LOCALIZING ACTIVE DYNAMICS IN SCIENTIFIC IMAGES: PHYSICS-GROUNDED VISUAL EXPLANATIONS FOR SEGMENTATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific image segmentation identifies structures of interest, but it does not necessarily reveal *where* the underlying physical or biological mechanism that produced them is currently *active*. We use *active dynamics* to denote image regions in which the governing mechanism is currently changing the observed system. These regions differ from the segmentation target. For example, a mask may identify material that has already crystallized, while the remaining transformation occurs in the untransformed regions between crystals. This creates a structural limitation for gradient-based explanation: it attributes the segmentation prediction to supporting image evidence, hence inherently drawn toward the segmented structure, even when the governing dynamics are active in a different region.

We introduce **PGAD-X** (Physics-Grounded Active-Dynamics eXplanation), a framework that produces an active-dynamics explanation alongside the segmentation mask. PGAD-X encodes the governing law and its characteristic spatial scale to condition shared visual features, and uses attention-based explanation to localize where the corresponding dynamics are expressed. Our key design decouples the explanation from the segmentation gradient while grounding both outputs in the same visual representation, enabling the model to localize physically active regions, rather than merely restate the predicted mask. A lightweight second stage aligns the explanation with automatically derived physics references while keeping the segmentation backbone fixed. Across three microscopy datasets, three governing mechanisms, and three segmentation backbones, PGAD-X improves active dynamics localization over gradient-based and latent-alignment baselines while preserving or improving segmentation accuracy. We release the physics references and evaluation protocol.

1 INTRODUCTION

Scientific image segmentation is a cornerstone task to quantify structures in *e.g.,* microscopy data, to help scientists recognize the transformed area of a material or the number and spatial distribution of cells. The resulting mask supports various downstream measurements and analysis. Nevertheless, it does not by itself indicate whether the model is grounded by genuine, scientifically meaningful evidence or whether its prediction is consistent with the mechanism that generated the observed structure. This limitation becomes critical when model predictions are used to support scientific hypothesis generation, experimental design, or when they disagree with established theory.

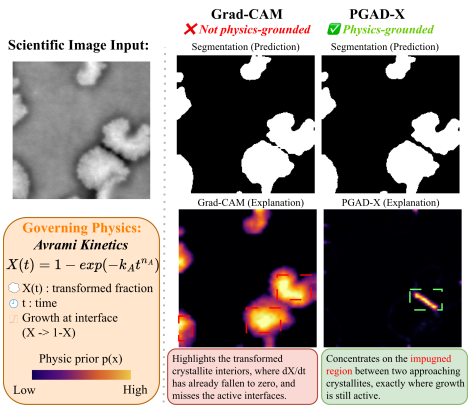

Figure 1: **Mismatch between segmentation and active dynamics.**

In many scientific imaging tasks, the region being segmented is *not* the same as the region in which the underlying mechanism is currently acting. We use

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.

active dynamics to denote image regions in which the governing physical or biological mechanism is currently *changing* the observed system. A segmentation mask typically records the current state of the system, whereas the active dynamics may occur at interfaces, in untransformed regions, or within only a subset of the segmented objects. Identifying these regions will help scientists assess whether a model captures the expected mechanism, diagnose discrepancies between data and theory, and interpret spatial variations in the ongoing transformation.

This distinction, with evident need for explainable and trustworthy scientific image segmentation, motivates us to investigate an important yet overlooked problem.

Physics-Grounded Active-Dynamics Explanation. Given a scientific image $\mathbf{x}$, a segmentation model f_ϕ, and a governing physical or biological law $\mathcal{G}_d$, the objective is to produce a spatial explanation that localizes the active dynamics described by $\mathcal{G}_d$ while remaining faithful to the model that produces the segmentation.

This objective differs from conventional visual attribution, which identifies image evidence supporting the predicted mask. When the segmentation target and the active-dynamics region do not coincide, an explanation derived from the segmentation objective can be well aligned with the prediction yet poorly aligned with the scientific mechanism of interest.

We illustrate this mismatch in a real-world example below.

Example 1: Crystallization: a mismatch between segmented state and active dynamics. Figure 1 shows an atomic-force microscopy image of a thin film during crystallization. The bright regions correspond to crystalline material that has already formed, and the segmentation model is trained to identify this transformed phase. The scientific quantity of interest is *different*: it is the location at which further transformation is occurring. In the nucleation and growth setting described by Avrami kinetics, continued transformation is associated with interfaces between transformed and untransformed material. As neighboring crystals approach one another, the remaining untransformed corridors indicate where additional conversion can occur (Avrami, 1939; Johnson, 1939). The segmentation target and the active-dynamics target are therefore spatially distinct.

Gradient-based methods such as Grad-CAM (Selvaraju et al., 2017) construct explanations from the segmentation objective. It consequently emphasizes visual features that support labeling a region as crystalline, rather than the interfaces at which transformation remains active (see the center column of Fig. 1). Such gradient explanation can be faithful to the segmentation prediction, while being systematically misaligned with the scientific quantity of interest. □

This mismatch extends to many domains whenever a segmentation target records an accumulated state, while the scientific objective is to localize where the underlying mechanism remains active. Examples include visible crack geometry versus the crack-tip process zone, and burned area versus the active fire front (Chen et al., 2024; Javaid et al., 2019; Ribeiro et al., 2023; Li et al., 2025).

Requirements. These examples motivate three requirements. An explanation should be (1) *model-faithful*, by using the representation that produces the segmentation; (2) *physics-aligned*, by highlighting regions where the governing mechanism is active; and (3) *mask-complementary*, by being able to differ from the predicted mask. These properties ensure that the explanation remains tied to the model while adding information beyond the segmentation. When active dynamics and the segmentation target differ, gradient-based attribution cannot generally satisfy all three.

While desirable, the problem is nontrivial. The main challenge is to extract a physics-aligned spatial signal from the segmentation model, without introducing a separate explainer or forcing it to follow the mask. The governing law must also be connected to visual features at an appropriate spatial scale. Evaluation is equally important, since physics references may be partly recoverable from intensity, boundaries, or the mask itself. We therefore evaluate only where the mechanism can be active, using threshold free metrics and explicit image- and mask-derived reference floors.

We introduce **PGAD-X** (*Physics-Grounded Active-Dynamics eXplanation*), a visual explanation framework that produces an active-dynamics explanation alongside the segmentation mask. Its design directly matches the three requirements: (1) Both outputs share the same visual representation; (2) a

compact descriptor of the governing law and its characteristic length conditions the relevant feature scales; (3) A gradient-independent cross-attention branch localizes physically active regions, allowing the explanation to differ from the predicted mask. A lightweight stage next aligns the map with mechanism-derived reference maps while keeping the backbone weights fixed.

Contributions. Our main contributions are: (1) We formulate active-dynamics explanation as a distinct problem in scientific image segmentation and identify a structural limitation of gradient-based attribution when the segmented state and physically active region differ. (2) We propose PGAD-X, which combines shared model features, physics-conditioned scale selection, and a gradient-independent cross-attention explanation branch, to satisfy all three aforementioned requirements: model faithfulness, physics alignment, and mask complementarity. (3) We construct and release mechanism-derived reference maps for three microscopy datasets, together with an evaluation protocol based on admissible regions, ranking metrics, and simple reference scores. Across three datasets and three backbones, PGAD-X improves explanation quality while maintaining segmentation IoU.

Related Work. We summarize related work as follows.

Scientific segmentation and explanation. General-purpose models such as SegFormer, SAM2, and DINOv2 Xie et al. (2021); Ravi et al. (2025); Oquab et al. (2023); Hu et al. (2022) have been adapted to scientific images. Existing attribution methods explain predictions through gradients, concepts, prototypes, or feature contributions Selvaraju et al. (2017); Sundararajan et al. (2017); Losch et al. (2019); Chen et al. (2019); Lundberg & Lee (2017); Gipiškis et al. (2024). They primarily identify evidence supporting the predicted mask. Our objective instead is to localize active dynamics specified by a governing law, which may occur outside the segmented region. Thus, an attribution map can be faithful to the prediction but misaligned with the governing mechanism.

Physics-guided representation learning. Physics-informed neural networks incorporate equations into representation learning Raissi et al. (2019); Li et al. (2020); Kang et al. (2023); Wang et al. (2024); Kapoor et al. (2024). In contrast, PGAD-X does not solve or enforce the complete governing equation. It encodes the governing mechanism and uses a measurable characteristic length to condition the appropriate visual feature scales. FiLM provides the feature conditioning operation Perez et al. (2018), while a dedicated cross-attention branch produces the spatial explanation.

