposure/gamma, illumination/shadow; geometry untouched, parameters frozen before scoring), a *false high* is a perturbed copy scoring above its matched clean clip, and

$$\Delta_{\text{ctrl}} \;=\; p_{\text{FH}}(\texttt{TGC}) - p_{\text{FH}}(\text{Relation}). \tag{8}$$

The frozen check is $\text{U}_{95}(\Delta_{\text{ctrl}}) < \delta$, $\delta{=}{+}5$ p.p., where U_{95} is the two-sided 95% paired-bootstrap interval's upper endpoint; a sample that cannot resolve the margin is reported unresolved, and the margin is never relaxed (Appendix E).

Freeze order throughout: development $\rightarrow$ formula freeze $\rightarrow$ label-blind extraction $\rightarrow$ rule freeze $\rightarrow$ single formal run, with post-hoc diagnostics labeled as such. No generator, clip, factor, threshold or baseline is removed because of observed performance; the same clause binds every fresh block.

4 Results

4.1 Primary result

Table 1 reports the held-out evaluation. Pooled AUROC rises from 0.6808 (Relation) to 0.7177 (TGC); the decision statistic is $+0.0367$ $[+0.0040, +0.0693]$ with $P(\Delta > 0) = 0.986$, so the interval excludes zero under the stricter inference. The gain persists on the same clips under additional

Table 2: Supporting cohorts (descriptive). Wan is the development set of the revised representation; marked rows (†) are Relation development sets; LTX is a supplementary cross-generator diagnostic.

Cohort	Relation	Normal	Depth	Joint	TGC / Δ [95% CI]
Wan, 40 clips, Q27	0.3599	0.5824	0.5275	0.5659	$0.6758, +0.3159\ [+0.1154, +0.5069]$
LTX, 39 clips, Q27	0.4158	0.6211	0.6526	0.6368	$0.7000, +0.2842\ [+0.0447, +0.5132]$
GeneVA exact†, 192	0.5563	0.5658	0.5885	0.5820	$0.6146, +0.0583\ [-0.0141, +0.1311]$
Physion†, 140	0.6451	0.6231	0.6259	0.6294	$0.6663, +0.0212\ [-0.0402, +0.0853]$

Q10 compression (equal-generator $+0.0429\ [+0.0073, +0.0781]$), a repeated-content robustness check rather than a second independent test.

Table 4 (Appendix D) splits the gain by generator; its per-generator rows also appear as open markers in Figure 3(a). All three generators have positive point estimates and the largest (vc2, 189 pairs) has the smallest, so the gain is not obviously concentrated in one generator; single-generator dependence is not ruled out, no single-generator interval excludes zero, and every leave-one-out point estimate stays positive while its interval widens.

4.2 Supporting cohorts

Table 2 lists the replay cohorts. Wan/LTX intervals are positive on Q27 but small, generator-specific and development-adjacent; GeneVA-exact and Physion have positive point estimates whose intervals span zero (marked rows are Relation development sets). These rows contextualize rather than add independent evidence.

4.3 Pre-encoding access does not explain the gain

On the 79 verified-label clips that carry matched lossless RGB and Q27 copies of the *same* generations, the gain over Relation is $+0.3042$ on RGB and $+0.3020$ on Q27; the paired difference $+0.0023\ [-0.0536, +0.0556]$ spans zero, and the same conclusion holds across the eleven compared variants of the frozen encoding sweep (largest paired difference 0.0098; Table 5). Encoding changes individual scores (component scores still move on identical content, Appendix Figures 6 and 7), but it does not change the cohort-level advantage. Thus the matched experiment does not support pre-encoding access as the explanation for the advantage, and it does not establish RGB–Q27 equivalence.

4.4 Ablations

Applying the rank transform to Relation alone does not explain the held-out gain. The unranked four-factor average reaches 0.6566 versus 0.7177 ranked; applying the same within-generator rank transform to Relation moves it only $0.6808 \rightarrow 0.6848$ with an interval spanning zero, and TGC still exceeds that rank-normalized baseline by $+0.0329\ [+0.0058, +0.0598]$.

The depth-derived factor is a predeclared comparator, not a replacement primary. The depth factor ϕ_{dep} alone reaches $0.7241/0.7830/0.7605$ on GeneVA/Wan/LTX and is the only factor whose point estimate exceeds Relation on all six evaluated pools. It was predeclared as a fresh-evaluation comparator, so none of these readings retroactively promote it; its role in Eq. 7 remains fixed. With it removed, the three-factor interval touches zero on GeneVA—a sensitivity of the ensemble rather than evidence that all four factors are necessary. The present evidence does not establish that the four-factor ensemble is necessary or superior to ϕ_{dep} alone; TGC remains the frozen primary candidate by protocol.

