SAFE4D: CAUSAL VIDEO PROTECTION AGAINST UNAUTHORIZED 4D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent feed-forward geometry models recover cameras, depth and dynamic point clouds from ordinary video, either offline from a complete clip or online as frames arrive, turning a video into a renderable 4D asset within minutes. However, this capability raises serious concerns about privacy and intellectual property, since anyone who records a published video can recover the shape and motion of the people, products and places it shows. Existing protection methods perturb images or complete videos before release, but live broadcasts and video calls are released *causally*: each frame is published as soon as it is captured and cannot be revised, while viewers can record the stream and reconstruct it. Protecting video against 4D reconstruction under such causal release remains unexplored and raises three challenges: (i) *temporal detachment*: a perturbation fixed in image coordinates drifts off the moving content it protects and is averaged out by motion-aligned purification; (ii) *purification uncertainty*: the attacker may reconstruct from the raw or the purified video, and a perturbation tuned to one input leaves the other unprotected; and (iii) *irrevocable release*: each frame is transmitted as an 8-bit image as soon as it is protected, so an update weakened by rounding cannot be corrected later. To address these challenges, we propose Safe4D, the first video protection method against 4D reconstruction under causal release. Safe4D transports the perturbation committed at the previous frame along the estimated scene motion, so that the protection stays attached to the scene points it perturbs. Moreover, we introduce a purification-aware objective on a frozen streaming surrogate, which maximizes the weaker of the depth discrepancies on the direct and the temporally averaged stream, and a verified commit that accepts a sparse 8-bit update only if the quantized frame to be released improves this objective. Empirically, quantitative and qualitative results on four dynamic benchmarks show that Safe4D effectively disrupts the geometry recovered by streaming 4D reconstruction and remains robust under both causal and non-causal temporal purification.

1 INTRODUCTION

Recent advances in feed-forward geometry models (Wang et al., 2024; Zhang et al., 2025) have made dynamic scene reconstruction accessible to anyone who holds a video. Offline models such as VGGT (Wang et al., 2025a) take a complete clip and predict cameras, depth and point maps for all of its frames in one pass, while streaming models such as StreamVGGT (Zhuo et al., 2025) and CUT3R (Wang et al., 2025b) keep a persistent state and update the same geometry as each frame arrives. Optimization-based pipelines such as Uni4D (Yao et al., 2025) further fuse video depth, dense point tracks and dynamic masks into camera poses, dynamic geometry and 3D motion fitted jointly to the whole clip. The recovered geometry initializes and supervises dynamic Gaussian representations (Kerbl et al., 2023; Wu et al., 2024), which turn a video into a renderable, editable 4D asset within minutes.

However, this capability raises serious concerns about the ownership of what a video shows. A published video is no longer only something to watch: it is an input from which a third party can recover the shape and motion of a performer, a product, a stage or a private room, re-render them from new viewpoints and reuse them without the creator's consent. The concern is most acute for video released while it is being captured, as in a sports broadcast, a concert or live-commerce stream,

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.

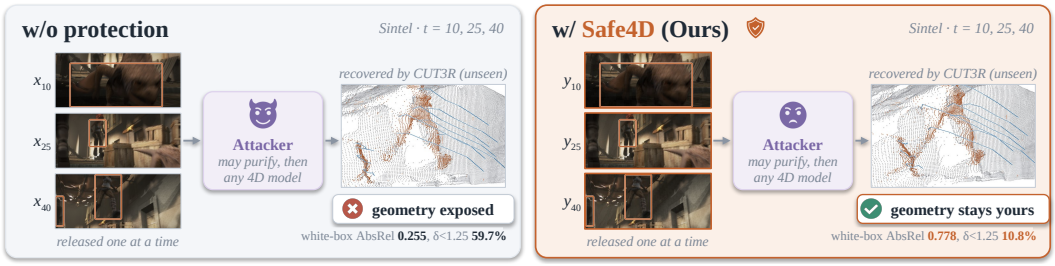


Figure 1: Causal video protection against 4D reconstruction. Frames are released one at a time and cannot be revised, while the attacker holds the whole stream and may run any model. Left: from the unprotected stream it recovers the scene and its moving figure. Right: Safe4D commits each frame from the past alone; the video looks the same, but the recovered geometry is distorted (unseen CUT3R, on a sequence where the protection transfers; Section 5.4).

or a video call: the creator wants the audience to watch in real time, yet viewers can record the stream and reconstruct it, either on the fly with a streaming model or from the complete recording once the stream has ended. Protecting such video from unauthorized 4D reconstruction is therefore a pressing concern.

Existing protection works add imperceptible adversarial perturbations before release. Image-level protection guards photographs against generative editing and personalization (Salman et al., 2023; Van Le et al., 2023) and, more recently, against feed-forward 3D reconstruction and avatar generation (Song et al., 2024; Meng et al., 2026; Song et al., 2026); video-level protection models the temporal structure of a clip to defend it against video customization (Huang et al., 2026; He et al., 2026). However, two assumptions of these methods fail in our setting. First, and most fundamentally, they optimize over the complete content before release. For a live stream the complete video does not exist at protection time: frame t must be committed from the frames observed up to t, and once viewers have recorded it, it cannot be revised. Second, they target generative pipelines rather than geometric recovery, so their objectives are not aligned with the depth and camera estimates that 4D reconstruction relies on. Protection that must be produced *causally*, against an attacker who is not bound by causality, remains unexplored; it must also hold whether the attacker reconstructs the stream online, as frames arrive, or offline, from the complete recording.

To this end, we study causal video protection against unauthorized 4D reconstruction and identify three challenges that distinguish it from existing image and video protection:

- **(i) Temporal detachment.** 4D reconstruction links scene points across frames by correspondence (Yao et al., 2025), and a temporal purifier aggregates observations along the same correspondences. A perturbation optimized at fixed image coordinates, such as a universal adversarial perturbation (UAP) repeated across frames (Moosavi-Dezfooli et al., 2017; Passananti et al., 2024; Huang et al., 2026), is anchored to the screen rather than to the scene: whether it restarts at every frame or is copied to the same pixels, it disagrees with itself at corresponding scene points, and motion-aligned aggregation cancels this disagreement (Section 5.3).
- **(ii) Purification uncertainty.** The attacker may reconstruct from the released video directly or after temporal purification, and the protector cannot know which. Tuning the perturbation to one input strengthens protection there and weakens it on the other (Appendix L), so optimizing either input alone leaves the other exposed.
- **(iii) Irrevocable release.** In a live stream each protected frame is encoded and transmitted to viewers as soon as it is produced: the protector commits frame t from $x_{\le t}$ and the already transmitted $y_{<t}$, and a transmitted 8-bit frame cannot be revised. A continuous update that is rounded at release need not realize the gain predicted for it, and a frame that has overshot cannot be taken back, so each update should be verified on the frame that is actually released.

To meet these challenges, we propose Safe4D, the first video protection method against 4D reconstruction under causal release (Figure 1). Safe4D targets the quantity that the reconstruction routes in our threat model share, the per-frame depth structure, and refines each frame against a frozen streaming geometry surrogate whose causal state mirrors the protector's information. To keep protection attached to the scene, it transports the perturbation committed at the previous frame along the optical flow between the two clean frames, so that the new frame starts from a perturbation

that agrees with its predecessor at corresponding scene points. To remove the attacker's least-cost option, it scores each candidate by the weaker of the depth discrepancies that the surrogate reports on the direct and on the temporally averaged protected stream. Because a released frame cannot be taken back, it changes a bounded number of code values per frame and commits an update only if the realized 8-bit frame improves the objective. Safe4D provides strong white-box protection on four complete benchmarks and retains over 90% of its depth damage under causal and non-causal temporal purification. Our contributions are as follows.

- We are the first to formulate causal video protection against unauthorized 4D reconstruction, in which the protector commits each frame from current and past observations while the attacker holds the complete video, may purify it and may run any model, and we identify its three challenges with evidence for each.
- We propose Safe4D, which transports the committed perturbation along scene motion, optimizes a purification-aware depth objective on a frozen streaming surrogate, and commits sparse 8-bit updates that are verified on the released frame.
- On four complete benchmarks (70 sequences), Safe4D provides strong white-box protection against streaming 4D reconstruction and retains most of it under causal and non-causal temporal purification.

2 Related Work

Protection against unauthorized generation and reconstruction. Adversarial perturbations that obstruct downstream use were first developed against generative editing and personalization (Salman et al., 2023; Van Le et al., 2023) and then extended to geometry: Geometry Cloak protects a single image against feed-forward Gaussian reconstruction (Song et al., 2024) and MVAP-G perturbs jointly available multi-view images against VGGT (Song et al., 2026); adversary-side attacks poison radiance fields and Gaussian splatting through their source views (Fu et al., 2023; Lu et al., 2025). For video, TC-UAP and ChronoLock protect identity and appearance against video customization and show that temporally inconsistent perturbations are erased by frame averaging (Huang et al., 2026; He et al., 2026). These methods optimize over content that is available in full before release; Safe4D commits each frame from the current and past frames alone, and it protects scene geometry, which moves across frames, against an attacker who may exploit temporal redundancy.

Adversarial attacks on video and geometry estimators. Sparse perturbations propagated across frames fool video recognition (Wei et al., 2019), and small perturbations disrupt optical flow (Ranjan et al., 2019) and monocular depth (Wong et al., 2020); online evasion attacks separate an unobservable future from an unmodifiable past (Gong et al., 2019). We take the protector's side under this constraint, score the depth structure recovered from the stream rather than a label, and use flow only to transport the perturbation.

3 Problem Formulation

Consider a video $X = \{x_t\}_{t=1}^{N}$ whose frames are released one at a time, where $x_t \in [0,1]^{3\times H\times W}$ is the 8-bit RGB frame captured at time t, scaled to $[0,1]$. At time t the protector has observed the clean frames $x_{1:t}$ and has already released the protected frames $y_{1:t-1}$; future frames do not exist yet, and released frames have been recorded by viewers and cannot be revised. The protector therefore commits the current frame from the past alone:

$$y_t = x_t + \delta_t, \qquad \delta_t = \mathcal{A}\big(x_{1:t},\, y_{1:t-1}\big), \qquad \|\delta_t\|_\infty \le \epsilon, \tag{1}$$

where $\mathcal{A}$ is the causal protector, δ_t is the realized perturbation of the published 8-bit frame, and ϵ bounds the change of each pixel value. Eq. 1 is the causal contract: y_t depends only on frames up to t, and $y_{1:t-1}$ enters as fixed inputs, not as variables. $Y = \{y_t\}_{t=1}^{N}$ denotes the protected video.

The *protector* has white-box access to a surrogate streaming geometry model f_s, which it queries with RGB frames only; it has no ground-truth 4D geometry of the scene (no depth maps, no camera poses) and no access to the attacker. The *attacker* records the released video Y and reconstructs it either online, running a streaming model as frames arrive, or offline, from the complete recording. It may first apply a purifier $\mathcal{T}$ that exploits temporal redundancy, e.g., by averaging or median-filtering frames

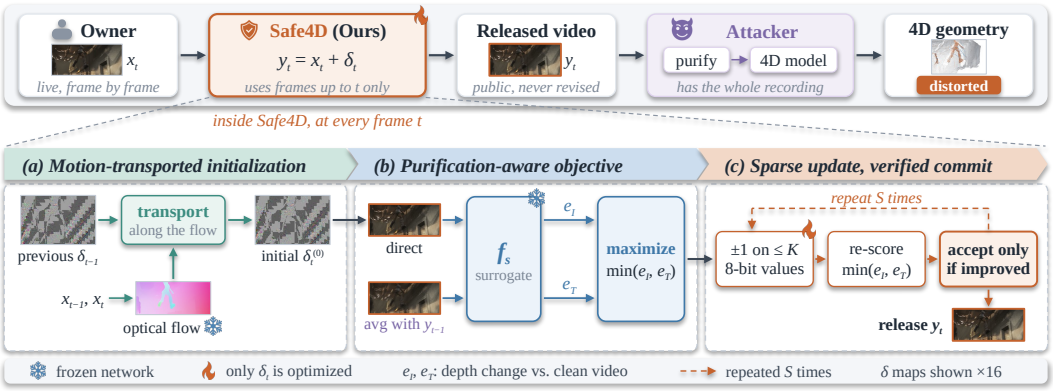


Figure 2: Overview of Safe4D. Top: owner → Safe4D → released video → attacker. Bottom, inside Safe4D at every frame t: (a) motion-transported initialization (Eq. 4), (b) purification-aware objective (Eqs. 5–6), (c) sparse update and verified commit (Eq. 7).

aligned by their motion; an offline attacker may also use future frames. It then runs a geometry model f_q, which may be the surrogate itself (white-box) or an unseen model with a different architecture, training data and preprocessing (transfer), and may fit a dynamic 4D representation from the result. The attacker's actions are not fed back to the protector.

Protection goal and evaluation. The protector aims to keep the video watchable while making the geometry recovered from it unreliable. Following the perturbation-based protection paradigm (Salman et al., 2023; Van Le et al., 2023; Huang et al., 2026), this goal is expressed as maximizing a surrogate geometric loss $\mathcal{L}_{\text{geo}}$ on the frozen surrogate under the budget of Eq. 1, using only the information available at time t:

$$\delta_t^\star = \arg\max_{\|\delta_t\|_\infty \le \epsilon} \mathcal{L}_{\text{geo}}\big(f_s;\, x_{1:t},\, y_{1:t-1},\, x_t + \delta_t\big). \tag{2}$$

Section 4 instantiates $\mathcal{L}_{\text{geo}}$ and the optimizer. On the attacker's side, protection is measured by how much worse the geometry recovered from Y is than that recovered from X under the same processing. For an error metric $m \in \{\text{AbsRel}, \text{ATE}\}$ (depth and camera pose) against ground truth z and a purifier $\mathcal{T}$ (possibly the identity) applied to both videos, we report the paired increment

$$\Delta_m(\mathcal{T}) = m\big(f_q(\mathcal{T}(Y)), z\big) - m\big(f_q(\mathcal{T}(X)), z\big), \tag{3}$$

which separates the effect of protection from the effect of the purifier itself, whose own cost on clean reconstruction we report alongside. Positive values mean that the protected video reconstructs worse. In brief (formulas in Appendix B), AbsRel is the mean relative depth error after the prediction is aligned to the truth by a fitted scale (and, per sequence, a shift in disparity), which removes the scale ambiguity of monocular depth, and ATE is the root-mean-square error of the camera centers after one similarity alignment (Umeyama, 1991); we also report the depth accuracy $\delta{<}1.25$, the fraction of pixels whose aligned depth is within a factor of 1.25 of the truth. Throughout, arrows point toward stronger protection: higher AbsRel and ATE, lower $\delta{<}1.25$. Ground truth is used only for evaluation, not by the protector.

4 Safe4D: Causal Geometry-Aware Video Protection

We propose Safe4D, whose three components answer the three challenge: a motion-transported initialization (temporal detachment), a purification-aware geometric objective (purification uncertainty), and a verified commit on the 8-bit lattice (irrevocable release). Figure 2 gives an overview; X and Y denote the clean and the protected video, $\mathcal{T}$ a purifier, and only y_t changes at time t.

Depth structure as the shared target. Whether the attacker runs a streaming model or an offline model, these routes pass through the same intermediate quantity, the depth structure predicted for each frame, which feed-forward models output directly, and 4D fitting takes as initialization and supervision. Safe4D therefore corrupts the relative depth structure recovered from the protected stream, so that the damage is inflicted before it reaches a downstream representation, and measures this corruption on a frozen streaming surrogate f_s whose causal state is built from the frames the protector has seen, so that the objective is available at every time step within the contract of Eq. 1.

4.1 MOTION-TRANSPORTED INITIALIZATION

4D Streaming reconstruction models and temporal purifiers aggregate evidence along cross-frame correspondences, so the effect of a perturbation depends on its agreement across the frames that see the same scene point. Under causal setiing, this agreement cannot be enforced by joint optimization, because earlier frames are fixed; what remains to decide is where the perturbation of the new frame starts. Restarting discards what earlier frames have built up, and copying the previous perturbation at the same pixels attaches it to whatever content now occupies them, so on moving regions perturbation and scene drift apart and a motion-aligned average $\frac{1}{2}(\delta_t + W\delta_{t-1})$ attenuates it (Appendix C). The perturbation should therefore be carried along the motion of the scene, one step at a time.

As shown in Figure 2(a), we estimate the bidirectional optical flow between the clean frames x_{t-1} and x_t with a frozen flow network (Teed & Deng, 2020). Let W_t gather the previous perturbation realized $\delta_{t-1} = y_{t-1} - x_{t-1}$ at the estimated correspondences, and let M_t mark the pixels whose correspondence lies inside the image and passes a forward–backward consistency check. The current frame is initialized as

$$\delta_t^{(0)} = M_t \odot W_t\delta_{t-1} + (1 - M_t) \odot \delta^{\mathrm{chk}}, \qquad y_t^{(0)} = Q_8\Big(\mathrm{clip}\left(x_t + \Pi_\epsilon\delta_t^{(0)}, 0, 1\right)\Big), \tag{4}$$

where Π_ϵ projects onto the ℓ_∞ box and Q_8 rounds to the 8-bit grid (gathering uses the nearest correspondence, so the integer perturbation is moved rather than blurred). Pixels without a reliable correspondence, and the first frame, start from a fixed deterministic pattern δ^{chk}, a ± 2-code checkerboard (Appendix A); a non-zero start is needed because the squared depth discrepancy below has zero gradient when the protected and the clean predictions coincide. Flow is used only to transport the perturbation, and the geometric objective refines the initialization.

4.2 PURIFICATION-AWARE GEOMETRIC OBJECTIVE

Persistence alone does not make a perturbation protective: it must change the geometry that the attacker recovers, so we score a candidate by how far the surrogate's depth prediction on the protected stream departs from its prediction on the clean stream, in a form that is invariant to the scale ambiguity of monocular depth and needs no ground truth. Which input to score is the second question: the attacker may purify the video by temporal aggregation, or may not, and optimizing for one input weakens the protection on the other (Appendix L). We therefore build a surrogate causal purifier, the two-frame average, into the objective, evaluate both inputs and maximize the weaker discrepancy, so that the attacker gains little by switching (Appendix C).

