SPARSE AUTOENCODERS REVEAL TEMPORALLY STABLE MEDICAL SEMANTICS IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse Autoencoders (SAEs) are useful for discovering sparse directions in generative-model activations, but diffusion activations vary strongly across denoising timesteps. An SAE trained on raw multi-timestep activations can therefore mix class-relevant medical structure with noise-stage effects, making the same sparse direction unstable over time. We introduce **Piecewise DoD-SAE**, a simple training recipe that reconstructs scale-normalized local Direction-of-Deviation (DoD) residuals between neighboring latent trajectory states rather than raw hidden states. The objective is motivated by residual-scale diagnostics along the denoising path and is designed to expose sparse directions whose medical responses persist across timesteps. We evaluate temporal stability with top-K latent intersection over union (IoU), image-space Semantic Steering Consistency, and matched classifier scoring of guided images. Across three liver MRI benchmarks, Piecewise DoD-SAE improves cross-timestep stability and steering consistency over SAE baselines under matched protocols. The results support a calibrated claim: piecewise DoD helps SAEs reveal temporally stable medical latent responses in diffusion models.

1 INTRODUCTION

Diffusion models have emerged as a dominant paradigm in generative AI, producing images through an iterative denoising process Ho et al. (2020); Song et al. (2021); Gong et al. (2025). In medical imaging, they are increasingly used for synthesis and data augmentation under limited-data regimes Kazerouni et al. (2023); Alfasly et al. (2025). Yet their internal generation mechanisms remain difficult to inspect, especially when one asks how class-relevant or lesion-related structure is constructed over many denoising steps Dalva and Yanardag (2024); Alblwi et al. (2025); Graikos et al. (2024); Xu et al. (2025). Understanding how diffusion models construct semantics, especially class-relevant semantics, across the multi-step generation process remains a fundamental challenge Park et al. (2026).

Sparse Autoencoders (SAEs) offer a natural tool for such inspection: they decompose model activations into sparse, reusable latent directions and have become standard tools for interpretable feature discovery Gao et al. (2025); Rajamanoharan et al. (2024). Most SAE-based interpretation settings, however, assume that the interpreted input distribution or activation space is effectively fixed, so a single sparse dictionary can be read as explaining a stable representation space. Diffusion models violate this assumption. Their activation geometry is strongly timestep-varying Kim et al. (2025); Tinaz et al. (2025), so the same sparse direction may reflect a mixture of class-relevant structure and noise-stage effects if an SAE is trained directly on raw multi-timestep activations. Prior work therefore differs in how it exposes, conditions on, or mitigates timestep dependence: some methods pool representations across timesteps into one dictionary Surkov et al. (2025); Cywiński and Deja (2025); Bohacek et al. (2026); Ijishakin et al. (2024); others use supervised concept alignment Cassano et al. (2026); He et al. (2026), time-conditioned SAEs Huang et al. (2026), or gradient-based tooling around sparse codes Shi et al. (2025). These strategies are useful, but they leave open how to learn a fixed sparse dictionary whose selected directions can be compared and intervened on across timesteps. We review related diffusion-SAE and controllable-direction methods in Appendix A.

Our core view is that diffusion interpretability should prioritize recovery: how a diffusion model maps noised states at different timesteps back toward a clean sample. Building on the Direction of Deviation (DoD) perspective of Beizaee et al. (2025), we interpret stable sparse directions through recovery-aligned deviations rather than raw timestep-specific activations. We first relate latent deviations

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

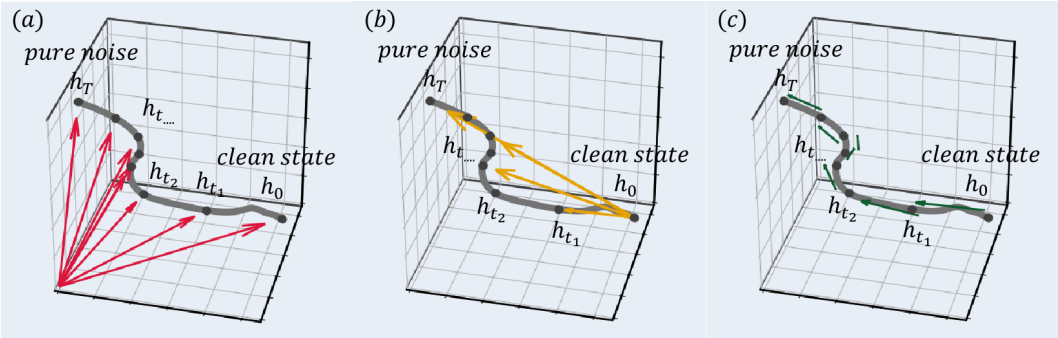


Figure 1: Training geometries between the *Standard SAE* and two variants of our method (i.e., *Global DoD-SAE* and *Piecewise DoD-SAE*). Latent activations along the denoising trajectory $(h_0, h_{t_1}, h_{t_2}, \ldots, h_T)$ evolve from clean states toward pure noise. **(a)** A *Standard SAE* treats each original intermediate representation h_t as an independent reconstruction target; **(b)** *Global DoD-SAE* models deviations from the clean state h_0 to every intermediate h_t; **(c)** *Piecewise DoD-SAE* composes the objective from short DoD-aligned steps between neighboring trajectory states.

to the clean reference h_0 as a global baseline (Figure 1(b)). We then use piecewise DoD between neighboring trajectory states (Figure 1(c)), which improves local linearity along the trajectory and is designed to reduce temporal interference and make structural semantics more stable across timesteps.

This setting also makes evaluation nontrivial: a low SAE reconstruction loss does not by itself show that the same medical semantics have been recovered across the denoising trajectory. An evaluation framework should therefore determine, first, whether selected sparse directions remain identifiable as the timestep changes; second, whether steering effects are consistent in image space rather than only reflecting ordinary reconstruction smoothness; and third, whether guided images preserve class-relevant evidence under a matched downstream probe.

Concretely, our contributions are threefold: **1). A piecewise DoD objective for diffusion SAE training.** We introduce *Piecewise DoD-SAE*, which trains SAEs on scale-normalized local latent residual directions to reduce timestep entanglement and improve extraction of trajectory-consistent sparse responses. **2). An evaluation framework for temporal SAE decomposition in diffusion models.** We evaluate dominant-direction stability with top-K latent intersection over union (IoU), image-space intervention stability with gain-controlled Semantic Steering Consistency (SSC), and class-relevant signal retention with matched classifier scoring. **3). Empirical evaluation with calibrated scope.** Across three liver MRI benchmarks, Piecewise DoD-SAE improves temporal latent-set overlap, gain-controlled steering consistency, and classifier-based preservation of class-relevant signal over SAE baselines under matched protocols. We additionally report lesion case studies to visualize whether the guided changes align with clinically relevant image regions.

2 Background

Diffusion models generate samples through a forward noising process and a learned reverse denoising process. In this section, we formalize the diffusion dynamics and the intermediate representations that motivate our interpretability objective. Let $x_0 \sim q(x_0)$ be a clean image sampled from the real data distribution. The forward diffusion process gradually injects Gaussian noise into x_0 over T discrete timesteps, producing a sequence of noisy states $x_1, \ldots, x_T$, which is defined as a Markov chain:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{1}$$

where $\beta_t \in (0,1)$ represents the predefined variance schedule. Using the reparameterization trick, the marginal distribution at any arbitrary timestep t can be directly sampled as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. During the reverse generation process, a neural network $\epsilon_\theta(x_t, t)$ is trained to approximate the true reverse transition $p(x_{t-1}|x_t)$ by predicting the noise ϵ.

To achieve this denoising objective, the neural network computes a series of internal representations. Let $h_t \in \mathbb{R}^d$ denote the intermediate representation at layer l and timestep t. Let m denote the SAE dictionary width. The SAE is a single linear encoder–decoder with learnable parameters $W_{\text{enc}} \in \mathbb{R}^{m\times d}$, $W_{\text{dec}} \in \mathbb{R}^{d\times m}$, $b_{\text{pre}}, b_{\text{dec}} \in \mathbb{R}^d$, and $b_{\text{enc}} \in \mathbb{R}^m$. A standard SAE computes sparse non-negative codes $f_t \in \mathbb{R}^m$ and decodes them back to $\hat{h}_t \in \mathbb{R}^d$ as

$$f_t = \text{ReLU}(W_{\text{enc}}(h_t - b_{\text{pre}}) + b_{\text{enc}}), \tag{3}$$

$$\hat{h}_t = W_{\text{dec}} f_t + b_{\text{dec}} = b_{\text{dec}} + \textstyle\sum_{i=1}^{m} f_{t,i} d_i \approx h_t. \tag{4}$$

Here $d_i \in \mathbb{R}^d$ denotes the i-th column of W_{dec} (the i-th dictionary atom). However, in diffusion models, h_t encodes highly entangled information: structural semantics and timestep-specific noise statistics co-exist in the same representation. As t changes, both noise scale and semantic granularity drift, so the effective feature basis is non-stationary across timesteps. A single SAE dictionary trained on raw intermediate representations must therefore absorb both class-relevant structure and timestep-dependent activation drift.

3 METHOD

Motivated by the timestep non-stationarity of h_t discussed in §2, we reformulate the SAE recipe to reconstruct semantically relevant directions. In §3.1, we introduce masked forward diffusion and define image-space Direction of Deviation (DoD), yielding a stabilized deviation signal η. In §3.2, we lift DoD to intermediate representations h_t, contrasting a *global* latent residual anchored at h_0 (Figure 1(b)) with a *piecewise* DoD-step target along trajectory segments (Figure 1(c)). Finally, in §3.3, we specify the evaluation protocol used throughout the experiments.

3.1 DIRECTION OF DEVIATION IN IMAGE SPACE

To extract time-consistent features, we formulate the problem as modeling the geometric deviation between representations at an arbitrary timestep t and the clean state at $t = 0$. Inspired by Beizaee et al. (2025), we inject random masking into the forward process so that visible patches remain fixed across timesteps and provide a spatial context, while only masked patches undergo diffusion and carry the timestep-dependent deviation. This explicitly partitions time-varying evolution from time-consistent context across image regions. Given an image sample $x_0 \sim q(x_0)$, we spatially divide x_0 into non-overlapping noisy patches x_0^n and visible patches x_0^v by applying a random mask with masking ratio r_{mask}; note that this random masking is applied only during training. During the forward process, only the noisy patches x_0^n are gradually diffused, while the visible patches x_0^v remain unchanged:

$$q(x_t^k | x_0^k) = \begin{cases} \mathcal{N}(x_t^n; \sqrt{\bar{\alpha}_t} x_0^n, (1-\bar{\alpha}_t) I), & \text{if } k = n \\ x_0^v, & \text{otherwise} \end{cases} \tag{5}$$

Consequently, the visible patches x_t^v provide fixed context across timesteps, whereas the noisy patches x_t^n contain the timestep-dependent residual that is modeled by DoD. We set the diffusion optimization objective to model the Direction of Deviation of the representations at timestep t. For the noisy patches in image space, the standard forward step can be explicitly rewritten to isolate the deviation from x_0^n:

$$x_t^n = x_0^n + \sqrt{1-\bar{\alpha}_t} \left(\epsilon - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} x_0^n \right). \tag{6}$$

In Eq. (6), the second term on the right-hand side is the exact deviation $x_t^n - x_0^n$. Factoring out the schedule-dependent scale $\sqrt{1-\bar{\alpha}_t}$, we define the normalized Direction of Deviation η as

$$\eta = \epsilon - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} x_0^n = \frac{x_t^n - x_0^n}{\sqrt{1-\bar{\alpha}_t}}. \tag{7}$$

This approach naturally encourages the model to spatially segregate deviations across different image regions. The interpretation of DoD (η) is made explicit by the rightmost expression in Eq. (7): it encapsulates the evolutionary trajectory of the representation, shifting the focus from instantaneous noise estimation to target-oriented reconstruction. Importantly, by dividing the difference by the scaling factor $\sqrt{1-\bar{\alpha}_t}$, η explicitly standardizes the magnitude of the time-dependent decay ($\bar{\alpha}_t$) introduced by different timesteps t. Our goal is to capture trajectory-consistent structure beneath these granularity shifts.

3.2 LATENT DOD FOR TIME-CONSISTENT SAE TRAINING

While Eq. (7) defines the deviation in image space, SAE-based interpretability operates on internal denoising representations. We therefore extend DoD to hidden activations. Let h_t^n denote the representation extracted at a fixed denoiser layer from the masked input x_t with timestep embedding t. A direct latent analogue compares this representation with the clean-state representation h_0^n and normalizes the residual by the same diffusion scale:

$$\eta_{hidden} = \frac{h_t^n - h_0^n}{\sqrt{1-\bar{\alpha}_t}}, t > 0. \tag{8}$$

This global latent DoD is an empirical transfer of the pixel-space normalization, not a closed-form identity. It can still be unstable because h_t^n and h_0^n are produced under different timestep embeddings and may lie in different local feature geometries, especially at large t. Rather than anchoring every latent residual to $t = 0$, we use a local trajectory segment as the reconstruction target. For an interval size s, the Piecewise DoD-SAE target compares the current hidden state with the preceding hidden state h_{t-s}^n:

$$\eta_{hidden}^{(s)} = \frac{h_t^n - h_{t-s}^n}{\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-s}}}. \tag{9}$$

The denominator is the diffusion-scale factor for the segment from $t - s$ to t, so the target is a scale-normalized local recovery direction rather than a raw hidden state. Appendix B.2 reports empirical residual-scale diagnostics across multiple denoising-model layers and datasets; these experiments show where the piecewise schedule better tracks hidden-state residuals than the global h_0 schedule, and where the two become similar near image-adjacent decoder layers. Unlike the pixel-space DoD identity, latent DoD is used as a reconstruction target rather than a closed-form diffusion equation; the piecewise form further reduces the mismatch caused by anchoring all hidden-state residuals to h_0. To train a robust dictionary, we uniformly sample intermediate timesteps and train the SAE to reconstruct $\eta_{hidden}^{(s)}$ rather than the raw activation h_t^n. This design encourages sparse codes to explain local recovery directions whose selected coordinates vary more slowly across timesteps. In our experiments, this objective is associated with higher top-K IoU and more category-concentrated latents. All SAE variants in the main comparison use the same frozen denoising model and masked-input protocol; only the SAE reconstruction target changes. For naming consistency, we use three terms throughout the remainder of the paper: **Standard SAE (S-SAE)**, **Global DoD-SAE (G-DoD-SAE)**, and **Piecewise DoD-SAE (P-DoD-SAE)**.

3.3 EVALUATION PROTOCOL

Let $\mathcal{D}$ denote the evaluation dataset. Our evaluation is organized around three concrete questions.

Q1. Are selected SAE directions stable as t changes? We independently form the noisy input x_t at each timestep t, **rank low-entropy SAE latents** (described in Appendix C), and denote by S_t^K the top-K selected latent indices at timestep t after aggregating over the evaluation dataset. Agreement across two timesteps t_i and t_j is measured by IoU in Eq. (10). A consistently high pairwise IoU indicates temporally stable dominant latents at the dataset level. We use this dataset-level form as the primary metric to keep comparisons fair across SAE parameterizations and sparsity mechanisms. In particular, per-sample top-K overlap becomes sensitive to the chosen K when architectures differ: for TopK-style SAEs or highly sparse codes, large K can include many shared inactive or zero-valued coordinates, inflating overlap without necessarily indicating matched semantic features. Dataset-level ranking reduces this dependence on the per-image active budget by comparing dominant coordinates after aggregation under the same selection rule.

$$\mathrm{IoU}(S_{t_i}^K, S_{t_j}^K) = \frac{|S_{t_i}^K \cap S_{t_j}^K|}{|S_{t_i}^K \cup S_{t_j}^K|}. \tag{10}$$

Q2. Are steering effects consistent beyond reconstruction alone? To compare steering behavior across timesteps, we define a sparse-intervention protocol at the SAE hook layer. For steering-based evaluations, we select a fixed dataset-global set S_{global}^K once by aggregating latent scores over the evaluation dataset and sampled timesteps, and use the same set for all timesteps. The timestep-specific sets S_t^K are used for the temporal-overlap diagnostic in Q1. For residual-space SAEs, h_b is computed

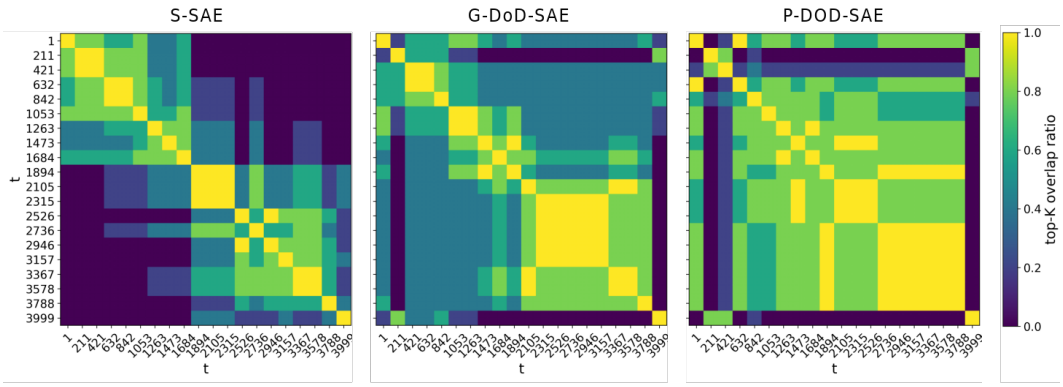


Figure 2: Temporal stability on LLD-MMRI, measured by pairwise IoU of top-K latent sets across timestep pairs. Higher IoU means that two timesteps share more selected latents. P-DoD-SAE maintains stronger cross-timestep overlap under the same $K{=}5$ selection budget.

by an additional frozen-denoiser forward pass at timestep b. Each method first maps the hook activation to the input space used by its SAE, amplifies the selected sparse coordinates in S^K_{global}, maps the decoded change back to the hook activation, and then resumes the denoising-model forward pass: We use the same fixed multiplicative gain r across methods. Given the steered denoiser prediction $\eta(t)$, we define the single-step reconstruction in Eq. (11):

$$\hat{x}_0(t) = x_t - \sqrt{1-\bar{\alpha}_t}\,\eta(t). \tag{11}$$

Using the corresponding steered reconstructions $\hat{x}_0(t_i)$ and $\hat{x}_0(t_j)$, we define Semantic Steering Consistency (SSC) as

$$\begin{aligned} \text{SSC}^m_{i,j} &= \mathbb{E}_{x\sim\mathcal{D}}\left[m(\hat{x}_0(t_i), \hat{x}_0(t_j))\right], \\ & m \in \{\text{SSIM}, \text{LPIPS}\}. \end{aligned} \tag{12}$$

Higher SSC^{SSIM} and lower $\text{SSC}^{\text{LPIPS}}$ indicate stronger cross-timestep image-space consistency under identical steering interventions. We examine SSC as a function of gain r to check whether consistency persists under controlled intervention strength.

