COSP: Consumer Oriented Second-order Pruning for Compact Implicit Neural Representations

Anonymous authors
Paper under double-blind review

Abstract

Implicit neural representations (INRs) encode a signal as a coordinate network. We study post-training structured compaction, which requires identifying hidden structures that can be removed with little functional damage. This is difficult because deleting a neuron changes adjacent affine tensors, whereas conventional group scores need not match the underlying functional perturbation and globally comparable second-order scores are costly. We propose COSP (Consumer Oriented Second-order Pruning) for structured INR compaction. COSP derives a module-wise second-order back-propagation rule and its diagonal propagation variant, which evaluate candidate groups without materializing a global parameter Hessian. It then defines a consumer-side group that captures the functional perturbation of removing its associated full dependency group, selects groups globally under a physical parameter budget, and applies ridge reconstruction after selection. We validate COSP through comprehensive experiments: (1) verifying its second-order derivative propagation and consumer-side saliency, (2) comparing with established pruning baselines, and (3) ablation studies on key design choices and computational cost. COSP achieves the best mean decoded quality for every evaluated modality and parameter reduction, and ranks first in 125 of 132 matched evaluation settings.

1 Introduction

Implicit neural representations (INRs) encode a signal as a continuous coordinate function, mapping spatial or spatiotemporal coordinates to signal values rather than storing a discrete sample grid. Recent surveys document their use across visual, geometric, medical-imaging, and broader signal-processing settings (Essakine et al., 2024; Jayasundara & Patel, 2026; Molaei et al., 2023; Zhu et al., 2025). INRs have also been used for video representation (Chen et al., 2021; 2023; Kayabasi et al., 2025), shape modeling (Sitzmann et al., 2020; De Luigi et al., 2023), and neural compression (Dupont et al., 2021). For a signal-specific INR, the trained coordinate network is itself the stored representation, making physical compaction important. We study one-shot structured compaction of chain-structured coordinate MLPs by selecting a structured mask and exporting a compact dense network.

Prior INR research pursues compact representations through signal-specific coding and scalable coordinate designs (Dupont et al., 2021; Girish et al., 2023; Li et al., 2024), sparse or adaptive INRs (Lee et al., 2021b; Jayasundara et al., 2025; Aldana et al., 2025), and video-specific representations (Chen et al., 2021; 2023; Kayabasi et al., 2025). These approaches improve the representation, coding scheme, or training procedure. Once a dense INR has been trained, further compaction requires determining which structures can be removed while retaining the information essential to the represented signal.

Traditional structured pruning removes complete neurons or channels. Its saliency criteria can be organized by the information used to estimate importance. Zero-order magnitude methods use parameter scale as a removable proxy (Han et al., 2015; Li et al., 2017; Lee et al., 2021a). First-order methods use gradients or first-order Taylor sensitivity (Lee et al.,

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.

2019; Wang et al., 2020; Tanaka et al., 2020). Second-order methods incorporate curvature through Hessian or Fisher approximations (LeCun et al., 1990; Liu et al., 2021; Kwon et al., 2022; Yu et al., 2022; Singh & Alistarh, 2020; Frantar & Alistarh, 2022). Activation-aware and structural methods use activation statistics or structural dependencies to estimate removable units (Liu et al., 2017; Lin et al., 2020; Ma et al., 2023; Fang et al., 2023).

For frozen INR compaction, a useful saliency should predict the loss change caused by deleting a candidate structure. Magnitude and activation statistics are indirect proxies, and first-order criteria omit curvature. Second-order methods often evaluate curvature locally within individual layers or modules, whereas directly accounting for global parameter interactions is costly. Evaluating second-order saliency for all candidate groups in a dense INR therefore remains challenging.

We propose COSP (Consumer Oriented Second-order Pruning) for this setting. COSP first derives a shared module-wise second-order reverse traversal and a linear-state diagonal approximation that score candidate groups without materializing a global parameter Hessian. It then defines a consumer-side scoring group that captures the functional perturbation of removing its associated full physical dependency group, selects low-damage groups globally under the physical parameter budget, and applies ridge reconstruction.

Controlled consistency experiments validate the propagated derivative states and show that consumer-side saliency predicts actual single-group removal damage. We evaluate real compact exports on images, biomedical volumes, signed-distance fields, and videos at 10%–70% physical parameter reduction. COSP achieves the highest mean decoded quality against the matched one-shot baselines on every modality and tested reduction, while preserving recognizable geometry, texture, and signal transitions at aggressive reductions. It ranks first among the seven evaluated methods in 125 of 132 matched settings.

Our main contributions are summarized as follows:

- We derive a module-wise second-order reverse-propagation rule that propagates shared first- and second-order states in one traversal for all candidate groups, together with a recursive diagonal approximation that avoids constructing a global parameter Hessian.
- We show that the consumer-side perturbation is functionally equivalent to removing a physical dependency group, and use it to align structured saliency evaluation with the functional effect of physical deletion.
- We globally rank dependency groups under the physical parameter budget for one-shot compaction of a frozen INR, export the resulting compact dense network, and apply fixed-mask ridge reconstruction.
- We verify the propagated derivatives and consumer-side saliency against exact perturbations, then evaluate COSP on images, biomedical volumes, SDFs, and videos. COSP achieves the best mean decoded quality for every task and tested reduction, and ranks first among seven methods in 125 of 132 matched evaluation settings.

2 Method

2.1 Pruning Model

As shown in Figure 1(a), pruning is formulated as a constrained functional optimization over structured parameter transformations. The dense INR is denoted by $f(\mathbf{x};\boldsymbol{\theta})$. Applying a structured pruning operator gives the pruned parameter set $\widehat{\boldsymbol{\theta}} = \mathcal{P}(\boldsymbol{\theta})$. Thus $\Delta\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ is the functional perturbation induced by the mask.