As shown in Figure 2(b), the frozen surrogate f_s maintains four causal streams with separate histories: the clean video X, the purified clean video $\mathcal{T}(X)$, the protected video Y and the purified protected video $\mathcal{T}(Y)$, where during generation $\mathcal{T}$ is the causal two-frame average of the original 8-bit frames (first frame unchanged). Let $\widehat{D}_{I,t}$ and $\widehat{D}_{T,t}$ be the current-frame depth predicted on Y and on $\mathcal{T}(Y)$, and $\widehat{D}^{\mathrm{ref}}_{I,t}$ and $\widehat{D}^{\mathrm{ref}}_{T,t}$ the predictions on the correspondingly processed clean streams X and $\mathcal{T}(X)$. On the mask $\Omega_{b,t}$ of valid clean depths, the discrepancy of branch $b \in \{I, T\}$ is the centered log-depth difference

$$e_b = \frac{1}{|\Omega_{b,t}|}\sum_{p\in\Omega_{b,t}}\Big[c_b\big(\log\widehat{D}_{b,t}\big)(p) - c_b\big(\log\widehat{D}^{\mathrm{ref}}_{b,t}\big)(p)\Big]^2, \qquad c_b(Z) = Z - \mathrm{mean}_{\Omega_{b,t}}\, Z, \tag{5}$$

where centering removes a global scale change so that only the relative depth structure counts, and comparing each protected branch with the identically processed clean stream keeps the purifier's own effect out of the score. The frame objective is the smooth lower envelope of the two branches,

$$J_t = -\tau\log\Big(e^{-e_I/\tau} + e^{-e_T/\tau}\Big) + \tau\log 2, \tag{6}$$

which approximates $\min(e_I, e_T)$ for a small temperature τ and whose gradient concentrates on the weaker branch, so that maximizing J_t pushes the recovered depth away from the clean solution whether or not the attacker purifies. Gradients reach only the current frame; committed frames and surrogate histories are fixed.

4.3 Sparse update and verified commit on the 8-bit lattice

The published object is an 8-bit image that cannot be revised: a continuous update that is rounded at publication need not realize the gain predicted for it, and a frame that has overshot cannot be taken back, so the decision to publish should rest on the frame that will actually be released. Moreover, the two branches may favor different directions, so the step should be chosen with both in view, and the transported initialization already carries the structure built up by earlier frames, so the update should add a bounded, sparse innovation rather than overwrite it.

As shown in Figure 2(c), let $\bar{y}_t$ be the current frame, initially $y_t^{(0)}$, and g_I, g_T the gradients of e_I, e_T with respect to it. For nine mixtures $g_\alpha = \alpha g_I + (1-\alpha) g_T$, $\alpha \in \{0, \frac{1}{8}, \dots, 1\}$, and the gradient direction of Eq. 6, a candidate y_t^α moves at most K eligible RGB coordinates of $\bar{y}_t$, those with the largest $|g_\alpha|$, by one code value in the direction of the gradient, where eligibility enforces the image range and the ϵ box. Each candidate is ranked, without a model call, by its linearized endpoint on the actual integer changes,

$$\hat{J}(\alpha) = \min_{b \in \{I,T\}} \left[e_b + \langle g_b,\, y_t^\alpha - \bar{y}_t \rangle \right]. \tag{7}$$

Only the best candidate is decoded, passed through both branches and re-scored on the realized 8-bit frame; it replaces $\bar{y}_t$ when the realized $\min(e_I, e_T)$ increases. This refinement step is repeated S times from the current $\bar{y}_t$ with all histories fixed, after which $\bar{y}_t$ is published. The published frame is read back from its bytes before it is appended to the protected histories, so the surrogate's state is built from the frames the attacker records. Causality and immutability hold by construction: the inputs to frame t are $x_{\le t}$ and $y_{<t}$, and each accepted step changes at most K code values by one level within the ϵ box. Algorithm 1 in Appendix A summarizes the per-frame procedure, which uses $4 + 6S$ surrogate calls per frame (two clean-reference and two commit calls, plus two base, two gradient and two re-scoring calls per step).

5 Experiments

5.1 Experimental setup

Datasets and models. We protect the complete evaluation splits of four dynamic benchmarks for video depth and pose (Yao et al., 2025): MPI Sintel (Butler et al., 2012) (23 sequences), Bonn RGB-D Dynamic (Palazzolo et al., 2019) (26), KITTI (Geiger et al., 2013; Uhrig et al., 2017) (13 drives) and TUM RGB-D dynamics (Sturm et al., 2012) (8, scored on camera pose), every sequence from its first frame; controlled analyses use the first 16 frames of three Sintel sequences, S1–S3 (Appendix A). The white-box surrogate is StreamVGGT (Zhuo et al., 2025); the unseen model is CUT3R (Wang et al., 2025b), a streaming model of different architecture and training data.

Purifiers, metrics and baselines. Generation uses the causal two-frame average of Section 4.2 as $\mathcal{T}$. The attacker may skip purification or apply purifiers that the protector does not see: a causal flow-aligned median of frames $t-2$, $t-1$ and t (*flow-median3*), which an online attacker can run, and two non-causal ones that use the next frame, *forward avg2* (average of frames t and $t+1$) and *centered flow-median3* (flow-aligned median of frames $t-1$, t, $t+1$); their own effect on clean reconstruction is reported with the results (Appendix D). Benchmark depth follows the public protocol (Yao et al., 2025) (one disparity scale and shift per sequence, AbsRel and δ<1.25 over valid pixels); TUM pose is scored by ATE and RPE; controlled analyses report the paired increments of Eq. 3 with per-frame scale alignment. Because existing protections require the complete video, we compare with the causal variants that differ from Safe4D only in the initialization: *same-grid* copies δ_{t-1} at fixed pixels and *reset* restarts every frame from δ^{chk}. All variants use $\epsilon = 16/255$, $K = \lfloor 3HW/32 \rfloor$, $\tau = 10^{-4}$, $S = 10$ refinement steps per frame ($S = 1$ in the single-step study) and frozen RAFT flow (Teed & Deng, 2020) (Appendix A).

5.2 Main results: white-box protection on complete benchmarks

Safe4D inflicts large white-box damage on every benchmark (Table 1). It raises the AbsRel of StreamVGGT from 0.229 to 0.478 on Sintel, from 0.0598 to 0.219 on Bonn and from 0.167 to 0.327 on KITTI (2.09×, 3.66× and 1.96×), lowers δ<1.25 by 34.4, 36.6 and 28.8 points, and on TUM raises ATE by 28.5% and the two relative pose errors by 29.4% and 39.2%. The damage is broad:

Table 1: White-box protection on four complete benchmarks (StreamVGGT, no purification). Depth: AbsRel and δ<1.25 (%); pose: ATE and RPE (m / m / deg). **Bold**/underline: best/second; green: change of *Ours* against the strongest baseline.

Input	Sintel		Bonn		KITTI		TUM dynamics		
	AbsRel ↑	δ<1.25 ↓	AbsRel ↑	δ<1.25 ↓	AbsRel ↑	δ<1.25 ↓	ATE ↑	RPE_t ↑	RPE_r ↑
Clean	0.229	67.7	0.0598	98.3	0.167	72.1	0.0775	0.0240	0.528
Same-grid	0.476	34.8	**0.219**	61.7	0.324	44.9	**0.108**	0.0307	**0.739**
Safe4D (Ours)	**0.478**↑0.539%	**33.3**↑4.27%	0.219↓0.358%	**61.7**↑0.0455%	**0.327**↑1.08%	**43.3**↑3.51%	0.0996↓7.89%	**0.0311**↑1.33%	0.735↓0.535%

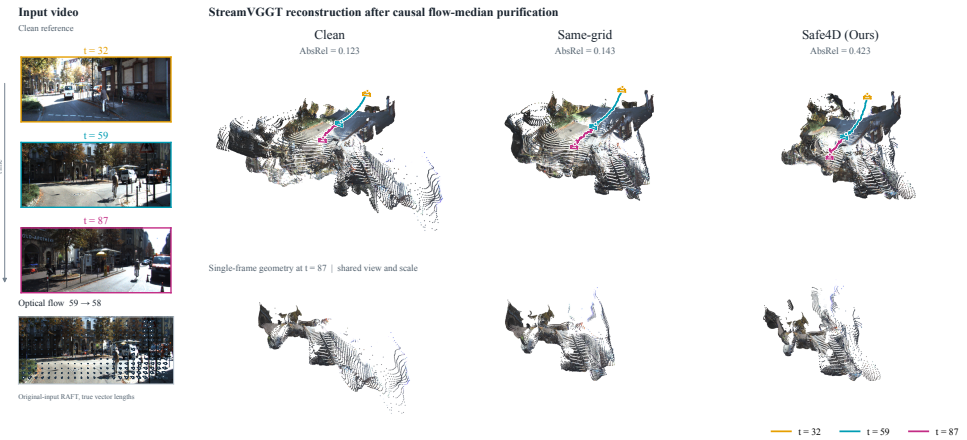


Figure 3: Reconstruction after the causal flow-median3 (StreamVGGT, a KITTI sequence). Top: point clouds of frames 32, 59 and 87 with the predicted cameras, colored by time; bottom: frame 87. AbsRel is computed over the full sequence.

AbsRel increases on 60 of the 62 depth sequences and δ<1.25 falls on 59 (Appendix E). Same-grid, which shares everything with Safe4D except the transport, reaches comparable white-box damage (AbsRel within 1.08%); Safe4D lowers δ<1.25 further on Sintel (1.49 points, 18 of 23 sequences) and KITTI (1.58 points, 8 of 13). Both variants are optimized against the evaluated model; they separate once the attacker purifies (Section 5.3).

5.3 Robustness to temporal purification

Table 2 applies the three purifiers to the protected videos of all 70 benchmark sequences (their cost on clean videos: Appendix D). Safe4D retains 91.3–108% of its unpurified depth increment under every purifier on Sintel, Bonn and KITTI, and 93.7% and 112% of its TUM ΔATE under the centered flow-median3 and forward avg2. The motion-aligned median separates the two carrying variants, as challenge (i) predicts: under the causal flow-median3, same-grid keeps 54.2% of its KITTI increment and 15.7% of its TUM ΔATE, whereas Safe4D keeps 91.3% and 44.9%, a 72.2% and 106% larger increment than same-grid. On Bonn the two variants stay within 1.60% of each other, and forward avg2, which averages frames at fixed pixels, keeps over 93% of the depth increment of both carrying variants; it is the motion-aligned median that separates them. Figure 3 shows the effect on a KITTI sequence: after the causal flow-median3 the same-grid reconstruction stays close to the clean one (AbsRel 0.123 → 0.143), whereas the geometry recovered from Safe4D is distorted (0.423).

Table 2: Temporal purification on the complete benchmarks (StreamVGGT): paired increments of AbsRel and, on TUM, of ATE ($\times 10^{-2}$ m). Green: change of *Ours* against same-grid.

Purification	Init.	Sintel	Bonn	KITTI	TUM
		ΔAbsRel ↑	ΔAbsRel ↑	ΔAbsRel ↑	ΔATE ↑
Direct	Same-grid	0.247	**0.160**	0.156	**3.06**
	Safe4D (Ours)	**0.249**↑1.04%	0.159↓0.492%	**0.160**↑2.25%	2.21↓27.8%
Forward avg2	Same-grid	0.230	**0.152**	0.161	2.21
	Safe4D (Ours)	**0.233**↑1.36%	0.151↓0.244%	**0.173**↑7.32%	**2.47**↑11.8%
Flow-median3	Same-grid	0.205	**0.158**	0.0847	0.482
	Safe4D (Ours)	**0.235**↑14.6%	0.155↓1.60%	**0.146**↑72.2%	**0.993**↑106%
Centered flow-median3	Same-grid	0.246	**0.158**	0.107	1.11
	Safe4D (Ours)	**0.252**↑2.36%	0.155↓1.51%	**0.168**↑57.2%	**2.07**↑85.9%

Table 3: Transfer to the unseen CUT3R (single-step videos of three Sintel sequences; paired increments).

Purification	Init.	ΔAbsRel ↑ ($\times 10^{-3}$) mean	S1	S2	S3	δ<1.25 drop ↑ (points)	ΔATE ($\times 10^{-3}$)
Direct	Reset	6.39	19.5	5.40	−5.71	1.93	+0.740
	Same-grid	5.29	20.2	2.49	−6.78	1.77	+1.19
	Safe4D (Ours)	**10.8**↑68.5%	**28.4**	**9.43**	**−5.54**	**3.51**↑81.3%	−1.10
Flow-median3	Reset	6.93	21.5	4.47	−5.15	3.09	+0.536
	Same-grid	6.20	20.9	2.67	−4.94	2.74	+0.209
	Safe4D (Ours)	**10.8**↑55.8%	**27.6**	**9.33**	**−4.54**	**4.31**↑39.4%	−1.22

Table 4: Native CUT3R geometry (ten-step videos): change of Chamfer distance (CD) and 3D displacement error (EPE) caused by Safe4D (%).

Seq.	Direct ΔCD ↑	Direct ΔEPE	Forward avg2 ΔCD ↑	Forward avg2 ΔEPE
S1	+13.2	+4.34	+11.2	−1.31
S2	+6.50	−2.38	+5.77	−5.71
S3	+9.15	+4.37	+10.1	+3.18

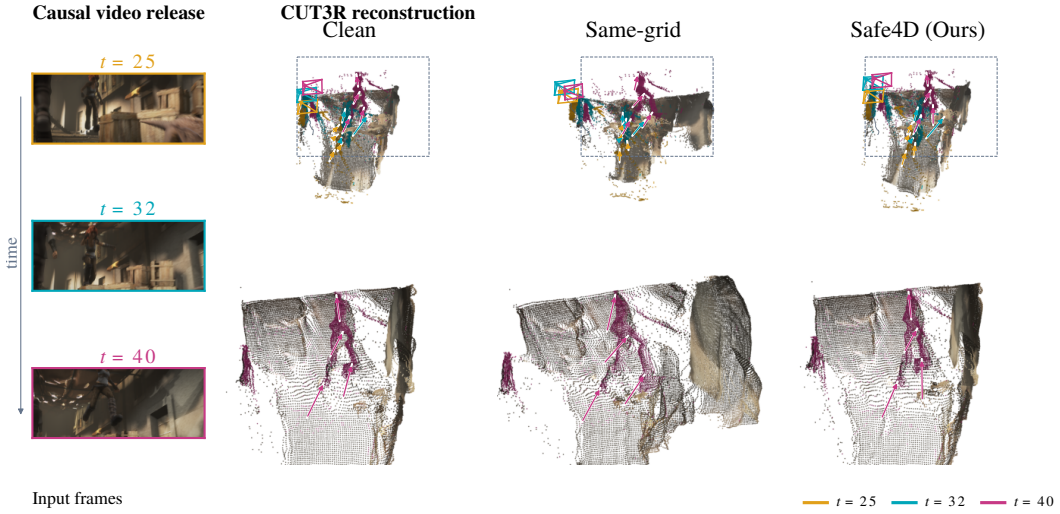


Figure 4: Causal release and 4D reconstruction by the unseen CUT3R. Left: frames released at $t = 25, 32, 40$. Top: recovered 4D scene with cameras and moving points colored by time; bottom: the dashed region at $t = 40$. The protected reconstructions stay recognizable, but the moving figure is torn and displaced.

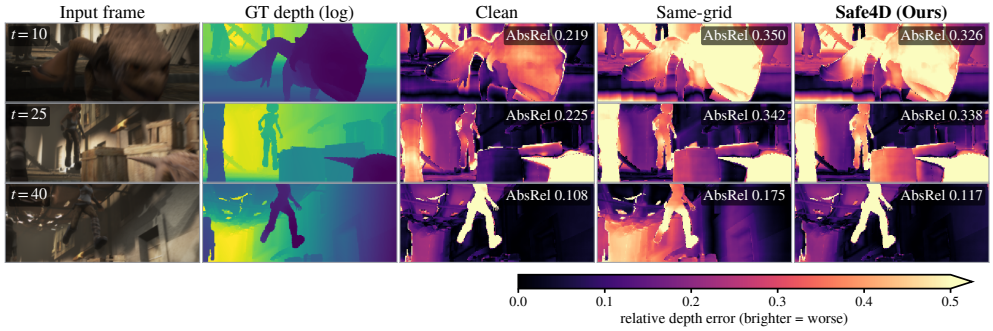


Figure 5: CUT3R depth error on the sequence of Figure 4 (insets: per-frame AbsRel).

5.4 Transfer to an unseen model

We score the protected videos of the three controlled sequences on CUT3R, which the protector does not query. With a single step (Table 3), Safe4D raises the mean AbsRel of CUT3R by 10.8×10^{-3} both without and with flow-median3, 68.5% and 55.8% more than the strongest baseline, and lowers δ<1.25 by 3.51 and 4.31 points. With ten steps its increment is 2.21–3.45× that of same-grid under all three treatments (Table 6). Figures 4 and 5 show a sequence on which the protection transfers: the moving figure is torn and displaced in the recovered 4D scene, and the depth error rises.

Beyond per-frame depth. Per-frame depth scores ignore whether the frames still fit together in 3D, so we also score the native point maps that CUT3R recovers from the ten-step videos in its own world coordinates (Table 4): Safe4D increases the Chamfer distance in all six sequence–purification conditions, by 5.77–13.2%, while the frame-to-frame 3D displacement changes by −5.71% to

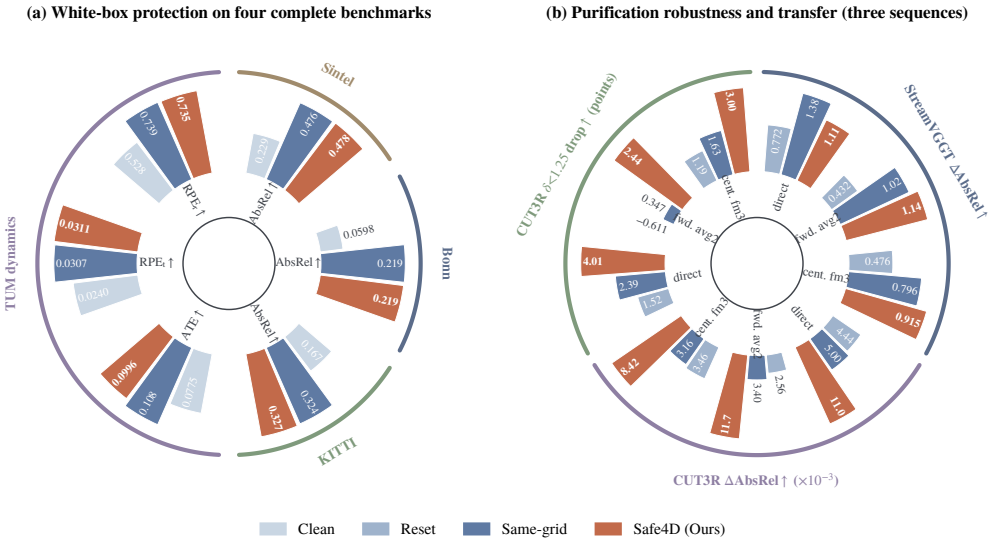


Figure 6: Results at a glance. (a) White-box protection on the four benchmarks (Table 1); (b) the ten-step study on three Sintel sequences (Table 6). Bars are scaled per metric; longer bars mean stronger protection.