Algorithm 1: Method-specific SAE hook steering

Require: $x_t, t, s, S^K_{\text{global}}, r$, method m, frozen denoiser

Cache hook state h_t
if $m \in \{\text{P}, \text{G}\}$ **then**
  $b \leftarrow \max(t-s, 0)$ if $m{=}$P else 0; cache h_b
  $\gamma_t \leftarrow \sqrt{1-\bar{\alpha}_t/\bar{\alpha}_b}$; $z_t \leftarrow (h_t - h_b)/\gamma_t$
else
  $\gamma_t \leftarrow 1$; $z_t \leftarrow h_t$
end if
$f_t \leftarrow \text{Enc}(z_t)$
$f'_{t,i} \leftarrow r f_{t,i}$ if $i \in S^K_{\text{global}}$; else $f'_{t,i} \leftarrow f_{t,i}$
$z'_t \leftarrow \text{Dec}(f'_t)$; $z^0_t \leftarrow \text{Dec}(f_t)$
$h'_t \leftarrow h_t + \gamma_t(z'_t - z^0_t)$
$\eta(t) \leftarrow \text{Denoise}_{h'_t \to out}(x_t, t)$

Q3. Do guided images preserve class-relevant signal? Finally, we evaluate whether SAE-guided images retain class-relevant evidence under matched classifier scoring. Starting from real images, we run the diffusion reverse process and inject the reconstructed signal from the fixed S^K_{global} latent set along the trajectory to obtain guided images. We then use a **guided-only** protocol: classifiers are trained only on guided images, with the number of training samples matched across compared settings, and evaluated by top-1 accuracy and macro-F1 using ConvNeXt Liu et al. (2022), VMamba Liu et al. (2024), Vision Mamba (ViM) Zhu et al. (2024), and OverLoCK Lou and Yu (2025). This classifier-based score tests whether class-relevant information remains recognizable at the image level after steering.

4 Experiments

4.1 Experimental Setup

Datasets. We validate our proposed framework on three clinical liver MRI datasets: LLD-MMRI Lou et al. (2025), HCC-ZSH Sun et al. (2022) and HCC-WCH Pang et al. (2025). These datasets provide disease labels and, for LLD-MMRI case studies, lesion ROI masks that allow us to connect sparse responses to lesion-related evidence rather than only to dataset-level labels.

Model & SAE Configurations. We train the denoising model over $T{=}4000$ diffusion timesteps and use DDIM sampling with a compressed 250-step denoising trajectory for image generation in

Table 1: LLD-MMRI temporal-stability and steering-validity summary. SSC is reported as LPIPS/SSIM; IoU measures top-K latent-set overlap. Lower LPIPS and higher SSIM/IoU are better.

	K=5 SSC / IoU		K=50 SSC / IoU		K=100 SSC / IoU	
Method	LPIPS / SSIM(%)	IoU(%)	LPIPS / SSIM(%)	IoU(%)	LPIPS / SSIM(%)	IoU(%)
SAeUron	49.6± 0.1 / 58.8± 2.6	30.1± 0.1	38.0± 0.2 / 67.1± 0.1	71.6± 0.5	37.9± 0.2 / 67.1± 0.1	71.9± 0.2
DiffLens	36.4± 0.2 / 69.0± 0.4	35.5± 1.3	35.8± 0.1 / 69.6± 0.1	61.8± 0.8	35.8± 0.1 / 69.6± 0.1	71.0± 0.6
G-DoD-SAE	23.1± 1.2 / 74.9± 1.4	72.1± 1.1	17.8± 1.0 / 77.4± 1.3	64.7± 0.4	16.3± 0.7 / 79.5± 0.8	73.1± 0.6
P-DoD-SAE	**19.8**± **0.9** / **77.2**± **0.9**	**88.5**± 1.3	**16.5**± **0.7** / **79.4**± **0.8**	**90.5**± 0.2	**15.7**± **0.3** / **79.8**± **0.6**	**78.7**± 0.2

§4.2. The SAE is trained on representations extracted from the `middle_block.1` layer to capture high-level structural responses. We configure the SAE with a $64\times$ expansion factor and utilize a standard ReLU activation function to compute the sparse non-negative codes. Unless otherwise specified, we set K=5 for the main top-K selections. We compare against SAeUron Cywiński and Deja (2025) and DiffLens Shi et al. (2025); we do not include He et al. (2026); Huang et al. (2026) as quantitative baselines due to the lack of publicly available code.

Training Implementation. The SAE is trained to reconstruct the piecewise DoD target $\eta^{(s)}_{hidden}$ with an interval step of $s = 500$ and a random masking ratio of $r_{mask} = 0.5$. Since the DoD is not defined at $t = 0$, we remove the $t = 0$ timestep when uniformly sampling timesteps during training. The interval is fixed across the liver datasets rather than tuned separately for each benchmark. The SAE is optimized using AdamW with a learning rate of 1×10^{-4} and a batch size of 128. Additional diffusion hyperparameters, extended SAE fields, and the data-split protocol are listed in Appendix F.

4.2 Evaluation results organized by question

We organize the main experimental results around the three questions introduced in §3.3: whether selected SAE directions are temporally stable, whether steering effects remain consistent beyond reconstruction smoothness alone, and whether guided samples preserve class-relevant evidence under matched classifier evaluation.

Q1: Are selected sparse directions stable across timesteps? Table 1 reports the main LLD-MMRI SSC/IoU summary across top-K budgets: P-DoD-SAE improves the main K=5 latent stability to 88.5% IoU, and remains strongest at K=50 and K=100. Figure 2 details the K=5 pairwise IoU matrices, where P-DoD-SAE expands the intermediate-timestep region with high set overlap under the same selection rule.

Q2: Are steering effects consistent beyond reconstruction alone? Figure 3 plots the SSC trajectories across timesteps. P-DoD-SAE and G-DoD-SAE generally stabilize earlier than SAeUron and DiffLens, while their relative performance varies across datasets and timestep ranges. Figure 4 further sweeps the intervention gain on LLD-MMRI under random, high-entropy, and low-entropy top-K latent selection, with no-steer reconstruction consistency shown as a baseline. The gain sweep separates base denoising consistency from intervention-dependent behavior: if SSC were dominated only by the reconstruction trajectory, changing the steering gain would not produce the method-specific separation observed in the curves. Across methods, SSC changes least under random top-K steering, whereas entropy-stratified selection produces larger differences; this pattern suggests that the measured steering consistency depends on which sparse directions are selected, not just on the base reconstruction. To further isolate the intervention from the base reconstruction path, Figure 5 evaluates effect-only consistency. For each timestep, we compute the steering effect $\Delta x_t = \hat{x}_0^{\text{steer}}(t) - \hat{x}_0^{\text{no-steer}}(t)$ and measure cosine similarity between effects across timesteps. This avoids applying perceptual image metrics directly to difference maps while testing whether the intervention direction itself remains aligned. Under the low-entropy setting used for semantic steering, P-DoD-SAE maintains stronger SSC and effect-direction consistency than SAeUron and DiffLens over the useful gain range, supporting the interpretation that its selected latents induce more stable category-associated image-space effects. Figure 6 presents representative LLD-MMRI case studies.

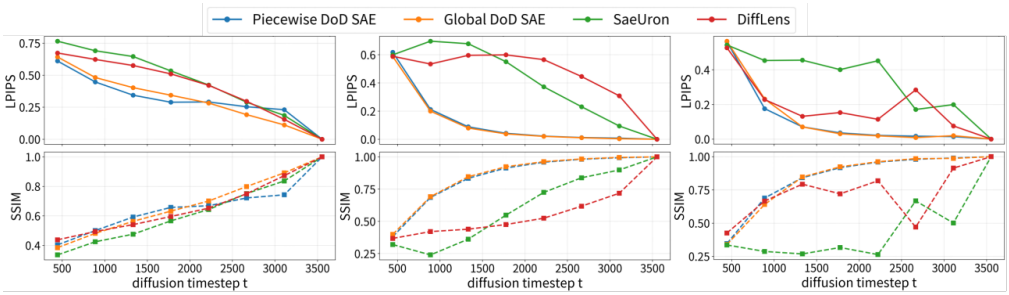


Figure 3: Semantic steering consistency across liver MRI benchmarks: LLD-MMRI (left), HCC-ZSH (middle), and HCC-WCH (right). Each panel compares SSC under the same intervention gain, measuring consistency between steered reconstructions at timestep t and the last sampled denoising step. The DoD-SAE curves stabilize earlier, suggesting that the proposed method reduces timestep-induced entanglement in the selected sparse responses.

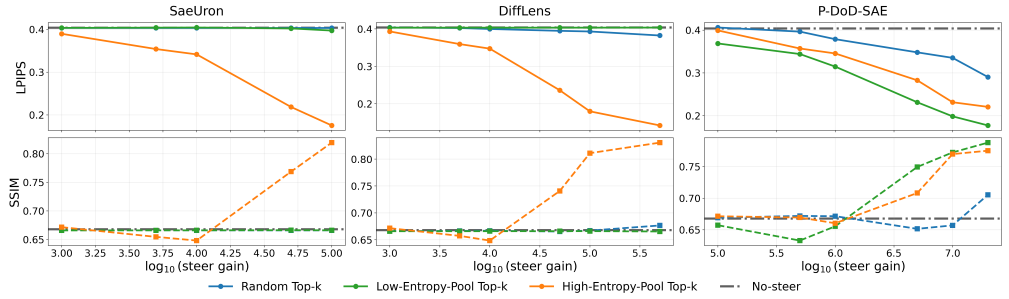


Figure 4: **LLD-MMRI:** SSC as a function of intervention gain for SAeUron Cywiński and Deja (2025), DiffLens Shi et al. (2025), and P-DoD-SAE, under random top-K, high-entropy top-K, and low-entropy top-K latent selection protocols. The no-steer baseline reports reconstruction consistency without SAE intervention. Curves correspond to $\mathrm{SSC}^{\mathrm{SSIM}}$ and $\mathrm{SSC}^{\mathrm{LPIPS}}$ as defined in §3.3.

Table 2 keeps the original HCC-ZSH and HCC-WCH results separate from the LLD-MMRI validity checks. The same ordering largely holds on both additional liver MRI datasets, although the absolute difficulty differs across cohorts.

Table 2: Cross-timestep SSC and mean top-K IoU (K=5) on the two additional liver MRI datasets under the unified evaluation protocol. Lower LPIPS and higher SSIM/IoU are better.

	HCC-ZSH			HCC-WCH		
Method	LPIPS(%)	SSIM(%)	IoU(%, K=5)	LPIPS(%)	SSIM(%)	IoU(%, K=5)
SAeUron	52.5 ± 7.7	51.8 ± 2.3	20.8 ± 1.6	30.4 ± 3.6	50.9 ± 7.3	6.3 ± 0.2
DiffLens	51.2 ± 9.2	54.2 ± 5.2	17.2 ± 0.7	16.6 ± 0.3	78.0 ± 0.5	14.6 ± 1.2
G-DoD-SAE	43.9 ± 0.9	58.3 ± 5.6	47.5 ± 5.1	30.4 ± 4.2	70.0 ± 4.5	25.8 ± 1.5
P-DoD-SAE	**43.1** ± 0.1	**59.4** ± 4.5	**49.8** ± 6.5	**16.1** ± 0.3	**81.4** ± 0.8	**47.4** ± 14.9

Q3: Do guided images preserve class-relevant evidence? Table 3 compares Real-only training with guided-only training across all SAE baselines under this protocol. P-DoD-SAE-only gives the strongest guided-only results in most LLD-MMRI and HCC-ZSH classifier settings, while the HCC-WCH results are more mixed under the main K=5 protocol. On HCC-WCH, although the per-classifier winners vary, P-DoD-SAE has the highest average guided-only score across the four classifiers (74.8 Acc / 72.4 F1), compared with DiffLens-SAE (72.8 / 70.1) and SAeUron-SAE (68.9 / 67.3). This more even HCC-WCH performance suggests that P-DoD-SAE steering preserves class-relevant evidence more stably, rather than benefiting only one downstream architecture. Appendix H provides additional **P-DoD-SAE** steering visualizations alongside the corresponding ground-truth (GT) images.

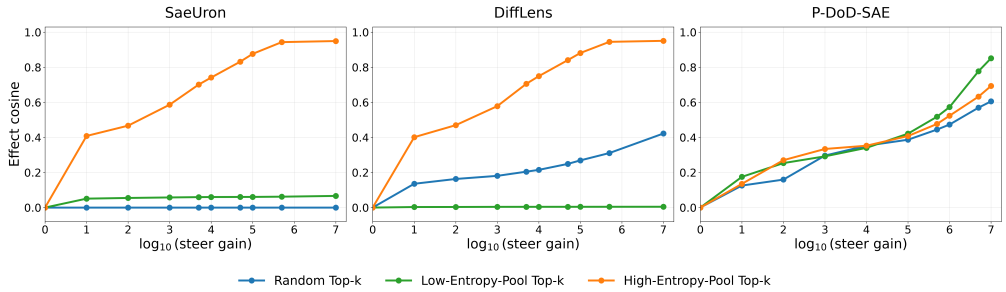


Figure 5: **LLD-MMRI effect-only steering consistency.** For each method and gain, we compute steering effects $\Delta x_t = \hat{x}_0^{\text{steer}}(t) - \hat{x}_0^{\text{no-steer}}(t)$ and report cosine similarity across timesteps. Higher values indicate that the image-space intervention direction remains more stable after removing the no-steer reconstruction baseline.

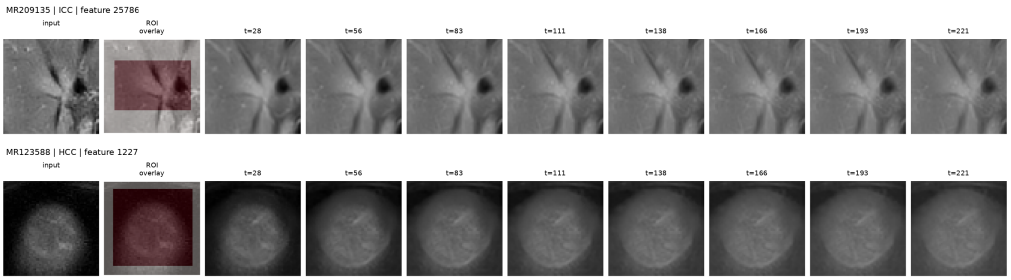


Figure 6: LLD-MMRI case studies for low-entropy P-DoD-SAE latents. The examples support lesion-related responses such as lesion-margin enhancement, heterogeneous tumor texture, and capsule-like boundary changes while preserving coarse anatomy.

4.3 Ablation Studies

To validate which components of our method are necessary, we conduct ablations on two factors: random masking in the diffusion forward process and piecewise normalization. Table 4 summarizes these results on LLD-MMRI. Removing random masking or replacing piecewise normalization with z-score or whiten normalization leads to clear drops in SSIM and IoU compared with the P-DoD-SAE reference, indicating weaker structural consistency and reduced temporal activation-set alignment. Figure 7 further evaluates normalization variants under gain-controlled SSC curves: z-score and whiten normalization favor high-entropy latents more often, while P-DoD-SAE preserves stronger low-entropy steering consistency over the useful gain range.

Table 5 reports step-size ablations. Smaller intervals preserve activation overlap but degrade pixel-space SSC, while larger steps cause a substantial IoU drop, plausibly because widening the piecewise segment strengthens global alignment toward x_0 and attenuates the local piecewise temporal consistency. In summary, random masking primarily contributes to IoU consistency, piecewise normalization is essential for SSC stability, and the step size controls the locality-globality trade-off. Appendix D further provides the corresponding layer-wise check at `output_blocks.7` and evaluates SAE-architecture extensibility by comparing the ReLU and TopK SAE parameterizations under the same temporal-stability protocol.

5 Conclusion

In this work, we presented a **piecewise DoD-SAE** training target, together with a temporal-stability protocol and medical imaging experiments. By shifting the SAE objective from time-entangled representations to piecewise DoD, the method *reduces* timestep-induced entanglement in our setting. Empirical results show more stable cross-timestep behavior, higher SSC, and stronger large-K stability than diffusion SAE baselines under matched protocols. The liver MRI case studies and

Table 3: Downstream liver MRI classification under matched data budgets. In SAE-only settings, images are obtained by feeding real images into the diffusion reverse process with latent steering. Each cell reports **Acc / F1** in %.