The goal is to remove groups while preserving the original network function on a calibration distribution $\mathcal{D}$. We measure the damage caused by a candidate pruning perturbation using

$$\mathcal{L}(\widehat{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[d\Big(f(\mathbf{x};\widehat{\boldsymbol{\theta}}), f(\mathbf{x};\boldsymbol{\theta})\Big)\right], \tag{1}$$

where $d(\cdot,\cdot)$ is an output-space discrepancy. Given a target physical parameter reduction ρ, the ideal pruning problem is

$$\min_{\widehat{\boldsymbol{\theta}}} \ \mathcal{L}(\widehat{\boldsymbol{\theta}}) \qquad \text{s.t.} \qquad \|\widehat{\boldsymbol{\theta}}\|_0 \le (1-\rho)\|\boldsymbol{\theta}\|_0, \tag{2}$$

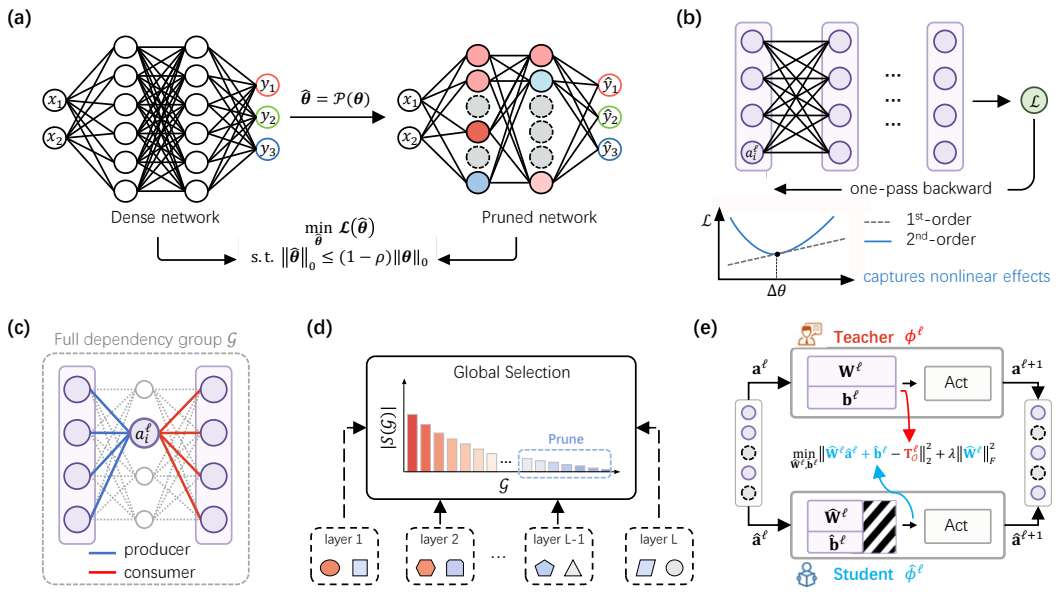

Figure 1: Overview of COSP. **(a)** Pruning transforms a trained dense INR into a compact model with minimal functional damage under a parameter budget. **(b)** A shared one-pass second-order back-propagation carries first- and second-order states from the output loss through the frozen INR. **(c)** Consumer-side scoring evaluates the consumer-side group $\mathcal{G}_{\text{con}}$, whereas physical compaction removes the full dependency group $\mathcal{G}$. **(d)** Candidates from all layers are globally ranked by predicted damage under the physical parameter budget. **(e)** With the mask fixed, closed-form ridge reconstruction fits each compact affine mapping to targets from the dense model while preserving the compact architecture.

where $\|\cdot\|_0$ counts nonzero parameters. Directly solving Eq. equation 2 is combinatorial. COSP therefore estimates the functional damage of each candidate group and removes groups with small predicted damage.

2.2 Second-order Taylor Estimation and Back-propagation

To estimate the functional damage of a candidate perturbation, we expand the loss in Eq. equation 2 at the dense solution:

$$\mathcal{L}(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}) \approx \mathcal{L}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\mathcal{L}^{\mathsf{T}}\Delta\boldsymbol{\theta} + \frac{1}{2}\Delta\boldsymbol{\theta}^{\mathsf{T}}\nabla^2_{\boldsymbol{\theta}}\mathcal{L}\,\Delta\boldsymbol{\theta}. \tag{3}$$

As illustrated in Figure 1(b), a shared layerwise back-propagation carries first- and second-order states from the output loss through the frozen INR. This avoids constructing a Hessian over all parameters while making the same local derivative state available to every candidate group at a layer. The challenge is therefore to obtain the local first- and second-order derivatives efficiently. At layer ℓ, the loss can be viewed as a composition function

$$\mathcal{L} = \psi^\ell(\mathbf{a}^{\ell+1}) = (\psi^\ell \circ \phi^\ell)(\mathbf{a}^\ell), \tag{4}$$

where $\mathbf{a}^\ell$ is the input to layer ℓ, ϕ^ℓ is the layer mapping, and ψ^ℓ contains the remaining network and the scalar calibration loss. Applying the scalar chain rule gives, for input coordinates i, j and output coordinates s, t,

$$\frac{\partial \mathcal{L}}{\partial a_i^\ell} = \sum_s \frac{\partial a_s^{\ell+1}}{\partial a_i^\ell}\frac{\partial \mathcal{L}}{\partial a_s^{\ell+1}}, \tag{5}$$

$$\frac{\partial^2 \mathcal{L}}{\partial a_i^\ell \partial a_j^\ell} = \sum_s \frac{\partial^2 a_s^{\ell+1}}{\partial a_i^\ell \partial a_j^\ell}\frac{\partial \mathcal{L}}{\partial a_s^{\ell+1}} + \sum_s\sum_t \frac{\partial a_s^{\ell+1}}{\partial a_i^\ell}\frac{\partial^2 \mathcal{L}}{\partial a_s^{\ell+1}\partial a_t^{\ell+1}}\frac{\partial a_t^{\ell+1}}{\partial a_j^\ell}. \tag{6}$$