$+4.37\%$, so the damage is specific to shape. Averaged over all 70 benchmark sequences, the error of CUT3R does not increase: the protection transfers strongly on some sequences, such as the one in Figures 4 and 5, but not on others (Appendix G).

5.5 Ablation Study

Temporal detachment. Reset, same-grid and Safe4D differ only in where the perturbation of a new frame starts, at nearly the same perturbation energy (RMS 2.01–2.06 codes). On three Sintel sequences the non-causal purifiers remove much of the damage of reset ($0.772 \rightarrow 0.432$ and 0.476), whose perturbation disagrees across frames, while Safe4D exceeds same-grid by 10.9% and 15.0% (Figure 6(b)); on the complete benchmarks transport gives the larger increment under the motion-aligned median (Table 2) and the lower δ<1.25 on Sintel and KITTI (Table 1), and on the three controlled sequences it transfers best to CUT3R (Table 3).

Purification uncertainty. In an earlier prototype, optimizing only the purified branch doubles the post-purification depth increment but lowers the unpurified pose increment by 28.9% (Appendix L); the final objective therefore takes the weaker of the two branches.

Irrevocable release. Every proposal was accepted in the short controlled runs, whereas on a 300-frame sequence a fraction of the later proposals survived re-scoring (Appendix N); the commit acts as a safeguard on long streams at a cost of $4 + 6S$ surrogate calls per frame.

6 Conclusion

In this work, we formulate causal video protection against unauthorized 4D reconstruction, where each frame is protected as it is released while the attacker holds the complete recording, and identify three challenges: temporal detachment, purification uncertainty and irrevocable release. We propose Safe4D, the first video protection method for this setting, which transports the committed perturbation along scene motion to keep the protection attached to the moving scene, maximizes a purification-aware depth objective on a frozen streaming surrogate so that the protection holds whether or not the attacker purifies, and releases each frame through a verified sparse 8-bit commit. Experiments on four dynamic benchmarks show that Safe4D effectively disrupts streaming 4D reconstruction and remains robust under causal and non-causal temporal purification, taking a step toward protecting live video from unauthorized 4D reconstruction. Extending the protection beyond the surrogate to other reconstruction models remains future work.

AI USE STATEMENT

Generative AI tools assisted with refining and polishing the text. They were not used to synthesize benchmark observations, ground truth, or reported experimental results. Numerical claims are checked against saved experiment records. The authors remain responsible for verifying the code, scientific claims, citations, and manuscript.

ETHICS STATEMENT

This work is motivated by creator-controlled protection against unauthorized geometric reconstruction of published video. Adversarial methods are dual use: the same procedure could be used to disrupt legitimate mapping, robotics or reconstruction services. The study uses bounded digital perturbations of public research datasets and public research models and involves no human subjects or private user data. A release will document the limitations reported here, include the clean-utility evaluation of purifiers, follow the licenses of all models, weights and datasets, and avoid packaging the method as a service-disruption tool. Because the protection varies across scenes, models and preprocessing choices, it should not be presented as a copyright or privacy guarantee.

REPRODUCIBILITY STATEMENT

The publication setting, threat model and metrics are defined in Section 3; the initialization, objective, update rule, acceptance test and per-frame call budget in Section 4 and Algorithm 1; and every experiment in Section 5 reports its data, models and purifiers. Appendix A lists implementation constants, preprocessing, benchmark splits, checkpoints and compute; Appendix D specifies the purifiers and their effects on clean reconstruction; Appendices E, F, H and J give per-sequence and per-asset numbers. Frozen protected videos, per-frame records, target predictions and the scripts that produce every figure and table from those records will be released with the code.

REFERENCES

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, pp. 611–625, 2012.

Yonggan Fu, Ye Yuan, Souvik Kundu, Shang Wu, Shunyao Zhang, and Yingyan Celine Lin. NeRFool: Uncovering the vulnerability of generalizable neural radiance fields against adversarial perturbations. In *ICML*, volume 202, pp. 10482–10493, 2023.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *IJRR*, 32(11):1231–1237, 2013.

Yuan Gong, Boyang Li, Christian Poellabauer, and Yiyu Shi. Real-time adversarial attacks. In *IJCAI*, pp. 4672–4680, 2019.

Jiaming He, Jiashu Zhang, Guanyu Hou, Shuhan Ye, Hanwei Zhu, Yi Yu, and Xudong Jiang. ChronoLock: Protecting videos from unauthorized text-to-video personalization. *arXiv preprint arXiv:2606.21146*, 2026.

Yuxin Huang, Ziming Hong, Mingming Gong, Wanyu Wang, Jing Zhang, and Tongliang Liu. Delving into the temporal challenges of unified video protection against image-to-video and fine-tuning-based customization. *arXiv preprint arXiv:2607.13336*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4), 2023.

Jiahao Lu, Yifan Zhang, Qiuhong Shen, Xinchao Wang, and Shuicheng Yan. Poison-splat: Computation cost attack on 3D gaussian splatting. In *ICLR*, 2025.

Lingzhuang Meng, Mingwen Shao, Xiang Lv, Mengyao Wu, Yuanjian Qiao, and Jie Zhang. AntiAvatar: Protect against unauthorized 3D head avatar generation via dual-space divergence. *AAAI*, 40(10):7990–7998, 2026.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, pp. 1765–1773, 2017.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *ICML*, volume 162, pp. 16805–16827, 2022.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. ReFusion: 3d reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In *IROS*, 2019.

Josephine Passananti, Stanley Wu, Shawn Shan, Haitao Zheng, and Ben Y Zhao. Disrupting style mimicry attacks on video imagery. *arXiv preprint arXiv:2405.06865*, 2024.

Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J. Black. Attacking optical flow. In *ICCV*, pp. 2404–2413, 2019.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *ICML*, 2023.

Qi Song, Ziyuan Luo, Ka Chun Cheung, Simon See, and Renjie Wan. Geometry cloak: Preventing TGS-based 3D reconstruction from copyrighted images. In *NeurIPS*, volume 37, 2024.

Qi Song, Ziyuan Luo, Haoliang Han, and Renjie Wan. Generating multi-view adversarial examples for visual geometry grounded transformer. *arXiv preprint arXiv:2608.20748*, 2026.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IROS*, pp. 573–580, 2012.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pp. 402–419, 2020.

Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant CNNs. In *3DV*, 2017.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, 13(4):376–380, 1991.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tran. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *ICCV*, pp. 2116–2127, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. In *CVPR*, pp. 21486–21499, 2026.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *CVPR*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, pp. 20697–20709, 2024.

Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su. Sparse adversarial perturbations for videos. *AAAI*, 33(01):8973–8980, 2019.

Alex Wong, Safa Cicek, and Stefano Soatto. Targeted adversarial perturbations for monocular depth prediction. In *NeurIPS*, volume 33, pp. 8486–8497, 2020.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian splatting for real-time dynamic scene rendering. In *CVPR*, pp. 20310–20320, 2024.

David Yifan Yao, Albert J. Zhai, and Shenlong Wang. Uni4D: Unifying visual foundation models for 4D modeling from a single video. In *CVPR*, pp. 1116–1126, 2025.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *ICLR*, 2025.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4D visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025.

Algorithm 1 Causal protection of frame t.

1: **Input:** current clean frame, fixed clean/protected histories, number of refinement steps S.
2: Initialize $\bar{y}_t \leftarrow y_t^{(0)}$ with motion transport and the reliability fallback (Eq. 4).
3: Advance the clean histories $X, \mathcal{T}(X)$ to obtain the two depth references and masks.
4: **for** $s = 1, \ldots, S$ **do**
5: Score $\bar{y}_t$ under the fixed histories of $Y, \mathcal{T}(Y)$; compute g_I, g_T.
6: Form bounded one-code candidates and select $\alpha^\star = \arg\max_\alpha \hat{J}(\alpha)$ (Eq. 7).
7: Re-score $y_t^{\alpha^\star}$; set $\bar{y}_t \leftarrow y_t^{\alpha^\star}$ only if its actual minimum discrepancy exceeds that of $\bar{y}_t$.
8: **end for**
9: Publish $\bar{y}_t$ as an RGB8 frame and commit the $Y, \mathcal{T}(Y)$ histories once; earlier frames are not revised.

A Implementation Details

Publication grid and budget. The three controlled Sintel sequences are denoted S1, S2 and S3, and the Sintel sequence of the qualitative figures Q; their identities are given in the released records and marked in Table 8. S1 served as the development sequence; S2 and S3 were fixed as confirmation sequences before the controlled runs (all three had been used in earlier development); all three are part of the complete Sintel split. Frames of the four benchmarks are resized with their full field of view to 518×224 (bilinear; Sintel from 1024×436) and published as lossless PNG; N3DV frames are protected at the 1352×1014 working resolution of the asset pipeline and fed to the surrogate at 518×392. The realized perturbation of every published frame satisfies $\|y_t - x_t\|_\infty \le 16/255$, equivalent to 16 RGB8 codes; this is verified on the decoded PNG bytes. Each refinement step changes at most $K = \lfloor 3HW/32 \rfloor$ scalar RGB coordinates (10,878 at 518×224), by one code ($1/255$) each. The step size and coordinate budget are distinct from the total perturbation bound. The full frame and the causal model history remain in use. The accumulated perturbation $y_t - x_t$ need not be sparse because transport carries earlier changes forward.

Benchmark splits and scoring. Split lists follow the public preparation of Yao et al. (2025): Sintel uses every sequence and frame of the `training/final` pass; Bonn uses frames 30–139 of each of the 26 sequences, with depth associated one-to-one by timestamp (difference below 20 ms) instead of the reference's independent slicing of the RGB and depth lists; KITTI uses the first $\min(110, n)$ annotated frames of each of the 13 validation drives, paired with the same-name RGB image; TUM uses every third of the first 270 frames of the eight sitting/walking sequences with the nearest valid ground-truth pose. Predicted depth is inverted to disparity, resized to the ground-truth grid with bicubic interpolation, aligned per sequence (Appendix B) and clipped at 70 m; ground truth is valid for $0 < d < 70$ m. Each sequence is generated and reconstructed separately for each arm; the clean predictions of the two arms give identical scores. Ground truth is read only by the scorer.

Initialization and transport. Frame 1 and all unreliable pixels use a fixed checkerboard of 32-pixel cells with amplitude ± 2 codes, identical on R, G and B. Flow is computed with the frozen torchvision RAFT-Large model (weights `C_T_SKHT_V2`) between the two clean frames in both directions; a pixel is reliable if its correspondence is in bounds and the forward–backward cycle error satisfies $\|\cdot\|^2 \le 0.01\|F\|^2 + 0.5$ (pixels^2). Gathering uses round-half-to-even nearest sampling without interpolating the integer perturbation. Same-grid carry uses the same mask and copies δ_{t-1} at identical coordinates; reset uses the checkerboard everywhere. In the three-arm experiments the evaluation purifier's active coverage exceeds 96% of the pixels on every scene, and the motion arm lowers that coverage relative to same-grid by at most 0.10.

Surrogate and objective. StreamVGGT is run in FP32 with memory-efficient attention, no mixed precision, and its complete causal key–value history. The inputs $X, \mathcal{T}(X), Y$ and $\mathcal{T}(Y)$ have separate histories of the same frozen weights. Each depth discrepancy in Eq. 5 uses its own clean-reference mask: finite depths greater than 10^{-3} from X for the direct branch and from $\mathcal{T}(X)$ for the purified branch, with at least 8 valid pixels. Masks remain fixed when scoring protected predictions. A prediction floor of 10^{-6} is used before the logarithm; candidates with non-finite predictions or non-positive predictions on the reference mask are rejected, and non-finite model graphs are not differentiated. The soft-min temperature is $\tau = 10^{-4}$. Gradients reach the current frame only; committed frames are rebuilt from PNG bytes and detached, and the purifier's rounding uses a straight-through derivative (1 at $t = 1$, $1/2$ afterwards, zero into history).

Update and acceptance. Nine mixtures $\alpha \in \{0, \frac{1}{8}, \ldots, 1\}$ and the exact soft-min direction are enumerated. Eligible scalar coordinates have a non-zero gradient and a one-code move within $[0, 255]$ and the ϵ-box. At most the K largest eligible $|g|$ values are moved, with stable ordering by flattened CHW index. Ties in the predicted endpoint score prefer the exact soft-min direction, then the mixture closest to its direct-branch weight, then the smaller α. The winner is accepted only if its realized $\min(e_I, e_T)$ strictly exceeds that of the current frame, which at the first step is the transported initialization. A rejected step leaves the current frame unchanged. All 48 candidates per arm were accepted in the single-step 16-frame experiments; the earlier single-branch prototype in Table 26 accepted 47–50 of 50. The selected PNG is read back before its model-history commit and re-scored to reproduce the selected score bit for bit. Each refinement step uses two base scores, two gradient calls and two candidate scores; with two clean-reference calls and two commit calls per frame, a frame costs $4 + 6S$ surrogate calls, i.e., ten for $S = 1$ and 64 for $S = 10$. Direction selection itself adds no model calls. Bidirectional flow is computed by the frozen auxiliary model, with no target-model query.

Targets and evaluation. VGGT and CUT3R (official 512-DPT checkpoint) use their own loaders. CUT3R resizes the 518×224 frames to 512×221 and center-crops to 512×208; ground truth is resampled with nearest neighbour to the evaluated grid. Every CUT3R branch starts with a fresh model and recurrent state and uses one native forward over the sequence in FP32. The earlier 50-frame VGGT study uses a fresh model per branch, the native crop loader and an FP16 outer autocast context. Depth AbsRel and $\delta{<}1.25$ use per-frame median scale alignment on the ground-truth validity mask; ATE is the camera-center error after Umeyama Sim(3) alignment (Umeyama, 1991). Evaluation code recomputes every metric from saved predictions on CPU, independently of the generation code (3,415 checks for the CUT3R run).

Compute. The controlled analyses used RTX 5090-class GPUs with large host memory; the benchmark runs used H200 and RTX 5090 GPUs, with each sequence and arm generated on a single GPU. The complete three-scene, three-arm, 16-frame study used 1,440 generation calls and 672 evaluation calls, taking 19.6 minutes for generation and evaluation on one GPU. Four 50-frame prototype arms (Table 26): 300 surrogate calls each (six per generated frame), four GPUs in parallel, 46 minutes including 12 VGGT evaluations; peak allocated memory 16.3–16.5 GiB per arm. CUT3R evaluation: 24 sequences, 164 s. The 300-frame N3DV sequence: 2,952 surrogate calls, 9.35 h, peak GPU memory 13.5 GiB allocated (22.7 GiB reserved) with 229 GiB of causal history offloaded to CPU. 4D Gaussian training: about 17 minutes per asset.

B Geometric Evaluation Details

Geometric errors. Depth absolute relative error (AbsRel) measures how far a predicted depth is from its true value relative to that true depth. Monocular predictions can have an arbitrary overall scale. We therefore multiply each predicted depth map $\widehat{D}_t$ by one scale factor before scoring:

$$s_t = \frac{\mathrm{median}_{p \in V_t} D_t^*(p)}{\mathrm{median}_{p \in V_t} \widehat{D}_t(p)}, \qquad \mathrm{AbsRel}_t = \frac{1}{|V_t|} \sum_{p \in V_t} \frac{|s_t \widehat{D}_t(p) - D_t^*(p)|}{D_t^*(p)}, \tag{8}$$

where D_t^* is ground-truth depth and V_t contains its valid pixels. This removes a uniform scale difference within a frame, not spatial errors or pixel misalignment; the controlled analyses use this per-frame alignment. The depth accuracy $\delta{<}1.25$ reports the fraction of valid pixels whose aligned prediction and true depth differ by a factor smaller than 1.25.

Benchmark alignment. The benchmark results of Table 1 follow the public video-depth protocol (Yao et al., 2025), which aligns a whole sequence at once and therefore also penalizes scale drift across frames. With predicted and true disparities $\widehat{\rho}_t = 1/\widehat{D}_t$ (resized to the ground-truth grid) and $\rho_t^* = 1/D_t^*$, one scale λ and one shift μ are fitted per sequence by least squares over all valid pixels of all frames,

$$(\lambda, \mu) = \arg\min_{\lambda, \mu} \sum_t \sum_{p \in V_t} \big(\lambda\, \widehat{\rho}_t(p) + \mu - \rho_t^*(p)\big)^2, \tag{9}$$

the aligned depth is $1/(\lambda\widehat{\rho}_t + \mu)$, clipped at 70 m, and AbsRel and $\delta{<}1.25$ are averaged over all valid pixels of the split ($0 < D^* < 70$ m), i.e., sequences are weighted by their numbers of valid

pixels. The reference scale-only variant sets $\mu = 0$ and λ to the ratio of the median true and predicted disparities of the sequence; Appendix E reports it together with sequence-equal means.

Camera absolute trajectory error (ATE) measures the root-mean-square distance between predicted and true camera centers. Since their coordinate systems can differ, we first fit a single similarity transform to the whole trajectory: one scale s, rotation R, and translation b (Umeyama, 1991):

$$\mathrm{ATE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N} \|sR\widehat{c}_t + b - c_t^*\|_2^2}. \tag{10}$$

This is camera-position error, not moving-object trajectory error. On TUM we additionally report the relative pose error (RPE), the translational and rotational error of relative camera motions after the same alignment. Ground truth is used for these evaluation procedures, not for optimizing protection.

