ENDORDH: CONSTRAINED REVERSIBLE INFORMATION TRANSPORT OVER NON-STATIONARY ENDOSCOPIC 3D GAUSSIAN REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic 3D Gaussian Splatting (3DGS) models are increasingly reused as persistent 3D assets, raising the need to embed ownership information without compromising the underlying representation. This is particularly challenging in endoscopic scenes, where clinically sensitive regions restrict admissible modifications and camera-coupled active illumination induces spatially non-stationary spherical harmonic (SH) carriers whose reliability varies across the scene. To address these issues, we introduce EndoRDH, a geometry-aware reversible transport framework that formulates reversible watermarking as constrained information transport over a non-stationary Gaussian representation, allocating information to perceptually low-cost and locally stable carriers while preserving an explicit inverse path to the original host. First, clinically admissible carrier projection restricts watermark updates to a prescribed editable SH subspace while freezing structural Gaussian parameters and excluding protected regions. Second, Riemannian photometric decoupling converts local photometric variation into a carrier-dependent transport geometry without assuming recovery of the underlying illumination physics. Within this geometry, an optimal transport inspired reversible flow adaptively redistributes the watermark payload toward low cost, locally stable carriers while retaining a key conditioned inverse for host recovery. Finally, dual-branch redundant coding distributes error corrected evidence across complementary carrier groups, improving robustness without increasing the per-carrier perturbation budget. Experiments on dynamic endoscopic 3DGS scenes show that EndoRDH achieves a favorable trade-off among representation fidelity, copyright verification, and authorized reversibility, while remaining robust to both rendering-domain and model-domain perturbations. These results suggest that modeling watermark embedding as constrained reversible transport offers a principled alternative to uniform or independently optimized carrier perturbations in non-stationary neural scene representations.

1 INTRODUCTION

Once learned 3D scenes are reused as persistent assets, ownership protection is no longer well described as a bit-insertion problem. It is more naturally posed as *constrained reversible information transport* over a structured representation. Given a host state, an ownership message should move only admissible coordinates, remain verifiable after plausible perturbations, preserve the rendered scene, and still admit an explicit route back to the original host for an authorized user. Existing Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS) watermarking methods establish that neural scene parameters and renderings can carry ownership evidence (Luo et al., 2023; Jang et al., 2024; Huang et al., 2024a; Chen et al., 2025b). Yet most formulations optimize perturbations directly in raw parameter or rendering coordinates, implicitly treating admissible carriers as comparable and reversibility as a secondary decoding objective. This misses the central structure of the problem: the message must be transported through a representation whose feasible directions are restricted, and the transport itself must preserve enough information to be inverted.

[0] Anonymous code and implementation details are provided in the supplementary material.

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.

Feasibility alone, however, does not determine how expensive an admissible modification is. A uniform ℓ_2 penalty imposes a stationary and isotropic geometry: equal coefficient displacements are assigned equal cost regardless of where the carrier lies or how strongly that carrier affects the rendered observation. Dynamic 3DGS violates this assumption. Spherical harmonic (SH) coefficients are attached to spatially distributed Gaussian primitives, so their effective reliability depends on visibility, geometry, and illumination. In endoscopic capture, this heterogeneity is amplified by camera-coupled active lighting, distance attenuation, shadowing, specular tissue response, and non-rigid motion; the same SH perturbation can therefore produce markedly different rendering effects at different carriers. The appropriate fidelity model is consequently not a single global norm, but a *non-stationary Riemannian carrier metric*. EndoRDH treats frozen Gaussian positions as base-space anchors and the editable SH coefficients as a vector-valued field defined over those anchors. Riemannian photometric decoupling (RPD) converts photometric non-stationarity into carrier-dependent edge costs and SH-order-aware local metrics, so that distortion is measured by the geometry of the representation rather than by coefficient magnitude alone.

Once the feasible set and the carrier geometry are fixed, watermark embedding becomes a problem of designing *invertible dynamics*. The desired transformation should transport information preferentially through low cost directions while remaining bijective on the editable subspace. EndoRDH realizes this principle as a geometry aware reversible flow rather than as a collection of independently optimized coefficient offsets. On the admissible SH state, the embedding map is organized as a composition

$$\mathcal{E}_{\mathbf{m},\mathbf{k}} = \Phi_{\mathrm{RPD}}^{-1} \circ F_\theta^{-1} \circ T_{\mathbf{m},\mathbf{k}} \circ F_\theta \circ \Phi_{\mathrm{RPD}}, \tag{1}$$

where Φ_{RPD} calibrates the non-stationary carrier geometry, F_θ is an additive-coupling reversible flow, and $T_{\mathbf{m},\mathbf{k}}$ is a bounded key-conditioned translation in transport coordinates. The optimal transport inspired component does not require a separate many-to-one quantization stage: it uses the metric-induced graph structure to bias message motion toward locally stable carriers, while the reversible coupling supplies the explicit inverse dynamics. This separation is important conceptually: the representation geometry determines *where transport is cheap*, whereas the invertible flow determines *how information can move without destroying host recoverability*.

Dynamic endoscopic 3DGS provides a particularly demanding instantiation of this framework because all three constraints appear simultaneously: protected regions restrict the feasible set, active illumination makes the carrier metric non-stationary, and a small per-carrier perturbation budget limits the robustness obtainable from any single coordinate. EndoRDH instantiates the full pipeline with clinically admissible carrier projection (CACP) to define the editable SH subspace while freezing structural Gaussian parameters, RPD to construct the carrier-dependent geometry, optimal-transport-inspired reversible flow (ORF) to realize geometry-aware invertible dynamics, and dual-branch redundant coding (DBRC) to aggregate error-corrected evidence across complementary carrier groups. In this view, endoscopic copyright protection is not treated as an isolated watermarking heuristic, but as a natural application of constrained reversible transport on a non-stationary neural representation.

Our contributions are:

1. We formulate reversible copyright protection for dynamic 3DGS as *constrained reversible information transport*: admissible edits are separated from the cost geometry used to measure them, and authorized host recovery is built into the transformation rather than delegated to a separately learned restoration network.

2. We develop a representation-geometric realization of this formulation. RPD induces a non-stationary, carrier-dependent metric over the editable SH field, while ORF transports the payload through an explicit composition of invertible maps whose inverse is available under the authorized key and codeword.

3. We instantiate the framework for dynamic endoscopic 3DGS with CACP and DBRC. Under the controlled two-sequence EndoGS protocol, the full EndoRDH configuration reaches 34.29 dB PSNR, 0.9663 SSIM, 0.0204 LPIPS, 98.47% rendering-domain accuracy, and 86.20% model-domain accuracy; attack-averaged accuracies are 94.69% and 83.69%, respectively.

2 Related Work

2.1 Dynamic endoscopic reconstruction and photometric non-stationarity

Dynamic endoscopic reconstruction must accommodate non-rigid tissue motion, occlusion, low texture, and strongly varying illumination. NeRF/surface methods such as EndoNeRF and EndoSurf (Wang et al., 2022; Zha et al., 2023) have been followed by a rapidly growing 3DGS line, including EndoGS, Endo-4DGS, Deform3DGS, EndoGSLAM, HFGS, foundation-model-guided reconstruction, SurgicalGS, and EndoRD-GS (Zhu et al., 2024; Huang et al., 2024b; Yang et al., 2024; Wang et al., 2024; Zhao et al., 2024; Liu et al., 2025; Chen et al., 2025a; Gao et al., 2026). These methods improve deformable-scene reconstruction and rendering, but their objective is scene modeling rather than determining how much watermark perturbation different SH carriers can tolerate. Because a camera-coupled light source, specular tissue, distance attenuation, and shadowing make local appearance sensitivity highly non-uniform, a watermarking method still requires an explicit carrier-dependent modification cost. This motivates RPD, which converts endoscopy-specific photometric variation into a location-dependent metric while retaining SH-order differences.

2.2 3DGS watermarking and reversible medical watermarking

NeRF copyright methods protect both learned representations and their renderings (Luo et al., 2023; Jang et al., 2024). For 3DGS, recent work spans uncertainty-aware parameter perturbation and SH-aware embedding (Huang et al., 2024a; Chen et al., 2025b), robust model/rendering watermarking and generalizable single-pass embedding (Jang et al., 2025; Li et al., 2025), and newer native, compression-aware, and policy-guided schemes (Qin et al., 2026; In et al., 2026; Li et al., 2026). Reversible data hiding and medical watermarking instead emphasize host preservation and recovery (Hu & Xiang, 2022; Taj et al., 2024). Invertible hiding networks explicitly couple forward embedding and backward recovery (Jing et al., 2021; Guan et al., 2023; Shang et al., 2023). Directly transferring these image-domain formulations to floating-point 3DGS parameters does not provide the parameter-level constraints or inverse structure required here. Dynamic endoscopy therefore leaves three coupled gaps: a clinically admissible editing domain is not explicitly enforced, learned parameter changes need not expose a deterministic inverse, and robustness under a bounded perturbation budget cannot rely only on stronger local signals. CACP, ORF, and DBRC address these requirements within a common EndoGS host.

2.3 Riemannian geometry, optimal transport, and reversible dynamics

Riemannian geometry allows local distances to vary over the carrier space, as exploited by recent manifold-aware diffusion and flow models (Huang et al., 2022; Chen & Lipman, 2024). We use this property to instantiate a photometrically conditioned metric rather than a global normalization. Optimal transport provides a complementary principle for preferring lower-cost state transitions under that metric (Tong et al., 2024), while invertible coupling supplies an explicit inverse that a transport objective alone does not guarantee. Accordingly, ORF uses ideas from optimal transport and Fokker–Planck dynamics as a geometric prior for preferred movement. It does not claim to solve physical light transport or a globally optimal transport problem.