Spatial and latent alignment. Spatial explanations are commonly evaluated using pointing or threshold-based localization scores Selvaraju et al. (2017). We instead evaluate within admissible regions using ranking metrics and scores from simple reference maps. I-JEPA Assran et al. (2023) aligns latent representations without explicit spatial localization.

2 Physics-Grounded Visual Explanation

2.1 Problem Setup and Governing Dynamics

Images and segmentation. Let $\mathcal{D}_d$ be a scientific image dataset. Each sample consists of an image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ and a segmentation mask $\mathbf{m} \in \{0,1\}^{H\times W}$. We denote the image grid as $\Omega = [H] \times [W]$, and the intensity at location $u \in \Omega$ as $I_{\mathbf{x}}$. A segmentation model f_ϕ predicts masks that best fit with $\mathbf{m}$: $H_{\text{seg}} = f_\phi(\mathbf{x}) \in [0,1]^{H\times W}$.

Active-dynamics explanation. We consider a dataset $\mathcal{D}_d$ that is associated with a *governing* (physical or biological) *law* $\mathcal{G}_d$ that characterizes how the observed system evolves. Our objective is to augment the segmentation model with a spatial explanation $H^*_{\text{vis}} \in [0,1]^{H\times W}$ whose values rank image locations according to the activity of the mechanism described by $\mathcal{G}_d$. Given $\mathbf{x}$ and $\mathcal{G}_d$, the model returns $\left(H_{\text{seg}}, H^*_{\text{vis}}\right)$, where H^*_{vis} localizes its active dynamics.

We do not require $\mathcal{G}_d$ to be differentiable, fully parameterized, or directly incorporated into the segmentation loss. Instead, we only require that it admit a (polynomial-time) computable grounding operator Γ_d that converts observable image information into a spatial activity reference: $p(\mathbf{x}) = \Gamma_d(\mathbf{x}, \mathbf{m}; \mathcal{G}_d) \in [0,1]^{H\times W}$. The map $p(\mathbf{x})$ provides supervision for the explanation alignment stage, rather than for the segmentation prediction. Γ_d is a mechanism informed rule, not a solution of $\mathcal{G}_d$:

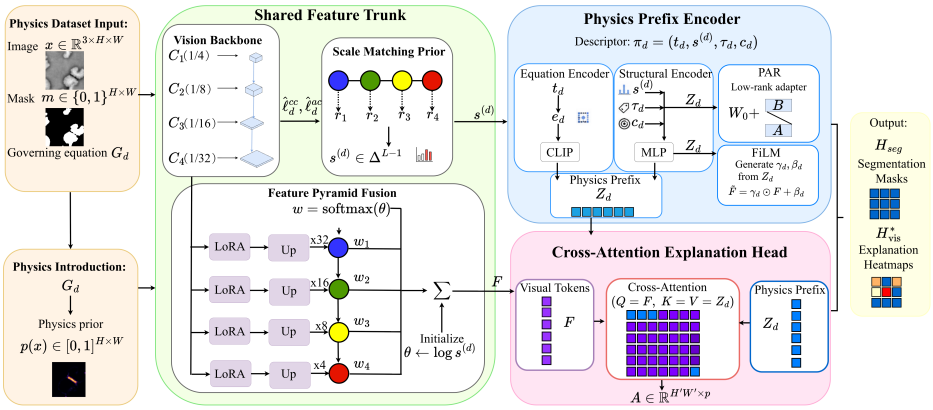

Figure 2: The PGAD-X framework: Major Modules and Architecture.

it encodes where $\mathcal{G}_d$ predicts activity given a single observed state, so $p(\mathbf{x})$ is a static proxy for the ongoing dynamics. When an independently constructed expert reference is available, we denote it by $p^\star(\mathbf{x})$ and reserve it for evaluation.

Scope of governing laws. PGAD-X supports "plug-able" governing laws, which can be user-defined and validated analytical or empirical equations, or a domain rule. Such laws characterize a spatiotemporal dynamic mechanism with a computational equation system. Specifically, we make case for PGAD-X with three established laws: Avrami kinetics Avrami (1939) for crystallization, logistic dynamics for cell proliferation Warne et al. (2017), and chromatin-condensation criteria for mitotic nuclei Nunes et al. (2026). These choices do not restrict the generalization of PGAD-X to other domains beyond microscopy analysis.

Example 2: Spatial grounding for crystallization. For the AFM crystallization scenario, Avrami kinetics describe the evolution of the transformed fraction, while the associated grounding operator assigns high activity to remaining untransformed corridors between approaching crystallites and zero activity to transformed interiors. Section 4 specifies the deterministic construction of $p(\mathbf{x})$. We found that the mean correlation between the segmentation mask and the dynamic area of interest ($p^\star(\mathbf{x})$) annotated by domain scientists is -0.10, quantitatively confirming that the active-dynamics target is not a surrogate for the segmentation mask. □

2.2 Requirements for Explanations

Physics-admissible domain. The governing law may rule out activity at large portions of an image. For example, crystallization is inactive inside fully transformed material, while chromatin condensation is undefined outside a nucleus. Including such locations in evaluation would reward an explanation for identifying regions that are trivially inactive. We therefore define the *physics-admissible domain* $\Omega_{\mathrm{adm}} \subseteq \Omega$ as the set of locations at which the mechanism described by $\mathcal{G}_d$ can be active. The requirements and evaluation of H^*_{vis} are restricted to Ω_{adm} (see Section 4 for its data-specific construction).

Explanation requirements. The distinction between the segmented state and its active dynamics motivates three specific requirements below.

Requirements for active-dynamics explanations.

D1: Model faithfulness. H^*_{vis} is derived from the visual representation used to produce H_{seg}, rather than from an independently trained predictor.

D2: Physics alignment. Within Ω_{adm}, H^*_{vis} ranks locations where the governing mechanism is active above inactive locations.

D3: Mask complementarity. H^*_{vis} can provide spatial information beyond H_{seg} and is not restricted to a monotone transformation of the predicted mask.

Unlike conventional explanation models, active-dynamics explanations remain connected to segmentation model, agrees with the governing dynamics, and can still depart from the predicted mask when the two spatial targets differ.

Remarks. We clarify why gradient attribution methods are limited for such needs. A gradient explanation has the form $H_{\text{vis}}^{\text{grad}} = \mathcal{A}\big(\Psi_\phi(\mathbf{x}), \nabla_{\Psi_\phi}\mathcal{L}_{\text{seg}}\big)$ where $\Psi_\phi(\mathbf{x})$ is an intermediate feature map and $\mathcal{A}$ aggregates activations and gradients spatially. It satisfies D1 because it is derived from the segmentation model, but it contains no independent signal from $\mathcal{G}_d$. When segmentation and active regions differ, this coupling prevents independent control of D2 and D3, motivating an explanation pathway that conditions shared features on the governing law.

3 PGAD-X Framework

3.1 Framework Overview

Given an image $\mathbf{x}$ and a governing law $\mathcal{G}_d$, PGAD-X produces a segmentation mask H_{seg} and an active-dynamics explanation H_{vis}^* (as illustrated in Fig. 2). Its main architectural novelty is to separate the *source* of the explanation from the segmentation gradient while retaining a shared visual representation. This design preserves model faithfulness, permits physics alignment, and allows the explanation to differ from the mask.

The framework comprises three modules. The **Shared Feature Trunk (SFT)** uses a hierarchical backbone and feature pyramid to construct multi-scale features, which are projected, fused, and adapted into a shared representation $\widetilde{F}$ for segmentation and explanation (D1). The **Physics Prefix Encoder (PPE)** encodes the governing law G_d and its characteristic-scale prior into prefix tokens Z_d, which condition visual features through lightweight adapters and FiLM layers to support physics alignment (D2). The **Cross-Attention Explanation Head (CAEH)** lets visual tokens from $\widetilde{F}$ attend to Z_d and uses the response magnitude to produce H_{vis}^*, while the segmentation head decodes the same enriched representation. This gradient-independent readout allows the explanation to localize regions spatially distinct from the predicted mask (D3).

The three modules are learned via a two-stage training phase. Stage 1 trains the visual trunk and segmentation head using segmentation supervision. Stage 2 freezes the backbone parameters and updates only the physics-conditioning pathway and explanation head to align H_{vis}^* with the mechanism-derived reference $p(\mathbf{x})$, while retaining the segmentation loss. In a nutshell, it adjusts how $\mathcal{G}_d$ modulates the learned representation without retraining the backbone as a new segmenter.

We next elaborate on each module, followed by the learning and inference process (Section 3.2).