C Design Motivation in Detail

Why the perturbation must follow scene motion (Section 4.1). 4D reconstruction is built on cross-frame correspondence: optimization-based pipelines track points densely over time and lift the tracks to 3D motion (Yao et al., 2025), a streaming reconstruction model accumulates evidence about each scene point from every frame in which it is visible, and a temporal purifier aggregates observations along the same correspondences. The effect of a perturbation therefore depends on its agreement across the frames that see the same scene point. Under causal release the protector cannot enforce this agreement by joint optimization, because earlier frames are already fixed; what remains to decide is where the perturbation of the new frame starts, and the two straightforward choices fall short. Restarting from scratch discards what earlier frames have built up, and copying the previous perturbation at the same pixel coordinates attaches it to whatever content now occupies those pixels, so on moving regions perturbation and scene drift apart. An idealized motion-aligned average of two frames with warp W shows why aggregation then attenuates it: its residual perturbation is $\frac{1}{2}(\delta_t + W\delta_{t-1})$, which is approximately δ_t if $\delta_t \approx W\delta_{t-1}$ but partially cancels if the two disagree at corresponding points. The perturbation should thus be carried along the motion of the scene, which the protector can only extend by one step as a frame arrives, from the two most recent clean frames.

Why both inputs are scored (Section 4.2). Persistence alone does not make a perturbation protective: it must change the geometry that the attacker recovers. We therefore score a candidate by how far the surrogate's depth prediction on the protected stream departs from its prediction on the clean stream, in a form that is invariant to the scale ambiguity of monocular depth and needs no ground truth. The second question is which input to score. Consecutive frames observe nearly the same scene, so averaging or median-filtering aligned frames acts as a low-pass filter along time that attenuates components inconsistent across frames while the content survives. The protector cannot enumerate the attacker's purifiers, but it can build the simplest member of this family into the objective, the causal average of the current and the previous frame: it is causal and can thus be evaluated within the protector's own information, and its effect on a perturbation is the residual analyzed above, so a perturbation that survives it retains the cross-frame consistency that temporal purifiers test. Finally, the attacker may equally well not purify, and optimizing only the purified input strengthens the protection there but weakens it on the direct input, and vice versa (Appendix L). We therefore evaluate both inputs and maximize the weaker discrepancy, so that the attacker gains little by switching from one to the other.

D Purifiers and Clean Reconstruction Quality

Causal two-frame average (avg2). $\mathcal{T}(y)_1 = y_1$ and $\mathcal{T}(y)_t = \mathrm{round}_{\mathrm{even}}\big((y_{t-1} + y_t)/2\big)$ for $t \geq 2$, computed in a wider integer type on the original (unpurified) parents; it is nonrecursive and reproducible. It is used inside the objective and, in Table 26 and Table 25, for evaluation.

Flow-aligned causal median of three (flow-median3). For $t \geq 3$, the two preceding original frames are warped into frame t with RAFT flow estimated *from the branch's own frames* (clean flow for clean branches, protected flow for protected branches; the protector's flows are not reused), bilinear RGB sampling, the same forward–backward cycle rule as the protector, and a channel-wise

median rounded once; pixels without a valid correspondence copy the current frame, and the first two frames are unchanged. The first-two-frame identity, the causal window and the branch-local flow make this purifier one that an online attacker can run.

Non-causal purification. An offline attacker (Section 3) can use future frames, so we also apply two purifiers that use the next frame to the same frozen single-step videos, after publication and without any feedback to the protector. *Forward avg2* averages the original frames t and $t+1$ with integer rounding to even (last frame unchanged) and uses no motion compensation. *Centered flow-median3* warps the original frames $t-1$ and $t+1$ directly into frame t with RAFT flow estimated from the branch's own frames and takes the channel-wise median with the current frame, rounded once; pixels without a valid correspondence copy the current frame, and the first and last frames are unchanged. Table 5 lists the paired increments on StreamVGGT, each protected video being compared with the identically purified clean video. Under the centered flow-median3, Safe4D keeps 92.0% of its unpurified increment (0.101 vs. 0.109), more than same-grid (90.7%) and reset (88.9%), and it leads both on every sequence. Forward avg2 is weaker still: the increments of both carrying arms grow (reset loses 2.47%), and same-grid carry, whose perturbation is fixed in image coordinates and thus survives an average taken at fixed pixels, slightly exceeds Safe4D on the mean (0.122 vs. 0.119) because of S2. Both purifiers change clean reconstruction little (Table 7).

Table 5: Non-causal purification of the single-step videos of Table 12 (StreamVGGT): paired $\Delta\mathrm{AbsRel}$ ($\times 10^{-3}$; purified protected minus identically purified clean) per sequence and three-sequence mean.

Purification	Init.	S1	S2	S3	mean
Direct	Reset	118	92.0	33.9	81.2
	Same-grid	133	119	63.3	105
	Safe4D (Ours)	**134**	**127**	**67.2**	**109**
Forward avg2	Reset	126	69.9	41.9	79.2
	Same-grid	140	**130**	94.9	**122**
	Safe4D (Ours)	**141**	118	**99.8**	119
Centered flow-median3	Reset	119	67.6	30.3	72.2
	Same-grid	135	103	48.1	95.4
	Safe4D (Ours)	**137**	**104**	**61.4**	**101**

Ten-step videos. The three development sequences were also re-generated for all three initializations with the benchmark setting $S = 10$ and scored, on StreamVGGT and on the unseen CUT3R, without purification and under the two non-causal purifiers (Table 6; per-sequence source increments in Table 27). Every protected video is compared with its own matching clean video processed identically on the same hardware; the two hardware paths used for the centered flow-median3 references differ by one code value on 297 of the $16{,}708{,}608$ channel values of the three clean clips, and no clean video is shared across paths. Two observations complement the single-step study. First, the ordering of the two carries on the surrogate depends on whether the attacker purifies: without purification same-grid carry has the higher mean because of S3, whereas under forward avg2 and the centered flow-median3 motion transport retains the largest increment (1.14 vs. 1.02 and 0.915 vs. 0.796), leading on every sequence under the motion-aligned purifier. Second, on CUT3R the increment of motion transport is 2.21–$3.45\times$ that of same-grid carry under all three treatments, whereas its mean $\Delta\mathrm{ATE}$ is negative under all three, so the transfer is a depth effect, not a pose effect. Purification also costs the attacker: forward avg2 lowers the mean clean $\delta{<}1.25$ of CUT3R from 79.4% to 76.0% (7.35 points on S2), which is why protection is measured by the paired increment.

Complete benchmarks. Table 2 applies forward avg2, the causal flow-median3 and the centered flow-median3 to the ten-step videos of all 70 benchmark sequences on StreamVGGT, each protected video being compared with the identically purified clean video. On the clean videos, the three purifiers change AbsRel by -5.52% to $+13.9\%$ and the TUM ATE by -0.331% to $+4.72\%$.

Clean-utility checks and reporting. The admissibility of a purifier (Section 5.1) is metered by the change of clean reconstruction it causes. The tolerances used are an error ratio of at most 1.25 (a 25% increase of clean AbsRel or ATE) and a $\delta{<}1.25$ drop of at most 0.05, and the clean cost of every purifier is reported. The current StreamVGGT study checks clean AbsRel ratio ≤ 1.25, $\delta{<}1.25$ drop ≤ 0.05 and mean flow-median3 coverage ≥ 0.5; it does not evaluate pose and has no

Table 6: Ten-step study (ten-step videos of three Sintel sequences; mean paired increments against the identically processed clean video; CUT3R ΔAbsRel and ΔATE in 10^{-3}, δ<1.25 drop in points; ΔATE is not a protection target and not ranked).

Purification	Initialization	StreamVGGT	CUT3R (unseen)		
		ΔAbsRel ↑	ΔAbsRel ↑	δ<1.25 drop ↑	ΔATE
Direct	Reset	0.772	4.44	1.52	+0.124
	Same-grid	**1.38**	5.00	2.39	+0.150
	Safe4D (Ours)	1.11 $_{\downarrow 19.9\%}$	**11.0** $_{\uparrow 121\%}$	**4.01** $_{\uparrow 67.6\%}$	−1.37
Forward avg2	Reset	0.432	2.56	−0.611	+0.692
	Same-grid	1.02	3.40	0.347	+0.526
	Safe4D (Ours)	**1.14** $_{\uparrow 10.9\%}$	**11.7** $_{\uparrow 245\%}$	**2.44** $_{\uparrow 604\%}$	−1.61
Centered flow-median3	Reset	0.476	3.46	1.19	−0.00785
	Same-grid	0.796	3.16	1.63	−0.0838
	Safe4D (Ours)	**0.915** $_{\uparrow 15.0\%}$	**8.42** $_{\uparrow 143\%}$	**3.00** $_{\uparrow 84.0\%}$	−1.04

ATE check. CUT3R checks clean AbsRel and ATE ratios ≤ 1.25 and δ<1.25 drop ≤ 0.05. The non-causal purifiers are checked for the same AbsRel ratio and δ<1.25 drop on StreamVGGT. The listed values in Table 7 satisfy the checks, as do both scenes of the earlier VGGT objective study.

Table 7: Clean-utility measurements of the purifiers (clean input only). Dashes: pose not measured on the source model; all applicable checks pass.

Purifier	Model	Seq.	AbsRel ratio	ATE ratio	δ<1.25 drop
avg2	VGGT	S1 (50 fr.)	0.945	0.936	−0.0076
		S3 (50 fr.)	1.13	1.17	0.0288
flow-median3	CUT3R	S1	1.03	0.975	0.00638
		S2	0.980	1.09	−0.00522
		S3	0.986	1.02	0.000143
flow-median3	StreamVGGT	S1	1.02	—	−0.0005
		S2	0.844	—	−0.0185
		S3	1.07	—	0.0119
forward avg2	StreamVGGT	S1	0.991	—	−0.0086
		S2	1.02	—	−0.0110
		S3	1.11	—	0.0098
centered flow-median3	StreamVGGT	S1	0.986	—	−0.0014
		S2	0.866	—	−0.0179
		S3	1.05	—	0.0070

E Complete Benchmark Results

Tables 8–11 list all 70 sequences behind Table 1. Under the primary protocol the effect of Safe4D varies widely across sequences, from a δ<1.25 drop of 80.3 points (97.5% → 17.2%) to 3.67 points, and the two Sintel sequences whose δ<1.25 improves under protection are those with the least accurate clean reconstruction. On Bonn every one of the 26 sequences is degraded by both arms, with δ<1.25 drops of Safe4D between 8.83 and 88.4 points; on KITTI 12 of the 13 drives are degraded and one drive improves under both arms; on TUM ATE increases on 5 (motion) and 6 (same-grid) of the 8 trajectories, the translational RPE on 6 and 7, and the rotational RPE on all 8 for both arms.

The ranking of the two arms depends on the benchmark and on the protocol. Under the primary alignment Safe4D leads same-grid carry in both metrics on Sintel and KITTI, with either weighting; on Bonn the two arms are within 0.358% in AbsRel and 0.0281 points in δ<1.25, same-grid marginally ahead in AbsRel; on TUM same-grid has the higher mean ATE, driven by one trajectory (0.327 vs. 0.274 m), while Safe4D has the higher translational RPE. Under the reference scale-only alignment, where both arms raise the Sintel AbsRel by more than an order of magnitude, Safe4D leads in AbsRel on Sintel (4.17 vs. 4.04) but not in δ<1.25 (10.9% vs. 10.2%), and same-grid carry leads in both metrics on Bonn and KITTI. The advantage of transport therefore appears under purification (Section 5.3) rather than in the white-box mean.

Table 8: All 23 Sintel sequences behind Table 1 (StreamVGGT, no purification, S=10). S1–S3: the controlled sequences; Q: the sequence of the qualitative figures. Upper block: primary protocol (one disparity scale and shift per sequence). Bottom rows: means under the primary protocol and under the reference scale-only alignment, weighted by valid pixels ("pixel") or equally by sequence ("seq.").

		AbsRel			δ<1.25 (%)		
Sequence	Frames	Clean	Same-grid	Safe4D (Ours)	Clean	Same-grid	Safe4D (Ours)
alley_1	50	0.173	0.854	0.844	76.8	3.65	3.07
alley_2 (S3)	50	0.115	0.467	0.478	87.7	26.6	26.4
ambush_2	21	0.517	0.794	0.809	35.7	11.7	8.61
ambush_4	33	0.338	0.734	0.780	45.6	13.1	11.9
ambush_5	50	0.439	0.489	0.487	43.2	32.7	32.5
ambush_6	20	0.210	0.618	0.589	65.0	9.72	11.9
ambush_7	50	0.0464	0.0914	0.0900	99.9	96.1	96.2
bamboo_1	50	0.195	0.506	0.507	74.1	12.3	12.3
bamboo_2	50	0.236	0.494	0.492	64.7	28.4	28.3
bandage_1	50	0.107	0.188	0.179	82.6	69.8	72.9
bandage_2	50	0.277	0.473	0.480	44.3	11.1	9.54
cave_2	50	0.553	0.660	0.633	21.2	29.6	27.2
cave_4	50	0.335	0.433	0.438	43.8	37.5	37.0
market_2	50	0.131	0.503	0.504	86.0	29.0	28.3
market_5 (Q)	50	0.255	0.786	0.778	59.7	10.4	10.8
market_6	40	0.352	0.731	0.736	46.2	12.2	11.2
mountain_1	50	0.0979	0.206	0.205	95.5	60.2	60.8
shaman_2	50	0.0678	0.525	0.514	97.5	24.8	17.2
shaman_3 (S1)	50	0.117	0.209	0.226	87.0	58.5	54.5
sleeping_1	50	0.0958	0.154	0.153	94.7	75.6	74.8
sleeping_2 (S2)	50	0.0461	0.159	0.200	98.9	73.9	63.4
temple_2	50	0.477	0.665	0.674	31.6	13.9	12.8
temple_3	50	0.527	0.511	0.516	15.8	24.9	23.3
Scale–shift, pixel	1,064	0.229	0.476	0.478	67.7	34.8	33.3
Scale–shift, seq.		0.248	0.489	0.492	65.1	33.3	32.0
Scale only, pixel		0.247	4.04	4.17	65.4	10.2	10.9
Scale only, seq.		0.266	3.72	3.81	63.5	13.0	13.6

Table 9: All 26 Bonn sequences behind Table 1 (StreamVGGT, no purification, S=10; frames 30–139 of every sequence; the prefix rgbd_bonn_ is omitted). Upper block: primary protocol (one disparity scale and shift per sequence). Bottom rows: means under the primary protocol and under the reference scale-only alignment, weighted by valid pixels ("pixel") or equally by sequence ("seq.").

		AbsRel			δ<1.25 (%)		
Sequence	Frames	Clean	Same-grid	Safe4D (Ours)	Clean	Same-grid	Safe4D (Ours)
balloon	110	0.0479	0.180	0.182	97.8	78.6	77.6
balloon2	110	0.0566	0.254	0.253	97.6	51.8	53.0
balloon_tracking	110	0.0554	0.170	0.167	98.8	69.8	71.8
balloon_tracking2	110	0.0817	0.197	0.192	97.9	69.2	69.6
crowd	110	0.0595	0.116	0.113	99.6	91.3	90.7
crowd2	110	0.0837	0.559	0.572	95.4	11.1	9.44
crowd3	110	0.0646	0.256	0.258	97.2	44.8	45.5
kidnapping_box	110	0.0692	0.208	0.208	99.1	63.9	62.2
kidnapping_box2	110	0.0486	0.209	0.201	98.8	56.2	57.2
moving_nonobstructing_box	110	0.0578	0.155	0.158	98.8	74.5	74.6
moving_nonobstructing_box2	110	0.0671	0.197	0.189	98.6	61.8	66.4
moving_obstructing_box	110	0.0665	0.245	0.241	97.6	46.4	46.3
moving_obstructing_box2	110	0.0492	0.134	0.135	98.7	84.5	84.6
person_tracking	110	0.0567	0.179	0.175	98.9	67.6	69.7
person_tracking2	110	0.0542	0.186	0.185	98.8	69.7	70.2
placing_nonobstructing_box	110	0.0597	0.164	0.163	98.5	72.0	72.3
placing_nonobstructing_box2	110	0.0536	0.170	0.168	98.7	77.7	77.4
placing_nonobstructing_box3	110	0.0473	0.164	0.161	99.1	78.8	78.6
placing_obstructing_box	110	0.0442	0.211	0.223	98.1	61.4	51.2
removing_nonobstructing_box	110	0.0609	0.175	0.174	98.9	63.5	64.3
removing_nonobstructing_box2	110	0.0588	0.205	0.204	99.0	62.8	62.9
removing_obstructing_box	110	0.0438	0.216	0.220	98.0	58.6	60.3
synchronous	110	0.0799	0.414	0.415	98.3	10.1	9.88
synchronous2	110	0.0643	0.259	0.261	97.1	51.3	50.2
static	110	0.0676	0.111	0.114	99.5	89.1	88.8
static_close_far	110	0.0574	0.444	0.420	96.4	26.8	27.8
Scale–shift, pixel	2,860	0.0598	0.219	0.219	98.3	61.7	61.7
Scale–shift, seq.		0.0599	0.222	0.221	98.3	61.3	61.3
Scale only, pixel		0.0667	1.45	1.38	97.7	11.0	13.6
Scale only, seq.		0.0668	1.45	1.39	97.7	11.0	13.6

Table 10: All 13 KITTI validation drives behind Table 1 (StreamVGGT, no purification, S=10; first $\min(110, n)$ annotated frames of each drive; drives are named by date and number). Rows and mean rows as in Table 9.