3 Method

EndoRDH is a constrained bijection over the spherical-harmonic (SH) part of a pretrained dynamic endoscopic 3DGS representation (Kerbl et al., 2023; Zhu et al., 2024). Figure 1 summarizes the pipeline. CACP defines the admissible SH subspace, RPD constructs a non-stationary carrier geometry, ORF realizes the keyed reversible transport, and DBRC distributes coded evidence across complementary carrier groups. The training, verification, and authorized-recovery procedures are specified in Section 3.6; metric construction, inverse details, and implementation parameters are deferred to Appendices A–D.

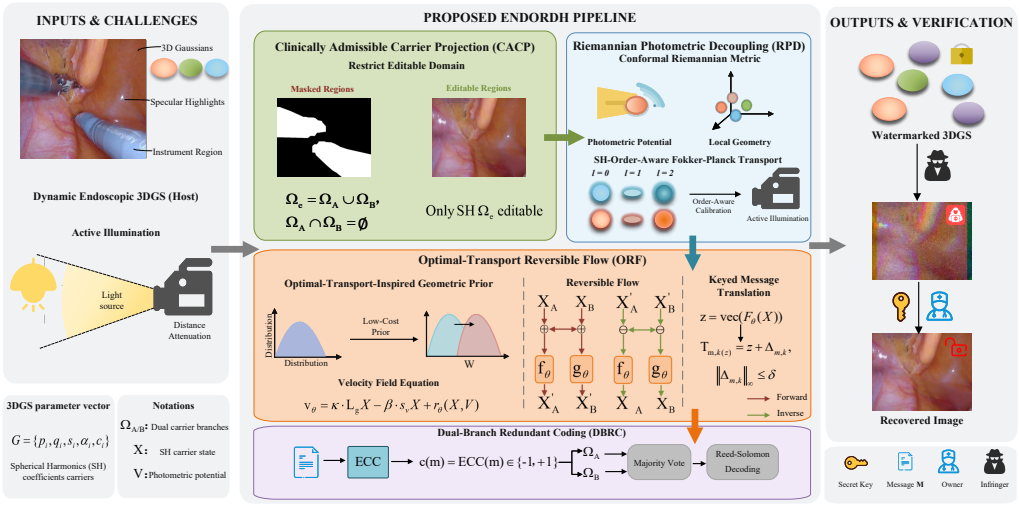


Figure 1: Overview of EndoRDH. CACP restricts watermark updates to admissible SH carriers, RPD defines carrier-dependent photometric geometry, ORF transports the keyed payload through an explicitly invertible map, and DBRC distributes redundant evidence for model- and rendering-domain verification. The authorized path removes the keyed transport displacement and applies the inverse maps to recover the host state.

3.1 Problem Formulation and Overall Framework

Let the pretrained host be $\mathcal{G}^h = \{(\boldsymbol{\mu}_i, \mathbf{R}_i, \mathbf{s}_i, \alpha_i, \mathbf{a}_i^h)\}_{i=1}^N$, where $\boldsymbol{\mu}_i \in \mathbb{R}^3$, $\mathbf{R}_i$, $\mathbf{s}_i$, and α_i are the Gaussian position, rotation, scale, and opacity, and $\mathbf{a}_i^h$ collects its RGB SH coefficients. A differentiable rasterizer $\mathcal{R}$ produces an image $\hat{I} = \mathcal{R}(\mathcal{G}, V)$ from camera V. Given a copyright message $\mathbf{m}$ and secret key $\mathbf{k}$, the watermark is allowed to modify only an editable SH subset Ω_e. Its complement Ω_f and all structural Gaussian parameters must remain unchanged.

We therefore seek a keyed map whose forward pass is expressive enough for robust verification but whose action on the admissible state is bijective. EndoRDH realizes this map by composing three transformations. RPD first maps the editable SH coefficients to photometrically calibrated coordinates, F_θ then maps these coordinates to reversible transport coordinates, and $T_{\mathbf{m},\mathbf{k}}$ inserts the coded message as a bounded translation in that transport space. The watermarked coefficients are

$$\mathbf{a}^w_{\Omega_e} = \Phi^{-1}_{\mathrm{RPD}} \circ F_\theta^{-1} \circ T_{\mathbf{m},\mathbf{k}} \circ F_\theta \circ \Phi_{\mathrm{RPD}}(\mathbf{a}^h_{\Omega_e}), \qquad \mathbf{a}^w_{\Omega_f} = \mathbf{a}^h_{\Omega_f}. \tag{2}$$

The F_θ^{-1} pullback is essential because $T_{\mathbf{m},\mathbf{k}}$ acts after the forward transport map. The translated state must therefore be returned to calibrated SH coordinates before applying Φ^{-1}_{RPD}. DBRC does not change this algebraic map. It determines how the message translation is assigned to carriers and how the resulting evidence is aggregated at verification time. For brevity, we suppress $\mathbf{k}$ and write $T_{\mathbf{m}}$ when the key is fixed.

This construction separates three objects that are often conflated in watermarking: the *feasible set* of coefficient changes, the *cost geometry* within that set, and the *invertible dynamics* used to carry information. CACP, RPD, and ORF instantiate these objects, respectively, while DBRC improves robustness without enlarging the allowed displacement of an individual carrier.

3.2 CACP: Clinically Admissible Carrier Projection

A rendering loss can discourage undesirable edits but cannot prevent the optimizer from using geometry, opacity, or protected regions as convenient watermark channels. CACP removes these degrees of freedom before watermark optimization. Let $\Delta\mathcal{G}$ denote a candidate parameter update. The

admissible update space is

$$\mathcal{A}_{\text{safe}} = \{\Delta\boldsymbol{\mu} = \Delta\mathbf{R} = \Delta\mathbf{s} = \Delta\alpha = 0,\ \Delta\mathbf{a}_i = 0\ (i \in \Omega_f)\}\,. \tag{3}$$

Thus only SH coefficients indexed by Ω_e can change. In practice, Ω_f contains carriers excluded by the supplied protected-region policy together with numerically unsuitable carriers, while the remaining candidates are filtered by host opacity and scale statistics. The exact thresholds are given in Appendix D. During optimization, the corresponding structural and frozen-SH gradients are masked, so Eq. 3 is enforced as a projection rather than as a penalty.

The editable set is further divided into two disjoint carrier groups,

$$\Omega_e = \Omega_A \dot{\cup} \Omega_B, \qquad \Omega_A \cap \Omega_B = \varnothing, \tag{4}$$

which are later used as complementary DBRC branches and as the two states in reversible coupling. This partition is fixed for a given key and host, so all subsequent modules operate on the same carrier indexing. CACP therefore provides a hard parameter-domain guarantee relative to the supplied protected-region policy and frozen-variable set: those variables are not updated by the watermarking optimization. This statement should not be interpreted as a clinical-safety certification of the resulting model or downstream procedure.

3.3 RPD: Riemannian Photometric Decoupling

After CACP, the remaining SH coefficients are editable, but they are not equally expensive to perturb. In endoscopic capture, the same coefficient displacement can produce different rendering changes because the attached Gaussians occupy different positions relative to the camera-coupled light and exhibit different local photometric stability. We encode this non-stationarity through a carrier metric rather than forcing all coefficients into one global Euclidean norm.

For editable carrier i, let V_i denote a fixed photometric potential computed from the pretrained host and let $\gamma_i > 0$ combine this potential with local host variation. We define the local base-space metric $\mathbf{G}_i = \gamma_i \mathbf{I}_3$ and the pairwise Riemannian distance

$$d_g^2(i,j) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \frac{\mathbf{G}_i + \mathbf{G}_j}{2} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \tag{5}$$

The Gaussian means are frozen anchors in Eq. 5 and are never transported by the watermark. The metric only determines how strongly SH carriers communicate. Specifically,

$$W_{ij} = \exp\left[-\frac{d_g^2(i,j)}{2\sigma_g^2}\right], \qquad P_{ij} = \frac{W_{ij}}{\sum_k W_{ik}}, \qquad (L_g\mathbf{X})_i = \sum_j P_{ij}(\mathbf{x}_j - \mathbf{x}_i), \tag{6}$$

where σ_g controls locality, $\mathbf{X} = [\mathbf{x}_i]_{i\in\Omega_e}$ is an SH-valued carrier field, and $L_g = P - I$ is the resulting graph diffusion operator. Hence, RPD lifts a metric defined on frozen 3D anchors to a neighborhood structure acting directly on editable SH states.

We additionally distinguish the direction-independent SH term from higher-order view-dependent terms. Writing $\mathbf{a}_i = [\mathbf{a}_i^0, \mathbf{a}_i^+]$, RPD uses the order-aware reparameterization

$$\Phi_{\text{RPD}}(\mathbf{a}_i) = [\mathbf{a}_i^0/s_i, \mathbf{a}_i^+], \qquad \Phi_{\text{RPD}}^{-1}(\mathbf{x}_i) = [s_i\mathbf{x}_i^0, \mathbf{x}_i^+], \tag{7}$$

where the positive scale s_i is computed once from frozen host statistics and clipped away from zero. Equivalently, this map induces the local SH metric

$$\mathbf{M}_i = \text{diag}(s_i^{-2}\mathbf{I}_{d_0}, \mathbf{I}_{d_+}), \qquad \|\Delta\mathbf{a}_i\|_{\mathbf{M}_i}^2 = \|\Phi_{\text{RPD}}(\mathbf{a}_i + \Delta\mathbf{a}_i) - \Phi_{\text{RPD}}(\mathbf{a}_i)\|_2^2, \tag{8}$$

with d_0 and d_+ denoting the DC and higher-order dimensions. Equations 5–8 separate *where* transport is expensive on the carrier graph from *which SH directions* are expensive at a carrier. RPD is therefore a carrier geometry consistent with local photometric stability. It is not an estimator of true illumination, intrinsic reflectance, or physical light transport.