The second term in Eq. equation 6 contains both diagonal curvature terms ($s = t$) and cross-state terms ($s \neq t$). To write Eqs. equation 5 and equation 6 compactly, define $\mathbf{g}^\ell = \nabla_{\mathbf{a}^\ell}\mathcal{L}$ and $\mathbf{H}^\ell = \nabla^2_{\mathbf{a}^\ell}\mathcal{L}$ as the loss gradient and Hessian at the layer boundary. For the layer mapping, define $\mathbf{J}^\ell = \nabla_{\mathbf{a}^\ell}\mathbf{a}^{\ell+1}$ and $\mathbf{K}^\ell_s = \nabla^2_{\mathbf{a}^\ell} a^{\ell+1}_s$ as its Jacobian and the local Hessian of its sth output, respectively. The matrix form is

$$\mathbf{g}^\ell = (\mathbf{J}^\ell)^\top \mathbf{g}^{\ell+1}, \tag{7}$$

$$\mathbf{H}^\ell = (\mathbf{J}^\ell)^\top \mathbf{H}^{\ell+1}\mathbf{J}^\ell + \sum_s g^{\ell+1}_s \mathbf{K}^\ell_s. \tag{8}$$

Equation equation 8 is the full second-order back-propagation rule: it retains every mixed partial derivative while propagating only a local hidden-state Hessian, rather than a Hessian over all network parameters.

To avoid propagating the quadratic-size matrix $\mathbf{H}^\ell$, COSP maintains only a diagonal-curvature state $\mathbf{h}^\ell$. Taking the diagonal part of Eq. equation 6 while retaining only diagonal downstream curvature gives

$$\mathbf{h}^\ell = \left(\mathbf{J}^\ell \odot \mathbf{J}^\ell\right)^\top \mathbf{h}^{\ell+1} + \sum_s g^{\ell+1}_s \operatorname{diag}(\mathbf{K}^\ell_s). \tag{9}$$

Thus, diagonal propagation carries one curvature value per hidden coordinate instead of a dense hidden-state Hessian. Since the approximation is applied at every layer, $\mathbf{h}^\ell$ is generally not the exact diagonal of the Full state. We use it as the default scoring mode.

We use Full for the dense local hidden-state curvature propagated by Eq. equation 8, Exact-Diag for its principal diagonal, and Diag for the linear-state approximation in Eq. equation 9. Full is verified against nested automatic differentiation, while Diag is evaluated against Exact-Diag as a low-cost approximation.

2.3 Consumer-side Saliency

Figure 1(c) distinguishes the consumer-side scoring group from the full physical dependency group removed to export a compact MLP. For hidden activation a^ℓ_i, the full dependency group and the consumer-side scoring group are

$$\mathcal{G}^\ell_i = \left\{W^{\ell-1}_{i,:},\, b^{\ell-1}_i,\, W^\ell_{:,i}\right\}, \qquad \mathcal{G}^\ell_{i,\mathrm{con}} = \left\{W^\ell_{:,i}\right\}. \tag{10}$$

The first two terms are producer-side parameters that create a^ℓ_i, while $W^\ell_{:,i}$ constitutes the consumer-side group that transmits it to the following affine mapping.

Removing a^ℓ_i is functionally equivalent to zeroing its consumer-side group. With $\Delta W^\ell_{:,i} = -W^\ell_{:,i}$, this induces the minimal affine perturbation $-a^\ell_i W^\ell_{:,i}$ at the next layer. Once $W^\ell_{:,i}$ is zero, a^ℓ_i cannot affect any subsequent computation, so removing $W^{\ell-1}_{i,:}$ and $b^{\ell-1}_i$ adds no functional perturbation and only realizes the same deletion as a compact dense MLP.

We assign the physical dependency group a second-order score by evaluating only its consumer-side perturbation in the Taylor model:

$$\mathcal{S}(\mathcal{G}^\ell_i) = \left(\nabla_{W^\ell_{:,i}}\mathcal{L}\right)^\top \Delta W^\ell_{:,i} + \frac{1}{2}(\Delta W^\ell_{:,i})^\top \nabla^2_{W^\ell_{:,i}}\mathcal{L}\, \Delta W^\ell_{:,i}. \tag{11}$$

The default diagonal score retains only the diagonal of the Hessian restricted to the consumer-side group:

$$\mathcal{S}_{\mathrm{Diag}}(\mathcal{G}^\ell_i) = \left(\nabla_{W^\ell_{:,i}}\mathcal{L}\right)^\top \Delta W^\ell_{:,i} + \frac{1}{2}\operatorname{diag}\left(\nabla^2_{W^\ell_{:,i}}\mathcal{L}\right)^\top \left(\Delta W^\ell_{:,i} \odot \Delta W^\ell_{:,i}\right). \tag{12}$$

The propagated states allow evaluating these consumer-side derivatives for every input activation of a layer in a single shared back-propagation, avoiding redundant producer-side perturbations that add no functional damage. Figure 2(c) shows that this consumer-side score predicts exact deletion damage more faithfully than scoring the full dependency group.