Method	Classifier	LLD-MMRI	HCC-ZSH	HCC-WCH
Real	ConvNeXt	69.7 ± 2.28 / 69.1 ± 2.49	64.7 ± 6.18 / 64.0 ± 6.74	67.7 ± 6.18 / 67.3 ± 6.03
	VMamba	64.6 ± 2.64 / 64.6 ± 2.68	85.5 ± 3.11 / 81.2 ± 3.76	67.2 ± 7.93 / 66.7 ± 7.77
	ViM	67.6 ± 2.27 / 64.5 ± 1.81	74.1 ± 2.51 / 67.5 ± 1.18	72.5 ± 8.14 / 72.7 ± 8.14
	OverLoCK	69.5 ± 1.58 / 69.6 ± 1.50	80.0 ± 3.46 / 77.1 ± 1.21	72.2 ± 2.55 / 69.5 ± 4.16
S-SAE	ConvNeXt	71.7 ± 1.28 / 71.6 ± 1.47	66.9 ± 1.36 / 68.0 ± 1.54	67.9 ± 3.29 / 64.6 ± 4.07
	VMamba	70.5 ± 1.54 / 70.5 ± 1.62	82.5 ± 6.94 / 79.0 ± 9.86	72.4 ± 5.31 / 67.2 ± 8.15
	ViM	65.7 ± 4.21 / 65.5 ± 4.43	82.5 ± 2.42 / 82.1 ± 2.45	69.2 ± 8.18 / 67.0 ± 9.60
	OverLoCK	69.5 ± 2.12 / 69.6 ± 2.25	76.1 ± 5.44 / 74.8 ± 5.43	68.3 ± 7.53 / 65.8 ± 6.70
P-DoD-SAE	ConvNeXt	71.2 ± 1.95 / 71.8 ± 1.94	**72.2** ± 4.82 / **69.6** ± 3.23	72.4 ± 6.69 / 69.8 ± 5.98
	VMamba	**71.1** ± 2.91 / **71.2** ± 3.06	**88.8** ± 3.51 / **86.2** ± 4.22	**77.5** ± 8.24 / **74.0** ± 9.38
	ViM	**71.8** ± 2.32 / **71.9** ± 2.32	80.0 ± 3.57 / 81.9 ± 3.81	74.2 ± 5.80 / **72.9** ± 5.27
	OverLoCK	**73.0** ± 1.94 / **73.1** ± 2.65	**85.0** ± 6.65 / **84.4** ± 6.09	74.9 ± 5.70 / 72.9 ± 5.34
DiffLens-SAE	ConvNeXt	72.3 ± 1.67 / **72.5** ± 1.73	67.2 ± 4.08 / 68.6 ± 4.74	**76.8** ± 4.47 / **76.4** ± 4.12
	VMamba	70.2 ± 2.68 / 70.1 ± 2.84	83.6 ± 6.04 / 81.5 ± 4.38	66.8 ± 3.72 / 61.5 ± 4.49
	ViM	67.0 ± 1.92 / 67.0 ± 1.84	79.1 ± 6.02 / 78.1 ± 5.47	**74.4** ± 10.10 / 72.7 ± 8.81
	OverLoCK	70.7 ± 2.74 / 70.4 ± 2.83	75.6 ± 1.61 / 75.3 ± 4.57	73.3 ± 7.15 / 69.6 ± 10.45
SAeUron-SAE	ConvNeXt	**72.4** ± 2.51 / 72.4 ± 2.58	58.6 ± 5.07 / 57.9 ± 5.51	67.7 ± 5.59 / 63.1 ± 5.78
	VMamba	70.0 ± 3.54 / 69.9 ± 3.65	80.2 ± 6.59 / 77.5 ± 5.79	65.6 ± 2.14 / 65.8 ± 2.30
	ViM	64.9 ± 1.84 / 64.9 ± 1.83	**84.4** ± 7.26 / **83.6** ± 7.10	67.0 ± 4.81 / 66.6 ± 4.44
	OverLoCK	68.8 ± 1.17 / 68.5 ± 1.83	82.2 ± 6.77 / 82.8 ± 5.99	**75.3** ± 7.35 / **73.5** ± 6.70

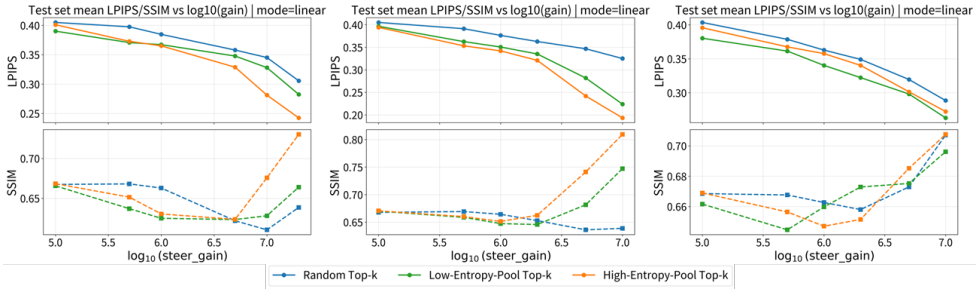


Figure 7: **Normalization ablation on LLD-MMRI:** SSC-gain curves under three alternative normalization variants (zscore, none, and whiten), each evaluated with random top-K, high-entropy top-K, and low-entropy top-K. To ensure fair comparison across normalizations, the gain discussion range is aligned by matched degradation intervals on the random top-K reference curve; e.g., under the random top-K setting, the LPIPS range consistently drops from 0.4 to 0.3.

Table 4: Component ablation on LLD-MMRI.

Setting	LPIPS(%)	SSIM(%)	IoU(%)
w/o mask	20.1	76.3	71.4
z-score norm	32.8	62.8	18.2
whiten norm	22.3	74.7	15.2
no norm	26.3	69.6	84.4
P-DoD-SAE	**19.8**	**77.2**	**88.5**

Table 5: Step-size ablation on LLD-MMRI.

Step	LPIPS(%)	SSIM(%)	IoU(%)
$s = 100$	36.1	60.9	85.9
$s = 200$	29.1	68.4	86.2
$s = 500$	19.8	77.2	88.5
$s = 1000$	41.9	58.8	65.7
$s = 2000$	22.5	75.9	65.6

matched classifier scoring further indicate that the resulting steering effects preserve class-relevant image evidence across classifier architectures.

AI USE STATEMENT

In this work, we used generative AI tools only for light language editing, readability polishing, manuscript formatting, and LaTeX compilation debugging. We have not used generative AI tools for generating synthetic data, formulating mathematical claims, assisting with proofs, designing the method, running experiments, or interpreting experimental results. The authors have reviewed all AI-assisted work 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 studies interpretability and steering behavior for medical-image diffusion models on previously documented liver MRI benchmarks. It does not claim regulatory clearance or clinical deployment readiness. The method may help audit generative models used in clinical-imaging research, but incorrect interpretation of steered outputs or attribution maps could mislead downstream users. Any use with patient data should follow the governance, privacy, and access restrictions of the corresponding datasets and institutions.

REPRODUCIBILITY STATEMENT

Sections 3 and 3.3 define the proposed objective and evaluation protocol. Section 4.1 specifies the main datasets, model configuration, SAE setup, and training choices. Appendix F reports diffusion and SAE hyperparameters, compute details, and patient-disjoint split protocol, while Algorithms 2 and 3 summarize the training procedures. Upon acceptance, we will publicly release the code and scripts needed to reproduce the reported experiments.

REFERENCES

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Nanxu Gong, Zijun Li, Sixun Dong, Haoyue Bai, Wangyang Ying, Xinyuan Wang, and Yanjie Fu. Sculpting Features from Noise: Reward-Guided Hierarchical Diffusion for Task-Optimal Feature Transformation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Amirhossein Kazerouni et al. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

Saghir Alfasly, Wataru Uegami, Md Enamul Hoq, Ghazal Alabtah, and Hamid Tizhoosh. Semantic and Visual Crop-Guided Diffusion Models for Heterogeneous Tissue Synthesis in Histopathology. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Abdalrahman Alblwi, Qile Wang, Norbert Zolek, Matthew Louis Mauriello, and Kenneth Barner. Reliable or Risky? Assessing Diffusion Models for Biomedical Data Generation. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance* (NeurIPS 2025 workshop).

Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned Representation-Guided Diffusion Models for Large-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8532–8542, 2024.

Meilong Xu, Saumya Gupta, Xiaoling Hu, Chen Li, Shahira Abousamra, Dimitris Samaras, Prateek Prasanna, and Chao Chen. TopoCellGen: Generating Histopathology Cell Topology with a Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20979–20989, 2025.

Yong-Hyun Park, Chieh-Hsin Lai, Satoshi Hayakawa, Yuhta Takida, Naoki Murata, WeiHsiang Liao, Woosung Choi, Kin Wai Cheuk, Junghyun (Tony) Koo, and Yuki Mitsufuji. Concept-TRAK: Understanding How Diffusion Models Learn Concepts through Concept-Level Attribution. In *International Conference on Learning Representations (ICLR)*, 2026.

Yusuf Dalva and Pinar Yanardag. NoiseCLR: A Contrastive Learning Approach for Unsupervised Discovery of Interpretable Directions in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24209–24218, 2024.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025.

Senthooran Rajamanoharan et al. Improving dictionary learning with gated sparse autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Farzad Beizaee, Gregory A. Lodygensky, Christian Desrosiers, and Jose Dolz. Correcting deviations from normality: A reformulated diffusion model for multi-class unsupervised anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 19088–19097, 2025.

Bartosz Cywiński and Kamil Deja. SAeUron: Interpretable Concept Unlearning in Diffusion Models with Sparse Autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-Step is Enough: Sparse Autoencoders for Text-to-Image Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Maty Bohacek, Thomas Fel, Maneesh Agrawala, and Ekdeep Singh Lubana. Uncovering Conceptual Blindspots in Generative Image Models Using Sparse Autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ayodeji Ijishakin, Ming Ang, Levente Baljer, Daniel Tan, Hugo Fry, Ahmed Abdulaal, Aengus Lynch, and James H. Cole. H-Space Sparse Autoencoders. In *NeurIPS 2024 Workshop on Safe Generative AI*, 2024.

Enrico Cassano, Riccardo Renzulli, Marco Nurisso, Mirko Zaffaroni, Alan Perotti, and Marco Grangetto. SAEmnesia: Erasing Concepts in Diffusion Models with Supervised Sparse Autoencoders. In *Forty-third International Conference on Machine Learning (ICML)*, 2026.

Zhenghao He, Guangzhi Xiong, Boyang Wang, Sanchit Sinha, and Aidong Zhang. CASL: Concept-Aligned Sparse Latents for Interpreting Diffusion Models. *arXiv preprint arXiv:2601.15441*, 2026.

Victor Shea-Jay Huang, Le Zhuo, Yi Xin, Zhaokai Wang, Fu-Yun Wang, Yuchi Wang, Renrui Zhang, Peng Gao, and Hongsheng Li. TIDE: Temporal-Aware Sparse Autoencoders for Interpretable Diffusion Transformers in Image Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40(1), pages 435–443, 2026.

Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, and Kan Ren. Dissecting and Mitigating Diffusion Bias via Mechanistic Interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8192–8202, 2025.

Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4659–4669, 2025.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and Evolution of Interpretable Concepts in Diffusion Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Calvin Yeung, Prathyush Poduval, Ali Zakeri, Zhuowen Zou, and Mohsen Imani. Residualized Temporal Sparse Autoencoders for Interpreting Diffusion Models. *arXiv preprint arXiv:2605.27813*, 2026.

Philipp Wesp, Robbie Holland, Vasiliki Sideri-Lampretsa, and Sergios Gatidis. Sparse Autoencoders for Interpretable Medical Image Representation Learning. *arXiv preprint arXiv:2603.23794*, 2026.

Krishna Kanth Nakka. Mammo-SAE: Interpreting Breast Cancer Concept Learning with Sparse Autoencoders. *arXiv preprint arXiv:2507.15227*, 2025.

Nhat Minh Le, Ciyue Shen, Neel Patel, Chintan Shah, Darpan Sanghavi, Blake Martin, Alfred Eng, Daniel Shenker, Harshith Padigela, Raymond Biju, Syed Ashar Javed, Jennifer Hipp, John Abel, Harsha Pokkalla, Sean Grullon, and Dinkar Juyal. Learning biologically relevant features in a pathology foundation model using sparse autoencoders. *arXiv preprint arXiv:2407.10785*, 2024.

Alireza Makhzani and Brendan Frey. k-Sparse Autoencoders. In *International Conference on Learning Representations (ICLR)*, 2014.

Meng Lou, Hanning Ying, Xiaoqing Liu, Hong-Yu Zhou, Yuqin Zhang, and Yizhou Yu. SDR-Former: A Siamese Dual-Resolution Transformer for Liver Lesion Classification Using 3D Multi-Phase Imaging. *Neural Networks*, 185:107228, 2025.

Bao-Ye Sun, Pei-Yi Gu, Ruo-Yu Guan, Cheng Zhou, Jian-Wei Lu, Zhang-Fu Yang, Chao Pan, Pei-Yun Zhou, Ya-Ping Zhu, Jia-Rui Li, et al. Deep-learning-based analysis of preoperative MRI predicts microvascular invasion and outcome in hepatocellular carcinoma. *World Journal of Surgical Oncology*, 20(1):189, 2022.

Shuting Pang, Yidi Chen, Xiaoshuang Shi, Rui Wang, Mingzhe Dai, Xiaofeng Zhu, Bin Song, and Kang Li. Interpretable 2.5 D network by hierarchical attention and consistency learning for 3D MRI classification. *Pattern Recognition*, 164:111539, 2025.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9592–9600, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A Comprehensive Benchmark for Sparse Autoencoders in Language Model Interpretability. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pages 29223–29264, 2025.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 103031–103063, 2024.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 62429–62442, PMLR, 2024.

Meng Lou and Yizhou Yu. OverLoCK: An overview-first-look-closely-next ConvNet with context-mixing dynamic kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 128–138, 2025.

A ADDITIONAL RELATED WORK

SAEs for diffusion-model interpretability. Sparse autoencoders (SAEs) matured as a mechanistic interpretability tool in language modeling Gao et al. (2025); Rajamanoharan et al. (2024) and have more recently been ported to diffusion-based generators. Representative lines train SAEs on internal U-Net activations for causal editing, steering, or unlearning Surkov et al. (2025); Cywiński and Deja (2025); use SAE-derived concept dictionaries to audit systematic failures of generative models Bohacek et al. (2026); and study sparse directions in latent h-spaces where localized interventions yield interpretable feature changes Ijishakin et al. (2024). Parallel threads tighten concept–feature associations through supervision for erasure or semantic alignment Cassano et al. (2026); He et al. (2026), or pair k-sparse codes with gradient-based attributions for targeted debiasing Shi et al. (2025). Beyond SAE-centric pipelines, related work also discovers steerable directions via contrastive objectives in noise space Dalva and Yanardag (2024) and attributes training-data influence at concept granularity Park et al. (2026).

SAE parameterization and evaluation. Fixed-budget k-sparse autoencoders provide an early basis for TopK-style parameterizations Makhzani and Frey (2014). In language-model interpretability, SAEBench further shows that unsupervised proxy metrics do not always track practical interpretability outcomes, motivating evaluation across complementary criteria Karvonen et al. (2025).

Medical SAE interpretability. In medical imaging, recent SAE studies decompose fixed foundation-model representations into clinically or biologically meaningful directions, including CT/MRI foundation-model embeddings, mammography vision–language features, and pathology representations Wesp et al. (2026); Nakka (2025); Le et al. (2024). These works motivate sparse medical-feature discovery, while our setting differs in that the interpreted representation comes from a denoising trajectory rather than a fixed encoder. Across these settings, internal activations remain high-dimensional and strongly coupled to the noise level, so a single global sparse dictionary trained on raw h_t^n must continually adapt as the denoising trajectory evolves.

Timestep dependence in diffusion representations. Prior work differs in how it exposes or mitigates timestep dependence. A common recipe pools activations across many denoising steps (sometimes spanning the full trajectory) and fits one reconstruction dictionary, trading simplicity for entanglement with t-varying statistics Cywiński and Deja (2025); Surkov et al. (2025); Bohacek et al. (2026). Another strand systematically sweeps layers and timesteps, using sparse codes together with probes or causal interventions to characterize when coarse layout versus fine detail emerges Kim et al. (2025); Tinaz et al. (2025). Recent residualized temporal SAEs explicitly model activation trajectories by separating initial activations from residual components not predicted by neighboring-timestep linear dynamics Yeung et al. (2026). Transformer-focused designs instead encode temporal evolution inside the SAE itself Huang et al. (2026). Supervised concept assignment concentrates semantics into a small set of coordinates, which reduces combinatorial search at inference but still leaves the underlying activation geometry coupled to t Cassano et al. (2026); He et al. (2026). Gradient-based tooling remains complementary yet path-sensitive relative to sparse interventions Shi et al. (2025). Our Direction-of-Deviation (DoD) construction (§3; Figure 1) targets a complementary axis by normalizing *piecewise* local latent deviations rather than fitting raw activations or trajectory residual predictors. Despite this growing toolbox, comparatively few studies foreground a *single* time-consistent sparse basis for standard U-Net diffusers as the central design goal rather than as an analysis by-product, and publicly released training or evaluation code remains uneven across papers—which complicates rigorous benchmarking and apples-to-apples comparisons at the intersection of diffusion models and SAEs. These gaps motivate our explicit geometric reframing.

B Detailed derivation of the direction of deviation (DoD)

In this appendix, we provide a step-by-step mathematical derivation for the Direction of Deviation (DoD) formulation introduced in §3. We start with the exact derivation in the pixel space and subsequently extend it to the piecewise formulation in the latent space.

