SPINSCI: MASK-CONDITIONED SPATIAL-TEMPORAL SPIN TRANSFORMER FOR VIDEO SNAPSHOT COMPRESSIVE IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video snapshot compressive imaging (SCI) recovers multiple frames from one spatiotemporally coded measurement. This paper proposes *SpinSCI*, a mask-conditioned Spatial-Temporal Spin Transformer whose *Mask-Conditioned Ising Refinement Block (MC-IRB)* combines attention logits and sensing information to form signed spin couplings. Finite mean-field updates gate residual feature refinement. Physics-guided initialization and final data consistency incorporate the measurement model. Compared with the state-of-the-art method HiSViT_9, SpinSCI achieves average PSNR gains of **0.50 dB**, **0.23 dB**, and **0.04 dB** on *small-scale gray, mid-scale color, and large-scale color benchmarks*, respectively, with **16.9–17.0% fewer parameters, 21.0–21.6% fewer FLOPs, 77.0–84.8% lower inference latency, and 77.0–78.0% lower peak memory**. Qualitative comparisons on the *real benchmark* further show recovery of fine details and textures. `If the paper is accepted, the code will be released by the camera-ready deadline.`

1 INTRODUCTION

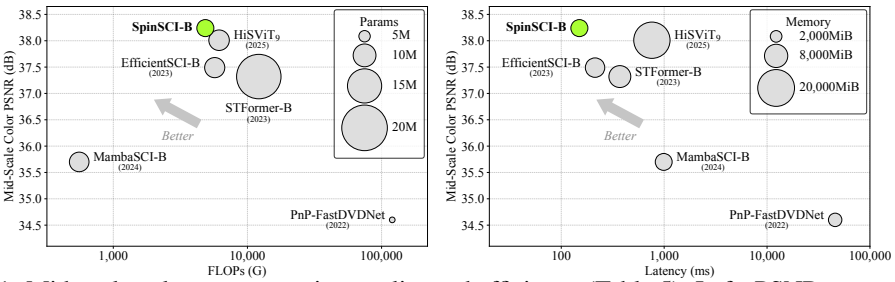

Figure 1: Mid-scale color reconstruction quality and efficiency (Table 5). Left: PSNR versus FLOPs; circle diameters are proportional to parameter counts. Right: PSNR versus latency; circle areas are proportional to peak memory.

High-speed imaging captures rapid motion but faces bandwidth, storage, and cost constraints (Song et al., 2025). Video snapshot compressive imaging (SCI) encodes multiple frames into one measurement using time-varying masks, as in Coded Aperture Compressive Temporal Imaging (CACTI) (Llull et al., 2013). Reconstruction is therefore a severely underdetermined inverse problem related to compressed sensing (Candes et al., 2006; Donoho, 2006).

Prior model-based and learned methods (Liu et al., 2019; Qiao et al., 2020; Wang et al., 2021; Yuan et al., 2022; Wang et al., 2023a;b;c; Cao et al., 2024; Pan et al., 2024; Wang et al., 2025) improve reconstruction, yet spatial masking and temporal integration complicate detail recovery and frame separation (Wang et al., 2025). In attention-based reconstruction, logits encode feature affinity, while coding masks specify each frame's contribution to the measurement. These cues motivate token refinement conditioned on sensing information.

SpinSCI coordinates this refinement through *Ising-model-inspired* interactions. The *Mask-Conditioned Ising Refinement Block (MC-IRB)* derives signed couplings from pre-softmax attention logits and sensing inputs: the coding masks, normalized initial estimate, and mask normalizer.

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

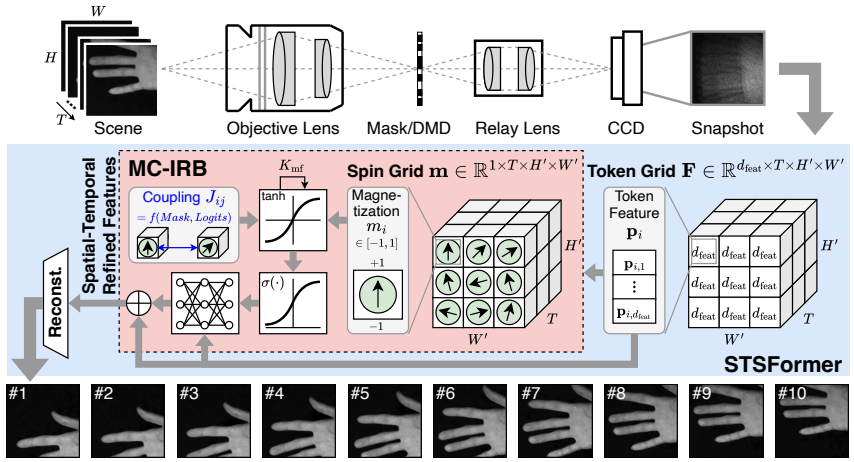

Figure 2: Illustration of the SpinSCI video SCI pipeline. A single snapshot encoded by CACTI (Llull et al., 2013) is input to SpinSCI. In the Spatial-Temporal Spin Transformer (STSFormer), token features $\mathbf{p}_i$ are extracted, and the Mask-Conditioned Ising Refinement Block (MC-IRB) performs differentiable mean-field inference for K_{mf} steps to estimate magnetizations $m_i \in [-1, 1]$, which are used to refine the spatial and temporal attention features. After post-processing, the video is reconstructed. Here, $d_{\mathrm{feat}} = 256$ denotes the token feature width. # denotes the frame number and a representative example from the Hand dataset (Yuan et al., 2022) ($512 \times 512 \times 40$) in the **real benchmark** is used at a compression rate of 10.

Local fields encode each token's evidence; positive and negative couplings favor agreement and opposition between latent spins. Finite mean-field updates produce magnetizations that gate learned residual corrections to the attention output.

MC-IRB refines the parallel spatial and temporal branches of a *Spatial-Temporal Spin Transformer (STSFormer)*. Physics-guided initialization and final data consistency incorporate the sensing model (Figure 2).