Shared Feature Trunk (SFT). The SFT module characterizes and learns the multi-scale visual features of an image with the following construction.

Multi-scale visual representation. The module works with a hierarchical backbone, which maps an image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ to feature maps $\mathbf{C}_k \in \mathbb{R}^{D_k\times H_k\times W_k}$, $H_k{=}\frac{H}{r_k}$, $W_k{=}\frac{W}{r_k}$, for $k \in [L]$, where D_k and r_k denote the channel width and spatial stride of level k, respectively.

We define the fusion stride as the finest available pyramid stride $r_{\text{f}} = \min_{k\in[L]} r_k$ and fuse all levels on the common grid

$$\Omega' = [H'] \times [W'], \qquad H' = \frac{H}{r_{\text{f}}}, \quad W' = \frac{W}{r_{\text{f}}}.$$

In our implementation, the pyramid strides are $\{4, 8, 16, 32\}$, so $r_{\text{f}} = 4$ and $\Omega' = [H/4] \times [W/4]$. Using the finest pyramid resolution preserves spatial detail for explanation localization without incurring the cost of full-resolution feature processing. Each level is projected to a common channel width D and resampled to Ω': $\mathbf{P}_k = \text{Resize}_{\Omega'}\Big(\text{Conv}_{1\times1}^{(k)}(\mathbf{C}_k)\Big) \in \mathbb{R}^{D\times H'\times W'}$. The fused pyramid determines which spatial scales contribute most strongly to the shared representation.

Since the active dynamics associated with $\mathcal{G}_d$ may be observable only at a characteristic scale, we use the governing law to initialize these fusion weights through a *physics-grounded scale prior*.

Physics-grounded scale prior. We use the characteristic length ℓ_d of the governing law to define a log-scale prior $\mathbf{s}^{(d)}$ over pyramid levels, assigning greater weight to feature resolutions commensurate with ℓ_d. We estimate ℓ_d independently from image autocorrelation and mask component sizes, and use $\mathbf{s}^{(d)}$ to initialize the trainable fusion weights. A dataset-specific affine adapter then produces the shared feature map $\widetilde{\mathbf{F}}$ used by both prediction and explanation (see details in Appendix A).

Physics Prefix Encoding (PPE). The physics prefix encoder PPE constructs the following.

Physics descriptor. Each dataset $\mathcal{D}_d$ is summarized once by $\boldsymbol{\pi}_d = \big(t_d, \mathbf{s}^{(d)}, \tau_d, c_d\big)$, where t_d describes the governing law $\mathcal{G}_d$ in natural language, $\mathbf{s}^{(d)} \in \Delta^{L-1}$ is the scale prior from Section 3.1, $\tau_d \in \{\text{BOUNDARY}, \text{DENSITY}\}$ indicates whether the dynamics are localized at interfaces or distributed within regions, and $c_d \in [0, 1]$ records descriptor confidence. We encode the structured terms as $\mathbf{q}_d = [\mathbf{s}^{(d)}; \text{onehot}(\tau_d); c_d]$.

Prefix encoding. The text component is embedded as $\mathbf{e}_d = \text{CLIP}_{\text{text}}(t_d) \in \mathbb{R}^{512}$. Textual and structured information are separately projected into n_p prefix tokens:

$$\mathbf{T}_d = \text{reshape}_{n_p \times D}(\mathbf{W}_t \mathbf{e}_d), \;\; \mathbf{S}_d = \text{reshape}_{n_p \times D}\big(\phi(\mathbf{q}_d)\big),$$

where $\mathbf{W}_t \in \mathbb{R}^{n_p D \times 512}$ and $\phi : \mathbb{R}^{L+3} \to \mathbb{R}^{n_p D}$ is a two-layer MLP.

The two token sets are fused row-wise: $\mathbf{Z}_d = [\mathbf{T}_d \| \mathbf{S}_d]\mathbf{W}_f \in \mathbb{R}^{n_p \times D}$, where $\mathbf{W}_f \in \mathbb{R}^{2D \times D}$. The tokens condition each backbone stage through Physics Attention Residual (PAR) and feature-wise linear modulation (FiLM). The adapters introduce physics-dependent residual updates to the attention layers, while FiLM scales and shifts feature channels according to $\mathbf{Z}_d$ [1].

Distinct conditioning tokens. Producing n_p distinct tokens is necessary for a spatially varying explanation. Projecting the descriptor to one vector and replicating it n_p times yields identical attention keys and values; each visual location then receives the same response, and H^*_{vis} becomes spatially constant. Such a map scores exactly at chance under our ranking metric. The encoder therefore generates multiple distinct conditioning directions rather than broadcasting a single descriptor vector. We set $n_p = 8$ by default in our system implementation. Appendix A verifies the above design.

Cross-Attention Explanation Head (CAEH). The physics prefixes encode *what* dynamics should be identified, whereas the shared feature map encodes *where* relevant visual evidence occurs. We connect them through cross-attention. Let $\mathbf{V} = \text{Flatten}(\widetilde{\mathbf{F}}) \in \mathbb{R}^{|\Omega'| \times D}$ denote the visual tokens, and let $\mathbf{P}_d = \mathbf{Z}_d \mathbf{W}_p \in \mathbb{R}^{n_p \times D}$ be the projected physics-prefix tokens. The visual tokens query the physics prefixes to obtain the cross-attention:

$$\mathbf{A} = \text{Attn}\big(Q = \mathbf{V},\, K = \mathbf{P}_d,\, V = \mathbf{P}_d\big) \in \mathbb{R}^{|\Omega'| \times D}. \tag{1}$$

Thus, $\mathbf{A}(u)$ measures the physics-conditioned response at location $u \in \Omega'$. The segmentation head decodes $\text{LayerNorm}(\mathbf{V} + \mathbf{A})$, so both outputs remain grounded in the same visual representation.

We define the reduced-resolution explanation by $h(u) = \|\mathbf{A}(u)\|_2$ and obtain the full-resolution map:

$$H^*_{\text{vis}} = \frac{\text{Up}_\Omega(h)}{\max_{v \in \Omega} \text{Up}_\Omega(h)(v)}, \tag{2}$$

where Up_Ω denotes bilinear upsampling from Ω' to Ω. Unlike gradient attribution, H^*_{vis} is generated directly from the interaction between $\mathbf{V}$ and $\mathbf{Z}_d$, rather than by differentiating $\mathcal{L}_{\text{seg}}$. It therefore preserves model faithfulness through the shared features while remaining independently alignable with the governing dynamics, satisfying D1– D3.

[1] For a stage-k feature map $\mathbf{C}_k$, FiLM applies $\text{FiLM}_k(\mathbf{C}_k; \mathbf{Z}_d) = \boldsymbol{\gamma}_k(\mathbf{Z}_d) \odot \mathbf{C}_k + \boldsymbol{\beta}_k(\mathbf{Z}_d)$, where $\boldsymbol{\gamma}_k$ and $\boldsymbol{\beta}_k$ are generated from a pooled representation of $\mathbf{Z}_d$. The low-rank adapter adds a descriptor-conditioned residual update to the attention projection.

Table 1: **Overall Performance over All Datasets.**

Method	Backbone	AFM Crystallites				LIVECell BV2				BBBC039			
		S1 IoU	S2 IoU	AUC	AP	S1 IoU	S2 IoU	AUC	AP	S1 IoU	S2 IoU	AUC	AP
GradCAM-only: gradient readout													
GradCAM Selvaraju et al. (2017)	SegFormer-B5	0.929	0.929	0.713	0.082	0.757	0.757	0.492	0.146	0.859	0.859	0.600	0.338
GradCAM Selvaraju et al. (2017)	SAM2 Hiera-L	0.917	0.917	0.722	0.086	0.745	0.745	0.485	0.143	0.879	0.879	0.610	0.367
GradCAM Selvaraju et al. (2017)	DINOv2	0.905	0.905	0.648	0.052	0.316	0.316	0.498	0.142	0.684	0.684	0.552	0.305
Physics-JEPA: latent-space alignment													
JEPA Assran et al. (2023)	SegFormer-B5	0.795	0.795	0.677	0.063	0.705	0.705	0.434	0.155	0.876	0.876	0.525	0.260
JEPA Assran et al. (2023)	SAM2 Hiera-L	0.858	0.858	0.709	0.067	0.703	0.703	0.439	0.158	0.879	0.879	0.528	0.263
JEPA Assran et al. (2023)	DINOv2	0.848	0.848	0.796	0.080	0.369	0.369	0.475	0.159	0.610	0.610	0.522	0.266
PGAD-X: no Stage-2 alignment													
PGAD-X (S1 only)	SegFormer-B5	0.929	0.929	0.755	0.091	0.757	0.757	0.459	0.139	0.859	0.859	0.521	0.277
PGAD-X (S1 only)	SAM2 Hiera-L	0.917	0.917	0.415	0.048	0.745	0.745	0.516	0.162	0.879	0.879	0.472	0.246
PGAD-X (S1 only)	DINOv2	0.905	0.905	0.706	0.086	0.316	0.316	0.525	0.162	0.684	0.684	0.528	0.266
PGAD-X (ours): cross-attention readout with Stage-2 alignment													
PGAD-X	SegFormer-B5	0.929	**0.942**	0.824	0.196	0.757	**0.754**	0.699	0.258	0.859	0.853	**0.758**	**0.516**
PGAD-X	SAM2 Hiera-L	0.917	0.933	0.830	0.195	0.745	0.748	0.561	0.173	0.879	0.871	0.712	0.466
PGAD-X	DINOv2	0.905	0.914	**0.837**	**0.204**	0.316	0.344	**0.719**	**0.273**	0.684	0.678	0.748	0.512