4.5 Appearance controls

The earlier descriptive Q27/RGB/Physion control rows were too wide to establish non-degradation; the pre-registered enlarged block then resolves the margin (Table 3, enlarged rows; Section 4.7).

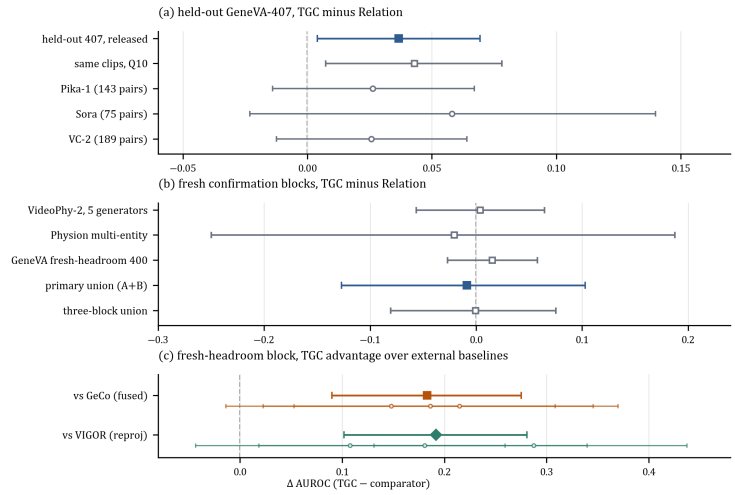


Figure 3: All decision statistics in one view: (a) held-out gain with per-generator rows; (b) fresh confirmation blocks and the two pre-registered unions (VideoPhy-2; Physion multi-entity; sealed GeneVA fresh-headroom; primary fresh union; three-block union); (c) sealed fresh-headroom advantage over the external geometry baselines, with thin markers for each generator sub-block. Filled squares are primary decision statistics, open markers are robustness or descriptive rows, and error bars are 95% paired-bootstrap intervals.

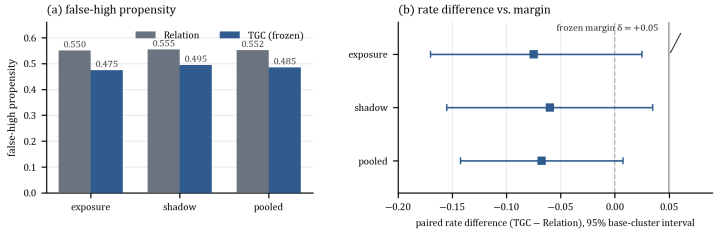


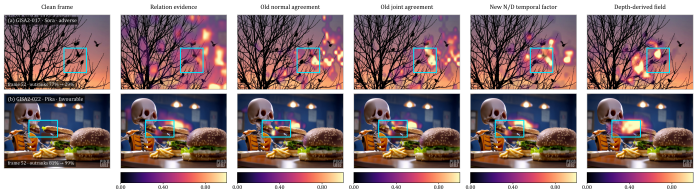
Figure 4: Pre-registered enlarged appearance-control block: false-high propensity for Relation (0.5525) versus TGC (0.485) over 200 clean bases and their exposure/shadow masters (400 paired comparisons; values above bars). The paired difference -0.0675 $[-0.1425, +0.0075]$ is judged against the frozen margin $\delta = +0.05$ (solid line; dashed marks zero).

4.6 Qualitative evidence

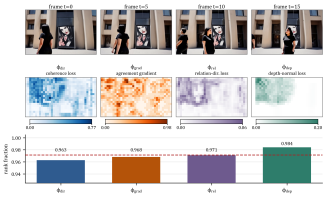
Figure 5 shows one ranking failure and one improvement (panel a; independent human boxes): both carry elevated component evidence inside the annotated region, with stronger region-to-background contrast in the favourable case. Panel (b) shows the four factor readouts and rank aggregation on an illustrative clip.

Table 3: Appearance-control readouts (a of n $[L, U]$): perturbed-above-clean count differences, TGC minus Relation, paired cluster intervals; no descriptive row excludes zero. The two enlarged rows are the pre-registered appearance-control block, whose frozen statistic $U_{95} = +0.0075 < \delta = +0.05$ meets the pre-specified margin.