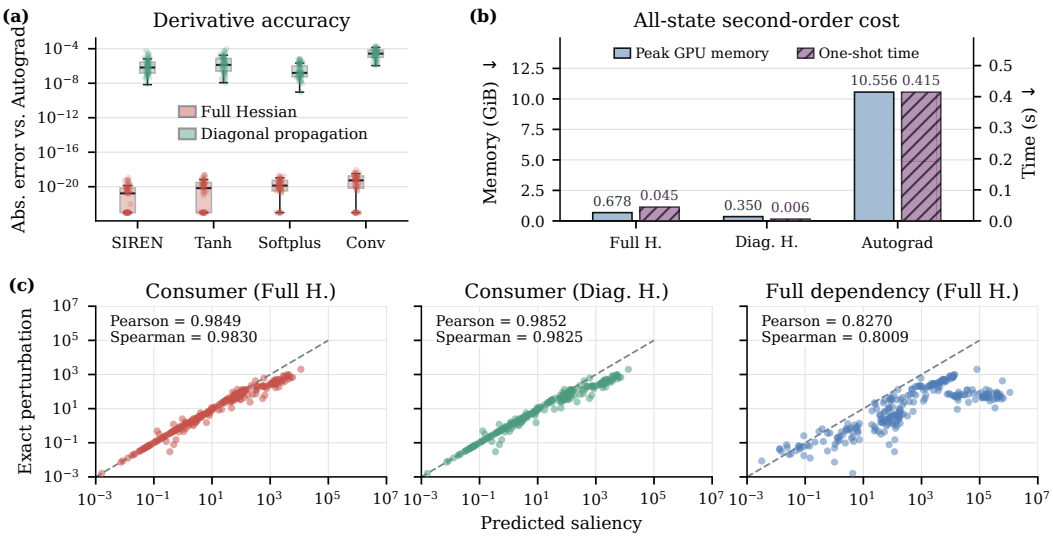

Figure 2: Consistency checks for local second-order pruning. **(a)** Full propagation agrees with nested Autograd, while Diag closely approximates Exact-Diag across the tested architectures. **(b)** For all hidden states of nine MLPs, both proposed derivative traversals require substantially less one-shot time and GPU memory than Autograd. **(c)** Consumer-side saliency from either full or diagonal curvature closely predicts the exact single-neuron perturbation, whereas scoring the full dependency group is less faithful.

2.4 Global Budgeted Selection

After computing $\mathcal{S}(\mathcal{G}_i^\ell)$ for every candidate dependency group, COSP pools candidates from all layers and ranks them globally by ascending predicted damage, as illustrated in Figure 1(d). It selects groups until the budget in Eq. equation 2 is met, then physically removes their full dependency groups.

This global allocation avoids preset layerwise pruning ratios: layers with lower predicted functional damage contribute more deletions. The budget is evaluated on the affine tensors of the resulting compact architecture, so parameters shared by overlapping dependency groups are counted only once, and the ranking is computed once from the frozen dense INR.

2.5 Ridge Reconstruction

After global selection and physical compaction, COSP reconstructs each affine mapping affected by the removed consumer-side groups. For an affected affine layer ℓ, let $\mathcal{O}$ denote the output channels retained by the compact mask. On the calibration set, the target from the dense model is

$$\mathbf{T}_{\mathcal{O}}^{\ell} = \left(W^{\ell}\mathbf{a}^{\ell} + \mathbf{b}^{\ell}\right)_{\mathcal{O}}, \tag{13}$$

where $\mathbf{a}^{\ell}$ is the layer input in the dense model. We fit the compact affine mapping to the activation $\widehat{\mathbf{a}}^{\ell}$ of the current compact prefix by

$$\min_{\widehat{W}^{\ell},\widehat{\mathbf{b}}^{\ell}} \quad \left\|\widehat{W}^{\ell}\widehat{\mathbf{a}}^{\ell} + \widehat{\mathbf{b}}^{\ell} - \mathbf{T}_{\mathcal{O}}^{\ell}\right\|_F^2 + \lambda\left\|\widehat{W}^{\ell}\right\|_F^2. \tag{14}$$

When multiple affine mappings are affected, reconstruction proceeds in forward order: each layer fits its dense-model target from the activations of the current compact prefix. We retain an update only if it does not increase the full-network calibration loss in Eq. equation 1, and otherwise the compact affine mapping is restored. This calibration-only step preserves the selected mask and the exported parameter count.

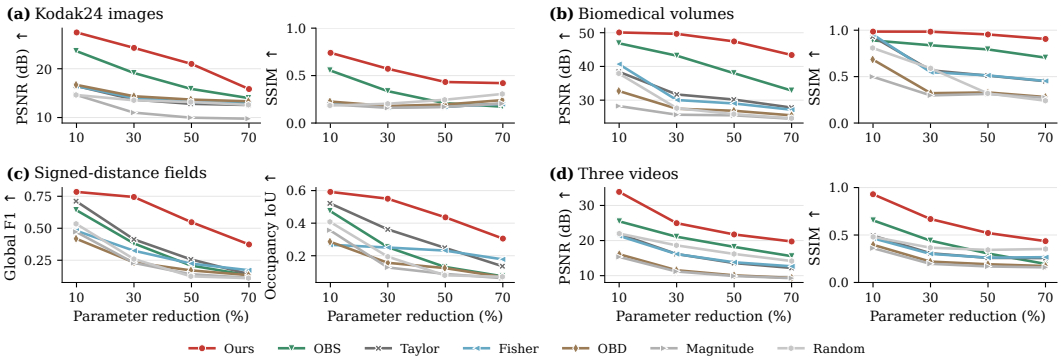

Figure 3: One-shot structured compression across four INR tasks. Each panel averages results within its task at matched physical parameter reductions.

Table 1: Deployment speedup and quality gap of compacted dense networks on A100. Gap is the mean margin over the strongest baseline (PSNR/SSIM or F1/IoU).