3.4 ORF: Optimal-Transport Reversible Flow

RPD specifies a non-stationary cost geometry, but a cost by itself does not provide a reversible write operation. Let $\mathbf{X} = [\Phi_{\text{RPD}}(\mathbf{a}_i)]_{i\in\Omega_e}$ denote the calibrated carrier state and introduce an auxiliary

transport coordinate $\tau \in [0, 1]$, unrelated to the physical time of the dynamic scene. We use the RPD geometry to bias a continuous transport field

$$\frac{d\mathbf{X}}{d\tau} = \mathbf{v}_\theta(\mathbf{X}, V), \qquad \mathbf{v}_\theta = \kappa L_g \mathbf{X} - \beta\, \mathbf{s}_V \odot \mathbf{X} + \mathbf{r}_\theta(\mathbf{X}, V), \tag{9}$$

where $\kappa, \beta \geq 0$ weight the geometry-induced terms, $\mathbf{s}_V$ summarizes local variation of the fixed photometric potential, and $\mathbf{r}_\theta$ is a learnable residual. The first two terms prefer locally coherent, lower-cost directions supplied by RPD. The residual adapts this prior to the carrier statistics. This drift-diffusion interpretation is inspired by Riemannian optimal transport and Fokker–Planck flow constructions (Huang et al., 2022; De Bortoli et al., 2022; Chen & Lipman, 2024; Tong et al., 2024). This connection serves only as a geometric inductive bias. Equation 9 is not claimed to solve physical light transport or a globally optimal transport problem.

To make the finite transformation exactly invertible, we do not numerically integrate Eq. 9 with an unconstrained solver. Instead, ORF discretizes the transport with additive coupling. For the two CACP branches, one block performs

$$\mathbf{X}'_B = \mathbf{X}_B + h f_\theta(\mathbf{X}_A, V_A), \qquad \mathbf{X}'_A = \mathbf{X}_A + h g_\theta(\mathbf{X}'_B, V_B), \tag{10}$$

where h is the step size, f_θ and g_θ are coupling subnetworks, and V_A, V_B are branch-specific RPD conditioning variables. The inverse follows analytically by reversing the update order,

$$\mathbf{X}_A = \mathbf{X}'_A - h g_\theta(\mathbf{X}'_B, V_B), \qquad \mathbf{X}_B = \mathbf{X}'_B - h f_\theta(\mathbf{X}_A, V_A). \tag{11}$$

Stacking such blocks defines the bijection F_θ and its exact algebraic inverse F_θ^{-1}. No separately learned recovery network is needed.

The copyright message is inserted after this reversible change of coordinates. Let $\mathbf{z} = \mathrm{vec}(F_\theta(\mathbf{X}))$ and let $\mathbf{\Delta}_{\mathbf{m},\mathbf{k}}$ be the key-indexed signed displacement generated from the coded message. We use

$$T_{\mathbf{m},\mathbf{k}}(\mathbf{z}) = \mathbf{z} + \mathbf{\Delta}_{\mathbf{m},\mathbf{k}}, \qquad T^{-1}_{\mathbf{m},\mathbf{k}}(\mathbf{z}) = \mathbf{z} - \mathbf{\Delta}_{\mathbf{m},\mathbf{k}}, \qquad \|\mathbf{\Delta}_{\mathbf{m},\mathbf{k}}\|_\infty \leq \delta. \tag{12}$$

The bound constrains the displacement magnitude. It does not clip, round, or snap the output to a message center. Such many-to-one operations would destroy the inverse required by Eq. 2.

3.5 DBRC: Dual-Branch Redundant Coding

The displacement budget in Eq. 12 limits how much evidence any single carrier can contribute. DBRC therefore improves robustness by repeating weak evidence rather than increasing the local perturbation. The message is first error-corrected, $\mathbf{c}(\mathbf{m}) = \mathrm{ECC}(\mathbf{m}) \in \{-1, +1\}^Q$, and the secret key defines a coordinate-to-symbol assignment $\pi_\mathbf{k}$ across Ω_A and Ω_B. For an assigned transport coordinate ℓ,

$$\Delta_{\mathbf{m},\mathbf{k},\ell} = \delta\, c_{\pi_\mathbf{k}(\ell)}(\mathbf{m}), \qquad \Delta_{\mathbf{m},\mathbf{k},\ell} = 0 \text{ for unassigned coordinates.} \tag{13}$$

Each coded symbol is represented by multiple local carrier groups in both branches. Model-domain verification makes carrier-level decisions, aggregates them by within-group majority voting and repetition fusion, combines the two branches, and then applies ECC decoding. Rendering-domain verification predicts the same codeword from rendered images after differentiable image perturbations, so the two verification paths share the message organization even though their observations differ.

The effect of repetition can be seen from a standard concentration bound. If one coded symbol s_k is represented by K weakly independent carrier decisions, each wrong with probability $p < 1/2$, majority voting obeys

$$\Pr(\hat{s}_k \neq s_k) \leq \exp\left[-2K\left(\frac{1}{2} - p\right)^2\right]. \tag{14}$$

This bound explains the design principle but is not used to predict empirical accuracy: model and rendering attacks can induce correlated errors. The relevant point for EndoRDH is that robustness is accumulated across carriers and complementary branches while the per-coordinate displacement remains bounded by δ.

3.6 JOINT OPTIMIZATION, TRAINING, AND INFERENCE

CACP and the RPD geometry are fixed before watermark optimization. Training then couples image fidelity, transport-domain message separation, rendering-domain detectability, and consistency of the explicit inverse. We minimize

$$\mathcal{L} = \mathcal{L}_{\mathrm{fd}} + \lambda_{\mathrm{2D}}\mathcal{L}_{\mathrm{2D}} + \lambda_{\mathrm{3D}}\mathcal{L}_{\mathrm{3D}} + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{inv}} + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{endo}}\mathcal{L}_{\mathrm{endo}}. \tag{15}$$

Here $\mathcal{L}_{\mathrm{fd}}$ measures host/rendering distortion over the admissible tissue region (RGB, depth, and local smoothness in our implementation). $\mathcal{L}_{\mathrm{3D}}$ supervises codeword separation in transport coordinates, and $\mathcal{L}_{\mathrm{2D}}$ trains the rendering-domain decoder on attacked watermarked renderings. $\mathcal{L}_{\mathrm{inv}}$ penalizes deviation along the authorized forward and inverse path, while $\mathcal{L}_{\mathrm{sem}}$ and $\mathcal{L}_{\mathrm{endo}}$ preserve semantic and endoscopy-specific stability. Exact weights, attack distributions, and schedules are reported in Appendix D.

At each training iteration, we ECC-encode the sampled message, assign its symbols to the two fixed CACP branches, evaluate Eq. 2, and render the resulting watermarked 3DGS. Model-domain perturbations act on the editable representation, whereas differentiable rendering-domain attacks act on the rendered images before decoding. Gradients update only the editable SH state and watermark-specific networks. The host geometry, opacity, protected SH carriers, and fixed RPD statistics remain unchanged. This ordering is important because it prevents the fidelity and robustness losses from relaxing the admissible set defined in Eq. 3.

At inference, ownership can be verified either directly from the model through DBRC voting/ECC decoding or from renderings through the 2D decoder. Authorized recovery uses the same key and coded message to remove the signed transport displacement. From Eq. 2,

$$\begin{aligned} \mathbf{z}^w &= F_\theta(\Phi_{\mathrm{RPD}}(\mathbf{a}^w_{\Omega_e})) = T_{\mathbf{m}}(F_\theta(\Phi_{\mathrm{RPD}}(\mathbf{a}^h_{\Omega_e}))), \\ \mathbf{z}^r &= T^{-1}_{\mathbf{m}}(\mathbf{z}^w), \\ \mathbf{a}^r_{\Omega_e} &= \Phi^{-1}_{\mathrm{RPD}}(F^{-1}_\theta(\mathbf{z}^r)) = \mathbf{a}^h_{\Omega_e}. \end{aligned} \tag{16}$$

CACP already preserves $\mathbf{a}_{\Omega_f}$ and all structural parameters, so Eq. 16 extends to the complete host state. The equalities characterize the structural limit of the reversible composition: zero recovery MSE, PSNR$\to +\infty$, and SSIM$\to 1$ follow only with the correct key/codeword, exact arithmetic, deterministic rendering, and no irreversible post-embedding processing. They are not reported as empirical experimental scores.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP AND EVALUATION PROTOCOL

We evaluate on the provided *cutting* and *pulling* endoscopic sequences. For each matched frame, GT denotes the observed ground-truth image and host denotes the rendering of the unwatermarked EndoGS model (Zhu et al., 2024). The supplied mask is first cropped to the image height. White pixels correspond to instrument regions and are excluded, so all image-quality metrics are computed over the remaining tissue ROI. Watermarked renderings are compared with GT. Authorized recovery uses the host as its reference because its objective is to invert the watermark transformation rather than reconstruct the camera observation. We first average all available frames within each scene and then report an equal-weight average over *cutting* and *pulling*. PSNR, SSIM, and spatial AlexNet LPIPS measure image quality, while 2DACC and 3DACC measure rendering-domain and model-domain watermark verification accuracy, respectively.