Block	Compression	Exposure/gamma	Illum./shadow	Texture/material
Q27 (n=24)	—	-4 $[-10, +2]$	-2 $[-7, +3]$	—
Lossless RGB (n=24)	—	$+3$ $[-4, +10]$	0 $[-7, +7]$	—
Physion (n=15)	0 $[-4, +4]$	$+1$ $[-2, +4]$	-1 $[-6, +4]$	-1 $[-7, +5]$
Enlarged exposure (n=200)	—	-15 $[-34, +5]$	—	—
Enlarged shadow (n=200)	—	—	-12 $[-31, +7]$	—



(a) One ranking failure and one improvement: cyan boxes are independent human annotations; heatmaps show component evidence, each colour bar on its own scale.



(b) Decoded frames, the four factor readouts, and rank fractions whose mean forms TGC (dashed).

Figure 5: Retrospectively selected held-out examples: (a) one ranking failure and one improvement; (b) four factor readouts and rank aggregation (y cropped to 0.925–1.005).

4.7 Confirmatory evaluation

On the sealed 400-clip GeneVA fresh-headroom block (Fresh C), TGC reaches 0.6868 pooled AUROC versus 0.5016 (GeCo)/0.4913 (VIGOR), the frozen paired differences excluding zero in its favour (Table 10); it is point-positive but unresolved against Relation ($+0.0151$ $[-0.0273, +0.0577]$), and C is a same-engine replication, not unseen-generator evidence. The VideoPhy-2 block (Fresh A; 170 pairs, 340 clips, five unseen generators) pools neutral (0.6390 vs. 0.6388 AUROC; decision statistic $+0.0036$ $[-0.0565, +0.0643]$); split by generator, the gain is positive in three (Hunyuan $+0.0036$, Ray2 $+0.0088$, VideoCrafter $+0.0667$) and negative in two (CogVideoX -0.0117, Cosmos -0.0496) (Agarwal et al., 2025; Luma AI, 2025), reported as positive and negative cases (Tables 8–9). The predeclared depth-only comparator stays unresolved ($+0.0196$ $[-0.0548, +0.0933]$ on Fresh A). The Physion multi-entity strict block (Fresh B; three fresh generators, 12 pairs) adds -0.0208 $[-0.2500, +0.1875]$. The pre-registered unions are -0.0086 $[-0.1270, +0.1028]$ (A+B, 182 pairs) and -0.0007 $[-0.0808, +0.0752]$ (A+B+C, 382 pairs; Table 7), so fresh superiority over Relation remains unresolved.

The enlarged appearance-control block completes at all 200 bases (400 paired comparisons, no build failures): false-high propensity 0.5525 for Relation versus 0.485 for TGC ($\Delta_{\text{ctrl}} = -0.0675$ $[-0.1425, +0.0075]$), and the frozen statistic $U_{95} = +0.0075 < \delta = +0.05$ meets the pre-specified margin (Table 3, Figure 4).

Limitations and future work. The appearance-control margin is met, yet the count-difference interval still includes zero; p_{FH} is a paired nuisance diagnostic, not a deployment false-positive rate, so the margin does not establish absolute robustness. Two of the five unseen VideoPhy-2 generators are net-negative and neither frozen fresh union resolves a gain over Relation. Future work targets independent multi-generator confirmation of the unchanged method (fixed-reference, generator-agnostic variants).

5 Conclusion

This study contributes a controlled diagnosis, a frozen scoring method, and a reproducible protocol. An encoding diagnostic localizes much of the encoding-induced score variation to the amplitude-agreement readout, where direct normal/depth agreement rerankers add no reliable value over Relation; a matched RGB/Q27 comparison shows no resolved advantage from pre-encoding access, so the gain belongs to the readout rather than the input format. The frozen four-factor TGC exceeds Relation by $+0.0367$ AUROC ($[+0.0040, +0.0693]$) on the GeneVA 407-pair set under equal-generator, prompt-cluster-aware inference; the gain is not explained by the rank transform of Relation alone, and all leave-one-generator-out point estimates stay positive. TGC stands clearly ahead of the external geometry baselines GeCo and VIGOR on the sealed fresh-headroom block (0.6868 versus 0.5016 and 0.4913), and the pre-registered appearance-control evaluation meets its frozen non-inferiority margin; a CPU-only fixture recomputes the reported tables. No frozen fresh block or union resolves a gain over Relation, so fresh superiority stays unresolved and cross-generator generalization remains an open limitation.

AI USE STATEMENT

Generative AI tools were used for drafting, editing and code-bootstrap assistance during manuscript and package preparation. All experiments, statistics and conclusions were designed, executed and verified by human authors; AI-assisted text was reviewed and corrected by the authors, who take responsibility for the final content.