		AbsRel			δ<1.25 (%)		
Sequence	Frames	Clean	Same-grid	Safe4D (Ours)	Clean	Same-grid	Safe4D (Ours)
09_26_0002	67	0.177	0.152	0.154	67.4	75.6	74.0
09_26_0005	110	0.173	0.424	0.414	75.4	27.8	28.4
09_26_0013	110	0.165	0.416	0.423	71.0	29.2	28.3
09_26_0020	76	0.0754	0.338	0.339	96.5	39.1	38.4
09_26_0023	110	0.243	0.410	0.411	52.9	29.3	29.4
09_26_0036	110	0.169	0.385	0.386	72.6	30.4	30.4
09_26_0079	90	0.143	0.371	0.377	81.0	35.9	35.0
09_26_0095	110	0.249	0.384	0.383	49.4	33.1	33.5
09_26_0113	77	0.101	0.372	0.372	95.8	36.2	36.0
09_28_0037	79	0.0812	0.367	0.366	90.7	34.0	34.2
09_29_0026	110	0.0399	0.0844	0.0579	98.5	92.5	96.0
09_30_0016	110	0.260	0.410	0.410	48.8	28.6	28.3
10_03_0047	110	0.129	0.143	0.191	84.7	81.3	65.5
Scale–shift, pixel	1,269	0.167	0.324	0.327	72.1	44.9	43.3
Scale–shift, seq.		0.154	0.327	0.329	75.7	44.1	42.9
Scale only, pixel		0.176	0.410	0.408	68.7	49.5	53.7
Scale only, seq.		0.163	0.427	0.409	72.8	47.9	52.4

Table 11: All 8 TUM dynamics sequences behind Table 1 (StreamVGGT, no purification, S=10; 90 frames per sequence; the prefix rgbd_dataset_freiburg3_ is omitted). ATE and RPE_t in m, RPE_r in degrees; one Sim(3) per trajectory, RPE on adjacent frames after that alignment. Bottom row: sequence-equal means of Table 1.

	ATE (m)			RPE_t (m)			RPE_r (deg)		
Sequence	Clean	Same-grid	Ours	Clean	Same-grid	Ours	Clean	Same-grid	Ours
sitting_halfsphere	0.0409	0.0590	0.0747	0.0188	0.0269	0.0272	0.507	0.616	0.640
sitting_rpy	0.0540	0.0546	0.0529	0.00861	0.0158	0.0146	0.606	0.810	0.884
sitting_static	0.0250	0.0138	0.0165	0.00468	0.00661	0.00673	0.375	0.637	0.587
sitting_xyz	0.0313	0.0682	0.0630	0.0131	0.0236	0.0293	0.370	0.622	0.702
walking_halfsphere	0.168	0.327	0.274	0.0860	0.0744	0.0781	0.653	0.873	0.792
walking_rpy	0.249	0.239	0.235	0.0315	0.0603	0.0593	0.911	0.981	1.04
walking_static	0.0123	0.0147	0.0155	0.00582	0.00675	0.00549	0.276	0.486	0.418
walking_xyz	0.0398	0.0881	0.0656	0.0236	0.0309	0.0279	0.526	0.888	0.821
Mean (seq.)	0.0775	0.108	0.0996	0.0240	0.0307	0.0311	0.528	0.739	0.735

F Complete Controlled and Transfer Results

Single-step controlled study. Before the ten-step videos of Table 6 were generated, the three initializations were compared on frozen single-step videos ($S = 1$) of the same three sequences, on the surrogate (Table 12) and on the unseen CUT3R (Table 3); Table 27 lists the per-sequence increments. On the surrogate, Safe4D raises the AbsRel of StreamVGGT from 0.0632 to 0.172, an increment 34.7% above that of reset and 3.97% above that of same-grid at nearly the same perturbation energy, with a positive increment on every sequence; the causal flow-median3 leaves clean reconstruction intact (0.0632 → 0.0626), Safe4D keeps 81.9% of its increment, and its lead over same-grid widens to 5.90% (Table 12). On CUT3R it leads both baselines on S1 and S2 (Figure 7); S3 does not transfer for any variant, and camera pose is not degraded (rotational RPE increments below 0.01° for every variant). Appendix G shows that the transfer of these three development sequences is not representative of the complete split. Tables 13 and 14 list the absolute CUT3R metrics of all 24 evaluations and the absolute source-side depth accuracy. The S2 ground truth contains 59 pixels (0.003%) with depth ≈ 10^{11} that dominate its RMSE; we report AbsRel and δ<1.25, which are unaffected.

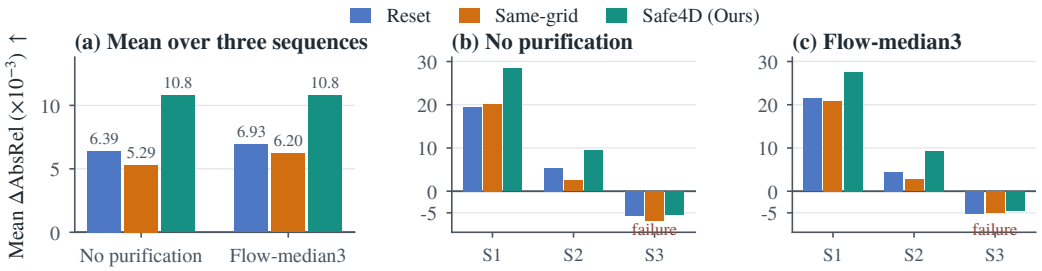


Figure 7: CUT3R depth-error increments of Table 3: (a) three-sequence mean, (b, c) per sequence. S3 does not transfer for any variant.

Table 12: Single-step study on StreamVGGT (three Sintel sequences): AbsRel and paired increment, directly and after the causal flow-median3. All variants have a perturbation RMS of 2.01–2.06 codes.

Input	Direct		Flow-median3	
	AbsRel ↑	ΔAbsRel ↑	AbsRel ↑	ΔAbsRel ↑
Clean	0.0632	–	0.0626	–
Reset	0.144	0.0812	0.130	0.0677
Same-grid	0.168	0.105	0.147	0.0846
Safe4D (Ours)	**0.172**$_{\uparrow 2.48\%}$	**0.109**$_{\uparrow 3.97\%}$	**0.152**$_{\uparrow 3.39\%}$	**0.0896**$_{\uparrow 5.90\%}$

Table 13: Absolute CUT3R metrics of all 24 single-step evaluations behind Table 3 (purified: causal flow-median3).

Seq.	Input	AbsRel	ATE (m)	δ<1.25
S1	Clean	0.116	0.00372	0.813
	Clean, purified	0.119	0.00363	0.807
	Reset	0.135	0.00708	0.771
	Reset, purified	0.141	0.00653	0.735
	Same-grid	0.136	0.00649	0.769
	Same-grid, purified	0.140	0.00538	0.742
	Safe4D (Ours)	0.144	0.00421	0.735
	Safe4D (Ours), purified	0.147	0.00473	0.714
S2	Clean	0.229	0.00890	0.654
	Clean, purified	0.225	0.00967	0.660
	Reset	0.235	0.00885	0.646
	Reset, purified	0.229	0.00928	0.647
	Same-grid	0.232	0.00896	0.652
	Same-grid, purified	0.227	0.00930	0.652
	Safe4D (Ours)	0.239	0.00872	0.639
	Safe4D (Ours), purified	0.234	0.00935	0.635
S3	Clean	0.0745	0.0128	0.915
	Clean, purified	0.0735	0.0130	0.915
	Reset	0.0688	0.0117	0.908
	Reset, purified	0.0684	0.0121	0.906
	Same-grid	0.0677	0.0135	0.908
	Same-grid, purified	0.0686	0.0122	0.905
	Safe4D (Ours)	0.0690	0.00918	0.904
	Safe4D (Ours), purified	0.0690	0.00854	0.902

Table 14: Absolute StreamVGGT AbsRel behind Table 12, directly and after the causal flow-median3 (purified).

Seq.	Clean		Reset		Same-grid		Safe4D (Ours)	
	direct	purified	direct	purified	direct	purified	direct	purified
S1	0.0636	0.0650	0.181	0.180	0.197	0.190	0.198	0.195
S2	0.0510	0.0430	0.143	0.102	0.170	0.131	0.178	0.136
S3	0.0748	0.0799	0.109	0.109	0.138	0.120	0.142	0.126

G Full-Split Transfer to CUT3R

Protocol. The 70 ten-step protected videos of Table 1 (same-grid and Safe4D, 5,913 frames per variant, all frames retained) and their matching clean videos are run through CUT3R with its official 512-resolution checkpoint, native RGB loader and one continuous inference per sequence, in FP32 and with a fresh state per condition; a clean video is evaluated once per distinct ordered input. Two receiver-side treatments are evaluated: the released video as is (direct) and forward avg2, applied identically to the clean video. Depth is mapped to the raw ground-truth grid on CUT3R's crop-aware support (bilinear disparity), and one disparity-affine alignment per sequence is fitted on the support that ground truth and the native crop define; a scale-only alignment is reported as a supplement (Table 17). TUM uses the frozen 90-frame mapping, one Sim(3) per trajectory and the ATE/RPE definitions of Appendix B. Every dataset mean weights sequences equally. This protocol differs from the full-frame cubic, valid-pixel-weighted protocol of Table 1, so absolute values are not comparable across the two models. All 420 evaluations and 280 paired scores are included. Tables 16–21 give the per-sequence direction of change and all per-sequence values.

Result. Averaged over the split, neither protected variant degrades CUT3R on the direct input: the mean AbsRel changes by -0.00224 to -0.00134 on Sintel, $+0.00233$ to $+0.00329$ on Bonn and -0.00630 to -0.00560 on KITTI, δ<1.25 rises, and the TUM ATE falls (Table 15); per sequence, AbsRel rises on roughly half of the sequences of every dataset (Table 16). Forward avg2 raises the mean error of both variants slightly on Sintel, Bonn and TUM; on KITTI it rises for same-grid and falls for Safe4D, and there is no consistent ordering between the two variants. With the scale-only alignment the Sintel means of the protected videos lie far below the clean mean (0.405–0.413 vs. 0.669 on the direct input) and the sign of the forward-avg2 change on Sintel reverses, because a few clean sequences fit poorly with a single scale; the affine alignment is therefore the primary protocol, and both are reported (Table 17). The single-sequence metrics of Q (direct: AbsRel $0.242 \rightarrow 0.382$

Table 15: Transfer to the unseen CUT3R on all 70 benchmark sequences (the videos of Table 1; CUT3R is not queried). AbsRel and δ<1.25 (%), one disparity-affine alignment per sequence; ATE (m), one Sim(3); sequence-equal means. **Bold**: stronger protected variant; green / gray: relative change of *Ours* against same-grid. The protocol differs from Table 1 (Appendix G); absolute values are not comparable across models.

Processing	Input	Sintel		Bonn		KITTI		TUM
		AbsRel ↑	δ<1.25 ↓	AbsRel ↑	δ<1.25 ↓	AbsRel ↑	δ<1.25 ↓	ATE ↑
Direct	Clean	0.334	48.9	0.0539	98.2	0.114	85.6	0.106
	Same-grid	**0.333**	**49.4**	0.0562	**98.3**	**0.109**	**87.6**	**0.0953**
	Safe4D (Ours)	0.332 ↓0.269%	50.8 ↓2.68%	**0.0572** ↑1.72%	98.3 ↓0.0612%	0.108 ↓0.636%	87.6 ↓0.00266%	0.0891 ↓6.58%
Forward avg2	Clean	0.346	47.1	0.0597	97.7	0.118	86.5	0.0808
	Same-grid	**0.359**	**47.5**	**0.0640**	**97.6**	**0.121**	**85.7**	0.0851
	Safe4D (Ours)	0.356 ↓0.647%	48.9 ↓2.94%	0.0621 ↓3.06%	97.9 ↓0.268%	0.116 ↓4.50%	87.0 ↓1.51%	**0.0951** ↑11.7%

for Safe4D and 0.282 for same-grid; δ<1.25 59.9 → 37.6 and 43.5) and Figures 4, 5 and 8 illustrate a sequence on which transfer occurs.

Table 16: Direction of the CUT3R change per sequence (Table 15): number of sequences on which the protected video raises AbsRel / lowers δ<1.25 (depth, affine alignment) or raises ATE (TUM), out of the sequences of each dataset.

Processing	Input	**Sintel** (23)	**Bonn** (26)	**KITTI** (13)	**TUM** (8)
Direct	Same-grid	13 / 12	16 / 11	8 / 7	4
	Safe4D (Ours)	12 / 11	14 / 8	8 / 9	6
Forward avg2	Same-grid	15 / 11	17 / 10	9 / 9	6
	Safe4D (Ours)	15 / 12	17 / 9	5 / 6	5

Table 17: CUT3R transfer with scale-only depth alignment (same evaluations as Table 15; one disparity scale per sequence, no shift). Sintel is dominated by a few sequences whose clean prediction fits poorly with a single scale.

Processing	Input	Sintel		Bonn		KITTI		TUM
		AbsRel ↑	δ<1.25 ↓	AbsRel ↑	δ<1.25 ↓	AbsRel ↑	δ<1.25 ↓	ATE ↑
Direct	Clean	0.669	48.1	0.0693	94.9	0.120	84.6	0.106
	Same-grid	0.405	**52.0**	**0.0721**	**94.9**	0.113	86.8	**0.0953**
	Safe4D (Ours)	**0.413** ↑2.00%	53.7 ↓3.21%	0.0716 ↓0.661%	94.9 ↓0.0158%	**0.115** ↑1.19%	**86.3** ↑0.594%	0.0891 ↓6.58%
Forward avg2	Clean	0.516	50.9	0.0749	94.1	0.128	84.3	0.0808
	Same-grid	**0.457**	**54.4**	**0.0788**	**94.0**	**0.131**	84.0	0.0851
	Safe4D (Ours)	0.440 ↓3.60%	55.5 ↓2.11%	0.0767 ↓2.67%	94.3 ↓0.300%	0.128 ↓2.04%	**83.7** ↑0.293%	**0.0951** ↑11.7%

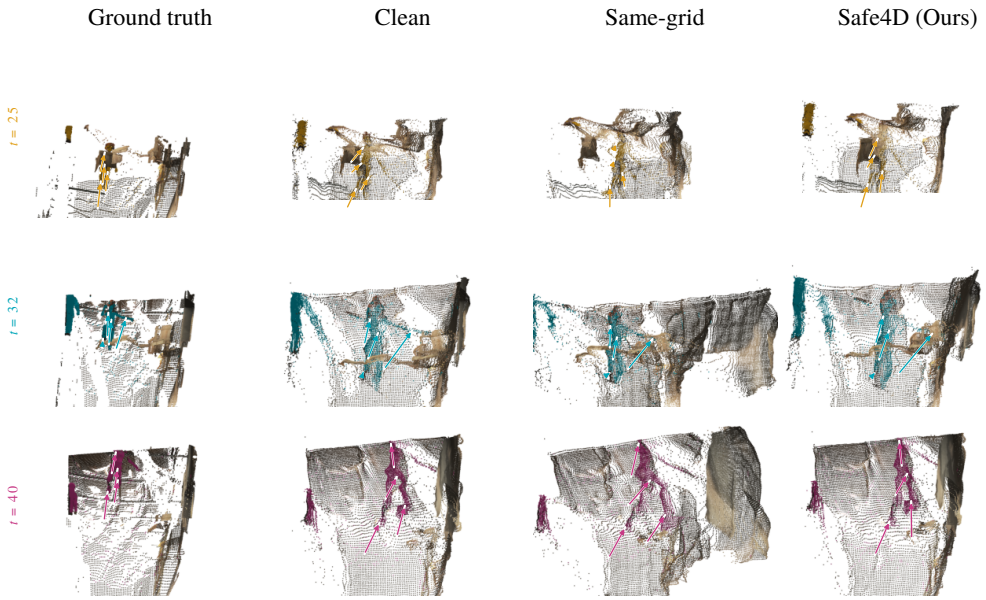


Figure 8: Per-frame dynamic geometry on the qualitative sequence Q (ten-step videos). Rows: frames 25, 32 and 40; columns: ground truth, clean, same-grid and Safe4D. All panels share the GT-defined frame-40 crop and camera; time colors mark GT-visible motion, and arrows join the same GT-matched native 3D points from $t-4$ to t without magnifying displacement. The alignment is shared with Figure 4.

Table 18: CUT3R per-sequence results on Sintel (affine alignment): AbsRel / δ<1.25 (%) of the clean video and of the two protected variants, direct and after forward avg2.

	Direct			Forward avg2		
Sequence	Clean	Same-grid	Ours	Clean	Same-grid	Ours
alley_1	0.351 / 37.8	0.403 / 38.6	0.437 / 33.4	0.369 / 34.3	0.508 / 27.3	0.522 / 31.1
alley_2 (S3)	0.121 / 86.6	0.124 / 85.3	0.122 / 84.8	0.133 / 84.8	0.142 / 82.6	0.147 / 82.3
ambush_2	0.642 / 21.6	0.603 / 31.3	0.488 / 26.0	0.641 / 20.2	0.725 / 16.0	0.647 / 18.3
ambush_4	0.762 / 12.6	0.738 / 11.4	0.733 / 12.4	0.759 / 12.6	0.767 / 9.83	0.763 / 10.3
ambush_5	0.427 / 37.4	0.444 / 39.3	0.438 / 41.5	0.454 / 36.0	0.437 / 38.5	0.444 / 38.1
ambush_6	0.381 / 32.1	0.518 / 17.5	0.545 / 14.1	0.415 / 34.3	0.506 / 25.5	0.442 / 30.6
ambush_7	0.0569 / 99.4	0.0743 / 98.7	0.0746 / 98.7	0.0598 / 99.0	0.0705 / 98.6	0.0705 / 98.4
bamboo_1	0.336 / 42.7	0.329 / 42.2	0.321 / 45.1	0.341 / 41.4	0.345 / 41.5	0.338 / 42.6
bamboo_2	0.304 / 50.0	0.301 / 51.6	0.299 / 52.6	0.302 / 51.2	0.304 / 51.6	0.302 / 52.7
bandage_1	0.118 / 82.6	0.145 / 77.2	0.156 / 76.6	0.161 / 71.3	0.175 / 70.1	0.170 / 71.5
bandage_2	0.474 / 6.78	0.222 / 53.2	0.202 / 75.0	0.320 / 24.8	0.277 / 39.0	0.341 / 21.5
cave_2	0.509 / 19.3	0.487 / 21.3	0.490 / 20.1	0.546 / 26.5	0.517 / 26.5	0.516 / 24.7
cave_4	0.343 / 36.0	0.380 / 37.7	0.365 / 40.8	0.370 / 40.6	0.386 / 35.4	0.392 / 38.4
market_2	0.266 / 47.6	0.250 / 52.0	0.254 / 51.2	0.274 / 49.2	0.267 / 51.4	0.259 / 53.3
market_5 (Q)	0.242 / 59.9	0.282 / 43.5	0.382 / 37.6	0.415 / 26.3	0.509 / 27.6	0.596 / 26.5
market_6	0.341 / 40.7	0.356 / 39.6	0.375 / 37.0	0.383 / 42.5	0.386 / 43.4	0.392 / 45.8
mountain_1	0.188 / 67.8	0.183 / 67.6	0.130 / 83.7	0.194 / 63.9	0.153 / 77.1	0.145 / 87.7
shaman_2	0.306 / 53.1	0.261 / 59.0	0.243 / 63.7	0.418 / 26.7	0.292 / 53.5	0.240 / 64.9
shaman_3 (S1)	0.0840 / 96.3	0.170 / 69.7	0.146 / 79.0	0.0953 / 94.5	0.204 / 58.0	0.175 / 70.8
sleeping_1	0.0746 / 97.1	0.0805 / 96.1	0.0815 / 95.8	0.0716 / 98.6	0.0829 / 97.1	0.0852 / 97.1
sleeping_2 (S2)	0.171 / 69.8	0.175 / 69.7	0.168 / 71.0	0.182 / 64.7	0.182 / 66.8	0.176 / 68.2
temple_2	0.660 / 7.82	0.739 / 10.2	0.744 / 10.1	0.620 / 26.9	0.652 / 23.2	0.634 / 26.0
temple_3	0.535 / 20.1	0.396 / 24.2	0.445 / 16.9	0.439 / 12.6	0.362 / 32.9	0.399 / 24.8

Table 19: CUT3R per-sequence results on Bonn (affine alignment): AbsRel / δ<1.25 (%) of the clean video and of the two protected variants, direct and after forward avg2.