All three configurations share the same reversible backbone, coding budget, training protocol, and frozen geometry and opacity. EndoRDH denotes the full configuration, while P-NODE and SFP-NODE are controlled ablations that differ only in photometric conditioning. This comparison reduces architectural confounding when assessing the SH-order-aware RPD design. For external baselines, 3D-GSW (Jang et al., 2025) and GS-Marker (Li et al., 2025) are *not* evaluated by directly running their original reconstruction codebases. Following the methods described in their papers and corresponding released implementations, we migrate their watermark embedding/extraction procedures onto the same pretrained EndoGS host used by our method. The compared methods therefore

share the EndoGS reconstruction, input sequences, ROI definition, rendering interface, and evaluation pipeline while retaining their method-specific watermark logic. The resulting numbers are controlled EndoGS transfer comparisons, not reproductions of the original papers' reconstruction benchmarks. Full architectural details, optimization settings, carrier selection, message parameters, baseline migration protocol, and attack configurations are provided in Appendix D.

4.2 Main results

Table 1: Equal-weight scene averages over tissue ROIs under the common EndoGS evaluation protocol. Watermarked rendering quality is measured against GT. Best results are in bold.

Method	PSNR↑	SSIM↑	LPIPS↓	2DACC↑	3DACC↑
GS-Marker	31.16	0.8981	0.1275	97.46	55.39
3D-GSW	32.09	0.9036	0.1532	97.99	56.03
P-NODE	31.94	0.8879	0.1380	95.28	55.41
SFP-NODE	33.96	0.9087	0.1149	97.16	80.42
EndoRDH	**34.29**	**0.9663**	**0.0204**	**98.47**	**86.20**

Table 1 shows that EndoRDH achieves the best performance across all reported primary metrics under the shared host and evaluation pipeline. Compared with SFP-NODE, SH-order-aware conditioning yields consistent improvements in PSNR and SSIM, reduces LPIPS, and improves 3DACC. The relatively limited gain in 2DACC suggests that both variants already decode reliably from renderings; order-aware carrier calibration therefore matters more for the more demanding model-domain verification task. Since the external baselines are controlled transfers to EndoGS, this protocol mitigates confounding due to different reconstruction backbones. The comparison should not be interpreted as a reproduction of the baselines' original benchmark settings. Figure 2 complements these aggregate results with matched render/residual panels. For visualization, the residuals are amplified with respect to the absolute RGB differences from GT and are masked to the tissue ROI; they serve as a spatial diagnostic rather than an additional metric.

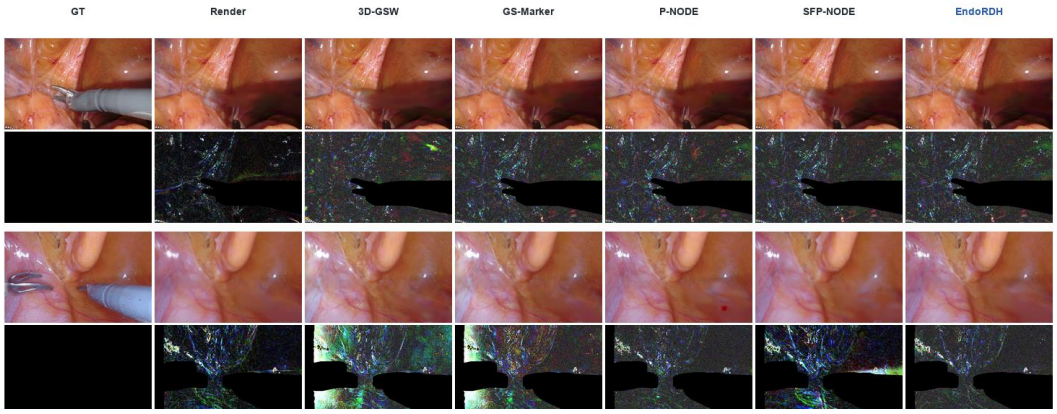


Figure 2: Matched rendering and residual panels for selected cutting and pulling views. Columns follow the evaluation order: GT, the unwatermarked host render, 3D-GSW, GS-Marker, P-NODE, SFP-NODE, and EndoRDH. The lower row for each view shows the absolute RGB residual amplified by $\times 10$ and masked to the tissue ROI.

4.3 Authorized recovery

Authorized recovery follows the explicit inverse of Eq. 2. With the correct key, the keyed message displacement is removed in transport coordinates, and the inverse ORF/RPD maps return the editable

SH state. Because CACP leaves the protected SH subset and all structural parameters unchanged, the complete host is recovered under exact arithmetic when no irreversible post-embedding modification has occurred. A deterministic renderer then yields zero recovery MSE, PSNR tending to $+\infty$, and SSIM tending to 1 in the ideal limit. Finite precision or irreversible attacks can leave nonzero residuals. The EndoRDH-only recovery panel is provided as Figure 5 in the supplement, while Appendix B gives the explicit inverse sequence and formal recovery limit.

4.4 Robustness to 2D and 3D attacks

Table 2: Verification accuracy (%) under clean conditions and averaged attacks. Overall is the mean of the 2D- and 3D-attack averages.

Method	Clean 2D	Avg. 2D attacks	Clean 3D	Avg. 3D attacks	Overall
GS-Marker	97.46	89.77	55.39	51.66	70.72
3D-GSW	97.99	89.75	56.03	52.06	70.90
P-NODE	95.28	87.99	55.41	52.59	70.29
SFP-NODE	97.16	93.27	80.42	75.73	84.50
EndoRDH	**98.47**	**94.69**	**86.20**	**83.69**	**89.19**

Table 2 summarizes clean and attack-averaged verification. EndoRDH reaches 94.69% average accuracy under rendering-domain attacks and 83.69% under model-domain attacks, compared with 93.27% and 75.73% for SFP-NODE. The larger separation in the model domain is consistent with the intended role of order-aware carrier calibration and redundant evidence when perturbations act directly on Gaussian parameters. Attack-specific values and the corresponding heatmap are reported in Appendix I (Figure 7).

5 Conclusion

We presented EndoRDH, a reversible watermarking framework for dynamic endoscopic 3DGS that separates four concerns that are otherwise entangled in a single perturbation objective: the admissible parameter subspace, the non-stationary cost of SH modification, the invertibility of message transport, and the aggregation of bounded verification evidence. CACP, RPD, ORF, and DBRC address these roles within a single reversible host transformation. Under the controlled two-sequence EndoGS protocol, EndoRDH improves both rendering fidelity and model-domain verification over the shared-backbone ablations and remains robust under the evaluated rendering- and model-domain perturbations. The exact recovery identity assumes the correct key, exact arithmetic, deterministic rendering, and an uncorrupted watermarked state. Finite precision and irreversible post-embedding changes can produce residual error. Broader validation across endoscope systems, protected-region policies, and downstream clinical tasks is therefore an important next step.

AI Use Statement

Generative AI tools were used for language polishing and literature discovery. All AI-assisted content and references were reviewed and verified by the authors, who take full responsibility for the final manuscript.

Reproducibility Statement

We provide detailed information to facilitate reproduction of EndoRDH. The main paper specifies the complete formulation of CACP, RPD, ORF, and DBRC. The supplementary material provides the RPD construction, reversible flow architecture and explicit inverse, carrier selection, message coding, optimization hyperparameters, attack configurations, and baseline adaptation protocols. All methods are evaluated using the same pretrained EndoGS host, input sequences, tissue ROI definition, rendering interface, and evaluation pipeline. Anonymous implementation code and additional implementation details are provided in the supplementary material.

REFERENCES

Jialei Chen, Xin Zhang, Mobarak I. Hoque, Francisco Vasconcelos, Danail Stoyanov, Daniel S. Elson, and Baoru Huang. SurgicalGS: Dynamic 3D gaussian splatting for accurate robotic-assisted surgical scene reconstruction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 572–582, 2025a. doi: 10.1007/978-3-032-05141-7_55.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, 2024.

Zixuan Chen, Guangcong Wang, Jiahao Zhu, Jianhuang Lai, and Xiaohua Xie. GuardSplat: Efficient and robust watermarking for 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16325–16335, 2025b.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Bingchen Gao, Jun Zhou, Jing Zou, and Jing Qin. EndoRD-GS: Robust deformable endoscopic scene reconstruction via gaussian splatting. *IEEE Transactions on Medical Imaging*, 45(2):528–541, 2026. doi: 10.1109/TMI.2025.3600253.

Zhenyu Guan, Junpeng Jing, Xin Deng, Mai Xu, Lai Jiang, Zhou Zhang, and Yipeng Li. DeepMIH: Deep invertible network for multiple image hiding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):372–390, 2023. doi: 10.1109/TPAMI.2022.3141725.

Runwen Hu and Shijun Xiang. Reversible data hiding by using CNN prediction and adaptive embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10196–10208, 2022. doi: 10.1109/TPAMI.2021.3131250.

Chin-Wei Huang, Milad Aghajohari, Avishek Joey Bose, Prakash Panangaden, and Aaron Courville. Riemannian diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2750–2761, 2022.

Xiufeng Huang, Ruiqi Li, Yiu-ming Cheung, Ka Chun Cheung, Simon See, and Renjie Wan. GaussianMarker: Uncertainty-aware copyright protection of 3D gaussian splatting. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-1040.