B.1 Derivation in the image space

Recall the marginal distribution of the forward diffusion process at an arbitrary timestep t, defined in Eq. (2):

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{13}$$

Our objective is to geometrically model the deviation of the current representation x_t from the clean reference state x_0. Subtracting x_0 from both sides of the equation yields:

$$x_t - x_0 = (\sqrt{\bar{\alpha}_t} - 1)x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon. \tag{14}$$

To isolate the noise scaling factor and reveal the direction of the semantic drift, we factor out $\sqrt{1-\bar{\alpha}_t}$ and rearrange the terms to obtain the exact formulation of the deviation:

$$x_t - x_0 = \sqrt{1-\bar{\alpha}_t}\left(\epsilon - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}x_0\right). \tag{15}$$

Adding x_0 to both sides yields the expression used in Eq. (6) for the noisy patches. This rearrangement provides a useful perspective shift, casting the diffusion trajectory as a geometric deviation from a baseline state:

$$x_t = \underbrace{x_0}_{\text{Base State}} + \underbrace{\sqrt{1-\bar{\alpha}_t}}_{\text{Magnitude}} \cdot \underbrace{\left(\epsilon - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}x_0\right)}_{\text{Direction}}. \tag{16}$$

In this formulation, the current representation x_t is explicitly modeled as a geometric step taken from the pristine anchor x_0, moving along a direction scaled by a time-dependent amplitude. By normalizing the absolute deviation $x_t - x_0$ by this scaling factor $\sqrt{1-\bar{\alpha}_t}$, we obtain a *normalized* direction component after removing the shared time-dependent scale, yielding the formal definition of DoD (η) in Eq. (7):

$$\eta = \frac{x_t - x_0}{\sqrt{1-\bar{\alpha}_t}} = \epsilon - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}x_0. \tag{17}$$

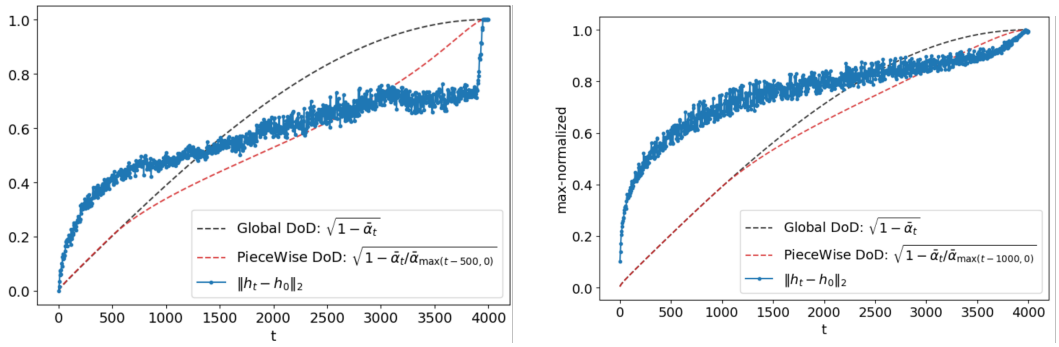
This formulation shows that DoD makes the time-dependent decay magnitude more comparable across timesteps, which yields a more stable reconstruction target for SAEs in our setup.

B.2 Extension to the latent space

It is important to contextualize our formulation within the existing literature. The concept of the Direction of Deviation was initially introduced by Beizaee et al. Beizaee et al. (2025), who reformulated the diffusion objective primarily by focusing on utilizing the isolated direction vector $\left(\epsilon - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}x_0\right)$ for anomaly detection. However, their work did not explicitly address or explore the physical significance of the temporal magnitude factor $\sqrt{1-\bar{\alpha}_t}$.

Empirically, we find that this magnitude tracks non-linear variation in the scale of hidden residuals across diffusion timesteps. In standard epsilon-prediction models, interpreting intermediate U-Net activations h_t^n is difficult because this varying amplitude is correlated with large changes in the local geometry of features as time evolves.

While the strict Gaussian addition applies mathematically to the pixel space, its transfer to hidden activations should be regarded as an empirical modeling choice rather than a closed-form derivation. Figure 8 places the `middle_block.1` latent-scale diagnostic side by side on LLD-MMRI and HCC-ZSH: normalized $\|h_t^n - h_0^n\|_2$ (blue) versus the global scale $\sqrt{1-\bar{\alpha}_t}$ (black dashed) and the piecewise scale $\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{\max(t-s,0)}}$ (red dashed), with diagnostic s=500 on LLD-MMRI and s=1000 on HCC-ZSH. The HCC-ZSH value is used only to visualize the residual-scale diagnostic and is not the interval selected for the main experiments. On LLD-MMRI the piecewise schedule



(a) LLD-MMRI (`middle_block.1`) **(b)** HCC-ZSH (`middle_block.1`)

Figure 8: Latent-scale diagnostic: normalized ℓ_2 latent residuals $\|h_t^n - h_0^n\|_2$ (blue) at U-Net `middle_block.1` versus global DoD scale $\sqrt{1-\bar{\alpha}_t}$ (black dashed) and piecewise scale $\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{\max(t-s,0)}}$ (red dashed). **(a)** LLD-MMRI (s=500): piecewise DoD tracks the residual trend better than global DoD, which motivates using latent residuals for SAE composition. **(b)** HCC-ZSH (s=1000 diagnostic only): the empirical residual progressively approaches the global and piecewise scales at later timesteps while remaining distinct earlier in the trajectory.

tracks the residual trend more closely than the global one; on HCC-ZSH the empirical residual and the two theoretical scales become progressively closer at later timesteps, although they remain separated during the early and intermediate stages.

To show that this alignment is not the outcome of a single convenient hook, we repeat the normalized residual-scale protocol on **LLD-MMRI** at four U-Net `output_blocks` locations with increasing index $\{1, 3, 5, 7\}$ (deeper toward the decoder output, same $s = 500$ diagnostic as in the main text). Figure 9 assembles these curves: moving toward the final decoder stages, the empirical residual profile (blue) increasingly resembles the *global* schedule (black dashed) over intermediate–late t; in contrast, the shallower `output_blocks.1` case tracks the *piecewise* reference (red dashed) more closely, in line with the `middle_block.1` pattern used for SAE training. This layer-wise trend indicates that a naive global h_0 residual normalization becomes progressively more self-consistent at deeper *image-adjacent* features, even though our piecewise construction remains a better match at the `middle_block.1` feature depth where we target time-aggregated, structure-level semantics.

B.3 Approximate piecewise latent DoD construction

Using $t = 0$ as an absolute baseline in the latent space introduces significant non-linear approximation errors when $t \to T$, as the U-Net heavily modulates its feature space conditioned on the timestep embedding. To mitigate this, we partition the trajectory into discrete intervals of size s. Consider the transition from an intermediate state $t - s$ to the current state t. By the Markov property of the variance schedule, the cumulative product of alphas satisfies $\bar{\alpha}_t = \bar{\alpha}_{t-s} \cdot \alpha_{t-s\to t}$, where the transition factor is exactly:

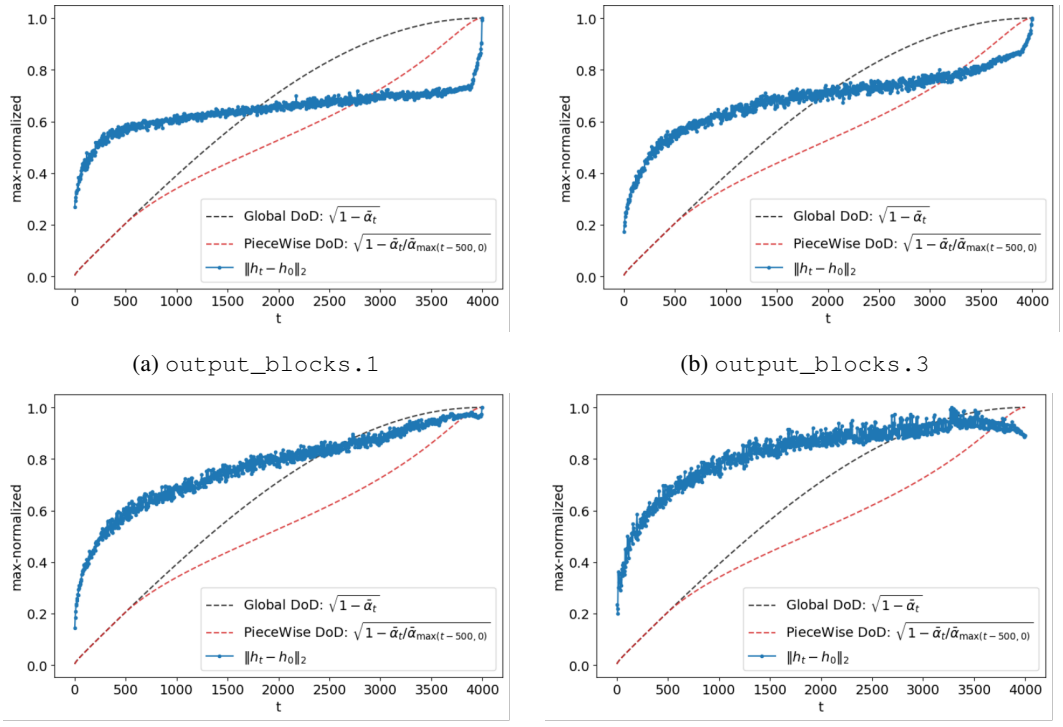
$$\alpha_{t-s\to t} = \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-s}}. \tag{18}$$

Assuming a locally linear feature evolution between temporally adjacent states h_{t-s}^n and h_t^n, the local segment can be approximated with the same schedule factor:

$$h_t^n \approx h_{t-s}^n + \sqrt{1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-s}}}\,\epsilon', \tag{19}$$

where ϵ' represents the effective noise injected during this interval. To define a scale-normalized direction for this trajectory segment, we divide the residual by the localized scaling factor, which directly yields our piecewise DoD optimization target (Eq. (9)):

$$\eta_{hidden}^{(s)} = \frac{h_t^n - h_{t-s}^n}{\sqrt{1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-s}}}}. \tag{20}$$



(a) `output_blocks.1` (b) `output_blocks.3`

(c) `output_blocks.5` (d) `output_blocks.7`

Figure 9: LLD-MMRI, layer-wise **latent-scale alignment** for different output_blocks: normalized $\|h_t^n - h_0^n\|_2$ (blue) vs. the global DoD scale $\sqrt{1-\bar{\alpha}_t}$ (black dashed) and the piecewise scale $\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{\max(t-s,0)}}$ (red dashed, $s = 500$). Deeper `output_blocks` indices (toward the decoder output) show residual scales that more closely track the *global* schedule in the mid-to-late denoising range, whereas shallower output blocks are closer to the piecewise schedule. This contextualizes the `middle_block.1` SAE hook in the main text as operating at a depth where the piecewise target meaningfully departs from a naïve global h_0 baseline.

This interval-based normalization mitigates accumulated trajectory mismatch and, in our experiments, improves SAE temporal stability relative to training on raw h_t^n.

Given the piecewise target $\eta^{(s)}_{hidden}$ above, we train the SAE with the same encoder–decoder parameterization as in §2 (Eqs. (3)–(4)), replacing the activation $h_t^{(l)}$ by $\eta^{(s)}_{hidden}$ and writing the reconstruction in the same space:

$$
\begin{aligned}
f_t &= \mathrm{ReLU}\Big(W_{\mathrm{enc}}\big(\eta^{(s)}_{hidden} - b_{\mathrm{pre}}\big) + b_{\mathrm{enc}}\Big), \\
\hat{\eta}^{(s)}_{hidden} &= W_{\mathrm{dec}} f_t + b_{\mathrm{dec}}, \\
\mathcal{L}_{\mathrm{SAE}} &= \big\|\eta^{(s)}_{hidden} - \hat{\eta}^{(s)}_{hidden}\big\|_2^2 + \lambda\|f_t\|_1.
\end{aligned} \tag{21}
$$

The first term is the reconstruction ℓ_2 loss; the second is an auxiliary ℓ_1 penalty on activations with coefficient λ (Table 10). For G-DoD-SAE, substitute η_{hidden} from Eq. (8) for $\eta^{(s)}_{hidden}$ and decode to $\hat{\eta}_{hidden}$ analogously; for S-SAE, the target and reconstruction reduce to $h_t^{(l)}$ and $\hat{h}_t^{(l)}$ as in Eq. (4).

B.4 Pseudocode for SAE-DoD framework

To facilitate reproducibility, we outline the detailed training procedures for our proposed framework. Algorithm 2 describes the forward diffusion training process utilizing the masked Direction of Deviation (DoD) in the image space, corresponding to §3.1. Algorithm 3 details the extraction of time-consistent concepts via the Latent DoD, covering both G-DoD-SAE and P-DoD-SAE as defined in §3.2. We first train the class-conditioned U-Net with the masked image-space DoD objective

(Algorithm 2; §3.1) and use the resulting checkpoints as frozen generators for all SAE experiments. We then **freeze** θ and train only the SAE on latent DoD-aligned reconstruction targets (Algorithm 3; §3.2), covering S-SAE, G-DoD-SAE, P-DoD-SAE, and SAE baselines.

Algorithm 2 Masked DoD Diffusion Training in Image Space

Require: $q(x_0)$, $\{\beta_t\}_{t=1}^T$, denoiser ϵ_θ, mask ratio r_{mask}

repeat
  Sample $x_0 \sim q(x_0)$, $t \sim \mathcal{U}(1,T)$, $\epsilon \sim \mathcal{N}(0,I)$
  Sample mask M with ratio r_{mask} // visible vs. noisy regions
  Split $x_0 \to (x_0^n, x_0^v)$ by M
  $x_t^n \leftarrow \sqrt{\bar{\alpha}_t} x_0^n + \sqrt{1-\bar{\alpha}_t}\epsilon$
  $x_t \leftarrow x_t^n \cup x_0^v$ // keep visible patches clean
  $\eta \leftarrow \epsilon - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} x_0^n$ // DoD target
  $\hat{\eta} \leftarrow \epsilon_\theta(x_t, t)$
  $\theta \leftarrow \theta - \lambda \nabla_\theta \|\eta - \hat{\eta}\|_2^2$
until converged

Algorithm 3 Time-consistent SAE Training (G-DoD-SAE and P-DoD-SAE)

Require: Frozen U-Net ϵ_θ, dataset $\mathcal{D}$, interval s, SAE params $(W_{enc}, b_{enc}, W_{dec}, b_{dec}, b_{pre})$

repeat
  Sample $x_0 \sim \mathcal{D}$, $t \sim \mathcal{U}(1,T)$, $\epsilon, \epsilon' \sim \mathcal{N}(0,I)$
  Sample mask M with ratio r_{mask} // visible vs. noisy regions
  Split $x_0 \to (x_0^n, x_0^v)$ by M
  if G-DoD-SAE **then**
    $x_t^n \leftarrow \sqrt{\bar{\alpha}_t} x_0^n + \sqrt{1-\bar{\alpha}_t}\epsilon$
    $x_t \leftarrow x_t^n \cup x_0^v$ // keep visible patches clean
    $h_0^n \leftarrow \text{Feat}(x_0),\ h_t^n \leftarrow \text{Feat}(x_t)$
    $\eta \leftarrow \frac{h_t^n - h_0^n}{\sqrt{1-\bar{\alpha}_t}}$
  else
    $t' = \max(t-s, 0)$ // piecewise DoD-SAE
    $x_{t'}^n \leftarrow \sqrt{\bar{\alpha}_{t'}} x_0^n + \sqrt{1-\bar{\alpha}_{t'}}\epsilon$
    $x_{t'} \leftarrow x_{t'}^n \cup x_0^v$
    $\gamma \leftarrow \sqrt{1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t'}}}$
    $x_t^n \leftarrow \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t'}}} x_{t'}^n + \gamma\epsilon'$
    $x_t \leftarrow x_t^n \cup x_0^v$
    $h_{t'}^n \leftarrow \text{Feat}(x_{t'}),\ h_t^n \leftarrow \text{Feat}(x_t)$
    $\eta \leftarrow \frac{h_t^n - h_{t'}^n}{\gamma}$
  end if
  $f \leftarrow \text{ReLU}(W_{enc}(\eta - b_{pre}) + b_{enc})$ // sparse code
  $\hat{\eta} \leftarrow W_{dec} f + b_{dec}$
  $\mathcal{L} \leftarrow \|\eta - \hat{\eta}\|_2^2 + \lambda \|f\|_1$
  Update SAE parameters by gradient descent on $\mathcal{L}$
until converged

C SELECTING AND VALIDATING LOW-ENTROPY SEMANTIC LATENTS

C.1 ACTIVATION STATISTICS FOR CANDIDATE STEERING LATENTS

This appendix explains how we decide which SAE latents are suitable for semantic steering. The selection problem is not simply to find large activations: a useful medical semantic latent should activate strongly enough to be steerable, fire sparsely enough to avoid representing global image style, and concentrate on a small number of diagnostic labels rather than spreading uniformly across classes. We therefore inspect the learned dictionary through three complementary statistics: activation significance, activation frequency, and label entropy. The scatter plot of significance versus frequency, colored by entropy, gives a practical way to identify candidate low-entropy latents before any image-space steering experiment.

We define these dictionary-level statistics below. Let $\mathcal{D}$ denote the evaluation dataset with image–label pairs $(x, y(x))$, where $y(x)$ takes values in a finite label set $\mathcal{C}$. Let $f_{t,i}(x)$ be the i-th SAE latent activation (non-negative code coordinate after the encoder nonlinearity) for image x at diffusion timestep $t \in \{1, \dots, T\}$, extracted at the same U-Net layer as in the main experiments. Figure 10 (scatter plots for every SAE variant) and the pairwise IoU analyses (e.g., Figure 2) use *different* marginalization over t for *significance*; we separate the two protocols below.