		30% reduction				50% reduction			
Task	Dense	Params	Forward	Decode	Gap	Params	Forward	Decode	Gap
Image	198.9K	139.1K	1.80×	1.49×	+5.14/+0.24	99.3K	1.94×	1.69×	+5.11/+0.19
Volume	790.5K	553.5K	1.30×	1.31×	+6.48/+0.15	395.4K	1.73×	1.73×	+9.41/+0.16
SDF	198.7K	139.0K	1.28×	1.27×	+0.33/+0.19	99.3K	1.47×	1.44×	+0.29/+0.19
Video	4.21M	2.94M	1.39×	1.39×	+3.86/+0.23	2.10M	1.87×	1.88×	+3.54/+0.18

3 Experiments

We evaluate COSP through three questions: whether the propagated derivative states and consumer-side score predict local deletion damage accurately and efficiently; whether one-shot structured compaction preserves decoded quality across four INR modalities; and how consumer-side grouping, curvature propagation, and ridge reconstruction affect the results.

We study the 24-image Kodak suite (Eastman Kodak Company, 1999), four biomedical volumes from BRIEF (Yang et al., 2024), the Living Room and Thai Statue SDFs from the official SIREN assets (Sitzmann et al., 2020), and three UVG clips (CityAlley, Jockey, and HoneyBee) (Ultra Video Group, 2016). Each task is evaluated at 10%–70% physical affine-parameter reduction after compaction of the dense model, and the reduction is measured from the affine tensors of the exported compact model.

All methods use the same dense checkpoint, calibration objective, physical budget, and evaluation grid, select one mask, and perform no gradient fine-tuning. Baselines retain their native scoring and update rules. Unless stated otherwise, COSP denotes Consumer-side/Diagonal scoring with fixed-mask ridge reconstruction. Table 3 applies the same reconstruction to every method. All comparisons use 65,536 calibration points and ridge coefficient $\lambda = 10^{-6}$. We compare Random, Magnitude (Han et al., 2015), Taylor (Molchanov et al., 2019), empirical Fisher (Liu et al., 2021), OBD (LeCun et al., 1990), and OBS (Hassibi & Stork, 1993). We report means in the main text and complete per-instance results in the supplementary material. All dense INRs are SIRENs (Sitzmann et al., 2020) trained with Adam (Kingma & Ba, 2015), at learning rates 10^{-1}, 10^{-2}, 10^{-4}, and 10^{-4} for 5K, 50K, 200K, and 320K updates on Image, Volume, SDF, and Video, respectively.

3.1 Does the local score predict deletion damage?

Figure 2 validates the sequence of approximations behind COSP. Independent tests over SIREN, Tanh, Softplus, and convolutional Tanh networks show that Full propagation agrees with nested PyTorch automatic differentiation, while Diag closely approximates Exact-Diag. Across nine MLP configurations, diagonal propagation reduces the all-state second-order cost from 0.045 to 0.006 s and from 0.678 to 0.350 GB on average, while nested Autograd requires 0.415 s and 10.556 GB. On a frozen Kodak23 SIREN, we evaluate 256 candidate

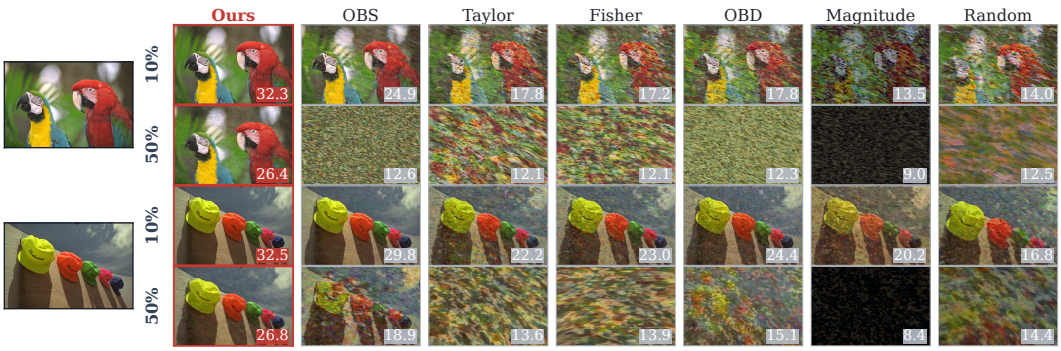

Figure 4: Representative Kodak reconstructions at 10% and 50% physical parameter reduction. Each row compares COSP with structured pruning baselines in PSNR.

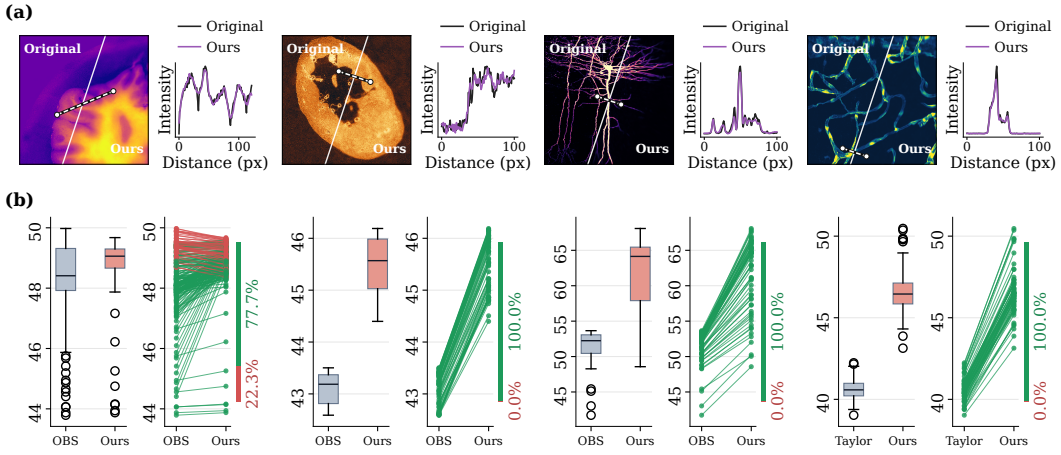

Figure 5: Biomedical-volume reconstruction at 30% physical parameter reduction. **(a)** Split views of the original and COSP reconstruction, with ROI intensity profiles. Brain and Kidney show central slices, Neurons and Vessels full-z maximum-intensity projections. **(b)** Slice-wise PSNR distributions and paired comparisons against the strongest baseline, with green and red marking COSP wins and losses and the bar showing their proportions. Each panel compares COSP with the strongest baseline for that instance.