This work makes the following key contributions:

- SpinSCI introduces MC-IRB, which uses *mask-conditioned positive and negative couplings* to favor agreement and opposition between latent spins, respectively. Finite mean-field inference converts these interactions into spatial and temporal gates that guide residual feature corrections to support the recovery of fine details and edges.
- SpinSCI achieves state-of-the-art reconstruction quality on the *small-scale gray, mid-scale color, and large-scale color benchmarks* and the **real benchmark**, with *low parameter counts, FLOPs, inference latency, and memory usage* (Figure 1, mid-scale color).

2 Related Work

Vision Transformers. Self-attention (Vaswani et al., 2017) extends to images through ViT (Dosovitskiy et al., 2021) and Swin (Liu et al., 2021), and to video through local windows (Liu et al., 2022) or space–time factorization (Bertasius et al., 2021). Building on non-local similarity (Buades et al., 2005; Dabov et al., 2007; Dong et al., 2014), STFormer (Wang et al., 2023b) separates spatial and temporal attention into parallel branches. HiSViT (Wang et al., 2025) uses sequential cross-scale attention with pooled keys/values, hierarchical branches, and a gated self-modulated feed-forward network (FFN). SpinSCI combines parallel branches and pooled-key/value attention with *mask-conditioned spin interactions for residual feature refinement.*

Mask-Aware Reconstruction and Mask Co-Design. MetaSCI (Wang et al., 2021) uses exposure-normalized, mask-modulated inputs and per-mask parameter adaptation. Res2former (Wang et al., 2023c) jointly learns structural masks and reconstruction to improve dynamic range under temporal multiplexing. SpinSCI requires neither per-mask parameter updates nor physical-mask optimization. It conditions refinement on the supplied mask, normalized initial estimate, and mask normalizer to *guide mask-aware feature refinement during reconstruction.*

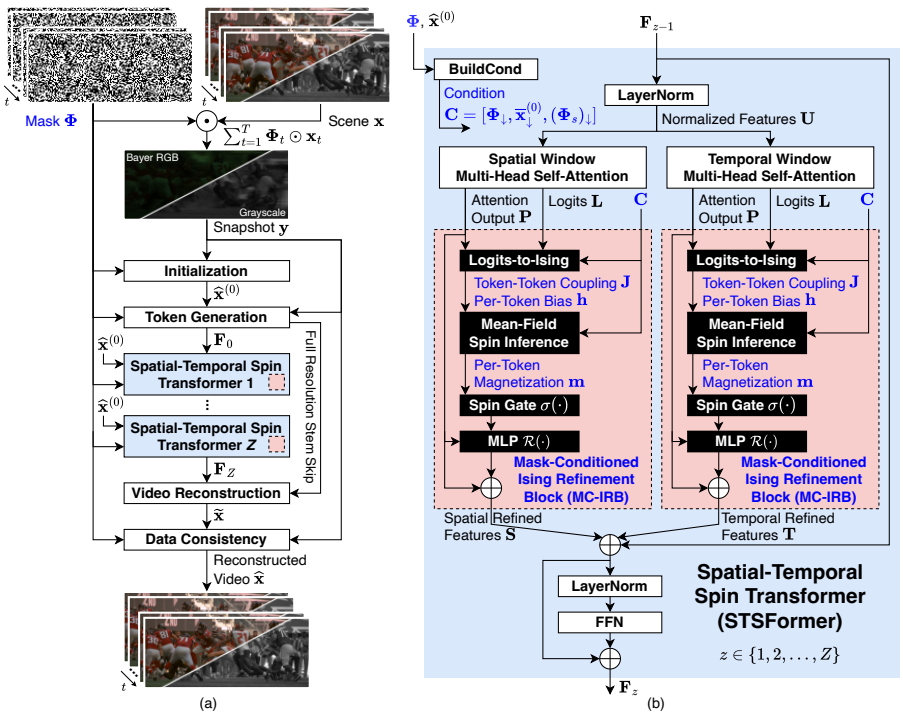

Figure 3: Overall architecture of SpinSCI. (a) The snapshot $\mathbf{y}$ and coding masks $\mathbf{\Phi}$ provide a physics-guided initialization $\widehat{\mathbf{x}}^{(0)}$. Token generation is followed by Z STSFormer blocks and video reconstruction, with a full-resolution stem skip supplied to the decoder. A final data-consistency correction maps the decoder output $\widetilde{\mathbf{x}}$ to the reconstructed video $\widehat{\mathbf{x}}$. (b) Spatial and temporal attention operate in parallel. MC-IRB uses the shared conditioning tensor $\mathbf{C}$ and attention logits $\mathbf{L}$ to form signed spin interactions and generate gates for residual feature corrections. The spatial branch combines native and pooled-key/value heads; only native-resolution heads supply its MC-IRB logits. The refined branch outputs are combined with the input features, followed by a residual feed-forward module.

3 Method

3.1 Overview

Video SCI. A snapshot $\mathbf{y} \in \mathbb{R}^{H \times W}$ encodes T frames $\mathbf{x}_t \in \mathbb{R}^{H \times W}$ through masks $\mathbf{\Phi}_t$:

$$\mathbf{y} = \sum_{t=1}^{T} \mathbf{\Phi}_t \odot \mathbf{x}_t + \mathbf{n}, \tag{1}$$

Here, H, W are the spatial dimensions, t indexes frames, $\odot$ denotes element-wise multiplication, and $\mathbf{n}$ is measurement noise.

Sensing and Learned Refinement. SpinSCI uses a learned prior to resolve temporal ambiguity. Initialization and final correction act in the measurement domain; MC-IRB conditions feature refinement on $\mathbf{\Phi}$, the normalized initial estimate $\overline{\mathbf{x}}^{(0)}$, and mask normalizer $\mathbf{\Phi}_s$. Throughout, $\epsilon = 10^{-8}$.

The initial estimate $\widehat{\mathbf{x}}^{(0)}$ and per-pixel mask normalizer $\mathbf{\Phi}_s$ are constructed as described in Section 3.3. MC-IRB follows four steps: *conditioning, signed-coupling construction, mean-field inference, and residual gating*. Sections 3.2 and 3.3 define the refiner and its backbone integration (Figure 3).