Scatter plots (Figure 10). For each SAE backbone, we estimate significance, activation frequency, and label entropy by **jointly** drawing evaluation images $x \sim \mathcal{D}$ and diffusion times $t \sim \mathcal{U}(1, T)$ (i.e., a fresh uniform timestep for each Monte-Carlo sample over the dataset), rather than enumerating every t for every image. *Significance* is the expected activation under this joint sampling:

$$\mathrm{Sig}_i \;=\; \mathbb{E}_{x\sim\mathcal{D},\, t\sim\mathcal{U}(1,T)}\big[\, f_{t,i}(x)\,\big]. \tag{22}$$

Activation frequency uses the same (x, t) distribution with threshold $\tau = 10^{-6}$ throughout:

$$\mathrm{Freq}_i(\tau) \;=\; \mathbb{E}_{x\sim\mathcal{D},\, t\sim\mathcal{U}(1,T)}\big[\, \mathbb{I}\big(f_{t,i}(x) > \tau\big)\,\big], \tag{23}$$

where $\mathbb{I}(\cdot)$ is the indicator function. For *label entropy*, we use the same t-marginalization *within* each image: let $\mathcal{D}_c = \{x \in \mathcal{D} : y(x) = c\}$. For each class $c \in \mathcal{C}$, define

$$\mu_{i,c} \;=\; \sum_{x\in\mathcal{D}_c} \mathbb{E}_{t\sim\mathcal{U}(1,T)}\big[\, f_{t,i}(x)\,\big] \;=\; \frac{1}{T}\sum_{x\in\mathcal{D}_c}\sum_{t=1}^{T} f_{t,i}(x). \tag{24}$$

Let $Z_i = \sum_{c'\in\mathcal{C}} \mu_{i,c'}$. If $Z_i = 0$, set $H_i = 0$; otherwise $p_{i,c} = \mu_{i,c}/Z_i$ and

$$H_i \;=\; -\sum_{c\in\mathcal{C}} p_{i,c}\,\log p_{i,c}, \tag{25}$$

with $0 \log 0 = 0$. Equivalently, $p_{i,c}$ is unchanged if each $\mu_{i,c}$ is scaled by the same positive constant; thus H_i matches the "sum over all (x, t) in class c" construction up to the global factor $1/T$, which drops out after normalization. Lower H_i indicates stronger class concentration under labels; higher H_i indicates a more uniform spread across $\mathcal{C}$. Because the liver MRI benchmarks do not exhibit extreme class imbalance (Appendix F, Figure 22) and our downstream splits are class-stratified, we do not apply additional class-prior reweighting when computing H_i. We use label entropy only as a candidate-selection heuristic before IoU/SSC evaluation; it is not used as a standalone proof that a latent is clinically semantic. For more imbalanced datasets, Eq. (24) can be replaced by class-conditional means to obtain a balanced-prior entropy. In practice, Sig_i and Freq_i for Figure 10 are estimated by Monte Carlo using i.i.d. draws of (x, t) as above; H_i can be computed either from the closed form in Eq. (24) or by the same random-t sampling scheme aggregated per class.

Significance under pairwise IoU. For set-level IoU (Eq. (10)), the top-K active sets $S_t^K(x)$ are formed at a **fixed** denoising time t using the largest activations $\{f_{t,i}(x)\}_i$ at that time (§3.3). When we report or rank latents by **significance** in this pairwise-IoU regime, we use a **timestep-specific** score that averages only over images, holding t fixed:

$$\mathrm{Sig}_i(t) \;=\; \mathbb{E}_{x\sim\mathcal{D}}\big[\, f_{t,i}(x)\,\big]. \tag{26}$$

Thus $\mathrm{Sig}_i(t)$ is computed **separately for each** t, and should not be confused with the time-marginalized Sig_i in Eq. (22) used for Figure 10. This timestep-conditioned score is the notion of "significance at each timestep" referenced when discussing pairwise IoU heatmaps (Figure 2).

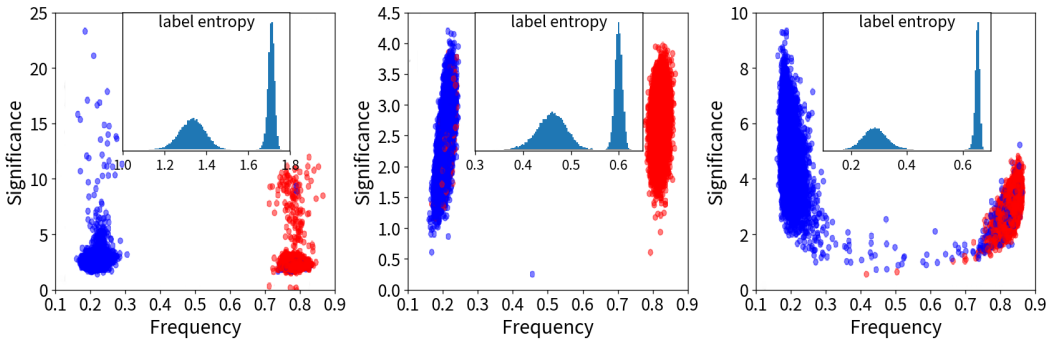


Figure 10: Distributional selection of candidate P-DoD-SAE latents on LLD-MMRI, HCC-ZSH, and HCC-WCH. Each scatter plot places latents by activation significance and activation frequency, with label-entropy histograms overlaid. Low-entropy latents with sufficient significance (blue) are selected as lesion-related candidates for downstream steering and temporal-stability checks.

C.2 Distributional selection of low-entropy latents

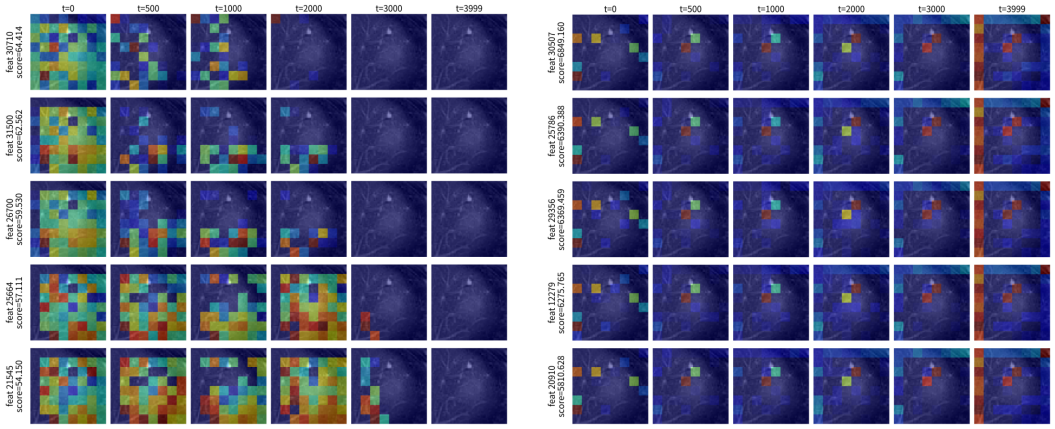
We next apply the statistics above to the learned P-DoD-SAE dictionaries on LLD-MMRI, HCC-ZSH, and HCC-WCH. Figure 10 plots each latent by activation significance and activation frequency, with label entropy overlaid. This view is used to decide which latent family should be guided: we prefer latents in the low-entropy regime because they are more label-concentrated, but we also require non-negligible activation significance so that the latent can produce a measurable intervention. The desired candidates therefore occupy the low-entropy, moderate-to-high-significance part of the scatter plot, rather than the high-frequency/high-entropy region that is more likely to encode common anatomy, scanner style, or global reconstruction factors. This criterion matches the medical prior that lesion-related evidence is spatially localized and label-skewed, not uniformly present across every generated slice.

Guided by this distributional view, the main analyses focus on the low-label-entropy candidate set and then rank latents by timestep-specific significance for top-K selection. The scatter plot itself, however, only establishes label concentration and activation prevalence. It does not by itself prove that the selected latents are temporally insensitive. We therefore perform a second validation step below: map the selected latent activations back to the image plane and inspect whether their activation strength and firing region remain stable across denoising timesteps and under DDIM step compression.

C.3 Heatmap validation across timesteps and DDIM compression

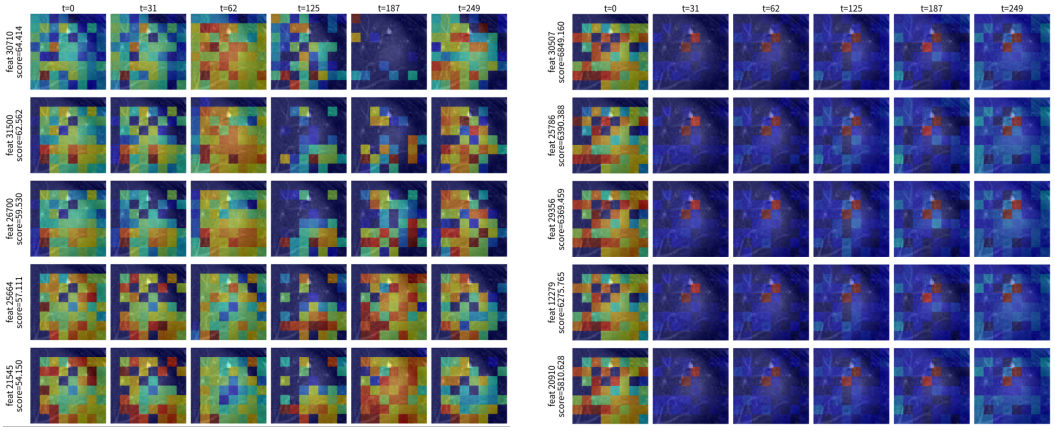
To validate that the low-entropy candidates are not artifacts of a particular timestep, we project their activations back onto the image plane. In these visualizations, heatmap intensity reflects the spatial distribution of the selected latent's activation strength after upsampling to the image resolution, while the persistence of a concentrated response across panels reflects stable activation frequency over the denoising path. We provide **side-by-side S-SAE** and **P-DoD-SAE** activation heatmaps on the *same* generated examples, using the same post-processing for each pair (Figures 11–16).

We first compare the selected latent under the original 4000-step diffusion timeline used for training and a 250-step DDIM sampling timeline. Figures 11 and 12 show that P-DoD-SAE keeps the response concentrated in a similar image region across both schedules, whereas the corresponding S-SAE activations are less visually stable. We further provide class-wise examples in Figures 13–16. Across these examples, the selected low-entropy P-DoD-SAE latents remain spatially concentrated after DDIM compression, supporting the intended interpretation: the distributional criterion identifies latents whose semantics are not strongly tied to a single denoising timestep. These heatmaps are qualitative validation of temporal insensitivity, not a quantitative benchmark against expert segmentations.



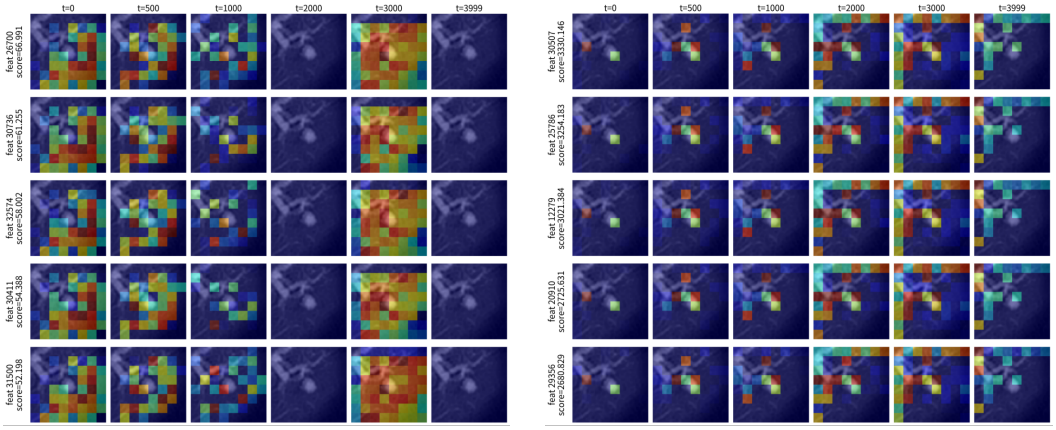
S-SAE P-DoD-SAE

Figure 11: Most time-consistent latent, original training schedule (T=4000). **Left:** S-SAE. **Right:** P-DoD-SAE.



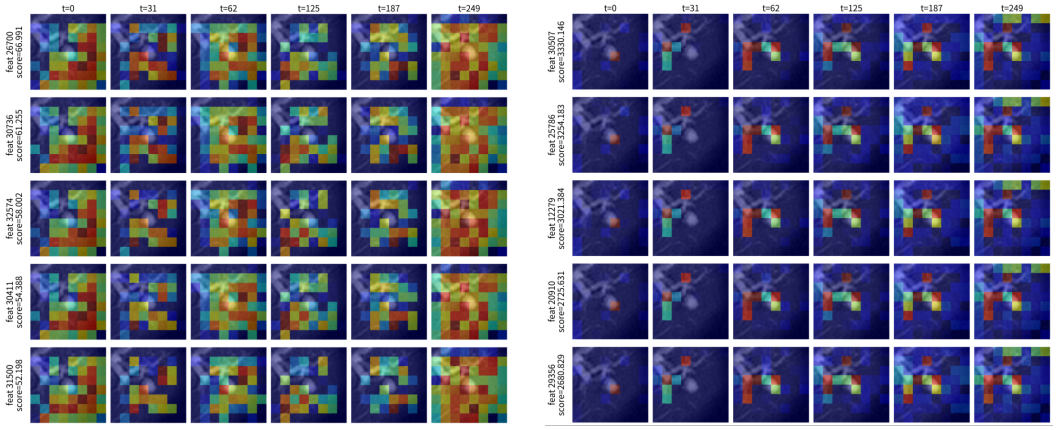
S-SAE P-DoD-SAE

Figure 12: Same latent as Figure 11 under DDIM-compressed sampling (T=250). **Left:** S-SAE. **Right:** P-DoD-SAE.



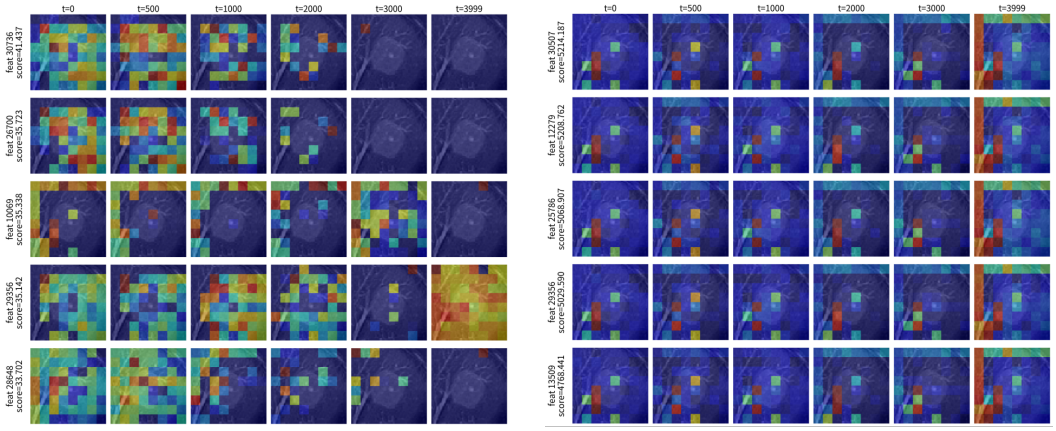
S-SAE P-DoD-SAE

Figure 13: Class-wise example, original training schedule (T=4000). **Left:** S-SAE. **Right:** P-DoD-SAE.



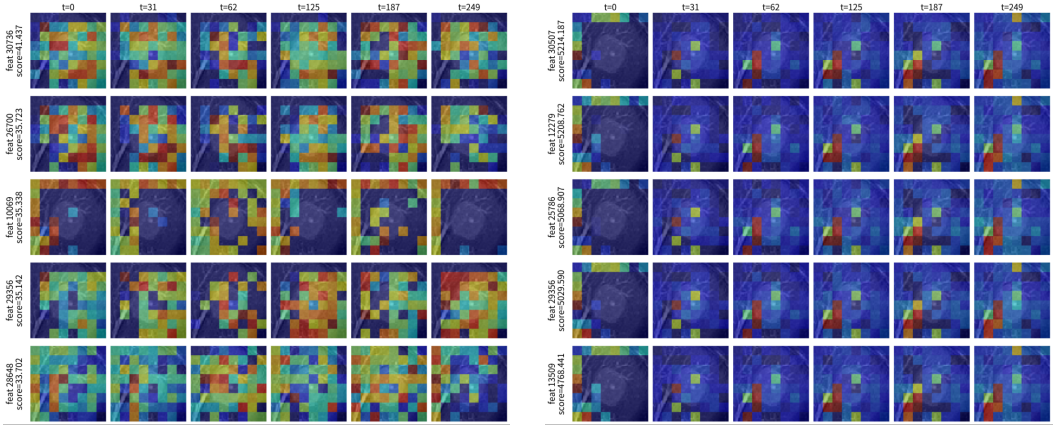
S-SAE P-DoD-SAE

Figure 14: Same class-wise setting as Figure 13 under DDIM-compressed sampling (T=250). **Left:** S-SAE. **Right:** P-DoD-SAE.



S-SAE P-DoD-SAE

Figure 15: Second class-wise example, original training schedule (T=4000). **Left:** S-SAE. **Right:** P-DoD-SAE.



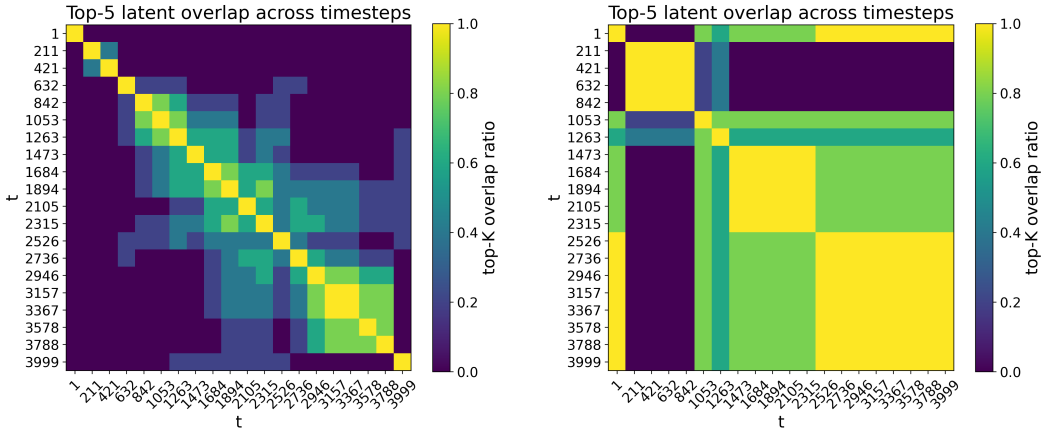
S-SAE P-DoD-SAE

Figure 16: Same class-wise setting as Figure 15 under DDIM-compressed sampling (T=250). **Left:** S-SAE. **Right:** P-DoD-SAE.