Sequence	Direct			Forward avg2		
	Clean	Same-grid	Ours	Clean	Same-grid	Ours
balloon	0.0468 / 98.2	0.0404 / 98.5	0.0498 / 98.5	0.0510 / 97.5	0.0456 / 98.0	0.0550 / 97.9
balloon2	0.0563 / 97.0	0.0644 / 97.3	0.0650 / 97.6	0.0815 / 95.5	0.0834 / 96.1	0.0868 / 95.9
balloon_tracking	0.0554 / 98.6	0.0467 / 99.1	0.0500 / 99.1	0.0565 / 98.4	0.0497 / 98.5	0.0515 / 98.5
balloon_tracking2	0.0597 / 96.8	0.0552 / 97.5	0.0537 / 97.2	0.0649 / 96.2	0.0660 / 96.4	0.0662 / 96.0
crowd	0.0318 / 99.6	0.0528 / 99.7	0.0401 / 99.6	0.0329 / 99.5	0.0466 / 99.6	0.0330 / 99.5
crowd2	0.0755 / 95.6	0.0840 / 95.1	0.0778 / 95.9	0.0895 / 93.3	0.112 / 91.9	0.0948 / 93.9
crowd3	0.0691 / 98.4	0.0686 / 98.3	0.0707 / 98.4	0.0806 / 97.4	0.0807 / 97.3	0.0835 / 97.3
kidnapping_box	0.0565 / 99.4	0.0621 / 99.0	0.0697 / 99.0	0.0652 / 97.7	0.0545 / 98.6	0.0651 / 99.1
kidnapping_box2	0.0442 / 99.1	0.0543 / 98.9	0.0752 / 98.2	0.0511 / 98.9	0.0596 / 98.7	0.0721 / 98.0
moving_nonobstructing_box	0.0529 / 99.0	0.0483 / 99.1	0.0476 / 98.9	0.0553 / 98.8	0.0486 / 98.8	0.0461 / 98.9
moving_nonobstructing_box2	0.0432 / 99.0	0.0463 / 98.9	0.0540 / 99.0	0.0498 / 98.4	0.0473 / 98.5	0.0606 / 98.2
moving_obstructing_box	0.0688 / 98.1	0.0566 / 98.3	0.0625 / 98.2	0.0783 / 97.1	0.0618 / 97.3	0.0793 / 96.9
moving_obstructing_box2	0.0313 / 98.9	0.0468 / 99.0	0.0415 / 99.0	0.0340 / 98.6	0.0551 / 98.6	0.0479 / 98.7
person_tracking	0.0448 / 99.0	0.0438 / 99.3	0.0423 / 99.3	0.0476 / 98.8	0.0766 / 95.0	0.0471 / 99.1
person_tracking2	0.0600 / 97.0	0.0583 / 97.7	0.0571 / 97.5	0.0600 / 96.8	0.0687 / 97.4	0.0543 / 97.6
placing_nonobstructing_box	0.0436 / 99.3	0.0446 / 99.4	0.0435 / 99.5	0.0474 / 98.8	0.0521 / 99.1	0.0474 / 99.0
placing_nonobstructing_box2	0.0391 / 99.0	0.0393 / 99.3	0.0383 / 99.1	0.0450 / 98.4	0.0427 / 98.8	0.0462 / 98.6
placing_nonobstructing_box3	0.0422 / 99.2	0.0681 / 99.1	0.0629 / 99.2	0.0538 / 98.7	0.0543 / 99.2	0.0666 / 98.8
placing_obstructing_box	0.0499 / 98.7	0.0533 / 98.3	0.0600 / 98.2	0.0528 / 98.4	0.0737 / 97.8	0.0508 / 98.5
removing_nonobstructing_box	0.0660 / 99.1	0.0482 / 99.3	0.0561 / 99.4	0.0640 / 98.9	0.0364 / 99.2	0.0496 / 99.1
removing_nonobstructing_box2	0.0393 / 99.4	0.0467 / 99.4	0.0420 / 99.4	0.0411 / 99.1	0.0584 / 99.1	0.0427 / 99.1
removing_obstructing_box	0.0443 / 99.1	0.0529 / 98.8	0.0502 / 98.9	0.0425 / 98.8	0.0621 / 98.2	0.0498 / 98.5
static	0.0804 / 96.6	0.0826 / 96.9	0.0791 / 97.1	0.0742 / 97.8	0.0867 / 96.6	0.0664 / 99.3
static_close_far	0.0778 / 92.4	0.0823 / 91.6	0.0756 / 93.3	0.0773 / 93.1	0.0847 / 93.6	0.0833 / 92.9
synchronous	0.0593 / 98.7	0.0607 / 98.8	0.0622 / 98.8	0.0833 / 97.5	0.0864 / 97.4	0.0907 / 96.9
synchronous2	0.0627 / 98.5	0.0541 / 98.4	0.0595 / 98.4	0.0734 / 97.8	0.0703 / 97.7	0.0767 / 97.7

Table 20: CUT3R per-sequence results on KITTI (affine alignment): AbsRel / δ<1.25 (%) of the clean video and of the two protected variants, direct and after forward avg2.

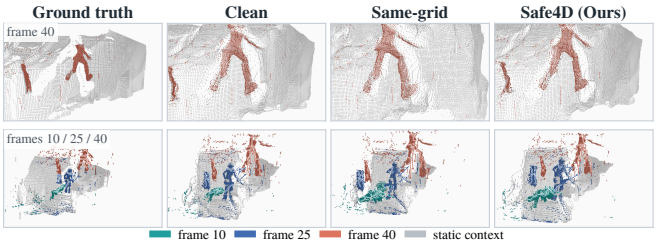
Sequence	Direct			Forward avg2		
	Clean	Same-grid	Ours	Clean	Same-grid	Ours
09_26/0002	0.0683 / 97.9	0.0688 / 97.9	0.0672 / 97.9	0.0833 / 95.0	0.0840 / 94.9	0.0813 / 95.2
09_26/0005	0.142 / 80.4	0.134 / 83.3	0.148 / 75.6	0.170 / 74.6	0.170 / 72.6	0.164 / 74.1
09_26/0013	0.104 / 88.3	0.120 / 84.4	0.119 / 83.3	0.121 / 84.5	0.133 / 82.3	0.131 / 82.2
09_26/0020	0.0813 / 94.2	0.0706 / 94.7	0.102 / 92.1	0.110 / 91.2	0.0883 / 94.0	0.123 / 89.8
09_26/0023	0.100 / 91.2	0.0875 / 92.6	0.0941 / 91.9	0.0940 / 90.6	0.0895 / 91.4	0.0901 / 91.8
09_26/0036	0.102 / 92.0	0.116 / 88.8	0.0941 / 93.4	0.115 / 87.5	0.151 / 80.8	0.0912 / 93.2
09_26/0079	0.0944 / 92.6	0.108 / 90.5	0.110 / 89.8	0.111 / 89.4	0.129 / 86.8	0.126 / 86.6
09_26/0095	0.101 / 92.3	0.105 / 91.0	0.102 / 92.0	0.131 / 85.1	0.133 / 84.9	0.126 / 86.6
09_26/0113	0.235 / 47.2	0.187 / 65.6	0.149 / 72.8	0.152 / 81.8	0.126 / 86.8	0.129 / 87.8
09_28/0037	0.107 / 87.2	0.109 / 86.9	0.113 / 85.7	0.124 / 83.4	0.125 / 83.3	0.129 / 82.7
09_29/0026	0.136 / 79.4	0.137 / 79.7	0.140 / 79.3	0.143 / 77.0	0.154 / 77.4	0.143 / 77.4
09_30/0016	0.0800 / 94.0	0.101 / 89.9	0.0910 / 90.5	0.101 / 89.3	0.107 / 89.1	0.0948 / 89.4
10_03/0047	0.137 / 75.5	0.0687 / 93.7	0.0757 / 94.5	0.0721 / 94.7	0.0833 / 89.7	0.0742 / 94.1

Table 21: CUT3R per-sequence pose on TUM dynamics (one Sim(3) per trajectory; RMS over the 90-frame sequence): ATE and RPE_t in m, RPE_r in degrees.

Processing	**Sequence**	ATE			RPE_t			RPE_r		
		Clean	Same-grid	Ours	Clean	Same-grid	Ours	Clean	Same-grid	Ours
Direct	sitting_halfsphere	0.202	0.198	0.211	0.0231	0.0241	0.0234	0.987	1.04	1.26
	sitting_rpy	0.0392	0.0428	0.0424	0.0109	0.0127	0.00995	0.724	1.02	0.740
	sitting_static	0.0221	0.0231	0.0229	0.00510	0.00498	0.00504	0.340	0.337	0.332
	sitting_xyz	0.0255	0.0257	0.0259	0.0104	0.0102	0.0106	0.349	0.357	0.367
	walking_halfsphere	0.291	0.190	0.144	0.0656	0.0486	0.0402	0.749	0.740	0.786
	walking_rpy	0.199	0.217	0.200	0.0429	0.0411	0.0393	1.33	1.32	1.36
	walking_static	0.0124	0.0114	0.0122	0.00471	0.00443	0.00401	0.295	0.302	0.295
	walking_xyz	0.0544	0.0541	0.0548	0.0193	0.0191	0.0182	0.600	0.618	0.576
Forward avg2	sitting_halfsphere	0.152	0.202	0.168	0.0257	0.0266	0.0221	1.05	1.24	1.04
	sitting_rpy	0.0352	0.0418	0.0440	0.0116	0.0109	0.0103	0.995	1.15	1.00
	sitting_static	0.0244	0.0245	0.0243	0.00490	0.00455	0.00475	0.447	0.441	0.459
	sitting_xyz	0.0239	0.0252	0.0249	0.0103	0.0111	0.0110	0.533	0.550	0.538
	walking_halfsphere	0.123	0.122	0.225	0.0409	0.0460	0.0609	0.986	1.00	1.03
	walking_rpy	0.224	0.187	0.211	0.0378	0.0430	0.0397	1.56	1.83	1.78
	walking_static	0.0113	0.0126	0.0118	0.00433	0.00442	0.00374	0.398	0.400	0.395
	walking_xyz	0.0524	0.0660	0.0517	0.0227	0.0220	0.0233	0.877	0.923	0.895

H NATIVE 3D GEOMETRY ON THE UNSEEN MODEL

Protocol. The frozen single-step videos of the controlled analysis (clean, same-grid and Safe4D, each directly and after flow-median3; three sequences, 16 frames) are re-evaluated on CUT3R with its official loader and checkpoint in FP32: 18 evaluation-only runs, with no feedback to the protector. We export CUT3R's native self-view point maps (208×512, sampled at a fixed stride of 2 without confidence filtering) and place them in world coordinates with the camera poses predicted in the same pass, as in the official demo. For every sequence and input, one Sim(3) transform fitted on corresponding ground-truth-valid samples of all 16 frames aligns the prediction to the ground-truth point maps, which are back-projected from ground-truth depth and cameras; no per-frame fitting is performed. The support is the same for every input ($0 <$ depth < 70 m). The Chamfer distance (CD) is the symmetric, unsquared nearest-neighbour distance $\frac{1}{2}(\bar{d}_{\text{pred}\to\text{GT}} + \bar{d}_{\text{GT}\to\text{pred}})$ and the F-score the harmonic mean of precision and recall at 1, 2.5 and 5 cm; the point error is the mean distance between the aligned prediction and the ground truth at the same pixel. The 3D displacement error ($\text{EPE}_{3\text{D}}$) links frame t to $t+1$ with the official ground-truth optical flow, samples the next point map bilinearly at the flow endpoint and compares the predicted displacement with the ground-truth displacement computed on the same lattice with the same operator; invalid, occluded and out-of-bounds samples are excluded, and scoring the ground truth as a prediction gives zero error on all 45 frame pairs. The focal error is the relative error of CUT3R's official focal estimate (fixed principal point, $f_x = f_y$), and ATE uses the trajectory alignment of Eq. 10. Frames (or frame pairs) are weighted equally within a sequence, and sequences are reported individually. This displacement error is conditioned on ground-truth correspondences.



Seq.	Direct		Forward avg2	
	ΔCD ↑	ΔEPE	ΔCD ↑	ΔEPE
S1	+13.2	+4.34	+11.2	−1.31
S2	+6.50	−2.38	+5.77	−5.71
S3	+9.15	+4.37	+10.1	+3.18

Figure 9: Native CUT3R geometry of the ten-step videos (one Sim(3) per sequence). Left, the qualitative sequence Q, on which the protection transfers: frame 40 (top, frontal GT camera) and frames 10/25/40 with GT-guided tracks (bottom, 12° orbit); same cameras, GT-defined crop and clean-RGB shading in every column, colors mark GT-visible 3D motion at the three frames, gray is static context. Right: Chamfer distance (CD) and 3D displacement error (EPE) of Safe4D relative to clean on the three Sintel sequences (%); positive = worse. Full-range views and all metrics: Appendix H.

Replay. The 18 runs are new evaluations of the frozen videos; compared with the archived CUT3R outputs of Table 13, across all 18, the mean relative depth difference from the archived predictions is at most 0.088% and the AbsRel difference at most 2.1×10^{-4}, and the ATE differs by at most 0.16 mm.

Results on the single-step videos. Table 22 lists every metric and Figure 13 plots the signed changes caused by protection; Figures 12, 15 and 16 show the three sequences. Same-grid carry also increases CD in all six conditions, and Safe4D's CD exceeds same-grid's in two of three direct and all three purified conditions, by less than 2%; without purification Safe4D has a lower displacement error and a higher 5 cm F-score than same-grid on all three sequences. Camera trajectories (Figure 14) are not consistently degraded: in this replay the mean ATE decreases by 1.15 and 1.16 mm without and with purification (archived run of Table 3: 1.10 and 1.22 mm), and the mean focal error decreases by 0.69 and 0.75 points. Absolute quality differs widely across sequences (clean 5 cm F-score of 86.2%, 14.3% and 0.46%), so we report sequences individually rather than as a mean in millimeters.

Ten-step videos. The same protocol is applied to the ten-step benchmark videos of the three development sequences, without purification and under forward avg2 (Table 23; the three-sequence means of the relative changes are +9.61% and +9.01% in CD and +2.11% and −1.28% in $\text{EPE}_{3\text{D}}$ for Safe4D, against +10.6% and +6.87% and +9.69% and +2.86% for same-grid carry). Both

carries increase CD, the point error and the loss of 5 cm-accurate points in all twelve conditions. The displacement error is the one quantity on which the two carries differ in sign: same-grid carry increases it in all six conditions, whereas Safe4D changes it by -5.71% to $+4.37\%$, so the transported perturbation distorts shape while leaving the frame-to-frame motion of the recovered points closer to the clean reconstruction. Camera ATE decreases under Safe4D in five of six conditions.

Q. Figures 9 and 10 show the qualitative sequence Q (50 frames, $S = 10$, original benchmark PNGs), chosen because Uni4D uses it as a qualitative example (Yao et al., 2025). Clean, same-grid and Safe4D inputs were each passed through CUT3R in one continuous 50-frame inference; every prediction receives one sequence-wide Sim(3) on the same ground-truth support. Both cameras are fixed from the ground-truth frame-40 pose (a frontal view and a $12°/4°$ orbit), the foreground crop of the main figure is defined by ground-truth moving points inside pre-declared clean-RGB windows with a common margin, and the same clean-RGB luminance shades every column, so rendered colors do not depend on the protected input. Points are colored when their ground-truth world displacement over the adjacent frame pair exceeds 2 cm; the trajectories sample the same official ground-truth correspondences (frames 10–15, 25–30 and 40–45, ten seeds per window chosen by spatial spread) in each native point map and are therefore GT-guided probes, not learned tracks. Figure 11 shows the archived clean-input RAFT flow that motion carry actually used on this sequence. The protected reconstructions remain recognizable but geometrically distorted; the animal at frame 10 and the persons at frames 25 and 40 are different objects.