ETHICS STATEMENT

The study uses publicly generated video content and public benchmark annotations, with additional internal annotations performed by an author; no user or demographic data are processed. The work concerns error triage for generated media; possible misuse (automated flagging of human-made video) is limited because the signals are designed and evaluated only on generator cohorts, a limitation stated in the paper.

REPRODUCIBILITY STATEMENT

The frozen scoring formulas, evaluation rule and table-reproduction scripts are released as a minimal CPU-only package with a self-test and an expected-output README (no model weights, no bundled videos; per-clip scores and ledgers are shipped, cached feature arrays are referenced by path). Appendix E summarizes the fixture and its one-command expected output, and Appendix D lists the exact reproducible cohorts. The method is training-free aside from frozen inference components; generation and feature-extraction pipelines are described with their freeze manifests.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9535–9545. IEEE, 2024.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, volume 2026, pp. 118456–118470, 2026.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320. IEEE, 2024.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22831–22840. IEEE, 2025.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10766–10776. IEEE, 2021.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Leslie Gu, Junhwa Hur, Charles Herrmann, Fangneng Zhan, Todd Zickler, Deqing Sun, and Hanspeter Pfister. Geco: Evaluating geometric consistency for video generation via motion and structure. In *European Conference on Computer Vision*, pp. 499–517. Springer, 2026.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Jenna Kang, Maria Beatriz Silva, Patsorn Sangkloy, Kenneth Chen, Niall L Williams, and Qi Sun. Geneva: A dataset of human annotations for generative text to video artifacts. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6174–6183. IEEE, 2026.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pp. 18–35. Springer, 2024.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labeling real videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502. IEEE, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Jiahao Lin, Weixuan Peng, Bojia Zi, Yifeng Gao, Xianbiao Qi, Xingjun Ma, and Yu-Gang Jiang. Brokenvideos: A benchmark dataset for fine-grained artifact localization in ai-generated videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13421–13427, 2025.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149. IEEE, 2024.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36:62352–62387, 2023.

Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25963–25973. IEEE, 2024.

Luma AI. Ray2 generative video model. Proprietary public model release, 2025. No accompanying technical report at the time of writing.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10106–10116. IEEE, 2024.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, pp. 538–554. Springer, 2022.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20087–20097. IEEE, 2023.

Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, et al. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*, 2024.

Jiaxin Wu, Yihao Pi, Yinling Zhang, Yuheng Li, and Xueyan Zou. Quantitative video world model evaluation for geometric-consistency. *arXiv preprint arXiv:2605.15185*, 2026.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381. IEEE, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Tengjiao Yin, Jinglei Shi, Heng Guo, and Xi Wang. Vigor: Video geometry-oriented reward for temporal generative alignment. *arXiv preprint arXiv:2603.16271*, 2026.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9009–9019. IEEE, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

A Statistical Details

Bootstrap construction. Clusters are sets of clips sharing a prompt key (same base prompt, including different camera conditions of that prompt, and clips derived from the same source video). Within each generator, clusters are resampled independently with replacement, stratified by class inside that generator; Relation and the candidate read identical views; generator-level AUROC deltas are averaged equally; 95% intervals are percentile intervals over three seeds of explicit paired bootstrap (200k draws for pooled descriptive intervals, 20k per seed for decision and control intervals). $P(\Delta > 0)$ is reporting shorthand for the bootstrap proportion of deltas above zero; no normal approximation is used. Cohort-relative ranks are computed once per full cohort before resampling and are reused by every replicate with identical views across methods; they are never recomputed per replicate. For the 407 pool, the released cluster mapping has all-distinct keys, so cluster resampling degenerates to clip resampling within generator as stated. Leave-one-out rows are the same estimator on the remaining generators.

Control bootstrap. A clean/perturbed pair stays in the same draw, and the two nuisance variants of one base clip move as one unit; the estimator is the perturbed-above-clean count difference (equivalently the false-high propensity difference of Eq. 8, with the same pairing).

B Negative Result and Encoding Diagnostic

B.1 The original verifier and its negative result

The incumbent verifier for candidate-issue triage is the *Relation* pipeline, which lifts a semantic-relation field over a fixed node grid sampled from tracked video content, and yields a clip-level score together with per-node field summaries. The original geometry rerankers aggregate over that grid a score of the form

$$s \;=\; \sum_{n=1}^{N} m_n \, u_n, \tag{9}$$

where m_n is the normalized relation node-drift mass at node n and u_n is the spatially smoothed agreement between the semantic assignment of the node and the normal/depth assignment of the same pipeline, i.e., how often the geometry channels "agree" with the semantic reading of that node under the relation support. Normal-, depth- and joint-variants differ only in which agreement channel enters Eq. 9.