deletions sampled uniformly over four hidden layers. Consumer-side Full and diagonal scores both closely predict their exact perturbations (Pearson 0.985), whereas jointly scoring all incident producer and consumer weights has lower agreement (Pearson 0.827). This supports the consumer-side group as both the functional deletion unit and the practical second-order target. Physical compaction agrees with the equivalent activation mask to 3.41×10^{-13}.

3.2 Structured compression across four modalities

Figure 3 summarizes the comparison across images, biomedical volumes, signed-distance fields, and videos. All methods prune the same frozen INR checkpoints and are matched by measured physical affine-parameter reduction. The figure reports the complete 10–70% trajectories, pairing decoded PSNR with SSIM for raster signals and Global F1 with occupancy IoU for SDF surfaces. COSP attains the highest mean task quality at every tested reduction on all four modalities. Across 132 matched settings comprising 96 image, 16 volume, 8 SDF, and 12 video evaluations, COSP ranks first among the seven methods in 125.

Figure 4 compares two representative Kodak images at 10% and 50% parameter reduction. At light compression, COSP retains fine texture and sharp color boundaries. After removing half of the affine parameters, it still preserves recognizable local structure, while the compared baselines exhibit texture noise, color mixing, or lost edges.

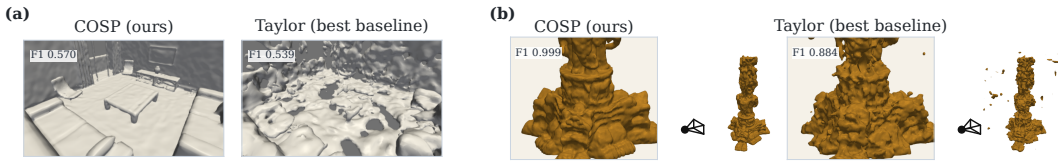

Figure 6: Fixed-view SDF reconstructions at 10% physical parameter reduction. **(a)** Living Room interior views. **(b)** Thai close-up and global views, linked by the projection glyph. COSP is shown before Taylor in each group.

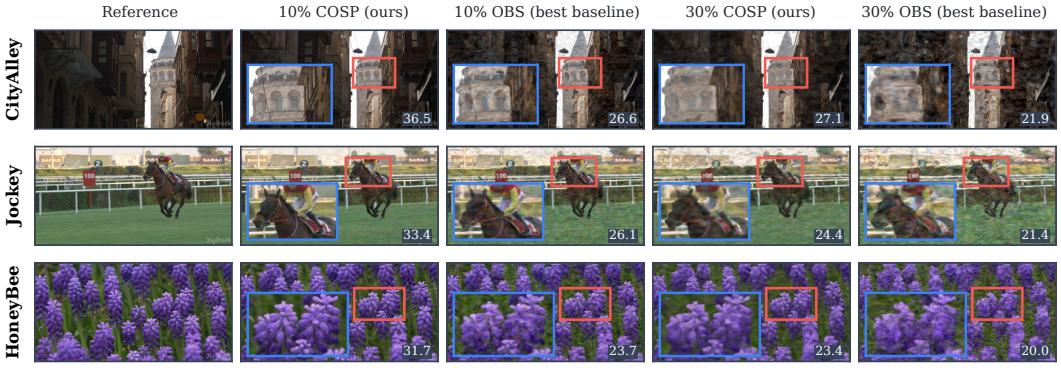

Figure 7: Frame-15 reconstructions from three UVG clips at 10% and 30% physical parameter reduction. Red boxes mark a shared ROI, and blue insets show its 2× enlargement.

Figure 5 examines the Brain, Kidney, Neurons, and Vessels volumes at 30% reduction. Central-slice or MIP views are paired with short ROI intensity profiles, directly testing whether the reconstruction follows the underlying density transitions. The paired z-slice comparisons evaluate every slice against the strongest matched baseline, and their connected win/loss segments and summary bars show that the visual fidelity is sustained across the volume rather than driven by a small set of favorable slices.

For Living Room and Thai Statue, we extract the complete zero surface and measure global F1 with a two-voxel tolerance. Figure 6 shows the reconstruction at low compression under fixed cameras, a common zero level, and rendering bounds taken from the dense model. The Room interior exposes disconnected Taylor artifacts, whereas the Thai detail and global views jointly reveal local geometry and the overall silhouette.

Figure 7 compares frame 15 from the CityAlley, Jockey, and HoneyBee UVG clips at 10% and 30% reduction. COSP retains coherent architectural edges, the horse structure, and flower-cluster texture more consistently, whereas OBS develops stronger blur and texture corruption at 30%. The full per-clip rate curves are provided in the supplementary material.

Materializing the compact selections as dense networks also speeds up deployment (Table 1). At 30% and 50% reduction on an A100 GPU, forward passes and full-signal decoding reach 1.27–1.94× speedup over the dense models. COSP exceeds the strongest baseline by 3.5–9.4 dB in PSNR, 0.15–0.24 in SSIM, 0.29–0.33 in Global F1, and 0.19 in occupancy IoU.