3.2 Two-Stage Training

Stage 1: Segmentation learning. We disable physics conditioning, replace $\mathbf{Z}_d$ with a learned dataset-agnostic prefix $\mathbf{Z}_0$, and train the backbone (with feature fusion, explanation head, and decoder):

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}}(H_{\text{seg}}, \mathbf{m}) + \mathcal{L}_{\text{Dice}}(H_{\text{seg}}, \mathbf{m}). \tag{3}$$

The training runs for at most 60 epochs with early stopping, after which the parameters are frozen.

Stage 2: Physics alignment. We then restore $\mathbf{Z}_d$ and train only the Physics Prefix Encoder, prefix-conditioned adapters, and FiLM layers:

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda \left[1 - \rho_{\text{soft}}(H^*_{\text{vis}}|_{\mathcal{U}}, p(\mathbf{x})|_{\mathcal{U}})\right], \tag{4}$$

where $\mathcal{U} \subseteq \Omega_{\text{adm}}$ is a dataset-adaptive set of admissible locations and ρ_{soft} is differentiable Spearman correlation. This aligns spatial rankings without requiring equal numerical scales; we use $\lambda = 1$. The backbone weights remain fixed, while the conditioning modules may modify their activations. Because these modules act on backbone activations and $\mathcal{L}_{\text{seg}}$ is retained, Stage 2 is also additional supervised segmentation training, and segmentation changes after Stage 2 should be read accordingly (Section 4). AFM alignment and evaluation references are constructed independently; see Appendix A.

Inference. The dataset-level descriptor and physics prefixes are computed once and cached. Each image then requires one forward pass to produce both H_{seg} and H^*_{vis}, without the backward pass required by gradient attribution. We provide a detailed algorithm and cost analysis in the Appendix.

4 Experimental Study

Datasets and Physics Priors. We evaluate on three real-world microscopy datasets spanning three modalities and governing laws, deriving each $p(\mathbf{x})$ from $\mathcal{G}_d$ by a fixed procedure on the image and its mask, with no extra annotation and no language model at inference, where the model sees only the image and predicts both outputs.

AFM crystallites Lu et al. (2023). Forty-six atomic-force images of a fluoroelastomer polycrystalline thin film, acquired as a time series through the phase transformation (224×224 px). The mask $\mathbf{m}$ marks the crystalline phase, whose area fraction rises monotonically from 0.22 to 0.92 and thus traces $X(t)$. Since Avrami kinetics place $\mathrm{d}X/\mathrm{d}t$ at the advancing front—the narrow corridors of untransformed matrix consumed last—we set $p(\mathbf{x}) = G_2 * [(\mathbf{m} \bullet B_r) - \mathbf{m}]$ ($\bullet$: morphological closing with a disc of radius $r = 11$; G_2: Gaussian of width 2 px; per-image max normalisation) and $\Omega_{\text{adm}} = \{\mathbf{u} : \mathbf{m}(\mathbf{u}) = 0\}$. The impingement annotations, produced and validated by a team of 22 materials scientists, serve as $p^\star(\mathbf{x})$ and are reserved for evaluation.

LIVECell BV2 Edlund et al. (2021). Phase-contrast images of BV2 microglia with binarised COCO polygons as masks. Here $\mathcal{G}_d$ is the logistic law $\dot{N} = \rho N(1 - N/N_\infty)$, whose active dynamics,

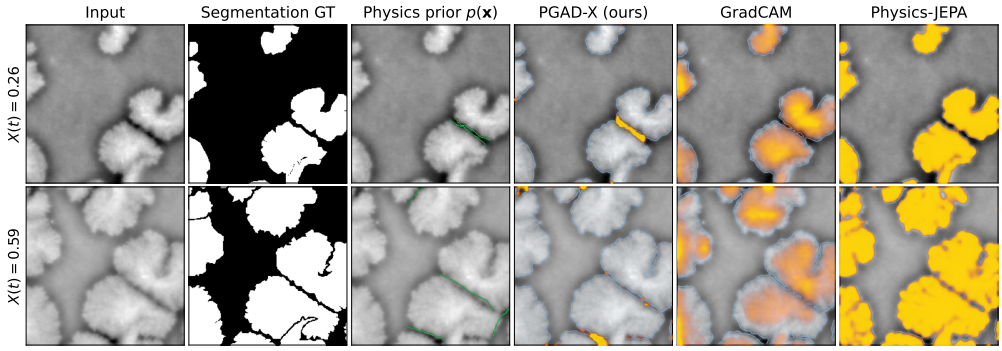

Figure 3: Case Study: Comparison on AFM crystallites at two stages of transformation.

proliferation, peak at intermediate density $N \approx N_\infty/2$ and vanish where cells are sparse or confluent, so the mask marks every cell but the active region is only the crowding fronts. We take $N = G_{18} * \mathbf{m}$, set N_∞ to its 95th percentile over occupied locations, and let $p(\mathbf{x}) \propto \exp(-[(N - N_\infty/2)/(N_\infty/12)]^2)$. A faithful explanation must therefore locate those fronts rather than the cell bodies the segmentation already marks; since cells may occur anywhere, $\Omega_{\text{adm}} = \Omega$.

BBBC039 Ljosa et al. (2012). Sixteen-bit DAPI images of U2OS nuclei with nucleus masks. Here $\mathcal{G}_d$ is the chromatin-condensation law $\mathrm{d}\,\mathrm{CV}(\kappa)/\mathrm{d}t > 0$: mitotic condensation raises a nucleus's intra-nuclear coefficient of variation $\mathrm{CV}(\kappa) = \mathrm{sd}_{\mathbf{u}\in\kappa} I(\mathbf{u})/\mathrm{mean}_{\mathbf{u}\in\kappa} I(\mathbf{u})$, so the active nuclei are a strict subset of the mask. We set $p(\mathbf{x}) = 1$ on nuclei with $\mathrm{CV}(\kappa)$ above the per-image 75th percentile, zero elsewhere and smoothed by $G_{1.5}$, with $\Omega_{\text{adm}} = \{\mathbf{u} : \mathbf{m}(\mathbf{u}) = 1\}$. A faithful explanation must therefore separate condensed from interphase nuclei, not nuclei from background.

Evaluation Protocol. All scores are computed on Ω_{adm}; locations outside it are trivially negative and, being numerous, would dominate any aggregate. With $y(\mathbf{u}) = \mathbf{1}[p(\mathbf{x})(\mathbf{u}) > \frac{1}{2}]$ we split the domain into positive and negative sets $\mathcal{P} = \{\mathbf{u} \in \Omega_{\text{adm}} : y(\mathbf{u}) = 1\}$ and $\mathcal{N} = \Omega_{\text{adm}} \setminus \mathcal{P}$; on AFM $p^\star(\mathbf{x})$ replaces $p(\mathbf{x})$. (1) **Segmentation.** We report mask IoU (with $\hat{\mathbf{m}} = \mathbf{1}[H_{\text{seg}} > \frac{1}{2}]$) after both stages, so any gain in explanation is read against its cost in accuracy. (2) **Explanation.** As H^*_{vis} is defined only up to a monotone rescaling, we score it without thresholding. Our primary measure is *AUC*, the probability that a random active location outranks a random inactive one, $\mathrm{AUC} = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{\mathbf{u}\in\mathcal{P},\,\mathbf{v}\in\mathcal{N}} \big(\mathbf{1}[h(\mathbf{u}) > h(\mathbf{v})] + \frac{1}{2}\mathbf{1}[h(\mathbf{u}) = h(\mathbf{v})]\big)$ with $h = H^*_{\text{vis}}$; it is invariant to increasing transformations of h and has chance level $\frac{1}{2}$, which a spatially constant map attains exactly through the tie term. Because $|\mathcal{P}|/|\Omega_{\text{adm}}|$ is small on some datasets, we also report *average precision*, the area under the precision–recall curve $\mathrm{AP} = \sum_k (\mathrm{Rec}(t_k) - \mathrm{Rec}(t_{k-1}))\,\mathrm{Prec}(t_k)$ over thresholds t_k; its chance level equals the positive rate $|\mathcal{P}|/|\Omega_{\text{adm}}|$ and varies by dataset, so AP is read within a column, not across. Both are computed per image and averaged.