3.2 Mask-Conditioned Ising Refinement Block (MC-IRB)

Each post-attention token $\mathbf{p}_i \in \mathbb{R}^{d_{\text{feat}}}$ carries a latent spin. A refinement group contains $n = 49$ spatial positions, including padding, or $n = T$ temporal positions. For spins $\mathbf{s} \in \{-1, +1\}^n$, pairwise

couplings J_{ij} and local fields h_i define

$$E(\mathbf{s}) = -\sum_{i<j} J_{ij} s_i s_j - \sum_i h_i s_i. \tag{2}$$

Mean-field inference represents spin means by magnetizations $\mathbf{m} \in [-1, 1]^n$, which determine refinement gates.

Physical Conditioning (BuildCond). Per-sample normalization controls the initial-estimate scale:

$$\overline{\mathbf{x}}^{(0)} = \frac{\widehat{\mathbf{x}}^{(0)}}{\max\left(\mathrm{mean}_{t,h,w} |\widehat{\mathbf{x}}^{(0)}|, \epsilon\right)}. \tag{3}$$

Nearest-neighbor resizing preserves mask values; bilinear resizing maps the estimate and exposure normalizer to the token grid. After resizing, the normalizer is repeated across frames and concatenated with the mask and normalized initial estimate to form

$$\mathbf{C} = \left[\mathbf{\Phi}_{\downarrow}, \overline{\mathbf{x}}^{(0)}_{\downarrow}, (\mathbf{\Phi}_s)_{\downarrow}\right], \qquad \mathbf{c}_i = [\phi_i, u_i, d_i] \in \mathbb{R}^3, \tag{4}$$

Here, $\downarrow$ denotes token-grid resizing, $u_i = \bar{x}_i^{(0)}$, and $d_i = \phi_{s,i}$. Shared across blocks, $\mathbf{C}$ supplies each token with its coding-mask value, normalized initial estimate, and spatial exposure normalizer.

Logits-to-Ising. The two native spatial heads or all four temporal heads form $\mathcal{H}$. The head-averaged logits are

$$\overline{\mathbf{L}} = \frac{1}{|\mathcal{H}|} \sum_{a \in \mathcal{H}} \mathbf{L}^{(a)}. \tag{5}$$

Reusing pre-softmax logits ties refinement to affinities already computed for attention. Differences in sensing information define a penalty applied before centering:

$$\Pi_{ij} = \lambda_\phi |\phi_i - \phi_j| + \lambda_u |u_i - u_j| + \lambda_d |d_i - d_j|. \tag{6}$$

For valid-pair indicators $v_{ij} = v_{ji} \in \{0, 1\}$ with $v_{ii} = 1$, conditioned logits and row means are

$$\widetilde{J}_{ij} = v_{ij}(\overline{L}_{ij} - \Pi_{ij}), \qquad n_i = \sum_j v_{ij}, \qquad \mu_i = \frac{1}{n_i} \sum_j \widetilde{J}_{ij}. \tag{7}$$

Centering measures relative affinity; remasking, symmetrization, and diagonal removal produce valid, reciprocal interactions without self-coupling:

$$J^{\mathrm{ctr}}_{ij} = v_{ij}(\widetilde{J}_{ij} - \mu_i), \qquad J_{ij} = \gamma v_{ij} \mathbf{1}[i \neq j] \frac{J^{\mathrm{ctr}}_{ij} + J^{\mathrm{ctr}}_{ji}}{2}. \tag{8}$$

The learned weights $\lambda_\phi, \lambda_u, \lambda_d$ and scale γ are clamped nonnegative. Diagonal entries enter the row means before removal. The resulting signed $\mathbf{J}$ is symmetric, with zeros on the diagonal and at invalid token pairs; its rows need not sum to zero.

Mean-Field Spin Inference. Learned affine projections encode token and sensing evidence into initial magnetization and local field:

$$m_i^{(0)} = \tanh\left(\mathbf{w}_m^\top [\mathbf{p}_i; \mathbf{c}_i] + b_m\right), \qquad h_i = \mathbf{w}_h^\top [\mathbf{p}_i; \mathbf{c}_i] + b_h, \tag{9}$$

Here, $[\cdot\,;\cdot]$ concatenates inputs, $\mathbf{w}_h, \mathbf{w}_m \in \mathbb{R}^{d_{\mathrm{feat}}+3}$, and $b_h, b_m \in \mathbb{R}$. The field retains unary evidence at each update; learned initialization supplies an input-dependent starting state for finite-step inference.

With $\mathbf{h}$ and $\mathbf{J}$ fixed, K_{mf} updates propagate interactions:

$$m_i^{(\ell+1)} = \tanh\left(\beta \left[h_i + \sum_j J_{ij} m_j^{(\ell)}\right]\right), \qquad \ell = 0, \ldots, K_{\mathrm{mf}} - 1, \tag{10}$$

The learned $\beta \geq 0$ is clamped nonnegative. Positive couplings favor agreement between latent spins; negative couplings favor opposite signs. Invalid pairs contribute zero, and $\tanh$ bounds every iterate

to $[-1, 1]$. The fixed update budget coordinates gates through pairwise evidence with predictable computation.

Spin Gate and Residual MLP. Final magnetizations control correction strength through

$$g_i = \sigma\left(\kappa m_i^{(K_{\mathrm{mf}})}\right), \qquad \sigma(-\kappa) \le g_i \le \sigma(\kappa), \tag{11}$$

where σ is sigmoid and the learned $\kappa \ge 0$ is clamped nonnegative. The residual MLP $\mathcal{R}$ uses LayerNorm–Linear–GELU–Linear with 64 hidden units. Zero initialization of its final affine layer makes MC-IRB initially the identity:

$$\mathbf{p}_i \leftarrow \mathbf{p}_i + g_i\, \mathcal{R}(\mathbf{p}_i). \tag{12}$$

The MLP supplies a learned feature correction, and the gate regulates its strength. Within each branch, signed spin interactions coordinate token-wise gates to support the recovery of fine details and edges. Residual addition preserves the attention output as the base. The correction magnitude $g_i\|\mathcal{R}(\mathbf{p}_i)\|$ depends on both gate and residual; $m_i = 0$ gives $g_i = 1/2$.