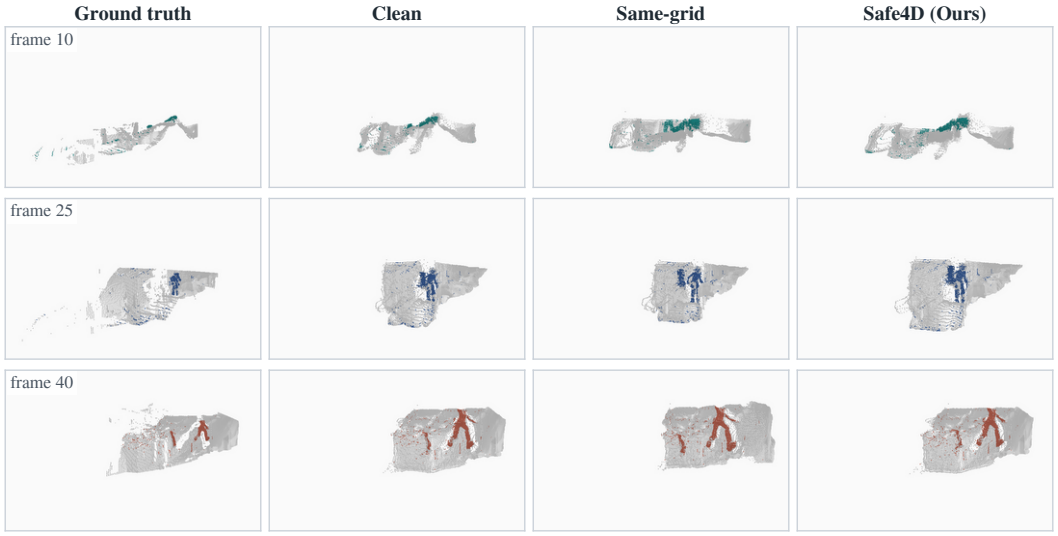


Figure 10: Full-range native CUT3R point clouds of the qualitative sequence Q at frames 10, 25 and 40 (frontal GT-fixed camera). The viewport retains every supported point of every branch over all 50 frames; teal/blue/coral mark GT-visible 3D motion at the three frames, gray is context. The top row of Figure 9 is the foreground crop of frame 40.

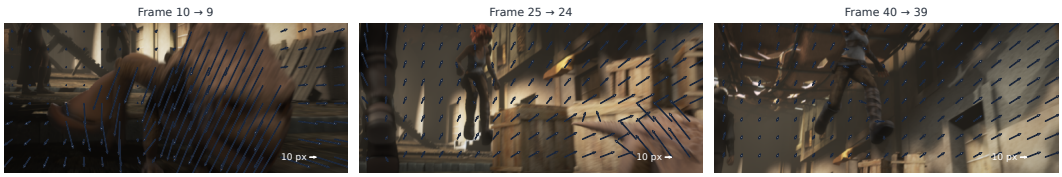


Figure 11: Optical flow used for transport on the qualitative sequence Q: archived clean-input RAFT flow of the original generation, frame $t \to t-1$ for $t = 10, 25, 40$, drawn on the clean frame with true pixel lengths (scale bar: 10 px). White arrows pass the forward–backward reliability test of Section 4.1; amber arrows fail it and start from the checkerboard.

Table 22: Native CUT3R geometry of the frozen single-step videos (three sequences, 16 frames; one Sim(3) per sequence and input). CD, point error, 3D displacement error (EPE_{3D}) and ATE in mm; F-scores (at 1, 2.5 and 5 cm) and focal error in %. Errors: lower = more accurate reconstruction; F-scores: higher = more accurate. Every input is scored against the same ground-truth support; median3: flow-median3.

Seq.	Purif.	Input	CD	F@1	F@2.5	F@5	Pt. err.	EPE_{3D}	Focal	ATE
S1	none	Clean	25.7	34.0	65.8	86.2	55.8	15.7	19.6	3.73
		Same-grid	30.3	28.6	57.1	80.0	66.2	18.1	18.6	6.32
		Safe4D (Ours)	30.1	28.5	57.4	80.4	66.6	17.3	17.5	4.21
	median3	Clean	26.3	32.8	63.9	85.7	56.4	18.9	20.0	3.62
		Same-grid	28.3	30.9	60.7	82.9	63.1	16.0	19.5	5.39
		Safe4D (Ours)	28.7	29.7	59.9	82.3	64.1	16.7	17.7	4.77
S2	none	Clean	183	1.35	5.97	14.3	382	33.6	18.4	8.89
		Same-grid	193	1.24	5.30	12.7	402	32.2	18.5	8.96
		Safe4D (Ours)	194	1.23	5.37	12.9	403	31.6	18.5	8.73
	median3	Clean	183	1.37	6.04	14.5	379	32.8	18.4	9.65
		Same-grid	190	1.32	5.53	13.2	395	32.2	18.5	9.30
		Safe4D (Ours)	191	1.31	5.63	13.3	395	31.0	18.5	9.35
S3	none	Clean	793	0.01	0.14	0.46	1502	172	8.60	12.9
		Same-grid	843	0.01	0.10	0.37	1584	178	8.50	13.4
		Safe4D (Ours)	856	0.01	0.10	0.38	1607	175	8.58	9.11
	median3	Clean	788	0.01	0.12	0.45	1493	171	8.76	12.9
		Same-grid	841	0.01	0.08	0.32	1576	176	8.77	12.1
		Safe4D (Ours)	850	0.01	0.08	0.32	1598	172	8.75	8.58

Table 23: Native CUT3R geometry of the ten-step benchmark videos (three development sequences, 16 frames; one Sim(3) per sequence and input on the same ground-truth support; no confidence or error filtering). Relative changes caused by protection in % of the clean value (protected−clean)/clean: Chamfer distance (CD), point error at corresponding pixels, F-score at 5 cm (negative = fewer accurate points), 3D displacement error along ground-truth correspondences (EPE_{3D}) and camera ATE. Absolute clean values in the last columns (CD and EPE_{3D} in mm).

Seq.	Purif.	Input	ΔCD	ΔPt. err.	ΔF@5	ΔEPE_{3D}	ΔATE	Clean CD	Clean EPE_{3D}
S1	direct	Same-grid	+21.7	+21.7	−8.42	+22.3	+68.8	25.7	15.7
		Ours	+13.2	+15.4	−4.99	+4.34	+10.1	25.7	15.7
	fwd. avg2	Same-grid	+8.56	+10.1	−4.71	+1.19	+37.1	26.9	15.8
		Ours	+11.2	+11.4	−5.03	−1.31	−23.2	26.9	15.8
S2	direct	Same-grid	+3.51	+3.53	−8.80	+0.815	−0.760	183	33.5
		Ours	+6.50	+6.28	−9.48	−2.38	−2.12	183	33.5
	fwd. avg2	Same-grid	+3.31	+2.93	−9.62	+3.17	−2.42	186	34.8
		Ours	+5.77	+5.03	−10.1	−5.71	−6.43	186	34.8
S3	direct	Same-grid	+6.74	+5.67	−12.9	+6.00	−16.0	792	172
		Ours	+9.15	+7.13	−21.7	+4.37	−33.6	792	172
	fwd. avg2	Same-grid	+8.75	+7.67	−13.5	+4.21	+2.83	834	188
		Ours	+10.1	+8.20	−21.2	+3.18	−23.8	834	188

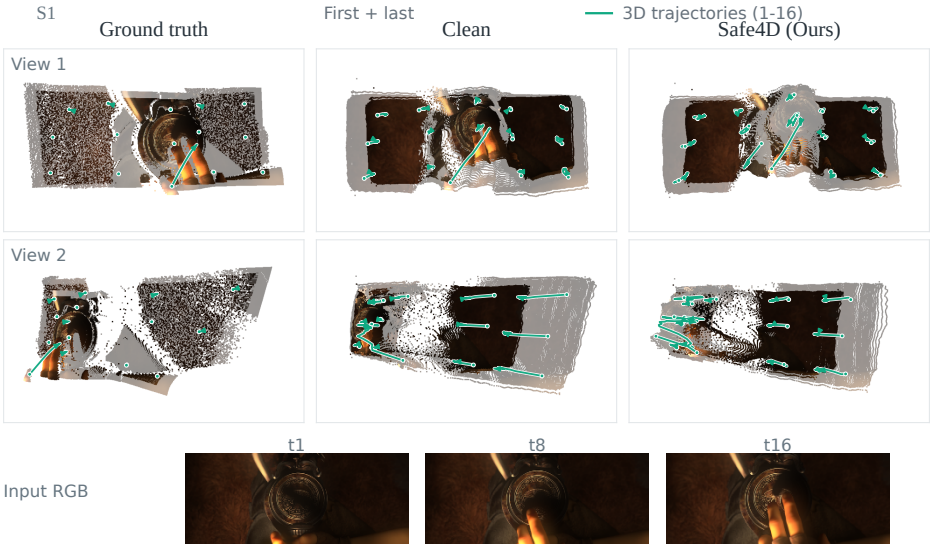


Figure 12: Native CUT3R point clouds and 3D trajectories on sequence S1 (single-step videos of Table 12; first and last point clouds, teal curves follow ground-truth correspondences over 16 frames; bottom: input frames).

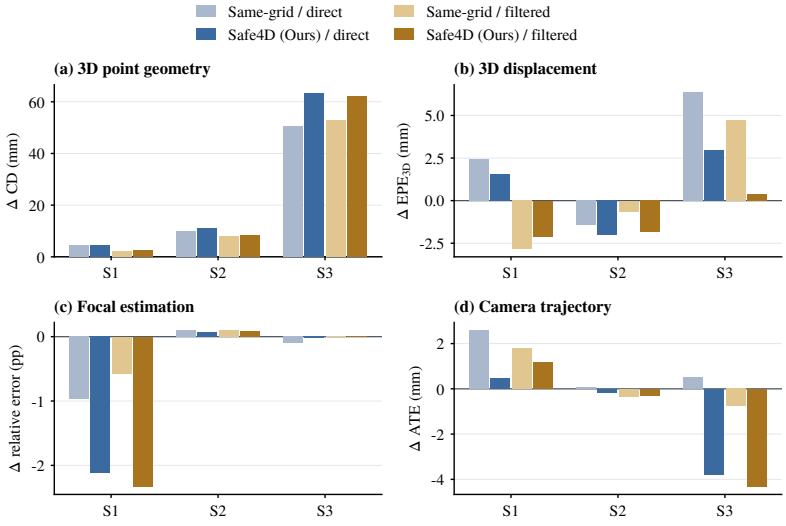


Figure 13: Signed error changes caused by protection on CUT3R (protected minus clean; positive = worse reconstruction), for same-grid and Safe4D, without purification (direct) and after flow-median3 (filtered). (a) CD of the native point clouds; (b) 3D displacement error along ground-truth correspondences; (c) focal error; (d) camera ATE.

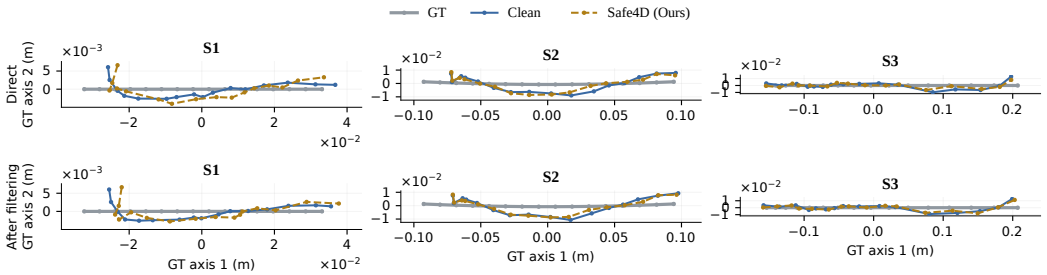


Figure 14: CUT3R camera trajectories of the clean and Safe4D inputs, without purification (top) and after flow-median3 (bottom), each aligned to the ground truth by one Sim(3) and projected onto the same ground-truth principal plane. Mean ATE does not increase under protection.

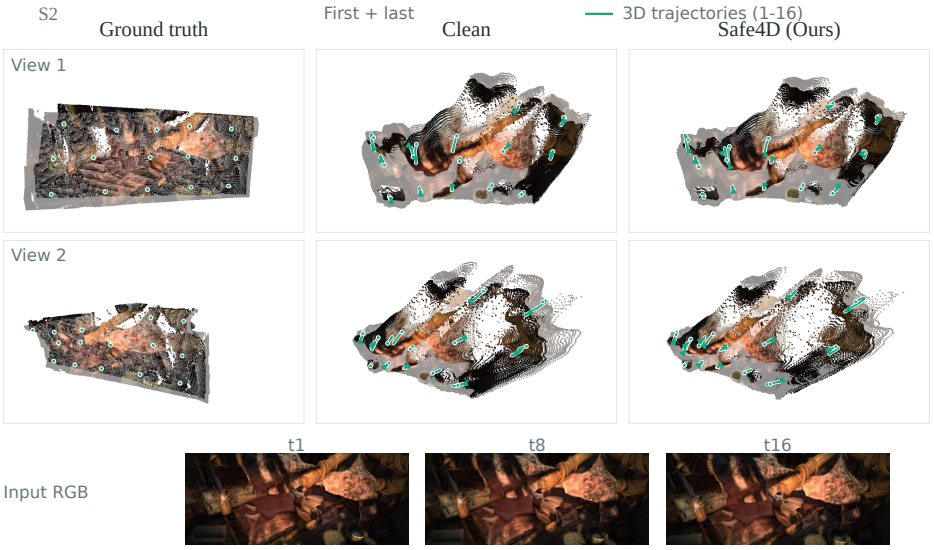


Figure 15: Native CUT3R point clouds and 3D trajectories on sequence S2 (protocol of Figure 12; bottom: input frames).

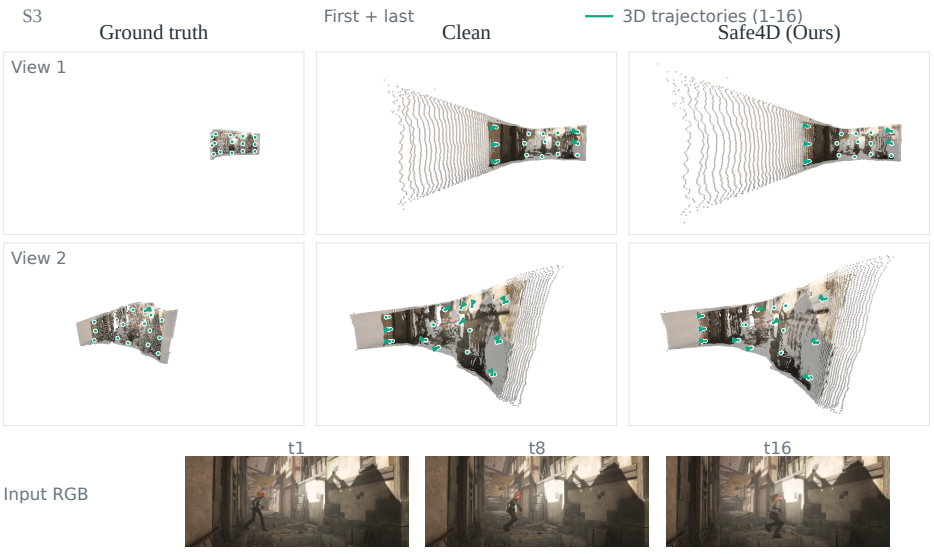


Figure 16: Native CUT3R point clouds and 3D trajectories on sequence S3 (protocol of Figure 12). The per-frame depth increment of this sequence is negative (Table 3), whereas its CD increases by 8.00% and 7.88%.

I MATCHED-SUPPORT DIAGNOSTIC

To examine sensitivity to the distribution and amount of perturbation, we built, after generation, a matched version of every arm: at each channel of each frame the three arms are reduced to the minimum absolute integer perturbation across arms while keeping each arm's own sign. The three matched videos have identical support, ℓ_0, ℓ_1, ℓ_2 and ℓ_∞; they were then re-run through fresh surrogate histories and fresh branch-local purification flows. Table 24 shows that the original advantage ($+6.41\%$ / $+8.32\%$ on the confirmation sequences) vanishes under this control (-2.90% / -0.0286%). Matching changes the generated perturbations, including their spatial support and magnitude distribution. This control therefore does not isolate a single mechanism behind the original gain.

Table 24: Confirmation-sequence $\Delta\mathrm{AbsRel}$ of the original and the support-and-magnitude-matched videos (single-step, StreamVGGT).

Videos	Purification	Same-grid	Safe4D (Ours)	Relative
Original	direct	0.0911	0.0970	+6.41%
	flow-median3	0.0643	0.0696	+8.32%
Matched	direct	0.0719	0.0698	−2.90%
	flow-median3	0.0496	0.0495	−0.0286%

J DOWNSTREAM 4D ASSET: PROTOCOL AND RESULTS

We test whether geometric damage reaches the rendered asset in a demanding setting: a real multi-view capture, the official 4D Gaussians pipeline, and 19 unprotected cameras owned by the attacker. Only the 300-frame stream of camera 01 is protected (same-grid variant, generated before motion transport was finalized), and assets are scored by PSNR, SSIM and LPIPS on the held-out camera 00 at all 300 times (Table 25). With calibrated cameras, the asset trained on purified protected inputs is 2.23 dB worse in PSNR than its clean counterpart at every held-out time, SSIM and LPIPS agree, and the drop reaches 5.74 dB on the moving animal (Figures 17 and 18). Without purification the protected asset is instead 1.94 dB better, and with VGGT-predicted geometry the purified drop shrinks to 0.0972 dB: the damage reaches a real dynamic asset, with a size that depends on the attacker's pipeline, most strongly on whether it purifies. This experiment covers a single scene with a single protected camera and an earlier same-grid video.

Protocol. The N3DV scene provides 21 synchronized cameras at 2704×2028; the official 4D Gaussians pipeline is used at its 1352×1014 working resolution with all 300 frames of the 20 training cameras and camera 00 held out at all 300 times. The protected input is the frozen 300-frame Safe4D output for camera 01 (dual-branch objective, same-grid carry, $\ell_\infty \leq 16$ RGB8 codes, per-channel RMS 2.39 codes, input PSNR 40.6 dB). The other 19 training cameras receive no protection perturbation; purified branches apply avg2 to all 20 training streams. Setting (a) uses the dataset's calibrated cameras and a bounded SIFT point-cloud initialization computed from each asset's own t=0 images (P and T(P) share the point cloud built from the protected t=0, C and T(C) the clean one). Setting (b) replaces calibrated training geometry with a joint 20-view VGGT prediction of intrinsics, extrinsics and depth at t=0 (40,000 back-projected points), computed from the asset's own inputs. After training, a single Sim(3) fitted using the predicted and calibrated centers of the 20 training cameras maps the held-out calibrated camera into the asset's frame for evaluation; no test image is used for fitting. Every completed asset uses the identical official schedule (coarse 3,000, fine 14,000 iterations).