3.3 Ablations and cost

Table 2 studies three components of the one-shot compact model: the scoring group (Consumer-side or Full dependency), the curvature readout (Diagonal, Exact-Diagonal, or Exact-Full), and ridge reconstruction, with the first row giving the default configuration.

The first two rows of Table 2 isolate ridge without changing the Consumer-side/Diagonal mask. Ridge lowers quality loss on all four tasks, with the largest change on Video. Comparing the ridge rows shows that Exact-Diagonal is close to Diagonal on Image, Volume, and Video, whereas Exact-Full has a clear additional benefit mainly on SDF. Full dependency is slightly better on SDF but worse on Image, Volume, and Video, so consumer-side scoring is more robust across modalities. Exact-Diagonal and Exact-Full cost about 7.5× more time

Table 2: Design ablation at 65K calibration points. Quality columns report the median decrease from the dense model to the compact model: ΔPSNR for Image, Volume, and Video, and ΔF1 for SDF, aggregated over each task's instances and native reduction budgets. UVG cost uses the three video INRs at batch size 512. Time includes ridge reconstruction when it is used. Each entry is the median of five repeats, then the median across videos.

Scoring rule			Quality degradation ↓				UVG cost ↓	
Group	Curv.	Ridge	ΔI (dB)	ΔV (dB)	ΔS (F1)	ΔVid (dB)	Time (s)	Peak GPU (GiB)
Consumer-side	Diagonal	Yes	18.559	0.313	0.095	21.401	25.98	4.03
Consumer-side	Diagonal	No	19.481	1.329	0.276	31.444	25.23	4.03
Consumer-side	Exact-Diagonal	Yes	18.559	0.317	0.097	21.401	198.98	8.02
Consumer-side	Exact-Full	Yes	18.559	0.293	0.054	21.415	193.43	8.02
Full dependency	Diagonal	Yes	19.285	0.409	0.064	23.182	81.28	4.03

Table 3: Ridge reconstruction control at 50% physical parameter reduction. Each entry is mean ± standard deviation across task instances. For each method and task, the two entries share an identical compact mask. Img., Vol., and Vid. use PSNR. SDF uses Global F1.

	w/o ridge				w/ ridge			
Method	Img.	Vol.	SDF	Vid.	Img.	Vol.	SDF	Vid.
COSP	**19.1±4.4**	**42.7±9.2**	**0.35±0.04**	**19.3±1.1**	**21.0±3.3**	**47.4±4.7**	**0.55±0.09**	**21.7±1.7**
OBS	15.8±3.3	40.0±4.0	0.21±0.07	18.2±0.2	17.1±3.1	38.2±6.5	0.29±0.03	19.4±0.3
Taylor	12.8±2.0	28.2±10.9	0.26±0.03	13.6±1.0	16.9±2.9	40.7±2.6	0.52±0.10	21.4±1.8
Fisher	13.1±1.6	29.2±11.0	0.22±0.02	13.8±1.3	17.0±2.7	44.0±3.6	0.53±0.09	21.3±1.6
OBD	13.6±1.8	26.8±10.4	0.17±0.04	10.1±1.4	17.2±2.5	36.0±3.5	0.33±0.01	21.1±1.1
Magnitude	9.9±2.6	25.5±10.2	0.14±0.07	9.9±1.4	15.1±2.2	36.0±4.8	0.36±0.02	20.3±1.1
Random	13.2±1.8	25.8±9.3	0.12±0.07	16.2±0.5	18.1±2.6	42.3±3.9	0.45±0.02	21.5±2.1

and 2.0× more peak GPU memory than Diagonal. Full dependency costs about 3.1× more time. We therefore use consumer-side Diagonal scoring as the default operating point.

To disentangle pruning quality from the effect of ridge reconstruction, Table 3 evaluates every method at 50% reduction before and after the same ridge reconstruction of its fixed compact mask. COSP achieves the highest mean quality across all four modalities both before and after the shared ridge reconstruction.

On a fixed 256-point calibration set, splitting accumulation into minibatches of 16–256 leaves the candidate ranking unchanged. Across the full compression experiments, increasing calibration size from 8K to 65K improves absolute quality while preserving the separation between the COSP curvature variants and Taylor. Detailed calibration curves and per-instance results are provided in the supplementary material.

4 Conclusion

We presented COSP, a one-shot method for converting a frozen INR into a physically smaller dense network. COSP scores each dependency group through the consumer-side functional perturbation equivalent to its removal, propagates diagonal hidden-state curvature in a shared module-wise reverse traversal, and allocates deletions globally under the exported parameter budget. After mask selection, closed-form ridge reconstruction improves fidelity without changing the compact architecture or performing gradient fine-tuning.

Across images, biomedical volumes, SDFs, and videos, COSP achieves the best mean decoded quality at every tested reduction and ranks first in 125 of 132 matched settings. Full-curvature variants can benefit SDFs but provide limited additional quality on the other modalities at substantially higher scoring cost. Together, these results show that aligning the saliency unit with the actual structural deletion, combined with efficient local curvature and fixed-mask reconstruction, provides an effective basis for compact INR deployment.

Our method targets chain-structured MLP INRs, and one-shot diagonal selection discards cross-coordinate curvature and interactions, so accuracy degrades at high reductions. Ridge reconstruction depends on the calibration set, and its acceptance guard certifies only the measured objective. Branched or residual architectures are left to future work.

AI use statement

Generative AI tools were used to assist with code drafting, figure-layout code, and language editing. They were not used to generate experimental measurements or replace execution of the reported evaluations. All AI-assisted code was run against frozen artifacts, numerical claims were checked against generated CSV tables, and the authors reviewed the manuscript and take responsibility for its content.