Appendix A of the supplementary material provides sensing and refinement proofs and specifies Lean verification coverage. Its sufficient convergence condition is not enforced.

3.3 SpinSCI Architecture

SpinSCI applies token generation, $Z = 9$ STSFormer blocks, decoding, and final data consistency to produce $\widehat{\mathbf{x}}$ from $\widehat{\mathbf{x}}^{(0)}$.

Initialization. For $\mathcal{A}(\mathbf{x}; \mathbf{\Phi}) = \sum_t \mathbf{\Phi}_t \odot \mathbf{x}_t$, an exposure-aware back-projection initializes the video:

$$\begin{aligned} \widehat{\mathbf{x}}^{(0)} &= \mathcal{A}^\top(\mathbf{y}; \mathbf{\Phi}) = \{\mathbf{\Phi}_t \odot \mathbf{y}\}_{t=1}^{T}, \\ \widehat{\mathbf{x}}^{(0)} &\leftarrow \widehat{\mathbf{x}}^{(0)} + \mathcal{A}^\top\left(\frac{\mathbf{y} - \mathcal{A}(\widehat{\mathbf{x}}^{(0)}; \mathbf{\Phi})}{\mathbf{\Phi}_s + \epsilon}; \mathbf{\Phi}\right). \end{aligned} \tag{13}$$

The normalizer $\mathbf{\Phi}_s$ is $\sum_t \mathbf{\Phi}_t$ with zeros replaced by one; it equals mask energy at binary-mask exposed pixels. Back-projection and exposure-normalized correction supply a sensing-informed starting point for learned reconstruction. Bayer initialization applies this operation independently on the four RGGB sub-grids before re-interleaving.

Token Generation. A 3D stem embeds $[\widehat{\mathbf{x}}^{(0)}, \mathbf{\Phi}, \mathbf{y}/(\mathbf{\Phi}_s + \epsilon)]$, with the normalized measurement repeated across frames, into 128-channel features. Spatial downsampling halves resolution and sets $d_{\mathrm{feat}} = 256$:

$$\mathbf{F}_0 \in \mathbb{R}^{d_{\mathrm{feat}} \times T \times H' \times W'}, \qquad H' = \frac{H}{2},\ W' = \frac{W}{2}. \tag{14}$$

Joint embedding exposes coding and measurement information before attention. Downsampling reduces attention cost while preserving temporal length; full-resolution stem features supply a decoder skip.

STSFormer. Following STFormer (Wang et al., 2023b), block z supplies $\mathbf{U} = \mathrm{LayerNorm}(\mathbf{F}_{z-1})$ to parallel spatial and temporal branches. The branches model complementary within-frame and across-frame dependencies before feature fusion.

The spatial branch uses four heads with key/value pooling factors $(1, 1, 2, 4)$ (Wang et al., 2025). Across successive blocks, both native heads alternate between unshifted 7×7 windows and windows with three-token cyclic shifts to exchange information across window boundaries; $\mathbf{M}$ excludes wrap-around pairs (Liu et al., 2021). For native head a,

$$\mathbf{L}^{(a)} = \frac{\mathbf{Q}^{(a)}(\mathbf{K}^{(a)})^\top}{\sqrt{d_h^{\mathrm{sp}}}} + \mathbf{B}^{(a)} + \mathbf{M}, \qquad \mathbf{P}^{(a)} = \mathrm{softmax}(\mathbf{L}^{(a)})\mathbf{V}^{(a)}, \tag{15}$$

Here, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are query/key/value projections and $\mathbf{B}$ is learned relative-position bias. The other heads retain queries at token-grid resolution and pool keys/values into 8×8 windows, broadening context with 64 keys per query and no cyclic shifts. Concatenated head outputs are projected to d_{feat}. Spatial MC-IRB uses only native-head logits, whose square matrices index the same tokens on both axes for pairwise spin interactions.

Temporal attention acts over $\mathbf{U}_{h,w} \in \mathbb{R}^{T \times d_{\text{feat}}}$, using four heads with total inner width $d_{\text{feat}}/2$ to reduce projection and attention cost. Thus $d_h^{\text{sp}} = 64$ and $d_h^{\text{tmp}} = 32$, with temporal logits

$$\mathbf{L}^{(a)} = \frac{\mathbf{Q}^{(a)}(\mathbf{K}^{(a)})^\top}{\sqrt{d_h^{\text{tmp}}}} + \mathbf{B}_{\text{tmp}}^{(a)}. \tag{16}$$

The temporal bias $\mathbf{B}_{\text{tmp}}^{(a)} \in \mathbb{R}^{T \times T}$ is learned; no shifted-window mask applies. Temporal MC-IRB uses logits averaged across all four heads and refines the attention output projected to d_{feat}.

Refined branch outputs $\mathbf{S}, \mathbf{T}$ fuse before the residual FFN:

$$\mathbf{F}_{\text{st}} = \mathbf{F}_{z-1} + \mathbf{S} + \mathbf{T}, \qquad \mathbf{F}_z = \mathbf{F}_{\text{st}} + \text{FFN}(\text{LayerNorm}(\mathbf{F}_{\text{st}})) \,. \tag{17}$$

The inverted residual FFN (Sandler et al., 2018) expands width to $2d_{\text{feat}}$, applies depthwise $3 \times 3 \times 3$ mixing, and projects back, with GELU after expansion and depthwise convolution. Pointwise layers mix channels, while depthwise convolution adds local space–time interactions at low cost.

Video Reconstruction. A transposed 3D convolution upsamples the final features; concatenation with the stem skip recovers local detail before grayscale/RGB decoding into $\widetilde{\mathbf{x}}$. A shared auxiliary head after blocks 3 and 6 supports optional intermediate reconstruction supervision, providing training signals to earlier blocks without inference overhead.

Data Consistency. Grayscale reconstruction corrects the decoder's measurement mismatch:

$$\widehat{\mathbf{x}} = \widetilde{\mathbf{x}} + \mathcal{A}^\top \left(\frac{\mathbf{y} - \mathcal{A}(\widetilde{\mathbf{x}}; \mathbf{\Phi})}{\mathbf{\Phi}_s + \epsilon}; \mathbf{\Phi} \right). \tag{18}$$

For exposed binary-mask pixels, this multiplies the measurement residual by $\epsilon/(\mathbf{\Phi}_s + \epsilon)$ in exact arithmetic; zero-exposure pixels cannot be corrected.