Baselines. We compare PGAD-X with the following representative methods. (1) **GradCAM-only** Selvaraju et al. (2017): A gradient-weighted activation map over the fused features $\tilde{F}$, read off the Stage-1 checkpoint with no Stage-2 training; it is the standard gradient-based attribution applied to the segmentation model. (2)**Physics-JEPA** Assran et al. (2023). A latent-space counterpart of our alignment: following I-JEPA, a predictor maps the pooled visual summary onto a frozen embedding of $p(\mathbf{x})$, and Stage 2 minimises $\mathcal{L}_{\text{seg}} + \beta\|\mathrm{pred}(\bar{v}) - \mathrm{enc}(p)\|^2$ ($\beta = 0.1$), adapting only the predictor. With no spatial injection or explanation head, its explanation falls back to H_{seg} and its spatial output is unchanged between stages. (3)**PGAD-X** (S1 only): a variant of PGAD-X, where the cross-attention readout is taken off with no alignment training. We compare it against PGAD-X to evaluate the contribution of Stage 2, and GradCAM to understand the contribution of the readout.

We next present our findings on effectiveness and case study (see more ablation tests in the Appendix).

Exp-1: Overall Effectiveness. Table 1 summarizes the overall performance of all the methods.

Segmentation preserved. We first evaluate whether alignment costs segmentation accuracy. Stage 2 keeps the backbone weights fixed, but its conditioning modules modulate backbone activations and

are trained with $\mathcal{L}_{\text{seg}}$ retained, so Stage 2 also acts as additional supervised segmentation training. Across the nine pairs, Stage 2 changes IoU by between -0.008 and $+0.028$. On BBBC039 this holds only when Stage 1 is evaluated with the 16-bit intensity scaling it was trained with; under the percentile scaling used for Stage 2 the Stage-1 IoU drops to 0.314–0.638, and Stage 2 recovers it by re-adapting the conditioning modules, as the same Stage 2 with $\lambda{=}0$ does equally well (0.861 with SegFormer, Fig. 5). PGAD-X optimizes segmentation and explanation jointly through a separate cross-attention response while maintaining segmentation accuracy.

Explanation quality. PGAD-X Stage 2 improves explanation quality across all nine dataset and backbone pairs. Against GradCAM read from the identical Stage-1 features, which sits at chance on LIVECell (AUC 0.485 to 0.498) and reaches only 0.552 to 0.610 on BBBC039, **PGAD-X** raises AUC by 30% on average on these two datasets and by 20% on AFM, and improves by 34% over latent-space Physics-JEPA across the nine pairs. GradCAM stays near chance because its gradient is pinned to the segmented phase, and Physics-JEPA aligns only a pooled embedding, so its predictor never reaches the spatial output. The margin reflects a readout that an alignment objective can move: the cross-attention response is not tied to the segmentation gradient, so Stage 2 can shift it onto the active region the mask omits (Table 4 isolates the readout).

Physical, not cosmetic. On AFM, GradCAM reaches AP 0.052 to 0.086 and Physics-JEPA 0.063 to 0.080, no better than ranking the crystalline phase, while **PGAD-X** reaches 0.195 to 0.204, over 170% higher; on BBBC039, against a positive rate of 0.253, GradCAM's 0.305 to 0.367 is only modestly above assigning the base rate everywhere, whereas PGAD-X reaches 0.466 to 0.516, an improvement of 49%. AP is the demanding measure here because it asks whether the strongest responses fall on the active region rather than whether most of the background is ranked low. The baselines respond to the transformed phase the mask already marks, since that is what drives the segmentation they differentiate, whereas the PGAD-X map concentrates on the untransformed corridors and triple junctions where the governing physics is active, exactly the region a physics-grounded explanation must find. On later AFM frames, the map also predicts where the film transforms next better than its alignment rule and a matched auxiliary head (AUC 0.865 vs. 0.778 and 0.773), though boundary distance does better still (Appendix A).

Additional experiments. We show that distinct prefixes prevent constant maps (Table 5) and rank alignment improves AP in all nine configurations, including 0.118 to 0.196 on AFM (Table 10). We outperform the auxiliary head in temporal evaluation AUC at $k = 3$ by 0.037 on SegFormer and 0.061 on SAM2 (Table 13). We obtain AUC standard deviations of at most 0.004 across three seeds on SegFormer and AFM AUC of 0.814–0.828 over closing radii 7–15 (Tables 14 and 16).

Exp-2: Case Study. We examine PGAD-X on AFM images governed by Avrami kinetics. While the transformed fraction $X(t)$ summarizes overall progress, assessing the well-mixed assumption behind $(1 - X)$ requires locating interfaces that still face untransformed material. In Fig. 3, these regions are the dark channels between bright crystallites, marked by the expert reference $p^{\star}(\mathbf{x})$.

PGAD-X concentrates on narrow untransformed corridors while remaining inactive inside crystallites. GradCAM highlights transformed interiors because they support segmentation, although transformation has ceased there. Physics-JEPA lacks a spatial explanation head and largely reproduces H_{seg}. Comparing $X(t) = 0.26$ and $X(t) = 0.59$ shows that, as crystallites impinge, the remaining transformable material becomes confined to these corridors. Localizing them reveals whether the remaining fraction is spatially accessible, rather than restating the segmented phase.

5 Conclusion

We introduced PGAD-X, a two-stage framework that produces an active-dynamics explanation alongside a scientific image segmentation. By combining shared visual features with a physics-conditioned, gradient-independent cross-attention branch, PGAD-X can localize where the governing mechanism is active even when this region differs from the predicted mask. A measurable characteristic length grounds the relevant feature scales, while a lightweight alignment stage adapts the physics-conditioning pathway with the backbone weights fixed. Across three microscopy datasets and three backbones, PGAD-X consistently improves physics-aligned localization while preserving or improving segmentation accuracy.

AI USE STATEMENT

We used generative AI primarily to polish the manuscript, including improving grammar, clarity, and readability. The research ideas, methodology, experimental design, analysis, and conclusions were developed by the authors. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the content of the manuscript.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Melvin Avrami. Kinetics of phase change. i general theory. *The Journal of chemical physics*, 7(12): 1103–1112, 1939.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Zhuangzhuang Chen, Zhuonan Lai, Jie Chen, and Jianqiang Li. Mind marginal non-crack regions: Clustering-inspired representation learning for crack segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12698–12708, 2024. doi: 10.1109/CVPR52733.2024.01207.

Christoffer Edlund, Timothy R Jackson, Nabeel Khalid, Nicola Bevan, Timothy Dale, Andreas Dengel, Sheraz Ahmed, Johan Trygg, and Rickard Sjögren. Livecell—a large-scale dataset for label-free live cell segmentation. *Nature methods*, 18(9):1038–1045, 2021.

Rokas Gipiškis, Chun-Wei Tsai, and Olga Kurasova. Explainable ai (xai) in image segmentation in medicine, industry, and beyond: A survey. *Ict Express*, 10(6):1331–1354, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Syed Saad Javaid, Wade R. Lanning, and Christopher L. Muhlstein. The development of zones of active plasticity during mode I steady-state crack growth in thin aluminum sheets. *Engineering Fracture Mechanics*, 218:106540, 2019. doi: 10.1016/j.engfracmech.2019.106540.

William A Johnson. Reaction kinetics in process of nucleation and growth. *Transactions of Transactions of the American Institute of Mining and Metallurgical Engineers*, 135:416–458, 1939.

Namgyu Kang, Byeonghyeon Lee, Youngjoon Hong, Seok-Bae Yun, and Eunbyung Park. Pixel: Physics-informed cell representations for fast and accurate pde solvers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 8186–8194, 2023.

Taniya Kapoor, Abhishek Chandra, Daniel M Tartakovsky, Hongrui Wang, Alfredo Nunez, and Rolf Dollevoet. Neural oscillators for generalization of physics-informed machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13059–13067, 2024.