Under the frozen evaluation rule (paired, class-stratified bootstrap with shared views, AUROC as the ranking measure), none of the three rerankers exceeded Relation on the GeneVA set of 407 local-drift/No-Issue pairs (Table 1): pooled AUROC $0.6132/0.6294/0.6313$ against 0.6808. We retain that negative conclusion for these rerankers throughout the paper.

B.2 What changes when only the encoding changes

Because the released videos are already encoded, the natural suspicion is compression: perhaps the geometry itself is gone, and no reranker can work on delivered video. We tested this with a bounded, pre-registered generator-direct run: fixed prompts, seeds and camera controls, each clip exported losslessly at generation time and re-encoded in matched copies (Q10–Q100 plus the delivered default). This isolates encoding from content because the two versions are *the same generation*. Three observations follow, each reproduced in the frozen ledger that accompanies the code package.

(a) Scores move, sometimes by a lot. On a full-pipeline single-clip scan (LTX R008, Figure 6), the incumbent channels are non-monotone across quality: 14 of 18 pooled channel series are non-monotone, 6 peak away from both endpoints, and endpoints split 8 downward, 7 upward, 3 flat, so no single scalar (e.g., "Q27 costs x AUROC") summarizes the response, and no fixed quality level reliably favors a channel.

(b) The sensitivity sits in the agreement term. Decomposing Eq. 9 between two encodings of the same clips, the agreement term carries 57–71% of total absolute variation and the support mass 14–25%; the score change correlates 0.90–0.98 with the agreement term versus -0.13 to $+0.51$ with the mass term. Across tested conditions, spatial field patterns remain highly correlated (0.79–0.999), while some temporal first differences show sign agreement of only 0.50–0.53.

(c) The same field, read differently, moves less. Reading the *level* (spatial mean) of the agreement field across an encoding ladder, the readout falls to 0.914–0.941 of its least-compressed value on the three pools with full ladders; reading the support-weighted *gradient energy* of the identical field, the same comparisons stay within 0.966–1.019. At the cohort-summary level, level readouts move more and structural summaries move less under the tested encodings.

What this does and does not show. These are statements about the measurement, not about the cause of local drift. They neither prove that geometry is a useful verification cue nor show that compression destroys it. They show that an amplitude-agreement readout over a shared support is an unstable place to put a decision, and they suggest the experiment that separates the hypotheses: score the same clip with the same candidate under lossless and encoded copies. On the 79 verified-label clips carrying both inputs, the candidate gain over Relation is $+0.3042$ $[+0.1559, +0.4487]$ on lossless RGB and $+0.3020$ $[+0.1497, +0.4503]$ on matched Q27; the paired difference is $+0.0023$ $[-0.0536, +0.0556]$ ($P = 0.54$). The matched experiment therefore provides no evidence that pre-encoding access is the source of the improvement; the earlier conditional hypothesis of a pre-encoding-RGB advantage is updated accordingly (Appendix D).

B.3 Design consequence

The diagnostic leaves two candidate explanations for the original failure—(i) the geometry channels carry little drift-relevant signal, or (ii) the readout discards it. The representation developed in Section 3 is designed to use only structural and directional *consistency* of geometry over time, to avoid agreement amplitudes, and it was frozen before the held-out labels were used; whether it adds discrimination is then answered in Section 4.

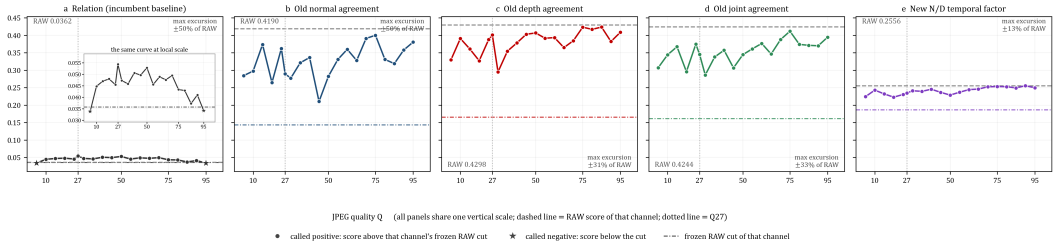


Figure 6: Single-clip encoding scan (LTX R008): the incumbent channels across twenty JPEG qualities plus lossless RAW. Markers show, per quality, whether a channel stays above its RAW-derived reference cut (circle) or falls below it (star); the pool-level RAW-to-Q27 flip counts are in the figure receipt.