Yiming Huang, Beilei Cui, Long Bai, Ziqi Guo, Mengya Xu, Mobarakol Islam, and Hongliang Ren. Endo-4DGS: Endoscopic monocular scene reconstruction with 4D gaussian splatting. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 197–207, 2024b. doi: 10.1007/978-3-031-72089-5_19.

Sumin In, Youngdong Jang, Utae Jeong, MinHyuk Jang, Hyeongcheol Park, Eunbyung Park, and Sangpil Kim. CompMarkGS: Robust watermarking for compressed 3D gaussian splatting. In *International Conference on Learning Representations*, 2026.

Youngdong Jang, Dong In Lee, MinHyuk Jang, Jong Wook Kim, Feng Yang, and Sangpil Kim. WateRF: Robust watermarks in radiance fields for protection of copyrights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12087–12097, 2024.

Youngdong Jang, Hyunje Park, Feng Yang, Heeju Ko, Euijin Choo, and Sangpil Kim. 3D-GSW: 3D gaussian splatting for robust watermarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5938–5948, 2025.

Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. HiNet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4733–4742, 2021.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

Lijiang Li, Jinglu Wang, Xiang Ming, and Yan Lu. GS-Marker: Generalizable and robust watermarking for 3D gaussian splatting. *arXiv preprint arXiv:2503.18718*, 2025.

Nan Li, Yike Zeng, Qian Zhang, Qi Zhang, Zhiyi Pan, Wei Feng, and Liang Wan. Write where it matters: Policy-guided watermarks for 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6580–6590, 2026.

Yifan Liu, Chenxin Li, Hengyu Liu, Chen Yang, and Yixuan Yuan. Foundation model-guided gaussian splatting for 4D reconstruction of deformable tissues. *IEEE Transactions on Medical Imaging*, 44(6):2672–2682, 2025. doi: 10.1109/TMI.2025.3545183.

Ziyuan Luo, Qing Guo, Ka Chun Cheung, Simon See, and Renjie Wan. CopyRNeRF: Protecting the CopyRight of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22401–22411, 2023.

Hao Qin, Yukai Sun, Luyuan Chen, Mengxu Lu, Feng Zhang, Ming Kong, Zhenhong Du, and Qiang Zhu. NGS-Marker: Robust native watermarking for 3D gaussian splatting. In *International Conference on Learning Representations*, 2026.

Fei Shang, Yuhang Lan, Jianhua Yang, Enping Li, and Xiangui Kang. Robust data hiding for JPEG images with invertible neural network. *Neural Networks*, 163:219–232, 2023. doi: 10.1016/j.neunet.2023.03.037.

Rizwan Taj, Feng Tao, Saima Kanwal, Ahmad Almogren, Ayman Altameem, and Ateeq Ur Rehman. A reversible-zero watermarking scheme for medical images. *Scientific Reports*, 14:17320, 2024. doi: 10.1038/s41598-024-67672-9.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Kailing Wang, Chen Yang, Yuehao Wang, Sikuang Li, Yan Wang, Qi Dou, Xiaokang Yang, and Wei Shen. EndoGSLAM: Real-time dense reconstruction and tracking in endoscopic surgeries using gaussian splatting. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 219–229, 2024. doi: 10.1007/978-3-031-72089-5_21.

Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 431–441, 2022.

Shuojue Yang, Qian Li, Daiyun Shen, Bingchen Gong, Qi Dou, and Yueming Jin. Deform3DGS: Flexible deformation for fast surgical scene reconstruction with gaussian splatting. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 132–142, 2024. doi: 10.1007/978-3-031-72089-5_13.

Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge. EndoSurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 13–23, 2023.

Haoyu Zhao, Xingyue Zhao, Lingting Zhu, Weixi Zheng, and Yongchao Xu. HFGS: 4D gaussian splatting with emphasis on spatial and temporal high-frequency components for endoscopic scene reconstruction. In *British Machine Vision Conference*, 2024.

Lingting Zhu, Zhao Wang, Jiahao Cui, Zhenchao Jin, Guying Lin, and Lequan Yu. EndoGS: Deformable endoscopic tissues reconstruction with gaussian splatting. In *MICCAI 2024 Workshops: ISIC, iMIMIC, EARTH and DeCaF*, pp. 135–145, 2024. doi: 10.1007/978-3-031-77610-6_13.

A CONCRETE RPD CONSTRUCTION AND SUPPLEMENTARY INTERPRETATION

The main text writes RPD through a positive scale γ_i induced by photometric potential and local variation. In the implementation, let $r_i = \|\boldsymbol{\mu}_i - \mathbf{c}\|_2$ and use $V_i = 1/\max(r_i, r_{\min})^2$ as a monotonic distance-related photometric potential. Denote local neighborhood variation by h_i, and define $g_i = V_i^2 + \lambda_h h_i + \varepsilon$ and $\mathbf{G}_i = g_i \mathbf{I}_3$. The distance in Eq. 5 then induces $W_{ij} = \exp[-d_g^2(i,j)/(2\sigma_g^2)]$, $D_{ii} = \sum_j W_{ij}$, and $L_g = D^{-1}W - I$.

The metric-weighted local first-order difference approximates the potential gradient as

$$\nabla_g V_i \approx \frac{1}{g_i} \sum_{j \in \mathcal{N}(i)} W_{ij}(V_j - V_i)(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i). \tag{17}$$

EndoRDH uses the clipped potential $\bar{V}_i = \mathrm{clip}(V_i, 0.01, 10)$ and sets $s_i = \bar{V}_i$ in Eq. 7. P-NODE uses $[\mathbf{a}_i^0, \mathbf{a}_i^+]$, SFP-NODE uses $[\mathbf{a}_i^0/\bar{V}_i, \mathbf{a}_i^+/\bar{V}_i]$, and EndoRDH uses $[\mathbf{a}_i^0/\bar{V}_i, \mathbf{a}_i^+]$. The three variants therefore share the downstream reversible backbone and coding pipeline and differ only in photometric conditioning.

A.1 EXISTENCE AND NUMERICAL STABILITY OF Φ_{RPD}^{-1}

Let d_0 and d_+ denote the dimensions of the DC and higher-order SH blocks, respectively, and write

$$\mathbf{D}_i = \mathrm{diag}\big(s_i^{-1}\mathbf{I}_{d_0}, \mathbf{I}_{d_+}\big), \qquad \Phi_{\mathrm{RPD}}(\mathbf{a}_i) = \mathbf{D}_i \mathbf{a}_i. \tag{18}$$

In EndoRDH, $s_i = \mathrm{clip}(V_i, s_{\min}, s_{\max})$ with $s_{\min} = 0.01$ and $s_{\max} = 10$. Hence $0 < s_{\min} \le s_i \le s_{\max} < \infty$ and every diagonal entry of $\mathbf{D}_i$ is nonzero. Therefore $\mathbf{D}_i$ is nonsingular and

$$\Phi_{\mathrm{RPD}}^{-1}(\mathbf{x}_i) = \mathbf{D}_i^{-1}\mathbf{x}_i = \big[s_i \mathbf{x}_i^0, \mathbf{x}_i^+\big]. \tag{19}$$

The global map over Ω_e is block diagonal, $\mathbf{D} = \mathrm{blkdiag}_{i \in \Omega_e} \mathbf{D}_i$, and is consequently a bijection as well. Moreover, $\kappa_2(\mathbf{D}_i) = \max(s_i, s_i^{-1}) \le \max(s_{\max}, s_{\min}^{-1})$, so clipping prevents a singular or unbounded rescaling. A second condition is equally important: s_i is computed from the frozen position $\boldsymbol{\mu}_i$ and fixed photometric statistics, rather than from the watermarked coefficient value itself. Thus embedding and recovery evaluate exactly the same scale. If the scale were recomputed from an edited SH state without being stored, the inverse in Eq. 19 would in general no longer be guaranteed.

A.2 FROM A POSITIONAL METRIC TO AN SH-SPACE TRANSPORT COST

Equation 5 is defined on the *base points* $\boldsymbol{\mu}_i \in \mathbb{R}^3$, whereas the editable variables are the SH coefficients. We do not identify these two spaces and do not map a 3D position vector into an SH vector. Instead, the frozen positions index a graph on which the SH coefficients form a vector-valued signal. The RPD construction lifts the base geometry to this coefficient field in two steps.

First, the base metric produces an edge cost $c_{ij}^{\mathrm{base}} = d_g^2(i,j)$ and its monotone conductance

$$W_{ij} = \exp\left(-\frac{c_{ij}^{\mathrm{base}}}{2\sigma_g^2}\right), \qquad P_{ij} = \frac{W_{ij}}{\sum_k W_{ik}}. \tag{20}$$

Thus, carriers that are close under the photometric position metric have large conductance, whereas geometrically or photometrically separated carriers interact weakly. Second, the SH-order map in Eq. 18 induces the local coefficient metric

$$\mathbf{M}_i = \mathbf{D}_i^\top \mathbf{D}_i = \mathrm{diag}\big(s_i^{-2}\mathbf{I}_{d_0}, \mathbf{I}_{d_+}\big). \tag{21}$$

For an SH-space flux/displacement $\mathbf{q}_{ij}$ assigned to edge (i,j), a corresponding local transport cost is therefore

$$c_{ij}^{\mathrm{SH}}(\mathbf{q}_{ij}) = d_g^2(i,j)\, \mathbf{q}_{ij}^\top \frac{\mathbf{M}_i + \mathbf{M}_j}{2} \mathbf{q}_{ij}. \tag{22}$$

This equation makes the dimensional roles explicit: d_g^2 determines *where* transport is expensive on the carrier graph, while $\mathbf{M}_i$ determines *which SH directions* are expensive at a carrier.