For RGB output, $\mathcal{B}$ samples RGGB sites and $\mathcal{B}^\top$ scatters corrections back:

$$\begin{aligned} \mathbf{r} &= \mathbf{y} - \mathcal{A}(\mathcal{B}\widetilde{\mathbf{x}}_{\text{rgb}}; \mathbf{\Phi}), \\ \widehat{\mathbf{x}}_{\text{rgb}} &= \widetilde{\mathbf{x}}_{\text{rgb}} + \mathcal{B}^\top \mathcal{A}^\top \left(\frac{\mathbf{r}}{\mathbf{\Phi}_s + \epsilon}; \mathbf{\Phi} \right). \end{aligned} \tag{19}$$

Only observed color sites change; other RGB values retain the decoder prediction.

3.4 Computational Complexity

For $N = TH'W'$ tokens, spatial attention uses 49 native or 64 pooled keys per query. Temporal attention uses T keys; each MC-IRB takes $K_{\text{mf}} = 3$ updates.

Sensing operations cost $O(THW)$. Fixed-width stem and decoder convolutions cost $O(Nd_{\text{feat}}^2)$.

Attention, residual MLPs, and mean-field interactions give the Z-block cost

$$O\big(ZN\left[d_{\text{feat}}^2 + 64d_{\text{feat}} + d_{\text{feat}}(49 + 64 + T) + (1 + K_{\text{mf}})(49 + T)\right]\big) \,. \tag{20}$$

For fixed windows, spatial cost is linear in pixel count, while temporal attention is quadratic in T. Counts include padding and full inference, excluding training-only heads; component costs appear in Appendix C of the supplementary material.

4 Experiments

4.1 Experimental Setup

Datasets. Training uses DAVIS2017 (Pont-Tuset et al., 2017). Evaluation covers six **small-scale gray**, six **mid-scale color**, and four **large-scale color** sequences (Yuan et al., 2022; Wang et al., 2023a;b; Pan et al., 2024). The **real benchmark** includes Hand (Yuan et al., 2022), Water Balloon, Domino (Qiao et al., 2020), and Pendulum Ball. Dimensions and protocols are in Appendix C of the supplementary material.

Evaluation Metrics. PSNR and SSIM (Wang et al., 2004) are computed using the evaluation code provided by Yuan et al. (2021); efficiency is assessed by parameter count, FLOPs, inference latency, and peak memory.

Efficiency Measurement. SpinSCI uses BF16 autocast, consistent with its mixed-precision training, and `torch.compile` on one NVIDIA GH200, batch 1, and $T = 8$. Latency is the CUDA-synchronized median of 20 runs after five warm-ups, excluding compilation; memory is peak allocated GPU memory in MiB. Each multiply–accumulate operation counts as one FLOP. Small-scale gray and mid-scale color tests use full frames; large-scale latency covers 24 tiles, with memory reported per tile. Baseline execution settings appear in Appendix C of the supplementary material.

Implementation Details. PyTorch training uses four NVIDIA H200 GPUs. SpinSCI-B uses the architecture in Section 3.3, $K_{\mathrm{mf}} = 3$, and a 64-unit residual MLP.

Training uses 200 epochs, with 256×256 gray crops (batch 10) and 512×512 color crops (batch 2). Muon (Jordan et al., 2024) updates eligible matrix weights and Adam (Kingma & Ba, 2015) the remainder, with cosine learning-rate decay from 2×10^{-4} to 10^{-6}. The MSE loss (Cao et al., 2024) is

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{T\, C_{\mathrm{out}}\, H\, W} \left\| \widehat{\mathbf{x}} - \mathbf{x} \right\|_F^2 , \tag{21}$$

where $C_{\mathrm{out}} = 1$ for gray and 3 for RGB. BF16/FP32 mixed precision (Micikevicius et al., 2018) accelerates training; gradient clipping (Pascanu et al., 2013) stabilizes optimization.

4.2 Ablation Studies

Ablations test spatial/temporal refinement, interacting gates, and explicit mask conditioning. Appendix B of the supplementary material provides additional comparisons, control definitions, and the repeated-run protocol.

Controlled Protocol. Variants share DAVIS2017 (Pont-Tuset et al., 2017), 80 epochs, batch 20, 128×128 crops, seed 2447, loss, optimizer, and schedule. EMA weights are tested on six **mid-scale color** sequences at compression rate 8. All models use nine blocks except the ten-block capacity control.

Table 1: Spatial and temporal MC-IRB ablation at 80 epochs, seed 2447. Efficiency uses 512×512 inputs, $T = 8$, and BF16 compiled inference (Section 4.1).

Spatial MC-IRB	Temporal MC-IRB	PSNR (dB)	Params (M)	FLOPs (G)	Latency (ms)	Memory (MiB)
-	-	36.68	6.86	4528	117	4280
-	✓	36.91	7.16	4690	134	4282
✓	-	36.84	7.16	4694	135	4282
✓	✓	37.07	7.47	4856	150	4284

Spatial and Temporal Refinement. Compared with the backbone without MC-IRB, spatial-only and temporal-only refinement improve PSNR by 0.16 and 0.23 dB; combining both yields a 0.39 dB gain (Table 1). This combined configuration increases parameters, FLOPs, latency, and memory by 8.9%, 7.2%, 28.6%, and 0.1%, respectively, relative to the same backbone.

Interactions, Gates, and Capacity. MC-IRB improves PSNR over fixed, conditioned, and field-only gates by 0.29, 0.20, and 0.23 dB (Table 2). Setting $\mathbf{J} = 0$ reduces PSNR by 0.21 dB; replacing $\mathbf{J}$ with $|\mathbf{J}|$ reduces it by 0.14 dB. These comparisons support adaptive signed interactions. The ten-block backbone remains 0.20 dB lower despite slightly higher parameter and FLOP counts.