Jianwei Li, Jiali Wan, Long Sun, Tongxin Hu, Xingdong Li, and Huiru Zheng. Intelligent segmentation of wildfire region and interpretation of fire front in visible light images from the viewpoint of an unmanned aerial vehicle (UAV). *ISPRS Journal of Photogrammetry and Remote Sensing*, 220: 473–489, 2025. doi: 10.1016/j.isprsjprs.2024.12.025.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637, 2012.

Max Losch, Mario Fritz, and Bernt Schiele. Interpretability beyond classification output: Semantic bottleneck networks. *arXiv preprint arXiv:1907.10882*, 2019.

Mingjian Lu, Sameera Nalin Venkat, Jube Augustino, David Meshnick, Jayvic Cristian Jimenez, Pawan K Tripathi, Arafath Nihar, Christine A Orme, Roger H French, Laura S Bruckman, et al. Image processing pipeline for fluoroelastomer crystallite detection in atomic force microscopy images. *Integrating Materials and Manufacturing Innovation*, 12(4):371–385, 2023.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Vanessa Nunes, Margarida Moura, Sara F Silva, Débora Vareiro, Nicolas Auduge, Nicolas Borghi, and Jorge G Ferreira. Chromosome condensation mechanically primes the nucleus for mitosis. *The EMBO Journal*, pp. 1–31, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Tiago F. R. Ribeiro, Fernando Silva, José Moreira, and Rogério Luís de C. Costa. Burned area semantic segmentation: A novel dataset and evaluation using convolutional networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:565–580, 2023. doi: 10.1016/j.isprsjprs.2023.07.002.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Zhuoyuan Wang, Reece Keller, Xiyu Deng, Kenta Hoshino, Takashi Tanaka, and Yorie Nakahira. Physics-informed representation and learning: Control and risk quantification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21699–21707, 2024.

David J Warne, Ruth E Baker, and Matthew J Simpson. Optimal quantification of contact inhibition in cell populations. *Biophysical Journal*, 113(9):1920–1924, 2017.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

A Appendix

A Module Design and Details

A.1 Physics-Grounded Scale Prior

Let $\ell_d > 0$ denote the characteristic spatial length, in input pixels, of the mechanism associated with $\mathcal{D}_d$. A structure of this extent is most directly represented at feature levels whose resolution is commensurate with ℓ_d, so we define a scale prior $\mathbf{s}^{(d)} \in \Delta^{L-1}$ in logarithmic scale,

$$s_k^{(d)} = \frac{\exp\!\big(-\delta_k^2/(2\sigma^2)\big)}{\sum_{j=1}^{L} \exp\!\big(-\delta_j^2/(2\sigma^2)\big)}, \qquad \delta_k = \log_2 r_k - \log_2 \ell_d, \tag{5}$$

where $\sigma > 0$ controls how broadly the prior spreads across neighbouring levels. Every governing law we consider carries such an intrinsic scale: Avrami kinetics through a nucleation density that fixes a mean impingement spacing, logistic proliferation through a contact-inhibition range set by the cell diameter, DNA-content imaging through the nucleus scale.

Properties. The bandwidth interpolates between the degenerate regimes: as $\sigma \to \infty$ the prior is uniform and the pyramid is physics-agnostic, while as $\sigma \to 0$ it becomes a hard one-hot selection, so σ reads as a confidence and is the quantity we sweep in Appendix A. The prior depends on the data only through ℓ_d: it is a one-parameter family, and two processes whose characteristic lengths agree to within the level spacing select the same level. Resampling the input by a factor η maps $\ell_d \mapsto \eta\,\ell_d$ and translates the prior by exactly $\log_2 \eta$ octaves, so the grounding is a property of the process rather than of the acquisition resolution.

Measuring ℓ_d. In deployment $\mathbf{s}^{(d)}$ is proposed by a language model that reads $\mathcal{G}_d$ and the dataset metadata. Because the prior is a one-parameter family, we can check that proposal against the data with two estimators that involve no language model and draw on disjoint information. The first reads only the image: with g the mean-removed grey channel of $\mathbf{x}$ and $\mathcal{A}(\mathbf{v}) = \mathcal{F}^{-1}\{|\mathcal{F}\{g\}|^2\}$ its autocorrelation, normalised so $\mathcal{A}(\mathbf{0}) = 1$,

$$\hat{\ell}_d^{\,\mathrm{ac}} = \mathrm{med}_{\mathbf{x}\in\mathcal{D}_d} \min\{\, \|\mathbf{v}\|_1 : \mathcal{A}(\mathbf{v}) < e^{-1} \,\}, \tag{6}$$

the radius at which spatial correlation has decayed by $1/e$. The second reads only the mask: with $\mathcal{K}(\mathbf{m})$ the connected components of $\mathbf{m}$ above a noise floor,

$$\hat{\ell}_d^{\,\mathrm{cc}} = \mathrm{med}_{\mathbf{x}\in\mathcal{D}_d}\, \mathrm{med}_{\kappa\in\mathcal{K}(\mathbf{m})}\, 2\sqrt{|\kappa|/\pi}, \tag{7}$$

the median equivalent diameter of the annotated structures. Their agreement provides an independent check that the selected scale is not an architectural assumption; Appendix A reports that they order the three datasets identically and select the same pyramid level in each case, agreeing with the level the descriptor up-weights. We therefore use $\mathbf{s}^{(d)}$ as proposed and treat the estimators as a check rather than a replacement.

Logit-space initialisation. The prior enters the fusion logits in logit space rather than probability space,

$$\boldsymbol{\theta} \leftarrow \log \mathbf{s}^{(d)} \quad \Longrightarrow \quad \mathrm{softmax}(\boldsymbol{\theta}) = \mathbf{s}^{(d)}. \tag{8}$$

The parameterisation is not incidental. Because the forward pass applies a softmax, storing $\mathbf{s}^{(d)}$ directly applies it twice, which realises a $3.5{:}1$ prior as $1.06{:}1$ and a $42.5{:}1$ prior as $1.27{:}1$, all but erasing the physical signal. The logits remain trainable thereafter, so the prior supplies a grounded initialisation without fixing the final feature weights. A dataset-specific affine adapter then produces

$$\widetilde{\mathbf{F}}(u) = \boldsymbol{\gamma}_d \odot \mathbf{F}(u) + \boldsymbol{\beta}_d, \qquad \boldsymbol{\gamma}_d, \boldsymbol{\beta}_d \in \mathbb{R}^D, \tag{9}$$

broadcast over $u \in \Omega'$. Both branches originate from $\widetilde{\mathbf{F}}$, which is what satisfies D1.

A.2 Implementation Details

We use SegFormer-B5 Xie et al. (2021), SAM2 Hiera-L Ravi et al. (2025) and DINOv2 ViT-L/14 Oquab et al. (2023) as backbones. LoRA Hu et al. (2022) (r=4, α=8) is injected into the attention

Algorithm 1 PGAD-X inference

1: **Input:** images $\mathcal{D}_d$ and descriptor $\boldsymbol{\pi}_d = (t_d, \mathbf{s}^{(d)}, \tau_d, c_d)$
2: **Output:** $\{(H_{\text{seg}}(\mathbf{x}), H^*_{\text{vis}}(\mathbf{x})) : \mathbf{x} \in \mathcal{D}_d\}$
3: $\mathbf{q}_d \leftarrow [\mathbf{s}^{(d)}; \text{onehot}(\tau_d); c_d]$ ▷ PPE
4: $\mathbf{e}_d \leftarrow \text{CLIP}_{\text{text}}(t_d)$
5: $\mathbf{T}_d \leftarrow \text{reshape}_{n_p \times D}(\mathbf{W}_t \mathbf{e}_d)$; $\mathbf{S}_d \leftarrow \text{reshape}_{n_p \times D}(g_{\text{str}}(\mathbf{q}_d))$
6: $\mathbf{Z}_d \leftarrow [\mathbf{T}_d \| \mathbf{S}_d]\mathbf{W}_f$; $\mathbf{P}_d \leftarrow \mathbf{Z}_d \mathbf{W}_p$
7: **for** $\mathbf{x} \in \mathcal{D}_d$ **do**
8: $\{\mathbf{C}_k\}_{k=1}^{L} \leftarrow \text{Backbone}(\mathbf{x}; \mathbf{Z}_d)$ ▷ SFT
9: $\mathbf{P}_k \leftarrow \text{Resize}_{\Omega'}(\text{Conv}^{(k)}_{1\times1}(\mathbf{C}_k)),\ k \in [L]$
10: $\mathbf{F} \leftarrow \sum_k \text{softmax}(\boldsymbol{\theta})_k \mathbf{P}_k$
11: $\widetilde{\mathbf{F}} \leftarrow \boldsymbol{\gamma}_d \odot \mathbf{F} + \boldsymbol{\beta}_d$; $\mathbf{V} \leftarrow \text{Flatten}(\widetilde{\mathbf{F}})$
12: $\mathbf{A} \leftarrow \text{Attn}(Q{=}\mathbf{V}, K{=}\mathbf{P}_d, V{=}\mathbf{P}_d)$ ▷ CAEH
13: $H_{\text{seg}} \leftarrow \text{Decoder}(\text{LayerNorm}(\mathbf{V} + \mathbf{A}))$
14: $h \leftarrow \text{Up}_{\Omega}(\|\mathbf{A}(\cdot)\|_2)$; $H^*_{\text{vis}} \leftarrow h / \max_{u\in\Omega} h(u)$
15: **store** $(H_{\text{seg}}, H^*_{\text{vis}})$
16: **end for**
17: **return** all stored output pairs