C Encoding Diagnostic Record

Single-clip and ladder statistics quoted in Section B are recorded in the figure receipt files shipped with the package: the R008 X-D scan ledger, per-quality decisions against the frozen RAW reference pool (38 strict-label LTX clips, R008 excluded), pool-level reads (level 0.914–0.941; gradient energy 0.966–1.019), and the agreement/mass decomposition (57–71% / 14–25%; correlations 0.90–0.98 / -0.13 to $+0.51$). LTX ladder rows reproduce `outputs/xd_scan_R008_scores.json` with maximum absolute difference 0; the decision markers use the RAW condition of the 96-clip generator-direct audit restricted to LTX and the strict labels (38 clips, 19/19), reproducing the strict-label LTX Q27 decision markers recorded in the frozen ledger exactly.

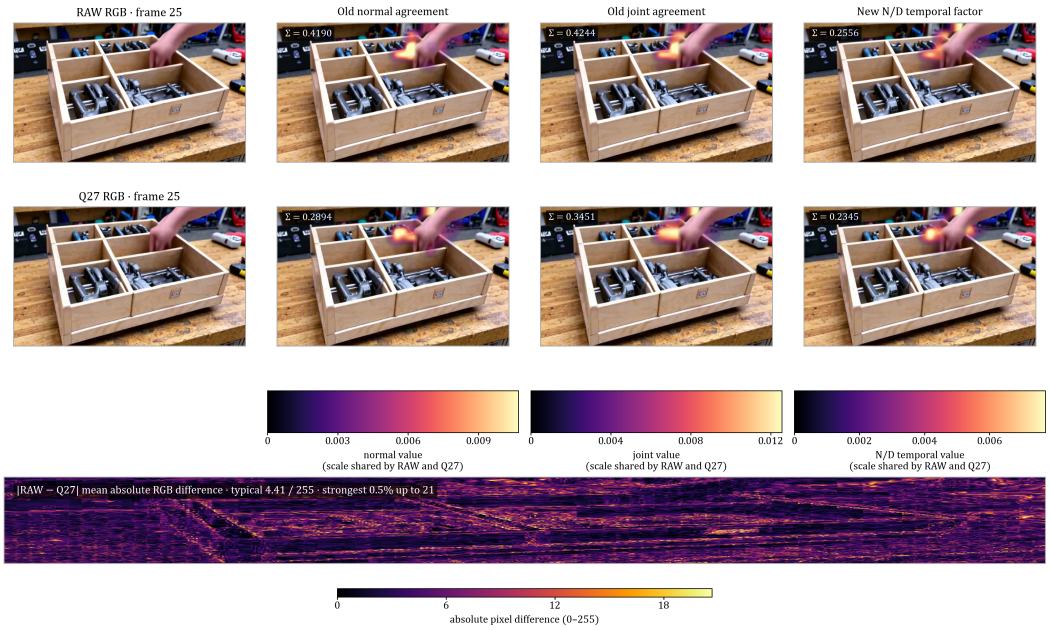


Figure 7: Where the old score loses evidence after re-encoding (LTX R008, RAW vs. Q27); the $|\text{RAW}-\text{Q27}|$ strip isolates the encoding-induced change.

D Additional Results and Ablations

Confirmatory ledger (full). The completed confirmatory blocks and the external-baseline context rows that accompany the VideoPhy-2 media are listed here for the record (Tables 6, 7, 8, 9 and 10). Discrimination rows (AUROC) for the external geometry baselines were pre-registered only on the sealed same-engine block; on the VideoPhy-2 block their frozen adapters contribute descriptive raw-readout rows, and no post-hoc AUROC is computed there.

Table 4: Generator-level and leave-one-generator-out views of the 407-pair gain; leave-one-out rows are sensitivity checks, not independent tests.

View	Pairs	Rel. AUROC	TGC AUROC	Δ [95% CI]	$P(\Delta > 0)$
pika1	143	0.7175	0.7438	$+0.0263$ $[-0.0140, +0.0670]$	.900
sora	75	0.6480	0.7060	$+0.0580$ $[-0.0231, +0.1399]$	.918
vc2	189	0.6747	0.7004	$+0.0257$ $[-0.0125, +0.0641]$	.905
Excl. pika1	264	—	—	$+0.0419$ $[-0.0031, +0.0867]$	.966
Excl. sora	332	—	—	$+0.0260$ $[-0.0017, +0.0538]$	.967
Excl. vc2	218	—	—	$+0.0422$ $[-0.0034, +0.0881]$	.965

Table 5: Lossless RGB vs. Q27 on the same 79 verified-label clips: gain over Relation per condition and the paired difference. The full eleven-variant encoding sweep, with largest paired difference 0.0098, is recorded in the shipped frozen ledger.