D EXTENSIBILITY CHECKS

This section collects experiments that test whether the proposed residual-based objective is tied to a single layer choice, downstream classifier setting, or SAE parameterization. We first evaluate whether the same residual-scale reasoning extends beyond the default hook layer. We then report a large-K downstream classifier check on LLD-MMRI. Finally, we evaluate SAE-architecture extensibility by comparing ReLU and TopK SAE parameterizations on LLD-MMRI under the same temporal-stability protocol.

Extension to other decoder layers. Following the analysis style of Figure 2 and the layer-wise diagnostics associated with Figure 9, we visualize pairwise IoU matrices of top-5 activations across timesteps t at the U-Net layer `output_blocks.7`. Figures 9 and 17 show that deeper decoder layers (e.g., `output_blocks.7`) align more closely with the global schedule, where Global DoD-SAE can perform well in terms of temporal stability. We interpret this as consistent with the known semantic hierarchy in U-Net diffusion models: bottleneck layers encode coarser, more nonlinear feature transitions that benefit from local piecewise approximation, whereas later decoder layers operate in a more linear regime where global normalization suffices. In practice, we recommend selecting the normalization variant based on the residual-scale diagnostic in Figure 9 applied at the target layer.



S-SAE G-DoD-SAE

Figure 17: Pairwise IoU matrices of top-5 activations across timesteps t at `output_blocks.7`.

Large-K downstream classifier check. Table 6 reports an additional LLD-MMRI guided-only classifier sweep with $K{=}100$ selected latents. This check complements the main $K{=}5$ classifier results by testing whether the downstream probe remains comparable when substantially more SAE coordinates are retained during steering.

SAE-architecture extensibility. Figure 18 compares the ReLU and TopK SAE parameterizations on LLD-MMRI. Each panel plots IoU between the top-K active latent set at timestep t and the reference set at $t{=}1$, for $K \in \{10, 50, 100\}$, with series for Standard SAE, Global DoD-SAE, Piecewise DoD-SAE, DiffLens Shi et al. (2025), and SAeUron Cywiński and Deja (2025). Across backbones and budgets, DoD-SAEs tend to maintain higher agreement over mid-trajectory timesteps than Standard SAE and the external baselines, while larger K yields smoother IoU traces with the same qualitative ordering.

Table 6: LLD-MMRI downstream classification with top-K=100 latent steering. Each entry reports mean ± std over folds.

Classifier	Method	Accuracy(%)	AUC(%)	F1(%)
ViM	P-DoD-SAE	**67.95**± 1.99	**74.42**± 2.06	**67.83**± 2.08
	DiffLens	66.93± 1.92	72.45± 1.92	66.33± 2.18
	G-DoD-SAE	64.79± 2.46	70.65± 3.18	64.27± 2.75
	SAeUron	64.62± 3.66	71.28± 3.88	63.88± 4.01
VMamba	P-DoD-SAE	**69.88**± 1.55	**76.50**± 2.10	**69.49**± 1.61
	DiffLens	67.47± 2.62	73.20± 2.97	66.63± 3.52
	G-DoD-SAE	66.66± 1.86	71.02± 1.94	66.29± 1.93
	SAeUron	64.00± 4.12	71.13± 3.89	63.22± 4.35
ConvNeXt	P-DoD-SAE	**69.46**± 1.12	**76.20**± 1.41	**69.27**± 1.14
	G-DoD-SAE	68.52± 1.47	75.40± 2.92	68.41± 1.42
	SAeUron	65.98± 1.80	71.79± 2.83	65.68± 2.07
	DiffLens	62.58± 3.50	69.72± 3.35	62.02± 3.71
OverLoCK	P-DoD-SAE	**67.72**± 1.74	**72.93**± 2.18	**67.49**± 1.93
	SAeUron	65.48± 5.36	69.88± 4.34	65.09± 5.91
	G-DoD-SAE	65.11± 2.20	70.42± 0.96	64.50± 3.16
	DiffLens	63.39± 3.03	69.31± 3.49	62.01± 3.35

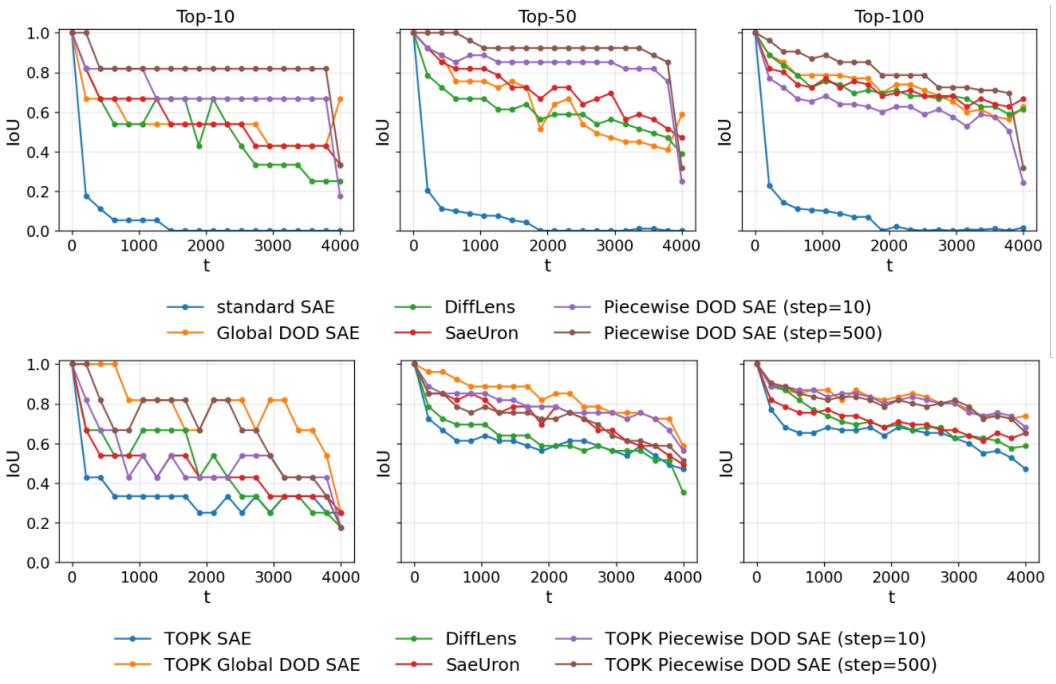


Figure 18: Complete temporal-stability comparison across two SAE backbones. **Top:** ReLU SAE. **Bottom:** TopK SAE. Curves report IoU of top-K active feature sets at timestep t versus reference t=1, with $K \in \{10, 50, 100\}$.

E GENERALIZATION TO MVTEC AD

This section tests diffusion-architecture and natural-image generalization on MVTec AD Bergmann et al. (2019) with an LDM-style backbone Rombach et al. (2022), a short 10-step DDIM sampler, and an additional DiT denoiser. These experiments are not intended to replace a full large-scale natural-image or transformer-backbone benchmark; instead, they check whether the main temporal-stability behavior persists outside the default liver-MRI U-Net setting.

Figure 19 shows a natural-image MVTec chip case in which the selected latent edits a localized abnormal pin pattern while largely preserving the surrounding layout. This case study uses an LDM-style diffusion backbone and a 10-step DDIM sampler, so it changes both the image domain and the sampling/backbone configuration relative to the main liver-MRI experiments.

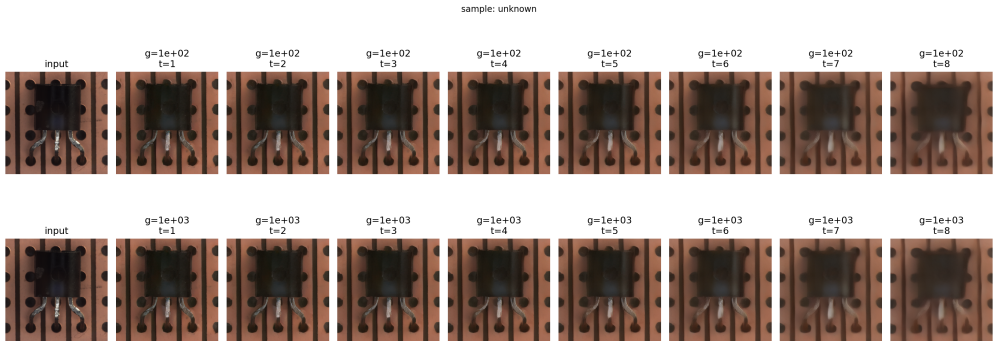


Figure 19: MVTec natural-image case study under an LDM-style backbone and 10-step DDIM sampler. Steering changes a localized abnormal chip-pin pattern while preserving the chip body and surrounding layout.

Table 7 summarizes the MVTec validation. P-DoD-SAE remains competitive under a non-medical anomaly dataset and a much shorter sampler, suggesting that the residual-based objective is not solely an artifact of the original liver-MRI setting.

Table 7: MVTec validation with an LDM-style diffusion backbone and a 10-step DDIM sampler. Metrics are reported in %.

Method / variant	Top-10 IoU(%)	Top-50 IoU(%)	Top-100 IoU(%)	LPIPS(%)	SSIM(%)
P-DoD-SAE, s=5	82.8	96.1	98.0	27.9	69.1
P-DoD-SAE, s=2	**83.4**	**96.2**	**98.0**	**27.9**	**69.1**
SAeUron	80.2	91.2	87.6	29.4	68.1
DiffLens	23.9	60.3	77.7	28.4	68.3

To further separate the objective from the denoiser architecture, we also train an MVTec diffusion model with a DiT denoiser and evaluate the same cross-timestep IoU and SSC protocols on the anomaly-only evaluation split. Table 8 shows that P-DoD-SAE improves temporal latent consistency and steering consistency over the SAE baselines after replacing the U-Net/LDM-style denoiser with a transformer-based backbone.

For interval selection on MVTec, we use latent-trajectory residual-scale diagnostics rather than selecting s only by final SSC or IoU. For hook-layer activations $h_t^\ell(x)$, let $m_\ell = (C_\ell H_\ell W_\ell)^{-1/2}$ and $a_s(t) = \max(t - s, 0)$. We estimate

$$\rho_s^\ell(t) = \mathbb{E}_{x\sim\mathcal{D}}\left[m_\ell \left\| h_t^\ell(x) - h_{a_s(t)}^\ell(x)\right\|_2\right], \tag{27}$$

Table 8: MVTec validation with a DiT denoiser under the same cross-timestep IoU and SSC protocols. Metrics are reported in %.

Method	Top-10 IoU(%)	Top-50 IoU(%)	Top-100 IoU(%)	LPIPS(%)	SSIM(%)
P-DoD-SAE	**56.2**	**62.1**	**67.8**	**35.5**	**54.8**
DiffLens	38.0	50.9	46.5	47.8	51.3
SAeUron	24.2	17.6	17.6	46.7	52.0

and compare it with the DoD normalization factor

$$\gamma_s(t) = \left(1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{a_s(t)}}\right)^{1/2}. \tag{28}$$

Intervals whose empirical residual-scale curve follows the normalization curve across timesteps are treated as plausible candidates before final evaluation. Figures 20 and 21 show this diagnostic on MVTec.

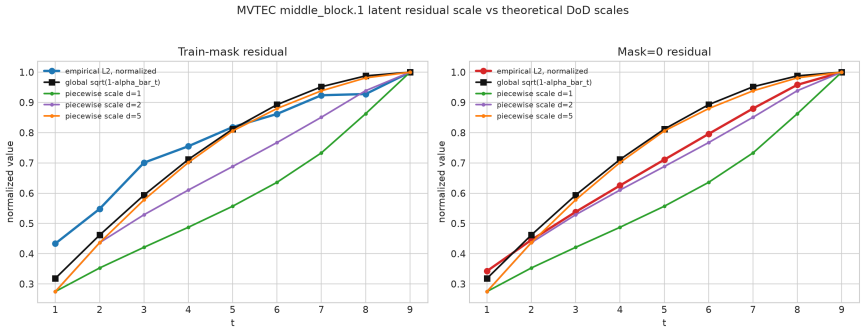


Figure 20: MVTec normalized latent residual-scale curves used to select plausible piecewise intervals before final metric evaluation.

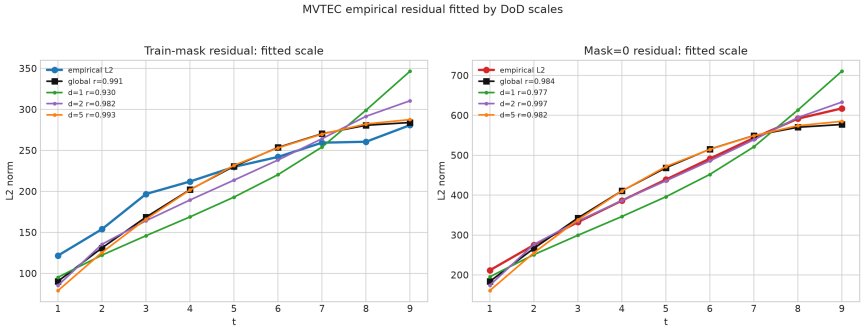


Figure 21: MVTec fitted latent residual-scale curves used to compare empirical hook-layer residuals with the DoD normalization factor.

F DIFFUSION AND SAE TRAINING DETAILS; DATASET SPLITS

Dataset splits and leakage control. For downstream liver MRI classification, we reserve 15% of subjects for a held-out *test* set. On the remaining 85%, we perform patient-disjoint 5-fold cross-validation for training and validation of the image classifiers: no patient appears in more than one fold across train/validation, and test subjects never overlap the folds used for model selection. Splits are constructed with *class-stratified* partitioning so that label distributions are similar across folds and between train/validation and the held-out test set. Table 9 lists the hyperparameters for the class-conditioned U-Net. Figure 22 summarizes the label distributions of the three liver MRI benchmarks used in the main experiments. The datasets contain nonuniform clinical category frequencies, but none of the main benchmarks is dominated by a single class to the degree that label-entropy selection is driven by an extreme class prior alone. Together with class-stratified subject splits, this supports using the unweighted label-entropy score in Appendix C as a practical candidate-selection heuristic for this study. For the DoD / masked-forward *mask ratio*, we adopt the same random-masking ratio strategy as Beizaee et al. (2025); their ablations suggest that results depend only weakly on this choice, so we keep the default value (r_{mask}=0.5 in Table 9) without further tuning. Unless otherwise stated in §4.1, the ReLU SAE follows Table 10.

Split-specific selection protocol. To avoid label leakage, the diffusion model and SAE dictionaries are trained on the training split, while label-informed latent selection and hyperparameter choices are made only on validation subjects. Specifically, label entropy, candidate latent ranking, K, steering gain, piecewise interval s, normalization choice, and hook-layer choice are determined without using held-out test subjects. For downstream classification, latent selection is performed only on the validation split, and generated samples are produced only from the training split. The test split is used solely for final IoU, SSC, and classifier evaluation.

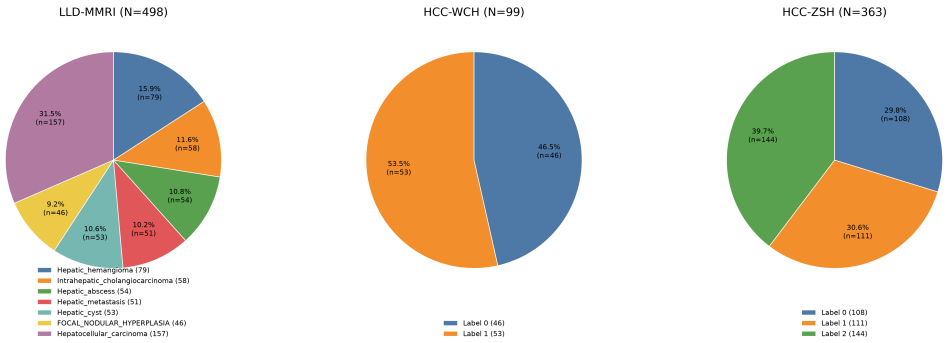


Figure 22: Label distributions for the liver MRI benchmarks used in the main experiments. The plots provide a dataset-level check that the label-entropy latent-selection heuristic is not operating under an extreme single-class-dominated regime; downstream classifiers use class-stratified, patient-disjoint splits.

Evaluation statistics and visualization details. For the IoU results in Tables 1 and 2, the reported standard deviation is computed over random repeats/seeds rather than over individual samples. In each repeat, we compute the Jaccard IoU between the top-K set at each sampled timestep and the reference top-K set, average over sampled timesteps, and then report the mean and standard deviation across repeats. We use five repeats by default, with seeds starting from `noise_seed`=2026 and increasing by one. For the medical diffusion experiments with T=4000, IoU uses 20 approximately uniform evaluation timesteps:

$$\{1, 211, 421, 632, 842, 1053, 1263, 1473, 1684, 1894,\\ 2105, 2315, 2526, 2736, 2946, 3157, 3367, 3578, 3788, 3999\}.$$

The reference timestep is t_{ref}=1; this yields 20 reference pairs including the self-pair $(1, 1)$, or 19 non-trivial pairs. For SSC, the LPIPS/SSIM standard deviation is also over repeats/seeds. Within each repeat, LPIPS/SSIM is first averaged over test samples with batch-size weighting and then averaged over sampled timesteps; the final mean and standard deviation are computed across the five repeats.