projections of SegFormer, whose two deepest stages are also unfrozen; the SAM2 and DINOv2 wrappers do not expose their attention projections in a form the injection routine accepts, so for those two backbones the encoder stays frozen and Stage 1 trains only the pyramid, the dataset adapter and the decoder. The three backbones therefore differ in trainable capacity, which accounts for the spread in Stage-1 IoU; explanation comparisons are made within a backbone and are unaffected.

Text is encoded with a frozen CLIP ViT-B/32. The prefix encoder uses $D{=}256$ and $n_p{=}8$; FiLM contributes one linear layer per pyramid level. Stage 1 trains for up to 60 epochs with AdamW at learning rate 10^{-4} and early stopping on a held-out split. Stage 2 adapts only the conditioning path, roughly 0.9M parameters, for 10 epochs at learning rate 10^{-4} with $\lambda{=}1$; the rank objective is computed on 4096 sampled locations per step with a soft-ranking temperature of 0.05.

On AFM, Stage 2 is trained on the morphologically derived corridors of the 32 frames that carry no expert annotation and evaluated on the 14 annotated frames, so the two image sets are disjoint and the two annotation procedures are not interchangeable (Appendix A). On the remaining datasets we use k-fold cross-validation over images with $k{=}5$ on LIVECell and $k{=}3$ on BBBC039, so every image is predicted exactly once by a model that did not see it.

B Algorithm Analysis

B.1 Line-by-Line Walkthrough

Lines 3–6: descriptor encoding (PPE). Line 3 assembles the structured terms of the descriptor into a single vector $\mathbf{q}_d \in \mathbb{R}^{L+3}$: the scale prior, a one-hot encoding of the prior type, and the confidence. Line 4 embeds the natural-language statement of $\mathcal{G}_d$ with a frozen text encoder. Line 5 projects the two sources separately into n_p prefix tokens each; the reshape is what produces distinct tokens rather than n_p copies of one vector, and Appendix A shows that replacing it with a broadcast collapses the readout to a spatially constant map. Line 6 fuses the two token sets row-wise and projects them into the key and value space of the explanation head. All four lines depend only on d and are computed once per dataset.

Lines 8–11: conditioned visual representation (SFT). Line 8 runs the backbone with the prefix tokens active, so each stage is modulated by PAR and FiLM. Line 9 projects every pyramid level to a common width and resamples it to the fusion grid Ω'. Line 10 combines the levels with weights $\text{softmax}(\boldsymbol{\theta})$, whose initialisation carries the scale prior of Eq. (8). Line 11 applies the dataset adapter and flattens the result into visual tokens. The map $\widetilde{\mathbf{F}}$ produced here is the single representation from which both outputs are read, which is what satisfies D1.

Lines 12–14: readout (CAEH). Line 12 lets the visual tokens query the physics prefixes; the result $\mathbf{A}(u)$ is the physics-conditioned response at each location. Line 13 decodes the segmentation from $\mathrm{LayerNorm}(\mathbf{V} + \mathbf{A})$, so the mask and the explanation share the same enriched features. Line 14 reads the explanation as the per-location magnitude of $\mathbf{A}$, upsampled and normalised. This is the step that separates the explanation from $\nabla\mathcal{L}_{\mathrm{seg}}$: it is a function of $\mathbf{Z}_d$, so perturbing the descriptor changes it in one step, without a gradient path through the segmentation objective.

B.2 Cost Analysis

Let $N_d = |\mathcal{D}_d|$, $M = |\Omega'| = HW/r_{\mathrm{f}}^2$, and let T_{bb} and T_{dec} denote the backbone and decoder costs. Encoding the descriptor (lines 3–6) costs $T_{\mathrm{CLIP}}(t_d) + \mathcal{O}(512 n_p D + n_p D^2)$, is performed once per dataset, and is cached, so it amortises to nothing. A single pass of the per-image loop (lines 8–14) costs

$$\begin{aligned} T_{\mathrm{img}} = T_{\mathrm{bb}} + T_{\mathrm{dec}} \\ + \mathcal{O}\Big(\underbrace{\textstyle\sum_k \frac{HW}{r_k^2} D_k D + LMD}_{\text{fusion, lines 9–11}}\Big) \\ + \mathcal{O}\Big(\underbrace{(M{+}n_p)D^2 + M n_p D}_{\text{cross-attn, line 12}}\Big) \\ + \mathcal{O}\Big(\underbrace{MD + HW}_{\text{readout, line 14}}\Big). \end{aligned} \tag{10}$$

Training cost. Estimating ℓ_d is a one-off preprocessing step: autocorrelation by FFT costs $\mathcal{O}(N_d HW \log HW)$ and connected-component analysis $\mathcal{O}(N_d HW)$, both cached with no per-iteration overhead. Each optimisation step then costs

$$T_{\mathrm{train}}^{\mathrm{step}} = \mathcal{O}(T_{\mathrm{img}}) \quad \text{for Stage 1 and Stage 2 alike.} \tag{11}$$

Stage 1 back-propagates through the whole model, whereas Stage 2 updates only the conditioning path (roughly 0.9M parameters). Because those modules are injected throughout the backbone, gradients still traverse the downstream operations, so Stage 2 keeps the same asymptotic per-step cost as Stage 1 while optimising far fewer parameters. The ranking loss is evaluated on at most 4096 sampled locations, so its cost is independent of image resolution.

Inference cost. Inference requires no backward pass. With the descriptor prefixes cached, the total cost over the dataset is

$$T_{\mathrm{infer}} = \mathcal{O}\big(n_p D^2\big) + N_d\, T_{\mathrm{img}}. \tag{12}$$

Three observations follow. First, because n_p is small and fixed ($n_p{=}8$) against $M{=}3136$ visual tokens, the cross-attention term is linear in M, unlike visual self-attention whose pairwise cost is quadratic. Second, the explanation adds only the $\mathcal{O}(MD + HW)$ norm and upsampling over what the segmentation branch already computes, since $\mathbf{A}$ is needed for the mask in any case; in our measurements this is under one percent of the forward pass. Third, no backward pass is required, unlike gradient attribution, which must differentiate the segmentation loss for every image it explains.

C Complementary Experiments

C.1 Validating the Evaluation Protocol

Trivial floors. Table 2 reports, for each dataset, the fraction of the frame retained by the admissible domain, the positive rate within it, and the score attained by maps carrying no information about the governing process. These are not competing methods but bounds on what is obtainable without any physical reasoning. Two of them, the predicted mask and the boundary distance transform, are computed from ground-truth segmentations, whereas our model observes only the image; the comparison is therefore conservative in their favour.

Which floor is strongest differs by dataset, and the pattern follows the physics. On BBBC039 the raw intensity leads at 0.637, since condensed chromatin is locally brighter. On AFM the inverted

Table 2: **Trivial floors** within Ω_{adm}, reported as AUC / AP. "dom" is $|\Omega_{\text{adm}}|/|\Omega|$ and "pos" the positive rate within the domain, which is also the chance level for AP.

	AFM	**LIVECell**	**BBBC039**
n images	14	39	197
dom	0.364	1.000	0.198
pos	0.043	0.153	0.253
Random	0.507 / 0.045	0.501 / 0.153	0.500 / 0.254
Raw intensity	0.185 / 0.031	0.497 / 0.148	0.637 / 0.400
Inverted intensity	0.815 / 0.151	0.503 / 0.149	0.363 / 0.196
Predicted mask	0.500 / 0.043	0.480 / 0.149	0.500 / 0.253
Boundary distance	0.740 / 0.066	0.562 / 0.157	0.438 / 0.224

intensity leads at 0.815: impingement corridors are thin dark lines, so a single negation of the grey channel recovers much of the prior. On LIVECell no floor separates from chance by more than 0.06. PGAD-X exceeds the strongest floor of each dataset on eight of the nine dataset and backbone pairs.