Condition	Relation	TGC	Δ [95% CI]	P
Lossless RGB	0.3843	0.6886	$+0.3042$ $[+0.1559, +0.4487]$	$>$ .9999
Matched Q27	0.3850	0.6869	$+0.3020$ $[+0.1497, +0.4503]$	.9999
RGB $-$ Q27 (paired)	—	—	$+0.0023$ $[-0.0536, +0.0556]$	.5395

Ablation row set (GeneVA 407): TGC 0.7177; unranked four-factor mean 0.6566; rank-transformed Relation 0.6848 (the change from Relation's 0.6808 spans zero); TGC minus the rank-normalized baseline $+0.0329$ $[+0.0058, +0.0598]$; the depth factor ϕ_{dep} alone 0.7241; the three-factor variant interval touches zero. Single-factor GeneVA deltas are shipped in the package's ablation output rather than quoted here.

CamI2V (Zheng et al., 2024) supplementary rows carry a weak relation baseline (absolute AUROC near chance), so their gains are supporting evidence, not main results. With the three-factor variant: RAW 0.5088 ($+0.2133$ $[+0.0484, +0.3793]$), Q27 0.5177 ($+0.1982$ $[+0.0219, +0.3757]$), Q70 0.4781 ($+0.1816$ $[+0.0083, +0.3569]$); Relation $0.2955/0.3195/0.2966$.

Qualitative (development) example: Wan R072 full-frame component fields are shown in Figure 8; no independent spatial annotation exists for this clip and Wan is the development set, so it illustrates mechanism, not evaluation.

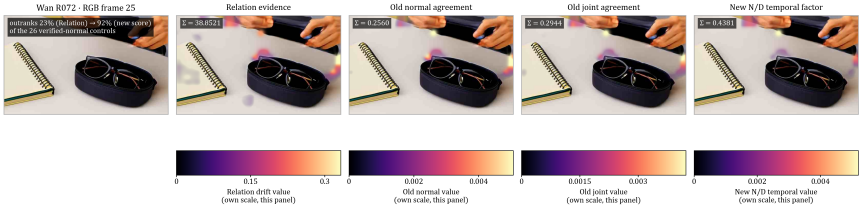


Figure 8: Wan development-set example of measured component fields for one clip; full-frame display, no independent spatial annotation.

E Frozen Protocols and Reproducibility Fixture

Final validation protocol (frozen, summary). Primary: TGC only, with formulas, weights and inputs unchanged. Comparators: Relation, the depth factor ϕ_{dep} alone, historical normal/depth/joint (negative references), GeCo, VIGOR, VBench selected dimensions; optional VideoPhy/PDI-Bench. Data: fresh generators distinct from all development cohorts; fixed candidate lists (prompts, source images, seeds, camera settings) sealed by manifest hash before scoring; one generation per task with a technical retry limit of one; human labels in four categories (strict local drift / fully normal / other issue / uncertain) assigned without score access, with other/uncertain excluded and reported; insufficiency of any label class is reported, never repaired by continued sampling. Statistics: equal-generator, prompt/source-cluster paired bootstrap, shared views, 95% interval, per-generator rows, leave-one-out sensitivity; all intervals and margins are pre-specified before scores are seen, and the manuscript reports them as computed rather than as binary verdicts. Anti-post-hoc clause: *no generator, clip, factor, threshold or baseline is removed because of observed performance.*.

Table 6: External geometry baselines on the same VideoPhy-2 block (frozen adapters, packaged release defaults; higher indicates more corruption). GeCo fused score; VIGOR mean reprojection error in pixels (pooled mean 8.2810; pooled median 4.5473). Descriptive context rows, not on the AUROC-gain scale; baseline discrimination was pre-registered on the sealed block only (Table 10).

Generator	Clips	GeCo fused	VIGOR reproj. err. [px]
cogvideo	96	0.1021	7.8708
cosmos	90	0.0395	8.0649
hunyuan	62	0.0183	5.7247
ray2	62	0.0721	8.6186
videocrafter	30	0.0493	14.8271
Pooled	340	0.0601	8.2810

Table 7: Fresh confirmation summary: per-block and union decision statistics (equal-generator, cluster-paired bootstrap) with the held-out reference; ϕ_{dep} is the predeclared depth-only comparator. External-baseline pooled AUROCs on the sealed block are in Table 10; the VideoPhy-2 external rows (Table 6) are descriptive, not on the gain scale.