The implementation uses the smooth conductance form rather than explicitly solving a global edge-flux optimization. If $\mathbf{X} = [\mathbf{x}_i]_{i\in\Omega_e}$ denotes the RPD-calibrated SH state, the graph term in ORF acts channel-wise as

$$(L_g\mathbf{X})_i = \sum_j P_{ij}(\mathbf{x}_j - \mathbf{x}_i). \tag{23}$$

Equivalently, for a raw-SH velocity field $\mathbf{U} = [\mathbf{u}_i]$, the implemented geometry is compatible with the positive quadratic action

$$\mathcal{C}_{\mathrm{RPD}}(\mathbf{U}) = \sum_i \|\mathbf{D}_i\mathbf{u}_i\|_2^2 + \frac{\lambda_g}{2}\sum_{i,j} W_{ij}\|\mathbf{D}_i\mathbf{u}_i - \mathbf{D}_j\mathbf{u}_j\|_2^2. \tag{24}$$

We use Eq. 24 as an interpretation of the metric induced by the implemented $\mathbf{D}_i$ and L_g, not as an additional optimization loss. In particular, $\boldsymbol{\mu}_i$ remains fixed throughout watermarking: the positional geometry only parameterizes the cost and coupling of edits made in the SH fibers.

B ORF INTERPRETATION, ARCHITECTURE, AND EXPLICIT INVERSE

In the Riemannian carrier geometry, $W_{2,g}^2$ can be interpreted as the minimum accumulated metric kinetic energy under the continuity equation. For a free energy consisting of potential and entropy terms, the associated gradient flow has the Fokker–Planck form $\partial_t\rho = \nabla_g\cdot(\rho\nabla_g V) + \kappa\Delta_g\rho$. Related manifold diffusion and transport constructions motivate this geometric reading (Huang et al., 2022; De Bortoli et al., 2022; Chen & Lipman, 2024; Tong et al., 2024). EndoRDH does not identify SH coefficients with a physical density and does not claim to solve the globally optimal transport problem. Instead, this structure is used as a geometric regularizer: the RPD metric defines locally expensive and inexpensive directions, while the learnable residual adapts the reversible dynamics to the observed carrier statistics.

In the implementation, the potential and branch state are concatenated and processed by a pointwise 1×1 convolution, GELU, and a second 1×1 convolution. The reversible backbone contains three additive-coupling blocks with step size $h = 1/3$. For one block, the inverse of Eq. 10 is

$$\begin{aligned}\mathbf{X}_A &= \mathbf{X}'_A - h g_\theta(\mathbf{X}'_B, V_B),\\ \mathbf{X}_B &= \mathbf{X}'_B - h f_\theta(\mathbf{X}_A, V_A).\end{aligned} \tag{25}$$

The full inverse evaluates these blocks in the reverse order, using exactly the same learned functions and branch-conditioned potentials as in the forward pass.

B.1 WHY THE KEYED MESSAGE OPERATOR $T_{\mathbf{m}}$ IS INVERTIBLE

For clarity, let $\mathbf{z} = \mathrm{vec}(F_\theta(\Phi_{\mathrm{RPD}}(\mathbf{a}_{\Omega_e}))) \in \mathbb{R}^D$ and let $\mathbf{c}(\mathbf{m}) \in \{-1,+1\}^Q$ denote the ECC-coded message. The secret key $\mathbf{k}$ determines the coordinate-to-symbol assignment $\pi_{\mathbf{k}}$ and the DBRC branch of each assigned coordinate. For an assigned coordinate ℓ, the implementation uses the signed codeword offset

$$\Delta_{\mathbf{m},\mathbf{k},\ell} = \delta\, c_{\pi_{\mathbf{k}}(\ell)}(\mathbf{m}), \qquad \delta = 0.15, \tag{26}$$

and sets $\Delta_{\mathbf{m},\mathbf{k},\ell} = 0$ for unassigned coordinates. The reversible message operator is the translation

$$T_{\mathbf{m},\mathbf{k}}(\mathbf{z}) = \mathbf{z} + \boldsymbol{\Delta}_{\mathbf{m},\mathbf{k}}, \qquad T^{-1}_{\mathbf{m},\mathbf{k}}(\mathbf{y}) = \mathbf{y} - \boldsymbol{\Delta}_{\mathbf{m},\mathbf{k}}. \tag{27}$$

We suppress $\mathbf{k}$ in the main paper and write $T_{\mathbf{m}}$. For fixed $(\mathbf{m},\mathbf{k})$, Eq. 27 is a bijection on $\mathbb{R}^D$: its Jacobian is the identity and its determinant is one. Boundedness refers to the fixed embedding amplitude, $\|\boldsymbol{\Delta}_{\mathbf{m},\mathbf{k}}\|_\infty \le \delta$, rather than to clipping the output state.

This distinction is necessary for reversibility. The branch medians p_b, target centers $p_b \pm \delta$, and thresholds used by DBRC are *decoding references and soft training targets*. The forward reversible path does not overwrite a coordinate by $p_b \pm \delta$, round it, or clip it. Such a hard projection would be many-to-one and would destroy exact invertibility. The 3D message loss instead encourages the translated coordinates to have sufficient decoding margin around these centers while preserving the residual carrier state. Authorized recovery subtracts the exact signed offset in Eq. 26 once the key and codeword are known. Therefore, exact algebraic recovery is conditional on the correct key and embedded codeword, either retained by the authorized party or decoded without error. After an attack that changes the decoded codeword, recovery is necessarily approximate.

B.2 AUTHORIZED RECOVERY AND ITS IDEAL LIMIT

Let $\mathbf{a}^h_{\Omega_e}$ denote the editable host SH coefficients and let $\mathbf{a}^w_{\Omega_e}$ be the output of Eq. 2. With the correct user key and the embedded codeword (or an error-free decoding of it), the carrier assignment and displacement used by $T_{\mathbf{m}}$ are reproduced exactly. Applying the forward coordinate maps to the watermarked state isolates this keyed translation, after which the inverse message operator and inverse coordinate maps give

$$\begin{aligned} \mathbf{z}^w &= F_\theta\big(\Phi_{\mathrm{RPD}}(\mathbf{a}^w_{\Omega_e})\big) = T_{\mathbf{m}}\big(F_\theta(\Phi_{\mathrm{RPD}}(\mathbf{a}^h_{\Omega_e}))\big)\,, \\ \mathbf{z}^r &= T^{-1}_{\mathbf{m}}(\mathbf{z}^w) = F_\theta\big(\Phi_{\mathrm{RPD}}(\mathbf{a}^h_{\Omega_e})\big)\,, \\ \mathbf{a}^r_{\Omega_e} &= \Phi^{-1}_{\mathrm{RPD}}\big(F^{-1}_\theta(\mathbf{z}^r)\big) = \mathbf{a}^h_{\Omega_e}. \end{aligned} \tag{28}$$

CACP keeps $\mathbf{a}_{\Omega_f}$ and all structural parameters unchanged throughout embedding, so the same identity extends to the complete host parameter set. If rendering is deterministic and the watermarked model has not undergone an irreversible post-embedding modification, exact arithmetic therefore gives identical host and recovered renderings. In that ideal case,

$$\mathrm{MSE}(I^r, I^{\mathrm{host}}) = 0, \qquad \mathrm{PSNR}(I^r, I^{\mathrm{host}}) \to +\infty, \qquad \mathrm{SSIM}(I^r, I^{\mathrm{host}}) \to 1. \tag{29}$$

This statement characterizes the structural limit of the inverse composition. Finite-precision arithmetic, quantization, imperfect estimation of a perturbed message displacement, or attacks that irreversibly modify the watermarked model can break exact equality and yield finite empirical recovery scores. Such deviations should not be interpreted as the theoretical ceiling of the reversible map.

Figure 5 in the main paper provides the qualitative counterpart to this inverse derivation. It should be read together with Eq. 28: exact identity is a property of the ideal keyed composition, whereas finite numerical or post-embedding perturbations can yield residual error.

C DBRC MESSAGE ORGANIZATION, DECISIONS, AND REDUNDANCY ANALYSIS

The default configuration uses a 32-bit copyright message and 4 Reed–Solomon ECC symbols. Each chunk uses $K = 16$ carriers and is repeated $M = 7$ times in each branch. For branch b, bit states are represented around a saved median p_b using centers $p_b \pm \delta$, with $\delta = 0.15$. Model-domain extraction performs carrier-level decisions, within-group majority voting, repetition fusion, cross-branch fusion, a three-point median filter, and Reed–Solomon decoding. Authentication succeeds when raw-bit decoding accuracy is at least 0.80.

If one bit is represented by K weakly independent carrier decisions and the per-carrier error rate is $p < 1/2$, majority voting satisfies

$$\Pr(\hat{s}_k \neq s_k) \leq \exp\left[-2K\left(\frac{1}{2} - p\right)^2\right]. \tag{30}$$

Real attacks typically induce correlated errors, so this bound only motivates the design principle that multiple weak pieces of evidence can be preferable to a single strongly perturbed carrier. It is not used as a quantitative predictor of measured accuracy.

For authorized recovery, $p_b \pm \delta/2$ is used only to infer the coded symbol when the codeword is not already supplied. The host is not recovered by snapping a coordinate back to p_b. Instead, after symbol inference, the exact signed offset $\Delta_{\mathbf{m},\mathbf{k},\ell}$ in Eq. 26 is subtracted coordinate-wise before applying inverse ORF coupling and the inverse RPD transform. Embedding, model-domain verification, and authorized recovery therefore share the same branch indices, medians, and codeword organization while preserving the bijective translation required by Eq. 27.