Table 2: MC-IRB compared with direct gates and additional backbone capacity at 80 epochs. Δ is full-model PSNR minus variant PSNR, in dB.

Variant	PSNR (dB)	Δ
Full MC-IRB ($K_{\mathrm{mf}} = 3$)	37.07	–
Residual MLP, fixed $g_i = 1/2$	36.78	+0.29
No MC-IRB, $Z = 10$	36.87	+0.20
Conditioned gate without mean-field	36.87	+0.20
Sigmoid gate on $\mathbf{h}$ only	36.84	+0.23
$K_{\mathrm{mf}} = 0$	36.81	+0.26
No interactions ($\mathbf{J} = 0$)	36.86	+0.21
Unsigned couplings (existing control)	36.93	+0.14
Entrywise $\lvert\mathbf{J}\rvert$ control	36.93	+0.14

Training Budget. Gains over the backbone without MC-IRB persist at 40/80 epochs (0.45/0.39 dB); two checkpoints do not establish convergence rate or training-time savings.

Table 4: Params, FLOPs, inference latency, peak memory, and PSNR/SSIM on **small-scale gray** video SCI benchmarks at compression rate 8. The best and second-best PSNR/SSIM results are shown in bold and underlined, respectively. SpinSCI's Params, FLOPs, latency, and memory are bolded to emphasize its efficiency.

Method	Params (M)	FLOPs (G)	Latency (ms)	Memory (MiB)	Aerial PSNR	Aerial SSIM	Crash PSNR	Crash SSIM	Drop PSNR	Drop SSIM	Kobe PSNR	Kobe SSIM	Runner PSNR	Runner SSIM	Traffic PSNR	Traffic SSIM	Average PSNR	Average SSIM
PnP-FastDVDNet	2.48	29605	2834	842	27.98	0.897	27.32	0.925	41.82	0.989	32.73	0.947	36.29	0.962	27.95	0.932	32.35	0.942
EfficientSCI-B	8.82	1424	50	1382	31.50	0.953	31.13	0.971	45.18	**0.995**	35.76	0.974	43.05	0.988	32.30	0.968	36.48	0.975
STFormer-B	19.49	3060	83	1883	31.56	0.953	31.06	0.970	45.08	**0.995**	35.53	0.973	42.64	0.988	32.15	0.967	36.34	0.974
HiSViT$_9$	8.99	1536	178	4904	31.67	0.957	31.62	0.976	45.55	**0.995**	**36.24**	**0.976**	43.84	**0.991**	33.06	**0.973**	37.00	0.978
SpinSCI-B	**7.46**	**1214**	**41**	**1130**	**33.47**	**0.967**	**32.82**	**0.978**	**45.66**	**0.995**	35.70	0.974	**43.85**	**0.991**	**33.49**	**0.973**	**37.50**	**0.980**

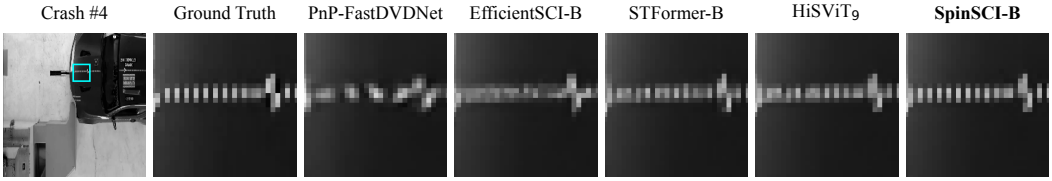

Figure 4: Qualitative comparison on frame 4 of the Crash sequence from the small-scale gray benchmark. The first panel shows the full ground-truth frame with the selected region marked in cyan. Subsequent panels show enlarged crops using nearest-neighbor interpolation.

Table 5: Params, FLOPs, inference latency, peak memory, and PSNR/SSIM on **mid-scale color** video SCI benchmarks at compression rate 8. The best and second-best PSNR/SSIM results are shown in bold and underlined, respectively. SpinSCI's Params, FLOPs, latency, and memory are bolded to emphasize its efficiency.

Method	Params (M)	FLOPs (G)	Latency (ms)	Memory (MiB)	Beauty PSNR	Beauty SSIM	Bosphorus PSNR	Bosphorus SSIM	Jockey PSNR	Jockey SSIM	Runner PSNR	Runner SSIM	ShakeNDry PSNR	ShakeNDry SSIM	Traffic PSNR	Traffic SSIM	Average PSNR	Average SSIM
PnP-FastDVDNet	2.48	120077	45907	2664	35.27	0.972	37.24	0.978	35.63	0.950	38.22	0.965	33.71	**0.969**	27.49	0.915	34.60	0.955
EfficientSCI-B	8.83	5702	213	5570	37.51	0.979	40.89	0.988	38.49	0.969	42.73	0.985	35.19	0.953	30.13	0.943	37.49	0.970
MambaSCI-B	8.59	557	990	4206	36.95	0.979	38.61	0.982	36.53	0.960	40.03	0.977	34.55	0.950	27.52	0.904	35.70	0.959
STFormer-B	19.49	12156	371	7081	37.37	0.981	40.39	0.988	38.32	0.968	42.45	0.985	35.15	0.956	30.24	0.939	37.32	0.970
HiSViT$_9$	8.99	6146	760	19462	37.75	0.981	41.50	0.990	39.29	0.972	43.27	0.986	35.49	0.958	**30.76**	**0.946**	38.01	0.972
SpinSCI-B	**7.47**	**4856**	**150**	**4284**	**38.04**	**0.982**	**42.05**	**0.991**	**39.30**	**0.973**	**43.51**	**0.987**	**35.83**	0.961	30.72	0.945	**38.24**	**0.973**



Figure 5: Qualitative comparison on frame 4 of the Bosphorus sequence from the mid-scale color benchmark. The first panel shows the full ground-truth frame with the selected region marked in cyan. Subsequent panels show enlarged crops using nearest-neighbor interpolation.

Explicit Mask Conditioning. Removing mask channels costs 0.13–0.19 dB on the training mask and 0.30–0.38 dB on the unseen binary mask (Table 3). The larger transfer gaps support explicit conditioning beyond the mask-dependent initialization and final correction retained by all controls.