SSC uses eight evaluation timesteps: the script first forms $t_{\text{list_num}} + 2$ uniformly spaced candidates and removes the first and last candidates. For feature visualizations, we take the absolute activation of the selected SAE latent at the hooked layer, reshape token activations to the spatial feature grid, bilinearly upsample to the image resolution, and min-max normalize the heatmap for overlay.

Table 9: Hyperparameters for diffusion (U-Net) pretraining.

Hyperparameter	**Value**
U-Net parameters	$\approx$ 174M
Training epochs	3000
Hardware	NVIDIA L40
Initialization checkpoint	`imagenet64_uncond_vlb_100M_1500k`
Diffusion timesteps T	4000
batch size	64(LLD-MMRI) / 32(HCC-WCH) / 24(HCC-ZSH)
Noise schedule	Cosine β-schedule
EMA decay	0.9999
Mask ratio (DoD / masked forward)	0.5
Optimizer	AdamW
Learning rate	5×10^{-5}
Wall-clock training time	$\approx$48 hours (single NVIDIA L40)

Matched SAE baseline configuration. All compared SAE methods use the same hooked denoising-model layer, input dimension d_{in}=512, dictionary size d_{sae}=512 $\times$ 64=32768, maximum training budget of 3000 epochs, and DDIM-250 evaluation sampler. For the default P-DoD-SAE, the ReLU encoder–decoder has approximately 33.6M trainable SAE parameters, uses AdamW, batch size 128, L1 coefficient 10^{-4}, dead-neuron window 64 steps, and dead-neuron threshold 10^{-6}; Appendix D also evaluates a TopK parameterization. Both diffusion and SAE training use sequential single-GPU jobs on NVIDIA L40 hardware.

Table 10: SAE and baseline configurations used for the matched LLD-MMRI experiments. Shared dimensionality and training-budget details are described in the surrounding text.

Config	**SAeUron**	**DiffLens**	**P-DoD-SAE**
SAE type	SAeUron SAE	DiffLens SAE	SparseAE (ReLU / TopK)
Active features	TopK k=32	TopK k=32	ReLU sparse / TopK k=32
Activation	batch TopK	TopK	ReLU+L1 / TopK
Learning rate	4×10^{-4}	1×10^{-4}	1×10^{-4}
Hook layer	`middle_block.1`	`middle_block.1`	`middle_block.1`
Diffusion eval	DDIM-250	DDIM-250	DDIM-250

G DIAGNOSTIC VARIANTS OF THE TEMPORAL METRICS

This appendix reports three diagnostic variants of the main temporal metrics on LLD-MMRI. The first is *per-sample IoU*: instead of selecting the top-K active latents after aggregating scores over the dataset, we compute a top-K set separately for each image and timestep, compare the sets across timesteps for that image, and then average over images. This directly tests whether the same image is represented by stable latent identities over the reverse trajectory. However, at large K, this protocol can be less fair to non-TopK and TopK-style SAEs in different ways. For TopK-style baselines, many unselected coordinates are exactly zero; when K approaches or exceeds the active budget, overlap can be inflated by shared inactive/zero structure rather than by genuinely matched semantic features. For this reason, we treat per-sample IoU as a supplementary diagnostic rather than the main IoU metric.

The second variant is *intervention-effect SSC*. For each timestep, we compute the difference between the steered and no-steer reconstructions, $\Delta x_t = \hat{x}_0^{\text{steer}}(t) - \hat{x}_0^{\text{no-steer}}(t)$, and then evaluate LPIPS/SSIM consistency across these difference images. This better isolates the intervention from the base reconstruction, but Δx_t is not a natural image. Applying perceptual image metrics such as LPIPS and SSIM to difference maps can therefore introduce its own bias, so we do not use this as the primary steering-consistency metric. Table 11 reports these variants only as supporting evidence alongside the main Table 1. For SAeUron, this diagnostic is marked invalid because several steered-minus-no-steer difference maps become nearly constant, yielding zero-variance SSIM terms and NaN aggregate values. Because cosine similarity measures direction but not magnitude, we additionally report the average effect magnitude $\|\Delta x_t\|_2$ under the same gain sweep in Table 12. This checks whether high effect-direction consistency is caused by near-zero interventions.

The third variant is *ROI IoU*, reported only for LLD-MMRI examples with available lesion-region annotations. We compare the support of steering-effect heatmaps with the provided ROI annotations. These ROI annotations are rectangular region labels rather than pixel-accurate lesion segmentation masks, so ROI IoU should be interpreted only as a coarse spatial sanity check: it tests whether guided changes concentrate near annotated lesion regions, not whether the method recovers precise lesion boundaries.

Table 11: Diagnostic variants of LLD-MMRI temporal metrics, reported in %. Per-sample IoU computes top-K sets separately per image before averaging; intervention-effect SSC measures LPIPS/SSIM consistency on steered-minus-no-steer difference maps; ROI IoU compares steering-effect heatmaps with rectangular lesion-region annotations. These diagnostics are not used as primary metrics because large-K per-sample IoU can be affected by shared zero structure in TopK-style SAEs, difference maps are not natural images, and the ROI annotations are rectangular region labels rather than pixel-accurate segmentation masks.

Method	Per-samp. IoU-50(%)	Per-samp. IoU-100(%)	Effect SSC(%)	ROI IoU(%)
SAeUron	87.4	91.8	invalid / NaN	36.1
DiffLens	93.5	96.6	17.9 / 75.2	34.3
P-DoD-SAE	**97.8**	**98.9**	**12.3 / 81.2**	**37.0**

Table 12: Effect-direction consistency and magnitude on LLD-MMRI. Effect cosine measures cross-timestep alignment of Δx_t, while effect L2 reports the average intervention magnitude.

	$\log r$=3		$\log r$=5		$\log r$=7	
Method	Cosine	L2	Cosine	L2	Cosine	L2
SAeUron	0.0577	0.0003	0.0608	0.0049	0.0664	0.1285
DiffLens	0.0041	0.0001	0.0044	0.0014	0.0047	0.0189
P-DoD-SAE	**0.2921**	**0.0026**	**0.4215**	**0.0316**	**0.8517**	**0.3474**

H COMPREHENSIVE VISUALIZATIONS OF LATENT STEERING ACROSS ALL CATEGORIES

To provide the highest level of transparency and facilitate a detailed inspection of our latent steering mechanism, we present full visualization grids for each of the seven pathological categories in the LLD-MMRI dataset Lou et al. (2025).

Let x_{orig} denote the real input slice and let x_{guided} denote the **DoD-guided enhanced image** obtained from x_{orig} using the same reverse-process latent-steering pipeline as §4.2 (top-K reconstructed latent injection along the trajectory). For each grid cell we visualize a **heatmap of the image-space deviation** between x_{guided} and x_{orig}. The *first column* additionally shows the reference anatomy for the category together with the **lesion ROI ground truth** from the dataset, so that large deviations can be read against an independent spatial prior. *Columns 2–5* increase the steering magnitude applied to the top-5 time-consistent SAE latents ($gain \in \{10, 20, 30, 50\}$). *Rows* correspond to different DDIM starting timesteps for the latent intervention ($st \in \{50, 100, 150, 200, 250\}$).

Across all categories, high-magnitude heatmap responses concentrate *within* the ROI-supplied lesion regions rather than drifting to unrelated hepatic parenchyma or vessels, which supports *spatially consistent*, *label-aligned* steering under the §4.2 generator.

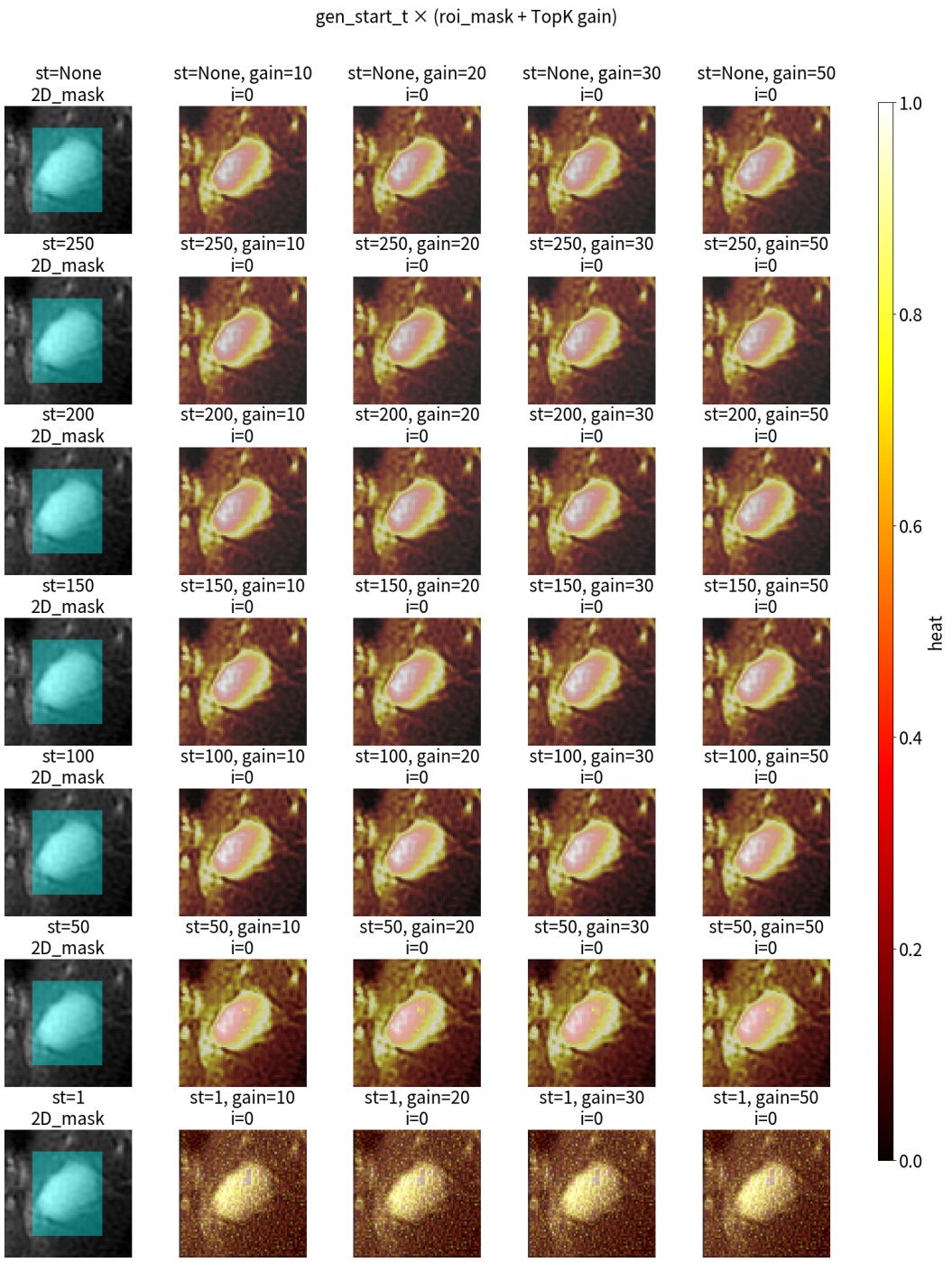


Figure 23: **Hepatic hemangioma (Benign):** heatmaps of image-space difference between DoD-guided x_{guided} and input x_{orig} (§4.2); first column includes anatomy with lesion ROI GT. Rows: $st \in \{50, 100, 150, 200, 250\}$; columns 2–5: $gain \in \{10, 20, 30, 50\}$.

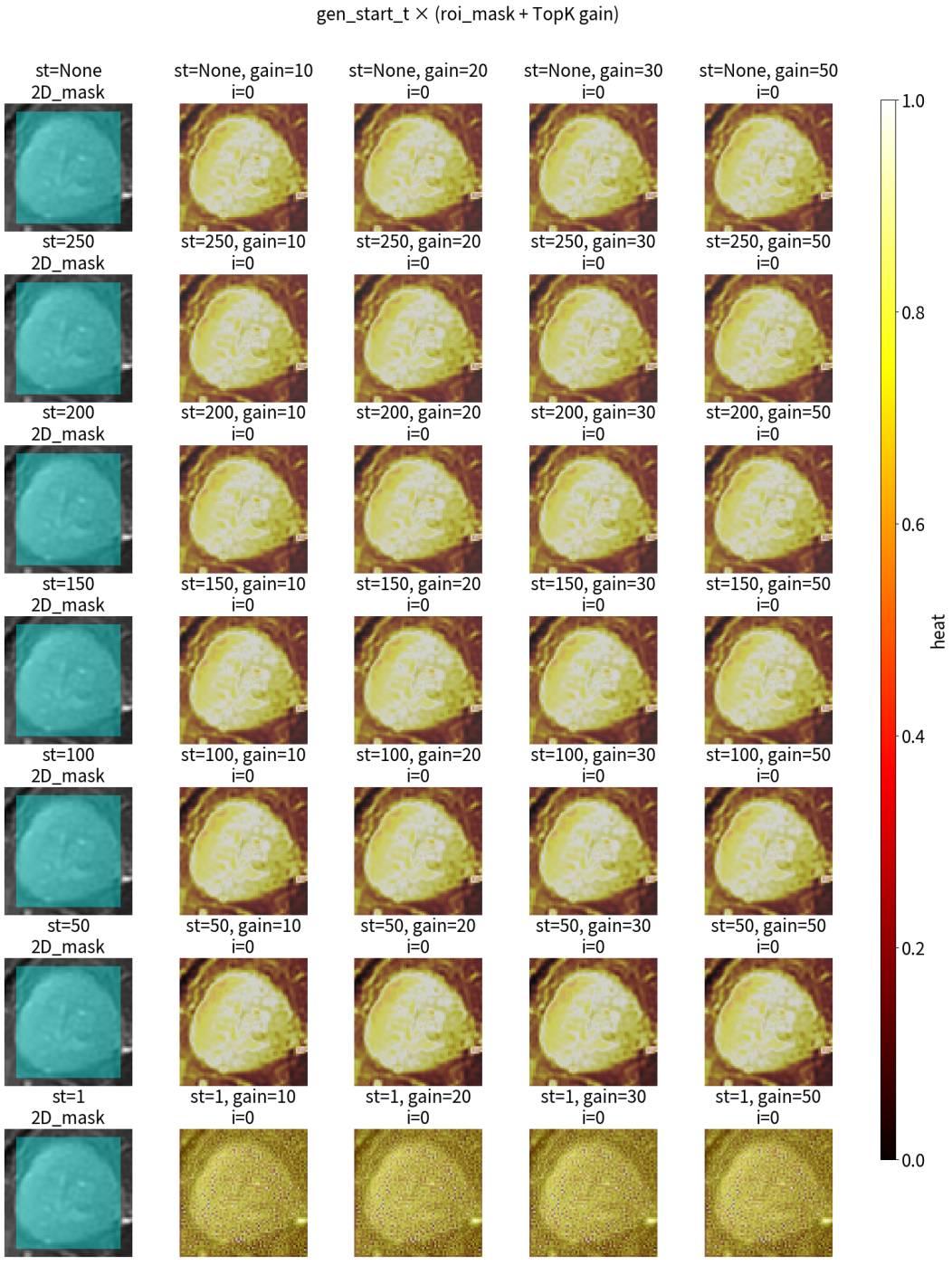


Figure 24: **Intrahepatic cholangiocarcinoma (Malignant):** $|x_{\text{guided}} - x_{\text{orig}}|$ heatmaps with ROI GT in column 1 (§4.2); rows vary st, columns 2–5 increase $gain$.

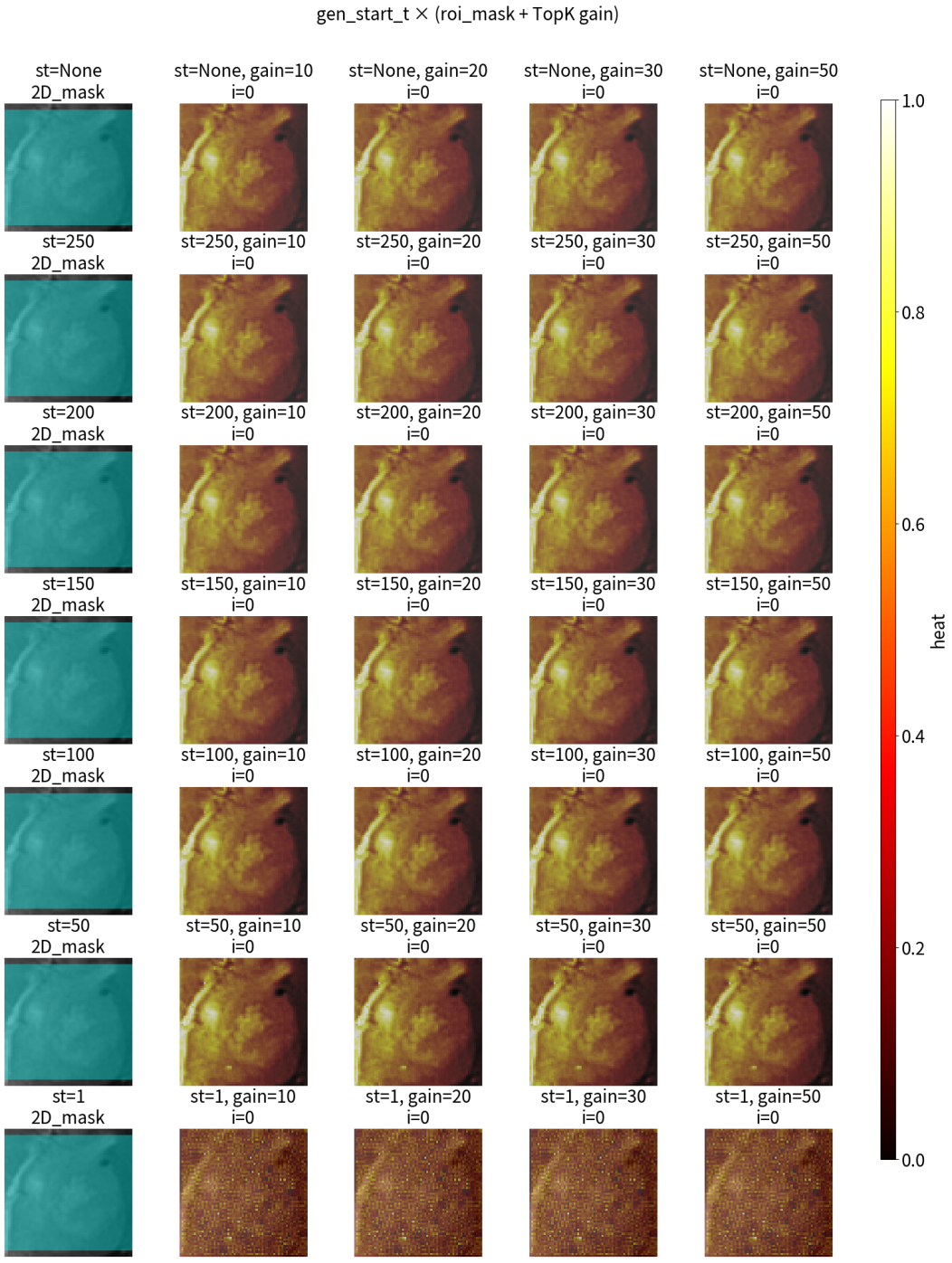


Figure 25: **Hepatic abscess (Benign):** difference heatmaps vs. x_{orig} with ROI GT (col. 1); layout as Figure 23.