D IMPLEMENTATION PARAMETERS, TRAINING OBJECTIVE, AND ATTACK CONFIGURATION

Controlled migration of external baselines. For 3D-GSW (Jang et al., 2025) and GS-Marker (Li et al., 2025), we use the corresponding papers and released code as the specification of the watermarking procedure, but replace their original reconstruction host with the same pretrained EndoGS

model used by our variants. Only the adaptations needed to interface each watermark mechanism with EndoGS parameters and rendering are introduced. The intended embedding and extraction logic of each baseline is retained. Consequently, sequence inputs, host reconstruction, ROI masks, rendering, and metric computation are common across methods. This protocol is designed to reduce confounding from different reconstruction systems, and the resulting comparison is explicitly limited to the transferred methods on EndoGS rather than claiming a verbatim reproduction of each paper's original benchmark.

CACP and carrier branches. When capacity is sufficient, the implementation retains non-frozen carriers whose sigmoid opacity is above 0.8 and whose scale volume satisfies $\prod_d \exp(s_{i,d}) > 10^{-6}$. If too few carriers remain, it falls back to the non-frozen candidates with the highest opacity. Candidates are then sorted by distance to the scene center and alternately assigned to complementary A/B branches. Geometry, scale, rotation, and opacity remain frozen during watermarking, and SH gradients corresponding to protected regions are set to zero.

Optimization and network. Experiments initialize from the provided 60k-iteration reconstruction model and use Adam with learning rate 10^{-4} to optimize the reversible flow and a four-layer convolutional image decoder. ORF uses the three coupling blocks described in Appendix B. During training, SH noise in $[0.001, 0.005]$ is applied with probability 0.5, and quantization with step size 0.001, 0.005, or 0.01 is applied with probability 0.4.

Exact loss weights. The base reconstruction objective over the tissue mask is $\mathcal{L}_{\text{base}} = \mathcal{L}_1 + 0.5\mathcal{L}_{\text{depth}} + 0.01\mathcal{L}_{\text{TV}}$. During the fine stage, Eq. 15 is instantiated as

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{base}} + 0.1q\,\mathcal{L}_{2D} + 0.05\,\mathcal{L}_{\text{sem}} + 0.1\,\mathcal{L}_{\text{endo}} \\ + 0.05\,\mathcal{L}_{\text{inv}} + (0.1 + 0.9q)\mathcal{L}_{3D},\end{aligned} \tag{31}$$

where $q \in [0, 1]$ denotes fine-stage training progress. $\mathcal{L}_{2D}$ is binary cross entropy after differentiable 2D attacks, $\mathcal{L}_{\text{sem}}$ is a masked semantic-feature distance, $\mathcal{L}_{\text{endo}}$ reweights decoding according to average depth and image variance, $\mathcal{L}_{\text{inv}}$ regularizes near-identity behavior and inverse recovery, and $\mathcal{L}_{3D}$ pulls transport-coordinate carriers toward the corresponding message states.

Attack configuration. Rendering-domain decoder training includes differentiable perturbations such as Gaussian noise, blur, erasing, brightness/contrast variation, and resizing. Model-domain stress tests include SH noise, primitive removal, SH quantization, compression, and 3D cropping. The main paper reports clean and attack-averaged results. Complete per-attack results are listed in Appendix I.

E SUPPLEMENTARY VISUAL COMPARISONS

The aggregate ROI metrics in Table 1 summarize rendering fidelity, but they do not show how residual differences are distributed across individual frames. Figures 3 and 4 provide frame-level render/residual panels under the shared EndoGS evaluation protocol. The panels are spatial diagnostics; method comparisons and robustness claims continue to rely on the reported PSNR, SSIM, LPIPS, 2DACC, and 3DACC values.

Figure 3 compares the remaining selected *cutting* views after the examples used in Figure 2. Each view is shown first as a rendering row and then as a matched residual row. The residual intensity is amplified only for visibility.

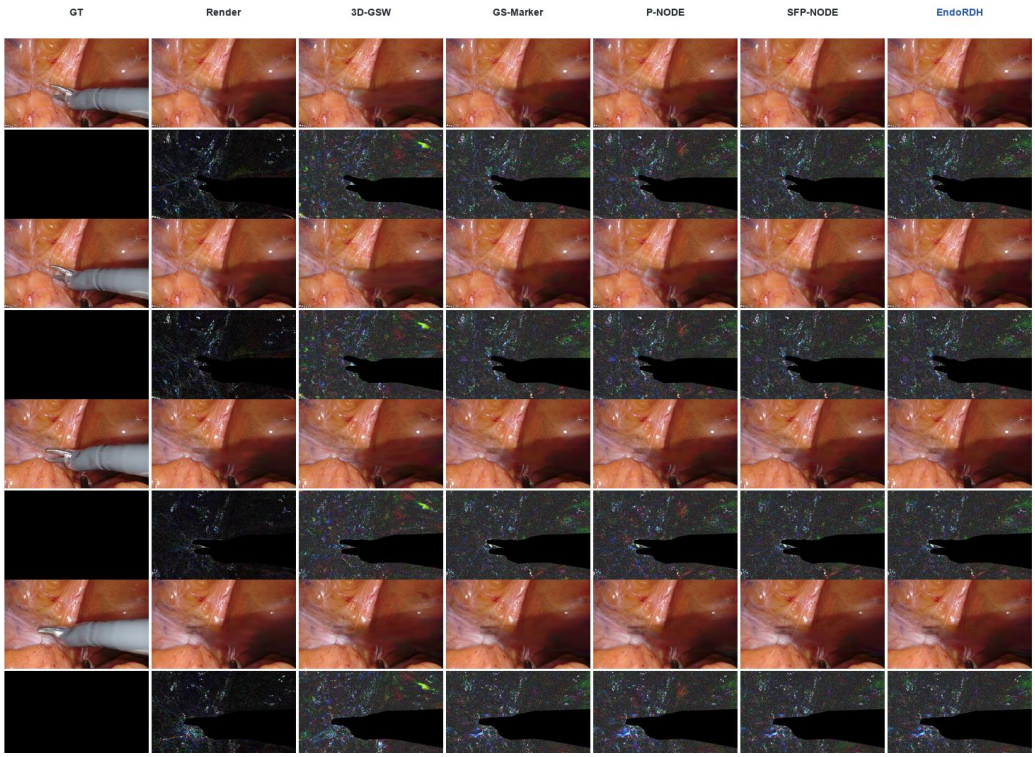


Figure 3: Cutting render/residual comparison over the remaining selected views. Columns follow the order GT, host render, 3D-GSW, GS-Marker, P-NODE, SFP-NODE, and EndoRDH. For each view, the upper row contains renderings and the lower row contains absolute RGB residuals amplified by $\times 10$ within the tissue ROI.

Figure 4 provides the corresponding comparison on *pulling* after excluding the examples shown in Figure 2. The second sequence changes tissue configuration and illumination, so this panel is a separate visual check rather than an independent quantitative evaluation.

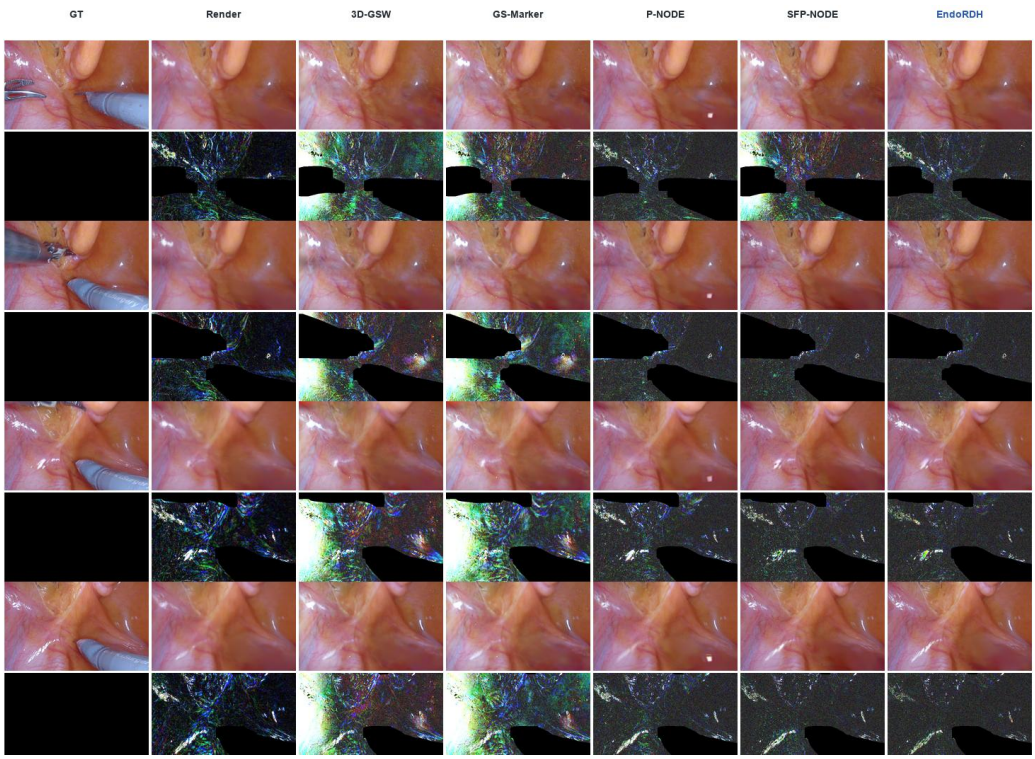


Figure 4: Pulling render/residual comparison over four selected views. Columns follow the order GT, host render, 3D-GSW, GS-Marker, P-NODE, SFP-NODE, and EndoRDH. For each view, the upper row contains renderings and the lower row contains absolute RGB residuals amplified by $\times 10$ within the tissue ROI.