Block / union	Pairs	TGC − Relation [95% CI]	ϕ_{dep} − Relation [95% CI]
Held-out GeneVA 407 (reference)	407	$+0.0367$ $[+0.0040, +0.0693]$	—
VideoPhy-2 (Fresh A, 5 unseen gen.)	170	$+0.0036$ $[-0.0565, +0.0643]$	$+0.0196$ $[-0.0548, +0.0933]$
Physion multi-entity (Fresh B, 3 gen.)	12	-0.0208 $[-0.2500, +0.1875]$	—
Sealed GeneVA headroom (Fresh C, same engine)	200	$+0.0151$ $[-0.0273, +0.0577]$	$+0.0227$ $[-0.0264, +0.0719]$
Primary fresh union (A+B)	182	-0.0086 $[-0.1270, +0.1028]$	—
Three-block union (A+B+C)	382	-0.0007 $[-0.0808, +0.0752]$	—

The optional VBench dimension rows and PDI-Bench adapters are recorded as unrun. The enlarged appearance-control block then completed under this protocol: 200 bases and their 600 built masters, of which the 400 paired comparisons under the two frozen nuisance families entered the analysis with no missing families; $\mathrm{U}_{95} = +0.0075 < \delta = +0.05$ meets the frozen margin (Table 3).

Appearance-control NI protocol (summary). Δ_{ctrl} and false-high are defined as in Eq. 8; the frozen check is $\mathrm{U}_{95}(\Delta_{\mathrm{ctrl}}) < \delta$, where U_{95} is the upper endpoint of the two-sided 95% paired-bootstrap percentile interval and the decision is therefore one-sided. The margin was fixed at $\delta = +5$ p.p. before control scores were seen: an increase above five percentage points—more than 20 additional appearance-induced ordering inversions on the 400 paired comparisons relative to Relation—was pre-specified as practically meaningful nuisance degradation, and pilot control variance was used only to size the block, never to choose the margin. If the enrolled block cannot resolve the frozen margin within the remaining compute budget the outcome is declared unresolved, and the margin is never relaxed or renegotiated; p_{FH} is a paired nuisance diagnostic, not a deployment false-positive rate, and meeting the margin does not establish absolute robustness; nuisance composition, sample plan, bootstrap unit and exclusions are frozen before scores are seen.

Fixture. The bundled package (self-test plus table reproduction) requires no models, network or videos: `python selftest.py` verifies semantics and numeric invariants, and `python reproduce.py --list` enumerates reachable cohorts, with `407id` reproducing the main table (pooled $0.6808 \to 0.7177$, $\Delta = +0.0369$ $[+0.0090, +0.0648]$) from the shipped score ledgers. Expected outputs are pinned in the package README; all six pooled endpoints were re-run on 2026-09-14 and matched the reported values to the fourth decimal. The bundle also ships the completed fresh-confirmation ledgers (VideoPhy-2, Physion multi-entity, the sealed GeneVA fresh-headroom block and both unions), the appearance-control ledger and build manifests, the frozen external-baseline adapters with the VIGOR failure rule, per-clip score ledgers, cohort membership and the

Table 8: VideoPhy-2 public-label evaluation, positive cases: the three unseen generators with a positive frozen TGC−Relation AUROC gain.

Generator	Clips	TGC − Relation
hunyuan	62	$+0.0036$
ray2	62	$+0.0088$
videocrafter	30	$+0.0667$

Table 9: VideoPhy-2 public-label evaluation, negative cases: the two unseen generators with a negative frozen TGC−Relation AUROC gain.

Generator	Clips	TGC − Relation
cogvideo	96	−0.0117
cosmos	90	−0.0496

Table 10: Primary comparison with external geometry baselines on the sealed 400-clip GeneVA fresh-headroom block: pooled AUROC, higher-ranks-better, with frozen paired intervals versus TGC (negative Δ means TGC ranks better); Relation is the ablation anchor. †: two VIGOR clips whose reprojection readout returned the upstream −1 sentinel (no usable tracked points) are excluded pairwise under the frozen failure rule, and all TGC–VIGOR rows use the same valid subset ($n = 398$); GeCo has no drops after the one-time OOM re-run clause.

Method	Pooled AUROC	Δ vs. TGC [95% CI]
TGC (primary)	**0.6868**	—
Relation (ablation anchor)	0.6655	—
GeCo fused	0.5016	−0.1830 [−0.2754, −0.0899]
VIGOR reproj. err.†	0.4913	−0.1918 [−0.2809, −0.1017]

ranking policy, extractor versions and freeze hashes, and the commands that reproduce every decision interval in this paper.