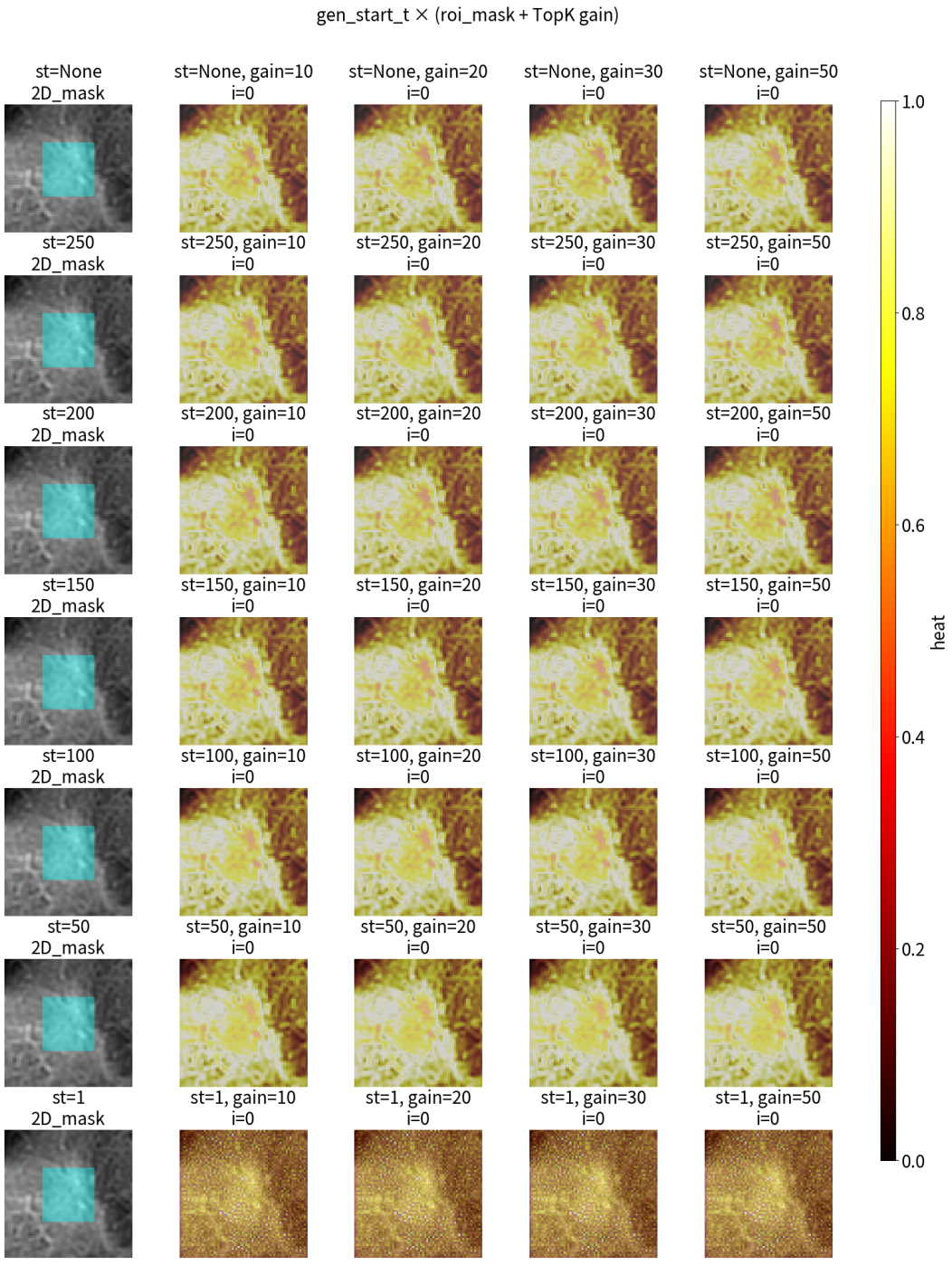


Figure 26: **Hepatic metastasis (Malignant):** guided-vs-original heatmaps (§4.2) with ROI GT reference; deviations concentrate on the metastatic ROI.

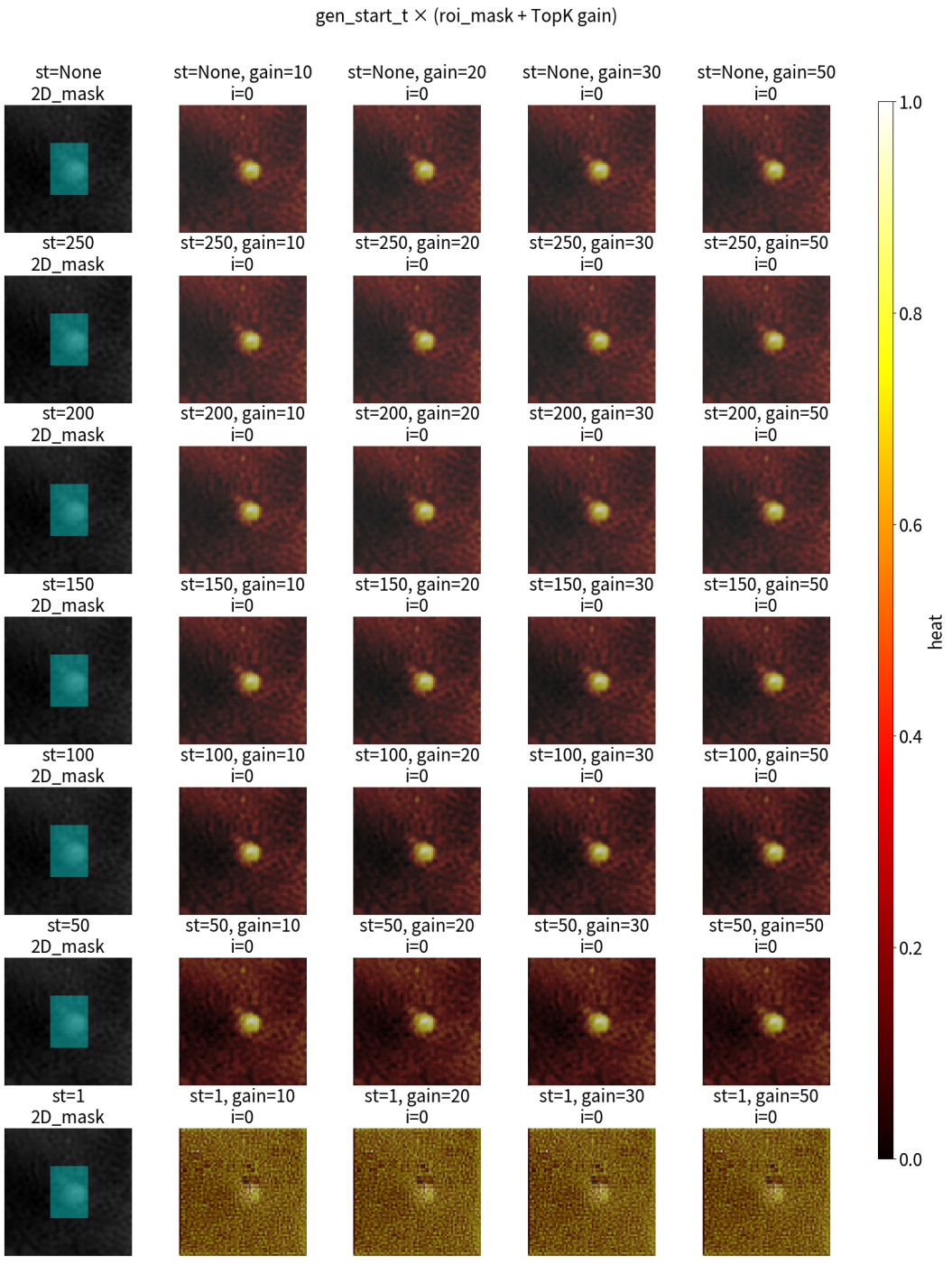


Figure 27: **Hepatic cyst (Benign):** same rendering as Appendix H; ROI GT col. 1, $|x_{\text{guided}} - x_{\text{orig}}|$ elsewhere.

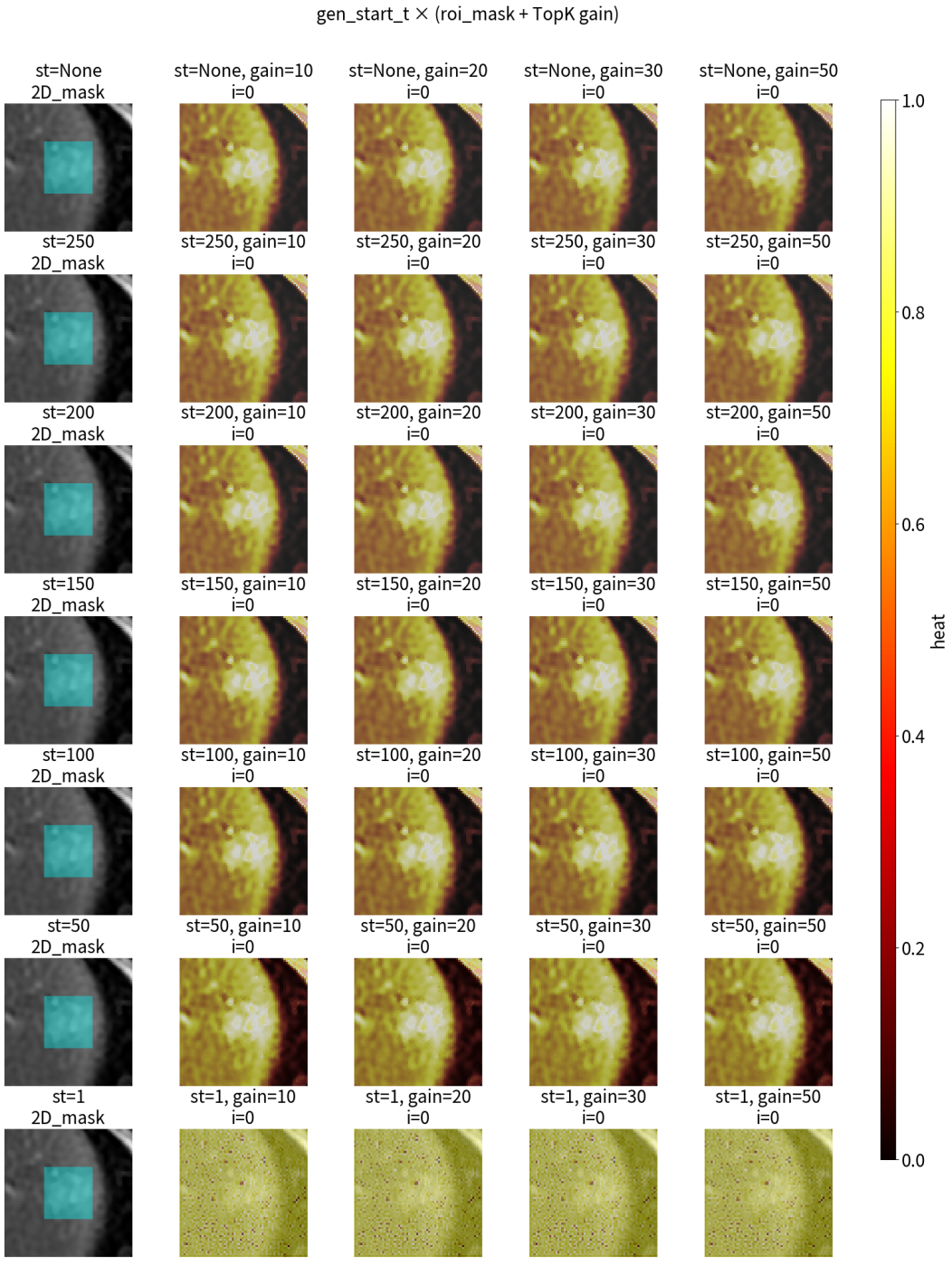


Figure 28: **Focal nodular hyperplasia (Benign):** heatmaps vs. original with ROI GT; FNH-associated differences align with the annotated region.

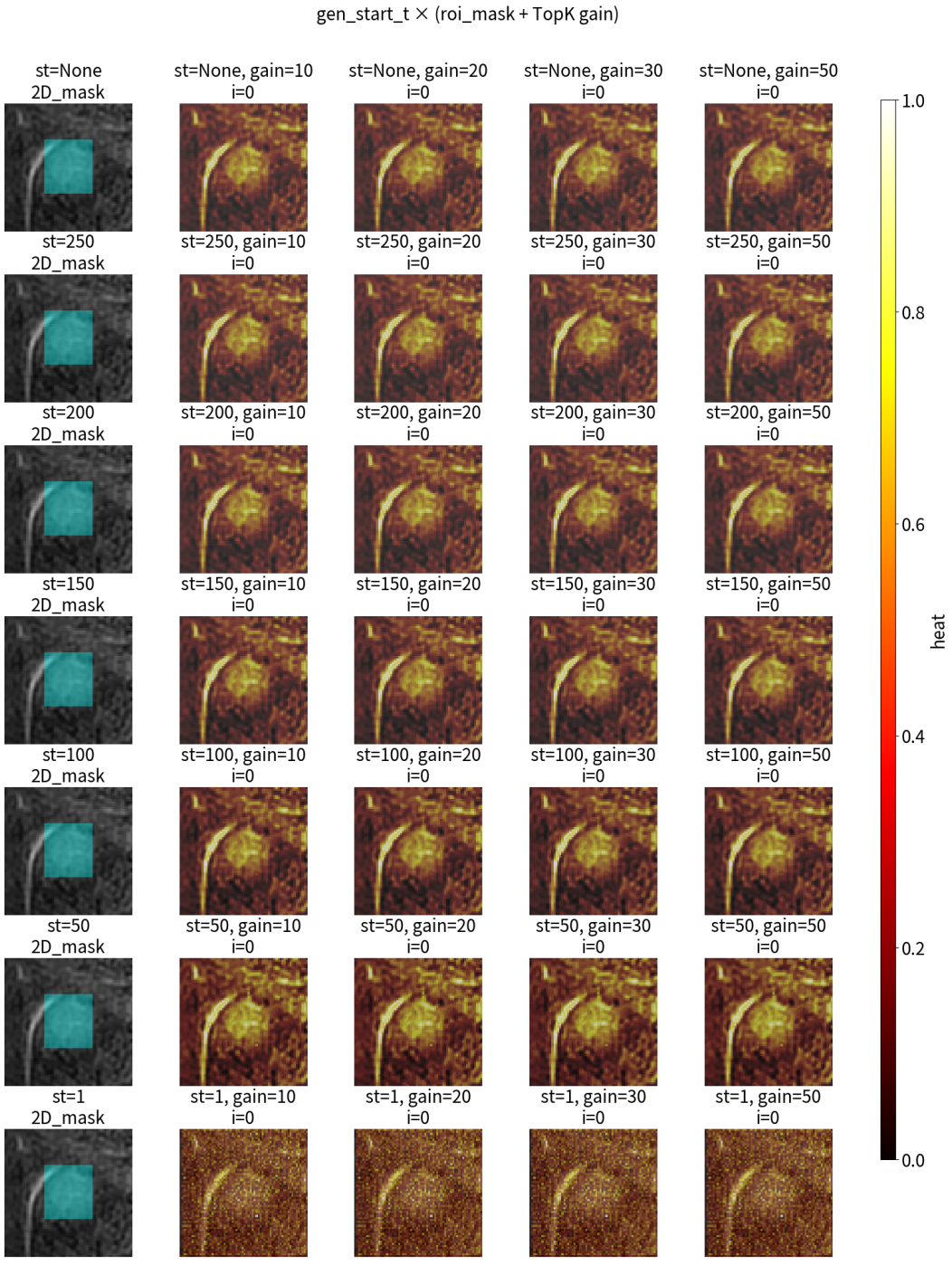


Figure 29: **Hepatocellular carcinoma (Malignant):** DoD-guided difference heatmaps (§4.2); rows st, cols. $gain$; col. 1 shows ROI GT. Supports lesion-localized amplification for augmentation.