Figure 5 shows authorized recovery for EndoRDH alone in a compact four-row grid. The rows are ordered as GT, host, watermarked, and recovered; the first four columns use the selected cutting views and the last four use the selected pulling views. It is intentionally separated from the cross-method render/residual panels so that recovery is read as an inverse-path diagnostic rather than as another baseline comparison.

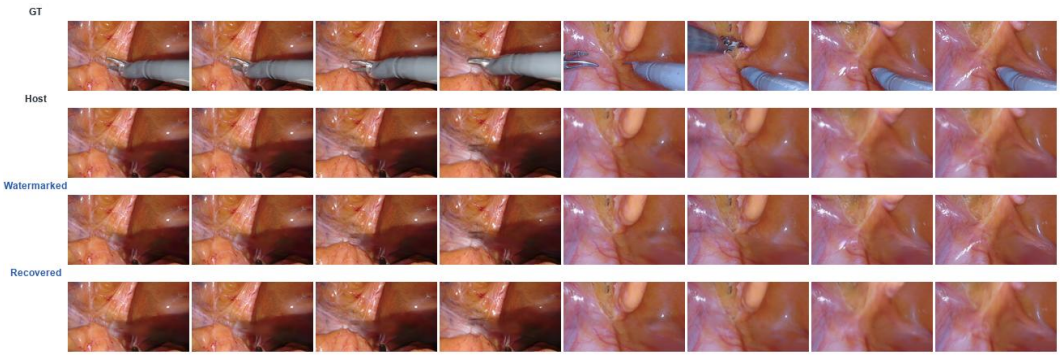


Figure 5: EndoRDH-only authorized recovery over eight selected views. Rows show GT, the unwatermarked host render, the EndoRDH watermarked render, and keyed recovery, respectively. The first four columns are cutting views and the last four are pulling views.

F SCOPE, ASSUMPTIONS, AND LIMITATIONS

The current evaluation covers only the provided *cutting* and *pulling* sequences and their tissue-ROI definitions. Different endoscope optics, tissue appearance, camera trajectories, or mask quality may change the observed performance. The photometric potential in RPD is a monotonic spatial prior rather than a calibrated radiometric model, and CACP enforces the supplied protected-region policy rather than certifying downstream clinical safety. The exact recovery identity holds for the uncorrupted watermarked state under the correct key and exact arithmetic. Practical recovery can deviate because of quantization, finite precision, imperfect displacement estimation, or irreversible post-embedding model drift. Future evaluation should test additional endoscope systems and directly measure whether downstream lesion analysis, navigation, or measurement performance is preserved before and after watermarking.

G SUPPLEMENTARY MODULE ABLATION

Table 3 preserves the available ablation summary while keeping its interpretation separate from the main benchmark. The records contain only two interventions: jointly removing CACP/RPD/ORF and removing DBRC. Consequently, the first row supports the importance of the constrained reversible-transport chain as a whole, but it does not permit independent numerical attribution to CACP, RPD, or ORF. The second intervention directly probes redundant evidence aggregation and shows a substantially larger verification drop when DBRC is removed. CACP additionally serves a constraint-enforcement role that is not fully reflected by a scalar fidelity or accuracy metric. Because these values originate from a separate ablation summary, the fidelity entries are used for within-table comparison only and should not be cross-calibrated in absolute value with Table 1.

Table 3: Supplementary ablation results from the provided summary. The joint CACP/RPD/ORF intervention is not used for factorized component attribution.

Configuration	PSNR↑	SSIM↑	LPIPS↓	2DACC↑	3DACC↑
w/o CACP/RPD/ORF	30.63	0.9121	0.0932	98.49	82.51
w/o DBRC	33.61	0.9327	0.0336	78.72	62.93
EndoRDH	34.29	0.9663	0.0204	98.47	86.20

H PAYLOAD SCALING

The payload sweep probes whether increasing message capacity changes the balance between rendering fidelity and verification reliability. These runs were produced as a separate controlled sweep and are therefore interpreted by within-sweep trends rather than by absolute comparison with identically named configurations in Table 1. Figure 6 plots the three provided payloads for P-NODE, SFP-NODE, and EndoRDH.

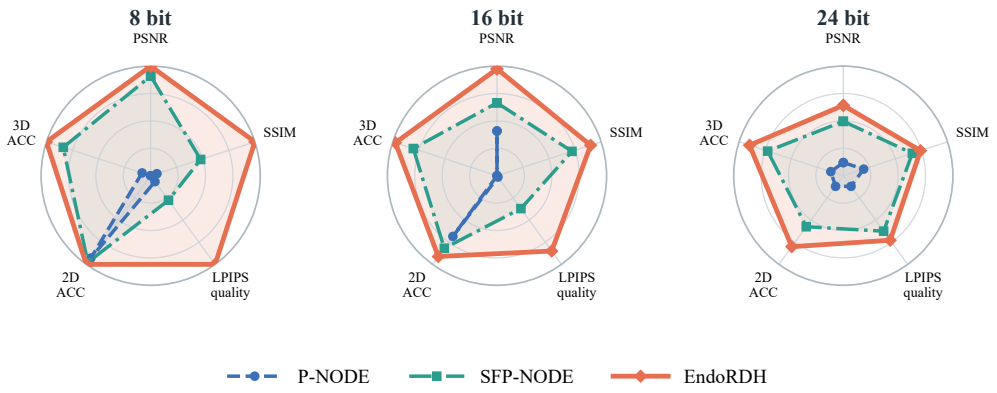


Figure 6: Payload scaling of P-NODE, SFP-NODE, and EndoRDH at 8, 16, and 24 bits. The curves visualize the within-sweep balance between fidelity and verification as capacity increases.

From 8 to 24 bits, all variants exhibit the expected pressure on verification as more information is encoded, but EndoRDH shows the smallest reported decrease in 3DACC over this range. The trend is consistent with using RPD to avoid disproportionately expensive carriers and DBRC to distribute evidence, although the sweep does not by itself isolate these mechanisms individually.

I PER-ATTACK ROBUSTNESS RESULTS

The attack averages in Table 2 summarize heterogeneous failure modes, so the attack-specific values are retained here to make the comparison auditable. Table 4 reports rendering-domain behavior under noise, blur, JPEG compression, resizing, cropping, and brightness changes.

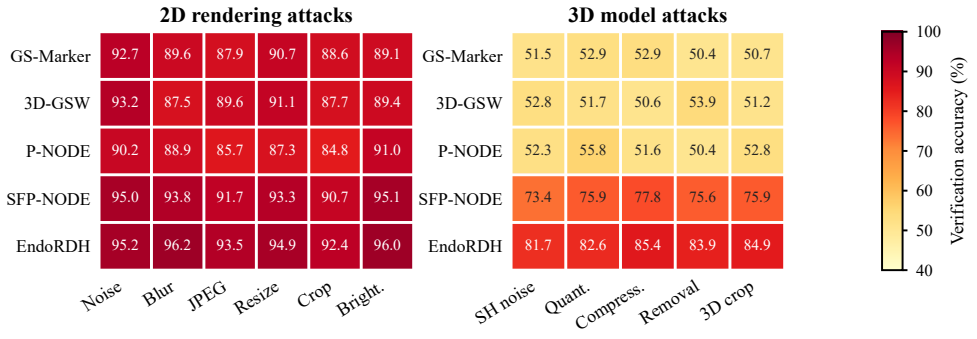


Figure 7: Attack-specific verification accuracy for rendering-domain and model-domain perturbations. The heatmap visualizes the values tabulated in Tables 4 and 5; Table 2 reports their main-paper averages.

Table 4: Rendering-domain verification accuracy under 2D attacks (%).

Method	Clean	Noise	Blur	JPEG	Resize	Crop	Brightness
GS-Marker	97.46	92.73	89.58	87.92	90.67	88.62	89.11
3D-GSW	97.99	93.21	87.47	89.56	91.14	87.73	89.37
P-NODE	95.28	90.16	88.93	85.74	87.26	84.81	91.02
SFP-NODE	97.16	95.03	93.84	91.72	93.27	90.66	95.11
EndoRDH	98.47	95.21	96.18	93.45	94.86	92.37	96.04

Several methods already decode reliably from clean renderings, so the rendering-domain gaps are comparatively modest. Model-domain perturbations are more discriminative because they directly alter or remove the SH carriers. Table 5 reports SH noise, quantization, compression, primitive removal, and 3D cropping. These values are the same ones visualized in the right heatmap of Figure 7.

Table 5: Model-domain verification accuracy under 3D attacks (%).

Method	Clean	SH noise	Quant.	Compression	Removal	3D crop
GS-Marker	55.39	51.47	52.86	52.92	50.36	50.71
3D-GSW	56.03	52.84	51.73	50.61	53.92	51.18
P-NODE	55.41	52.26	55.83	51.62	50.42	52.84
SFP-NODE	80.42	73.36	75.91	77.83	75.64	75.92
EndoRDH	86.20	81.74	82.57	85.36	83.93	84.86

Across the provided model-domain attacks, EndoRDH retains the highest reported verification accuracy in each column. The stronger separation under direct carrier corruption is consistent with calibrated carrier geometry and redundant evidence, while the controlled-transfer protocol limits this conclusion to the common EndoGS host.