Full-model PSNR is 37.07 dB on both masks. This supports transfer to one unseen realization; broader generalization requires evaluation across additional masks. Gaussian and calibrated-mask results appear in Section B.2 of the supplementary material.

Table 3: Effect of explicit mask conditioning on reconstruction with the training mask and an unseen binary mask. Values are PSNR in dB. All variants retain the correct masks in initialization and final data consistency.

Explicit mask inputs	Training mask	Unseen binary mask
Full model	37.07	37.07
$\mathbf{\Phi}, \mathbf{\Phi}_s$ removed from $\mathbf{C}$	36.94	36.77
No explicit mask channels in $\mathbf{C}$ or stem	36.88	36.69

Repeatability. Across five additional training seeds, the full model obtains 37.07 ± 0.10 dB (mean $\pm$ sample SD). Paired gains over field-only gating, $\mathbf{J} = 0$, and $|\mathbf{J}|$ are 0.23, 0.21, and 0.14 dB, with pointwise 95% intervals above zero. Section B.4 of the supplementary material details the paired protocol and interval assumptions; these intervals do not establish simultaneous superiority across controls.

Table 6: Params, FLOPs, inference latency, peak memory, and PSNR/SSIM on **large-scale color** video SCI benchmarks at compression rate 8. The best and second-best PSNR/SSIM results are shown in bold and underlined, respectively. SpinSCI's Params, FLOPs, latency, and memory are bolded to emphasize its efficiency.

Method	Params (M)	FLOPs (G)	Latency (ms)	Memory (MiB)	Football PSNR	Football SSIM	Hummingbird PSNR	Hummingbird SSIM	Messi PSNR	Messi SSIM	Swinger PSNR	Swinger SSIM	Average PSNR	Average SSIM
PnP-FastDVDNet	2.48	956866	369528	891	37.42	0.982	38.51	0.979	33.54	0.979	28.26	0.930	34.43	0.967
EfficientSCI-B	8.83	45100	1634	1841	40.33	0.989	40.41	0.974	34.85	0.974	29.45	0.940	36.26	0.969
MambaSCI-B	8.59	4405	7673	1309	35.91	0.974	31.67	0.641	29.09	0.923	27.01	0.900	30.92	0.859
STFormer-B	19.49	96071	2616	2416	40.64	0.990	40.43	0.974	34.87	0.976	29.59	0.943	36.38	0.971
HiSViT$_9$	8.99	48977	8377	6503	**42.25**	**0.993**	41.48	0.989	36.39	0.986	**30.43**	**0.954**	37.64	**0.980**
SpinSCI-B	**7.47**	**38402**	**1270**	**1474**	41.68	0.992	**41.99**	**0.992**	**37.07**	**0.987**	29.98	0.948	**37.68**	0.979

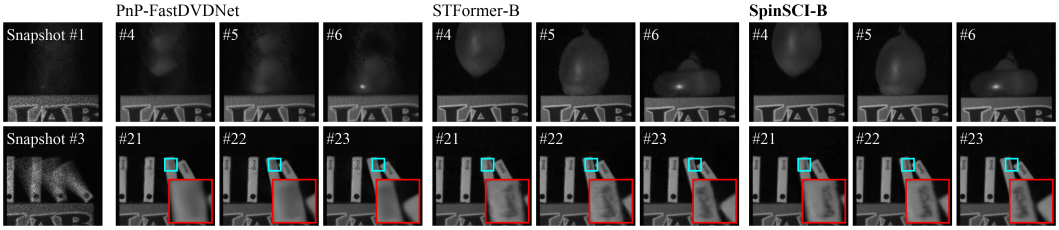

Figure 6: Qualitative comparison on Water Balloon (Qiao et al., 2020) ($512 \times 512 \times 10$) and Domino (Qiao et al., 2020) ($512 \times 512 \times 60$) in the **real benchmark** at compression rate 10. # denotes the frame number.

4.3 Comparison with State-of-the-Art Methods

Baselines. Comparisons include PnP-FastDVDNet (Yuan et al., 2022), EfficientSCI-B (Wang et al., 2023a), MambaSCI-B (Pan et al., 2024), STFormer-B (Wang et al., 2023b), and HiSViT$_9$ (Wang et al., 2025). MambaSCI (Pan et al., 2024) is evaluated only on the color benchmarks, consistent with its design for quad-Bayer color video reconstruction.

Quantitative and Qualitative Results. Tables 4 to 6 report per-sequence and average PSNR/SSIM alongside efficiency measured under the protocol in Section 4.1. Small-scale gray and mid-scale color baseline PSNR/SSIM values come from published results, except for MambaSCI-B. MambaSCI-B and the large-scale baselines use official-checkpoint evaluations. Appendix C of the supplementary material documents mask differences and shared-protocol checks. SpinSCI-B achieves 37.50 dB PSNR and 0.980 SSIM on small-scale gray using BF16 compiled inference, a 0.50 dB gain over HiSViT$_9$. On mid-scale and large-scale color, SpinSCI-B achieves $38.24/0.973$ and $37.68/0.979$ PSNR/SSIM, respectively, exceeding HiSViT$_9$ by 0.23 and 0.04 dB.

SpinSCI-B uses fewer parameters and FLOPs than HiSViT$_9$ on all three benchmarks. Its BF16 compiled latency and peak memory are reported in Tables 4 to 6; these runtime comparisons reflect the execution settings in Appendix C of the supplementary material.

Figures 4 to 6 show representative reconstructions. Figures D.20–D.22 of the supplementary material cover all gray/color sequences using the checkpoints evaluated in Tables 4 to 6.

Full-frame reconstructed videos for datasets in the **real benchmark** at compression rates of 10 and 50 are shown as frame sequences in Sections D.1–D.2 of the supplementary material.

5 Conclusion

SpinSCI combines visual affinity and sensing information through signed spin interactions. Finite mean-field inference coordinates residual refinement within the spatial and temporal branches, while physics-guided initialization and final data consistency incorporate the sensing model. Formal analysis establishes coupling symmetry, bounded magnetizations and gates, and binary-mask measurement-residual reduction. Benchmark comparisons show PSNR gains over the state-of-the-art method HiSViT$_9$ with fewer parameters and FLOPs, lower inference latency, and lower peak memory.