Reproducibility statement

The main paper specifies the objective, group definition, recurrences, score, budget, reconstruction, datasets, metrics, and comparison protocol. The supplement contains full-state derivations, per-instance results, calibration and negative-result ablations, and artifact provenance. An anonymized package will include the pruning implementation, frozen manifests, table/figure generators, and commands needed to reproduce every reported result from the checkpoints.

References

Diana Aldana, João Paulo Lima, Daniel Csillag, Daniel Perazzo, Haoan Feng, Luiz Velho, and Tiago Novello. AIRe: Adaptive training of INRs via pruning and densification. *arXiv preprint arXiv:2510.23943*, 2025.

Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser-Nam Lim, and Abhinav Shrivastava. NeRV: Neural representations for videos. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Hao Chen, Matthew Gwilliam, Ser-Nam Lim, and Abhinav Shrivastava. HNeRV: A hybrid neural representation for videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Luca De Luigi, Adriano Cardace, Riccardo Spezialetti, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Deep learning on implicit neural representations of shapes. *arXiv preprint arXiv:2302.05438*, 2023.

Emilien Dupont, Adam Golinski, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet. COIN: COmpression with implicit neural representations. *arXiv preprint arXiv:2103.03123*, 2021.

Eastman Kodak Company. Kodak lossless true color image suite. `http://r0k.us/graphics/kodak/`, 1999.

Amer Essakine, Yanqi Cheng, Chun-Wun Cheng, Lipei Zhang, Zhongying Deng, Lei Zhu, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Where do we stand with implicit neural representations? a technical and performance survey. *arXiv preprint arXiv:2411.03688*, 2024.

Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16091–16101. IEEE, 2023.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4475–4488, 2022.

Sharath Girish, Abhinav Shrivastava, and Kamal Gupta. Shacira: Scalable hash-grid compression for implicit neural representations. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17467–17478. IEEE, 2023.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems*, volume 5, 1993.

Dhananjaya Jayasundara and Vishal M Patel. Implicit neural representations: A signal processing perspective. *arXiv preprint arXiv:2604.15047*, 2026.

Dhananjaya Jayasundara, Sudarshan Rajagopalan, Yasiru Ranasinghe, Trac D. Tran, and Vishal M. Patel. SINR: Sparsity driven compressed implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3061–3070, 2025.

Alper Kayabasi, Anil Kumar Vadathya, Guha Balakrishnan, and Vishwanath Saragadam. Bias for action: Video implicit neural representations with bias modulation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27999–28008. IEEE, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2015.

Woosuk Kwon, Sehoon Kim, Michael W. Mahoney, Joseph Hassoun, Kurt Keutzer, and Amir Gholami. A fast post-training pruning framework for transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24101–24116, 2022.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2, 1990.

Jaeho Lee, Sejun Park, Sangwoo Mo, Sungsoo Ahn, and Jinwoo Shin. Layer-adaptive sparsity for the magnitude-based pruning. In *International Conference on Learning Representations*, 2021a.

Jaeho Lee, Jihoon Tack, Namhoon Lee, and Jinwoo Shin. Meta-learning sparse implicit neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11769–11780, 2021b.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. SNIP: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2019.

Hao Li, Aryaan Kadav, Igor Durdanovic, Hanan Samet, Hans Peter Graf, Su Zheng, and Peyman Milanfar. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, 2017.

Jason Chun Lok Li, S. Luo, Le Xu, and Ngai Wong. ASMR: Activation-sharing multi-resolution coordinate networks for efficient inference. In *International Conference on Learning Representations*, 2024.

Mingbao Lin, Rongrong Ji, Yan Wang, Yichen Zhang, Baochang Zhang, Yonghong Tian, and Ling Shao. HRank: Filter pruning using high-rank feature map. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1529–1538, 2020.

Liyang Liu, Shilong Zhang, Zhanghui Kuang, Aojun Zhou, Jing-Hao Xue, Xinjiang Wang, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Group fisher pruning for practical network compression. In *International Conference on Machine Learning*, pp. 7021–7032, 2021.

Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *IEEE International Conference on Computer Vision*, pp. 2736–2744, 2017.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-Pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21702–21720, 2023.

Amirali Molaei, Amirhossein Aminimehr, Armin Tavakoli, Amirhossein Kazerouni, Bobby Azad, Reza Azad, and Dorit Merhof. Implicit neural representation in medical imaging: A comparative survey. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2373–2383. IEEE, 2023.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11264–11272, 2019.

Sidak Pal Singh and Dan Alistarh. Woodfisher: Efficient second-order approximation for neural network compression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18098–18109, 2020.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7462–7473, 2020.

Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6377–6389, 2020.

Ultra Video Group. Ultra video group test sequences. `http://ultravideo.fi/#testsequences`, 2016.

Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020.

Runzhao Yang, Tingxiong Xiao, Yuxiao Cheng, Anan Li, Jinyuan Qu, Rui Liang, Shengda Bao, Xiaofeng Wang, Jue Wang, Jinli Suo, Qingming Luo, and Qionghai Dai. Sharing massive biomedical data at magnitudes lower bandwidth using implicit neural function. *Proceedings of the National Academy of Sciences*, 121(28):e2320870121, 2024. doi: 10.1073/pnas.2320870121.

Shixing Yu, Zhewei Yao, Amir Gholami, Zhen Dong, Sehoon Kim, Michael W. Mahoney, and Kurt Keutzer. Hessian-aware pruning and optimal neural implant. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3880–3891, 2022.

Yanjie Zhu, Yuanyuan Liu, Yihang Zhang, and Dong Liang. Implicit neural representation for medical image reconstruction. *Physics in Medicine & Biology*, 70(12):12TR01, 2025.