Region and temporal diagnostics. Two fixed rectangular foreground regions (a cooking region, 17.9% of the image, and the region of the moving animal, 8.8%) were chosen on clean contact sheets before scoring. In setting (a) the purified PSNR drop is 5.74 dB on the animal region and -0.00593 dB on the cooking region; without purification both regions improve (-3.19 and -0.372 dB). In setting (b) all four region drops are small and positive ($+0.0314$ to $+0.145$ dB). The temporal error of consecutive held-out renders changes by at most 0.660% between clean and protected assets. Figure 18 shows the held-out view at $t = 225$ for setting (a).

Table 25: 4D Gaussian reconstruction of the N3DV scene with one of 20 training cameras protected, scored on the clean held-out camera at all 300 times. C/P: clean/protected inputs; T: avg2. "Drop" is the quality loss caused by protection (positive = worse). (a) Calibrated cameras and SIFT initialization; (b) VGGT-predicted cameras and depth.

Asset	(a) Calibrated + SIFT				(b) Predicted geometry			
	PSNR ↑	SSIM ↑	LPIPS-V ↓	LPIPS-A ↓	PSNR ↑	SSIM ↑	LPIPS-V ↓	LPIPS-A ↓
C	30.5	0.935	0.160	0.0565	16.1	0.617	0.441	0.387
P	32.4	0.944	0.158	0.0528	16.1	0.619	0.440	0.385
T(C)	32.3	0.946	0.159	0.0543	16.1	0.618	0.442	0.386
T(P)	30.0	0.938	0.162	0.0593	16.0	0.616	0.443	0.389
Drop (no purif.)	−1.94	−0.00814	−0.00214	−0.00377	−0.00990	−0.00182	−0.00188	−0.00246
Drop (avg2)	**+2.23**	**+0.00793**	**+0.00341**	**+0.00507**	**+0.0972**	**+0.00202**	**+0.000596**	**+0.00276**

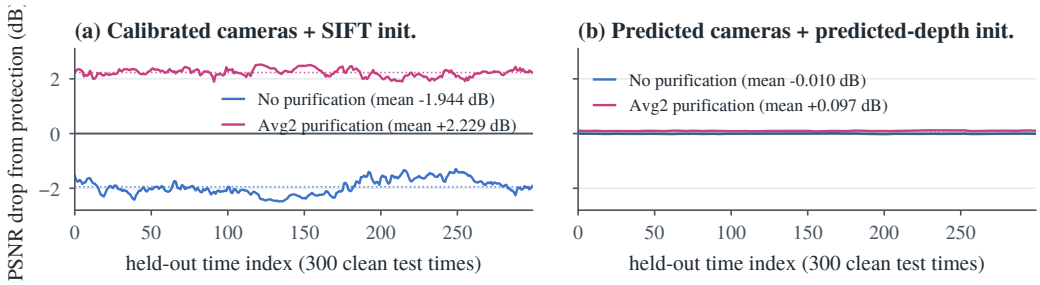


Figure 17: Held-out PSNR drop of the 4D Gaussian asset at all 300 times (positive means lower protected quality; dotted lines show means). The calibrated purified condition has a 2.23 dB mean drop, versus 0.0972 dB with predicted geometry. The unpurified conditions have negative mean drops. Each curve describes one scene over time.

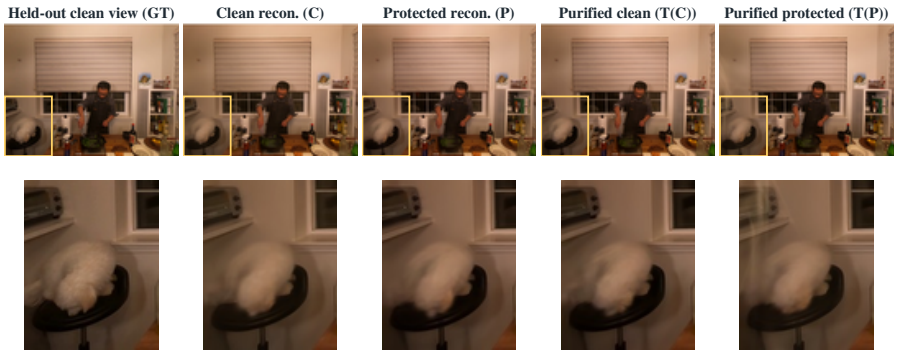


Figure 18: Held-out renderings at $t = 225$ (calibrated + SIFT; enlarged region: the moving animal). After purification, protection causes ghosting and loss of detail (5.74 dB regional PSNR drop).

K Additional Qualitative Results

Figures 19–21 show the per-frame CUT3R depth error of the ten-step videos of the three controlled sequences on the direct input, with the protocol of Figure 5: CUT3R's native depth is aligned to the ground truth by one scale per frame, the map shows the relative error and the inset its mean. On S1 and S2 the error grows on the moving foreground under both carrying variants, most under Safe4D; on S3, the failure case, the clean error concentrates near the bright far-field doorway and the protected videos slightly *reduce* the frame-mean error. The camera-space depth of each branch is recovered from its Sim(3)-aligned point map by a per-frame least-squares pose fit with the fitted intrinsics (median reprojection residual ≤ 1.3 pixels).

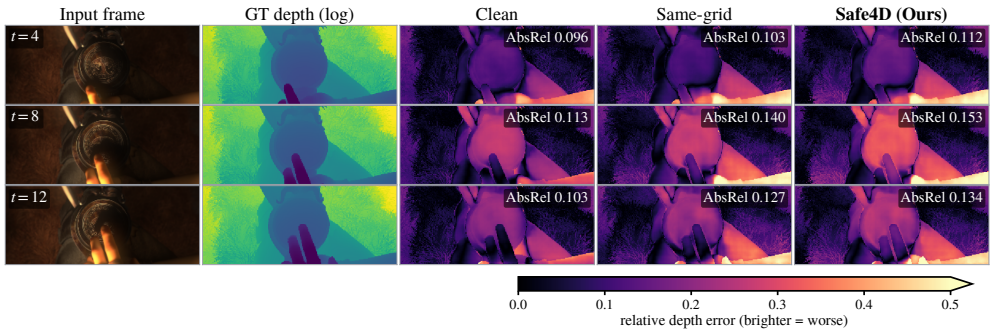


Figure 19: CUT3R depth error on sequence S1 (ten-step videos, direct input; insets: per-frame AbsRel).

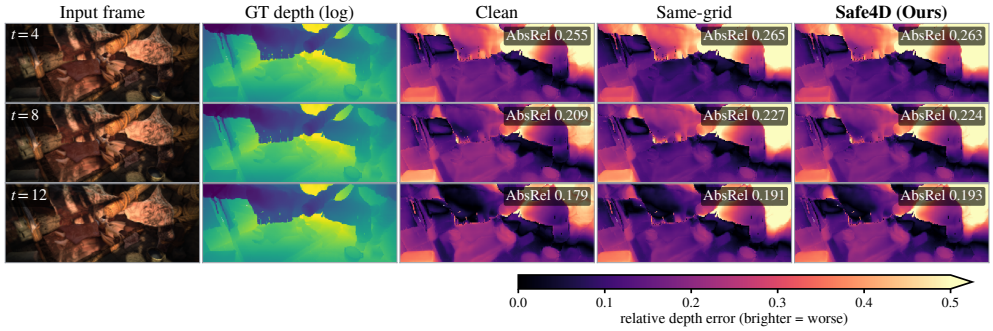


Figure 20: Sequence S2: CUT3R depth error (protocol of Figure 19).

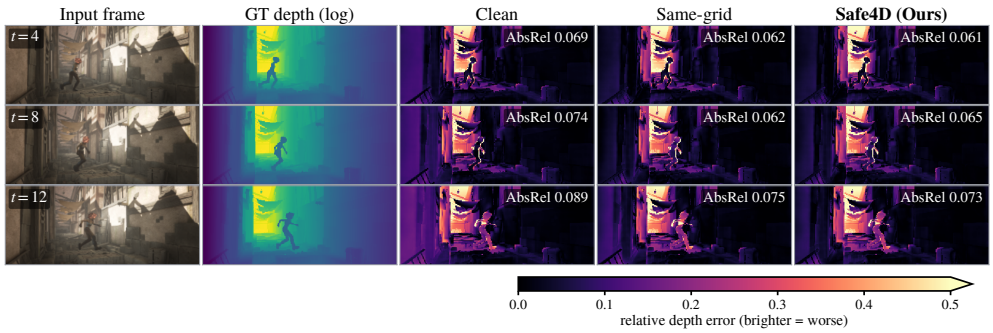


Figure 21: Failure case, sequence S3: clean error is concentrated near the far-field doorway, and frame-mean AbsRel decreases slightly under protection.

Table 26: Purification-aware objective (earlier prototype, VGGT, two complete 50-frame videos). Paired increments ($\times 10^{-3}$) without and with causal avg2 purification, for the direct-only objective and the purified-branch objective (shaded). Optimizing the purified branch doubles post-purification depth damage but weakens unpurified pose damage.

Metric	Purification	S1		S3		Mean		Rel. change
		direct	purified	direct	purified	direct	purified	
ΔATE ↑	none	**18.7**	16.8	**6.79**	1.34	**12.7**	9.05	−28.9%
	avg2	7.43	**14.7**	9.51	**16.0**	8.47	**15.4**	+81.3%
ΔAbsRel ↑	none	82.8	**135**	**43.5**	21.1	63.2	**78.3**	+23.9%
	avg2	90.5	**192**	18.1	**27.4**	54.3	**110**	+102%

L Earlier Prototype and the Purified-Branch Study

The study of Section 5.5 used an earlier prototype that differs from the current method in three respects. (i) Transport used the surrogate's clean depth and camera predictions to back-project and re-project the preceding perturbation, without a z-buffer. Projection-invalid pixels received random initialization, while the subsequent gradient update could edit feasible coordinates across the whole image. The

current method uses two-frame optical flow and a deterministic fallback. (ii) The earlier objective used the implementation's `gauge_balanced_ate` objective and its reference-observability rules, including a legacy depth/geometry warm start at frame index 0 and a relative warm start at index 1. It was not a fixed soft-min of depth and pose throughout the sequence. Each arm optimized either direct input against the clean reference or avg2 input against the avg2-clean reference; the current method uses Eq. 5 on both branches. (iii) The prototype formed one global Top-K one-code proposal and accepted it when the actual score gain exceeded $10^{-6}\max(1, |\text{base score}|)$. A nonempty proposal used six surrogate calls; an empty proposal skipped candidate scoring and used five. The completed runs used 300 calls per 50-frame arm.

The two arms of Table 26 have the same perturbation energy to within 0.2% (105M code^2 on S1 and 142M on S3), and the same fraction of it survives the causal average (98.1% on S1 and 80.3%–80.4% on S3). Scored on CUT3R, the same frozen videos give equal-weight mean increments ΔATE of $13.9/16.0$ ($\times 10^{-3}$, unpurified/avg2) for the direct-only objective and $11.2/18.6$ for the purified-branch objective, and $\Delta\mathrm{AbsRel}$ of $56.3/76.0$ vs. $55.5/74.3$, i.e., the post-purification pose gain ($+16.5\%$) and the unpurified pose loss (-19.7%) transfer to a model that neither arm had seen.

M Refinement Steps

Refinement steps. Table 27 repeats the controlled study with $S = 10$ and $S = 20$ refinement steps per frame, each step re-scoring both branches and recomputing their gradients on the current frame, with the same budget, surrogate and fixed histories, and one commit per frame; the baselines receive the same step budget. Ten steps raise the white-box increment of every initialization about tenfold over a single step ($9.51\times$ reset, $13.1\times$ same-grid, $10.1\times$ motion); the mean input PSNR of the ten-step Safe4D videos is 41.2 dB, and they keep 82.6% of their increment under the centered flow-median3. Twenty steps add little: in steps 11–20 only 40 of 1,440 proposals are accepted, because a rejected step leaves the frame unchanged and the deterministic proposal is then repeated, and five of the nine twenty-step videos are byte-identical to their ten-step counterparts. The plateau therefore reflects the acceptance rule rather than convergence, and a proposal that changes after a rejection is a natural refinement.

With ten steps the ranking of the initializations on these short clips without purification becomes sequence-dependent: motion transport leads on S1, reset on S2 and same-grid carry on S3. S3 is a failure case for motion transport on the surrogate: its same-grid increment (3.22) is the largest of the study and dominates the three-sequence mean (same-grid 1.38 vs. motion 1.11); it is also the sequence whose per-frame depth protection does not transfer to CUT3R (Table 3). The reversal is confined to the unpurified surrogate: under both non-causal purifiers motion transport again has the largest ten-step increment (Tables 27 and 6), and on CUT3R its increment is 2.21–$3.45\times$ that of same-grid carry. On the complete 50-frame sequences of the benchmark, which use the same $S = 10$, motion transport exceeds same-grid carry on all three development sequences in both AbsRel and $\delta{<}1.25$ (Table 8), and on the whole split it gives the lower $\delta{<}1.25$ (Table 1).

Table 27: Refinement steps per frame on the development sequences (StreamVGGT, 16 frames, per-frame scale alignment; same budget, surrogate and fixed histories, one commit per frame): paired $\Delta\mathrm{AbsRel}$ per sequence and three-sequence mean. Fwd. avg2 and cent. fm3 are the non-causal purifiers of Appendix D; the $S{=}1$ columns repeat Tables 12 and 5. Dashes: purified $S{=}20$ videos were generated for motion transport only.

		$S=1$				$S=10$				$S=20$			
Purif.	Init.	S1	S2	S3	mean	S1	S2	S3	mean	S1	S2	S3	mean
direct	Reset	0.118	0.0920	0.0339	0.0812	0.295	1.41	0.612	0.772	0.310	1.48	0.612	0.800
	Same-grid	0.133	0.119	0.0633	0.105	0.366	0.561	3.22	1.38	0.366	0.561	3.22	1.38
	Safe4D (Ours)	0.134	0.127	0.0672	0.109	0.379	0.687	2.26	1.11	0.379	0.530	2.26	1.06
fwd. avg2	Reset	0.126	0.0699	0.0419	0.0792	0.317	0.397	0.581	0.432	—	—	—	—
	Same-grid	0.140	0.130	0.0949	0.122	0.381	0.438	2.26	1.02	—	—	—	—
	Safe4D (Ours)	0.141	0.118	0.0998	0.119	0.235	0.718	2.46	1.14	0.235	0.592	2.46	1.09
cent. fm3	Reset	0.119	0.0676	0.0303	0.0722	0.248	0.605	0.576	0.476	—	—	—	—
	Same-grid	0.135	0.103	0.0481	0.0954	0.298	0.347	1.74	0.796	—	—	—	—
	Safe4D (Ours)	0.137	0.104	0.0614	0.101	0.359	0.568	1.82	0.915	0.359	0.457	1.82	0.878

N Cost and Long-Sequence Generation Diagnostics

The method uses $4+6S$ surrogate calls (ten at $S=1$, 64 at $S=10$) and bidirectional flow estimation for each new frame. The complete three-scene, three-arm, 16-frame generation and source-evaluation study took 19.6 minutes on one RTX 5090. The four earlier 50-frame prototype arms used six calls per nonempty proposal and ran in parallel on four such GPUs in 46 minutes including target evaluation. The 300-frame N3DV same-grid sequence at the surrogate's 518×392 input resolution took 9.35 hours on one GPU with four causal histories offloaded to CPU memory (229 GiB of key–value state at the end); per-frame time grows with history from 26 s to over 200 s. It did not use the current motion initialization. Table 28 summarizes this generation run. The surrogate scores (e_I, e_T) are centered log-depth discrepancies, not ground-truth errors. Their values across time involve different frames and model histories and are not directly comparable.

After frame 100, only 8 of 200 proposals were accepted. The recorded linearized endpoints frequently overestimated the realized change of the quantized candidates (mean predicted $+2.01$ vs. realized -0.30 per candidate); 269 of the 292 selected directions had $\alpha = 0$ (pure purified-branch direction). These observations point to a mismatch between proposal scoring and realized response. Rejected proposals retain the transported initialization. An accepted update improves over that frame's initialization. Smaller or history-dependent steps and a windowed surrogate state are possible follow-ups.

Table 28: 300-frame generation on the N3DV scene, camera 01 (frames 0–7 replay a frozen prefix).

Frames	mean e_I	mean e_T	mean $\min(e_I, e_T)$	accepted / new	s per frame
0–7 (replay)	0.0796	0.0815	0.0792	—	1.34
8–49	1.19	1.31	1.19	21/42	26.0
50–99	4.21	4.47	4.20	20/50	57.9
100–149	5.47	5.55	5.46	4/50	93.5
150–199	5.25	4.99	4.99	0/50	123
200–249	5.32	5.12	5.09	3/50	170
250–299	5.37	5.39	5.33	1/50	206
all 300	4.44	4.44	4.35	49/292	—

O Limitations and Future Work

Transfer. Averaged over the 70 benchmark sequences, the depth and pose errors of the unseen CUT3R do not increase, and per sequence they rise at chance level, although individual sequences transfer strongly and the three sequences of the controlled study transfer for every variant (Table 15). A perturbation shaped by one frozen surrogate is therefore largely specific to that surrogate, and the protection established here is white-box protection of the streaming model the protector can query. Optimizing against several surrogates at once, hedging the objective over pose and over downstream fitting, and understanding which sequences transfer are the directions we consider most promising.

Coverage. Purification robustness is measured on the complete benchmarks for the surrogate and, together with transfer and native geometry, on three 16-frame Sintel sequences; the downstream asset experiment uses one scene with one of twenty cameras protected by an earlier same-grid video, where the asset loses quality when the pipeline purifies its input (Appendix J). Broader coverage of purifiers (including learned purification, image-level purification (Nie et al., 2022) and codec re-encoding) and of reconstruction models, including the offline VGGT and models designed for dynamic content such as VGGT-Ω (Wang et al., 2026), is the natural next step, as is an ablation of the per-frame update budget K.

Scope of the objective. The objective targets depth structure: the recovered surface geometry is degraded, whereas camera pose, focal length and frame-to-frame 3D displacement change less consistently. In the white-box setting without purification the two carrying variants reach similar mean damage, same-grid has the higher TUM ATE, and the margin of motion transport varies across sequences (Appendix M); its advantage appears under purification (Table 2). The verified commit accepted every proposal in the single-step controlled runs and a fraction of them on a long sequence (Appendix N), so its value as a safeguard remains to be isolated.

Cost. The causal contract is an information constraint: the current implementation keeps the surrogate's complete history and does not reach real-time throughput on long streams (Appendix N), which a windowed state could address.