AI USE STATEMENT

Generative AI tools were used to assist with writing and checking formatting compliance. The authors reviewed all AI-assisted changes and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses public video snapshot compressive imaging datasets and standard benchmark protocols and does not involve human subjects or private data. The authors have read and commit to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The model architecture, mask-conditioned refinement computations, and data-consistency operations are specified in Section 3. Section 4.1 describes the datasets, training settings, loss, evaluation metrics, and efficiency protocol; Sections 4.2 and 4.3 describe the ablation and comparison settings. Appendices A–D of the supplementary material provide proofs and verification coverage, additional ablations and statistical reporting, implementation details, and qualitative comparisons, respectively. If the paper is accepted, the official code will be released by the camera-ready deadline.

REFERENCES

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *ICML*, pp. 813–824, 2021.

A. Buades, B. Coll, and J. M. Morel. A Review of Image Denoising Algorithms, with a New One. *Multiscale Modeling & Simulation*, 4(2):490–530, 2005. doi: 10.1137/040616024.

E.J. Candes, J. Romberg, and T. Tao. Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory*, 52 (2):489–509, 2006. doi: 10.1109/TIT.2005.862083.

Miao Cao, Lishun Wang, Mingyu Zhu, and Xin Yuan. Hybrid CNN-Transformer Architecture for Efficient Large-Scale Video Snapshot Compressive Imaging. *IJCV*, 132(10):4521–4540, 2024. doi: 10.1007/s11263-024-02101-y.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE TIP*, 16(8):2080–2095, 2007. doi: 10.1109/TIP.2007.901238.

Weisheng Dong, Guangming Shi, Xin Li, Yi Ma, and Feng Huang. Compressive Sensing via Nonlocal Low-Rank Regularization. *IEEE TIP*, 23(8):3618–3632, 2014. doi: 10.1109/TIP.2014.2329449.

D.L. Donoho. Compressed Sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. doi: 10.1109/TIT.2006.871582.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An Optimizer for Hidden Layers in Neural Networks, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. doi: 10.48550/arXiv.1412.6980.

Yang Liu, Xin Yuan, Jinli Suo, David J. Brady, and Qionghai Dai. Rank Minimization for Snapshot Compressive Imaging. *IEEE TPAMI*, 41(12):2990–3006, 2019. doi: 10.1109/TPAMI.2018.2873587.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*, pp. 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *CVPR*, pp. 3192–3201, 2022. doi: 10.1109/CVPR52688.2022.00320.

Patrick Llull, Xuejun Liao, Xin Yuan, Jianbo Yang, David Kittle, Lawrence Carin, Guillermo Sapiro, and David J. Brady. Coded Aperture Compressive Temporal Imaging. *Opt. Express*, 21(9): 10526–10545, May 2013. doi: 10.1364/OE.21.010526.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed Precision Training. In *ICLR*, 2018. doi: 10.48550/arXiv.1710.03740.

Zhenghao Pan, Haijin Zeng, Jiezhang Cao, Yongyong Chen, Kai Zhang, and Yong Xu. MambaSCI: Efficient Mamba-UNet for Quad-Bayer Patterned Video Snapshot Compressive Imaging. In *NeurIPS*, 2024. doi: 10.52202/079017-2544.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training Recurrent Neural Networks. In *ICML*, pp. 1310–1318, 2013. doi: 10.48550/arXiv.1211.5063.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv:1704.00675*, 2017. doi: 10.48550/arXiv.1704.00675.

Mu Qiao, Ziyi Meng, Jiawei Ma, and Xin Yuan. Deep learning for video compressive sensing. *APL Photonics*, 5(3):030801, 03 2020. doi: 10.1063/1.5140721.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *CVPR*, pp. 4510–4520, 2018.

Yunfeng Song, Mengyuan Liu, and Xin Yuan. Deep learning based video snapshot compressive imaging of Leidenfrost droplets. *iOptics*, 1(2):100020, 2025. doi: 10.1016/j.iopt.2025.100020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, volume 30, 2017.

Lishun Wang, Miao Cao, and Xin Yuan. EfficientSCI: Densely Connected Network with Space-time Factorization for Large-scale Video Snapshot Compressive Imaging. In *CVPR*, pp. 18477–18486, 2023a. doi: 10.1109/CVPR52729.2023.01772.

Lishun Wang, Miao Cao, Yong Zhong, and Xin Yuan. Spatial-Temporal Transformer for Video Snapshot Compressive Imaging. *IEEE TPAMI*, 45(7):9072–9089, 2023b. doi: 10.1109/TPAMI.2022.3225382.

Ping Wang, Lishun Wang, and Xin Yuan. Deep Optics for Video Snapshot Compressive Imaging. In *ICCV*, pp. 10612–10622, 2023c. doi: 10.1109/ICCV51070.2023.00977.

Ping Wang, Yulun Zhang, Lishun Wang, and Xin Yuan. Hierarchical Separable Video Transformer for Snapshot Compressive Imaging. In *Computer Vision – ECCV 2024*, pp. 104–122. Springer, 2025. doi: 10.1007/978-3-031-73004-7_7.

Zhengjue Wang, Hao Zhang, Ziheng Cheng, Bo Chen, and Xin Yuan. MetaSCI: Scalable and Adaptive Reconstruction for Video Compressive Sensing. In *CVPR*, pp. 2083–2092, 2021. doi: 10.1109/CVPR46437.2021.00212.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE TIP*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Xin Yuan, David J. Brady, and Aggelos K. Katsaggelos. Snapshot Compressive Imaging: Theory, Algorithms, and Applications. *IEEE Signal Processing Magazine*, 38(2):65–88, 2021. doi: 10.1109/MSP.2020.3023869.

Xin Yuan, Yang Liu, Jinli Suo, Frédo Durand, and Qionghai Dai. Plug-and-Play Algorithms for Video Snapshot Compressive Imaging. *IEEE TPAMI*, 44(10):7093–7111, 2022. doi: 10.1109/TPAMI.2021.3099035.