Effect of the admissible domain. Restricting the score to Ω_{adm} removes locations at which $\mathcal{G}_d$ is inactive by definition. Scored on the full frame instead, the floors rise sharply, because a large trivially negative region is added to every comparison.

Table 3: **Full frame versus admissible domain** (AUC). The mask-derived floors use ground-truth segmentations.

	AFM		**BBBC039**	
Map	full	Ω_{adm}	full	Ω_{adm}
Raw intensity	0.073	0.185	0.934	0.637
Inverted intensity	0.927	0.815	0.066	0.363
Predicted mask	0.345	0.500	0.884	0.500
Boundary distance	0.892	0.740	0.778	0.438

On BBBC039 the effect is stark. Scored on the full frame, using the raw image as the heatmap attains 0.934 and copying the segmentation mask attains 0.884, because both separate the nuclei from a background that occupies four fifths of the image and contains no DNA at all. Within the nuclear interior the same two maps fall to 0.637 and to exactly 0.500: the mask is constant there and cannot distinguish condensed from interphase nuclei. Without the domain restriction a method could report a score above 0.9 while expressing nothing about chromatin condensation.

C.2 The Explanation Readout

Gradient versus cross-attention. Table 4 isolates the readout by reading both heads off the same Stage-1 checkpoint with no alignment training, so any difference is attributable to the head rather than to optimisation.

The alignment stage improves the readout on all nine pairs, by between 0.045 and 0.415. The claim of Section 2.2 is not that the cross-attention response is intrinsically a better detector of physical activity, but that it is a readout an alignment objective can move, whereas a gradient map is pinned to the segmentation objective and cannot be moved without moving the segmentation with it.

Distinct conditioning tokens. Section 3.1 projects the descriptor to n_p distinct prefix vectors rather than replicating one. Replication makes the prefixes identical, the attention over them uniform, and the readout spatially constant. Table 5 also truncates $\mathbf{Z}_d$ to its first k rows at inference: $k{=}1$ reproduces the degenerate case exactly, and beyond $k{=}2$ the readout is stable, so what the descriptor must supply is several distinct query directions rather than a particular number of them.

The replicated variant attains exactly the chance score, 0.500 in AUC and 0.253 in AP against a positive rate of 0.253, the signature of a spatially uniform map under our ranking metric. The same

Table 4: **Explanation head**. The first two columns are zero-shot from the Stage-1 checkpoint.

Dataset	Backbone	Grad CAM	Cross attn	After Stage 2
AFM	SegFormer	0.713	0.755	**0.824**
AFM	SAM2	0.722	0.415	**0.830**
AFM	DINOv2	0.648	0.706	**0.837**
LIVECell	SegFormer	0.492	0.459	**0.699**
LIVECell	SAM2	0.485	0.516	**0.561**
LIVECell	DINOv2	0.498	0.525	**0.719**
BBBC039	SegFormer	0.600	0.521	**0.758**
BBBC039	SAM2	0.610	0.472	**0.712**
BBBC039	DINOv2	0.552	0.528	**0.748**

Table 5: **Prefix multiplicity** (BBBC039, SegFormer-B5, zero-shot). Pairwise cosine is measured between the n_p=8 prefix vectors; spatial std is over the 224×224 heatmap. Absolute values come from a separate evaluation path and may differ from Table 1; the comparison of interest is between rows.

Setting	Pairwise cos	Spatial std	AUC	AP
Replicate $1 \to n_p$	1.000	0.0000	0.500	0.253
Distinct n_p vectors	0.476	0.1268	0.532	0.287
$k = 1$	–	0.0000	0.500	0.253
$k = 2$	–	0.1163	0.559	0.305
$k = 4$	–	0.1347	0.515	0.275
$k = 8$	–	0.1153	0.528	0.276

map would score 0.253 under a thresholded overlap measure, which binarises it to the full frame; that value is indistinguishable from a sparse but correctly placed map, and the collapse would go unnoticed.

Descriptor content. We separate the two descriptor channels by retraining Stage 2 with each replaced in turn, holding everything else fixed. We report AFM because Table 7 shows the scale prior to have a large zero-shot effect there.

Table 6: **Descriptor channels** (AFM, SegFormer-B5, AUC / AP within Ω_{adm}). Every row retrains Stage 2 from the same Stage-1 checkpoint with n_p=8 distinct prefixes.

Descriptor	AUC / AP
Full (ours)	0.824 / 0.196
Text from another dataset	0.827 / 0.197
Non-scientific text	0.829 / 0.198
Text embedding zeroed	0.812 / 0.184
Scale prior flattened	0.818 / 0.189
Both channels removed	0.808 / 0.181

The six settings span 0.021 in AUC, within the spread between backbones in Table 1. What the descriptor supplies to the trained model, and what Table 5 shows to be necessary, is a set of mutually distinct query directions. Their particular values matter most before alignment: Table 7 shows a zero-shot spread of 0.318 to 0.790 on this dataset. Since the pyramid logits are themselves trainable, the prior determines where optimisation starts rather than where it ends, which is what makes it valuable when the alignment stage is short or the annotated data are few.

C.3 Scale Grounding

Sensitivity to the prior. We sweep $\mathbf{s}^{(d)}$ over four single-level-dominant settings and a uniform control, holding the descriptor and all trained parameters fixed, and read the explanation out zero-shot. If the prior were inert, all settings would coincide.

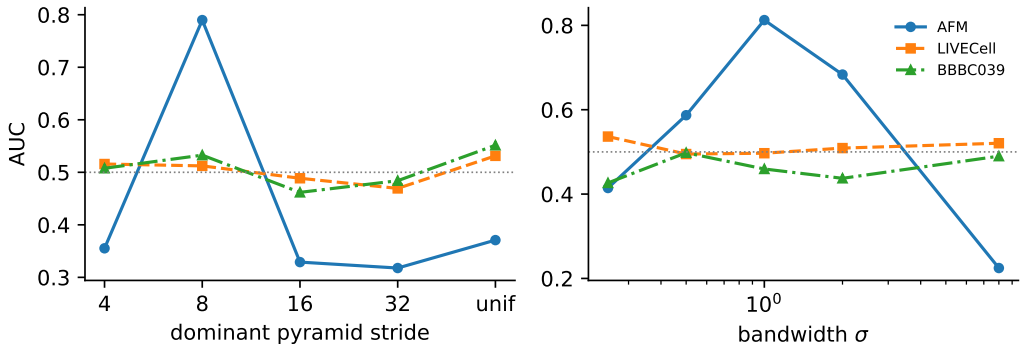

Figure 4: **Scale sensitivity** (AUC within Ω_{adm}, SegFormer-B5, zero-shot). Left: each probe places 0.70 of the mass on one pyramid level. Right: bandwidth σ of Eq. (5) with ℓ_d fixed at its measured value. The dashed line marks the chance level. On AFM the choice of level spans 0.318 to 0.790 and the bandwidth has a clear optimum at $\sigma{=}1$.

Table 7: **Scale sensitivity probe** (AUC within Ω_{adm}). "Matched" uses Eq. (5) with the measured ℓ_d of Table 8.

Dataset	s4	s8	s16	s32	unif	Matched
AFM	0.355	**0.790**	0.329	0.318	0.371	0.759
LIVECell	0.516	0.512	0.489	0.469	**0.531**	0.495
BBBC039	0.508	0.533	0.462	0.484	**0.551**	0.468

On AFM, scale selection has a substantial zero-shot effect. The matched smooth prior performs strongly, although the best single-level probe occurs at stride 8 rather than at the nominal stride-16 peak.The effect is present where the structures of interest are large relative to the finest pyramid stride, as the 13.4 px grains are, and smaller where they are close to it, as the 3.4 px cells and 7.1 px nuclei are. These are zero-shot measurements, so they characterise the starting point the prior supplies to Stage 2 rather than a constraint that persists through it.

Measured characteristic length. The prior reduces to a single scalar, which we measure with the two estimators of Eqs. (6)–(7). The two estimators order the datasets identically, and the ordering

Table 8: **Measured characteristic length**, computed on $224{\times}224$ inputs.

Dataset	Autocorr.	Mask size	Matched level	Descriptor peak
LIVECell	3.4 px	6.4 px	stride 4	stride 4
BBBC039	7.1 px	11.4 px	stride 8	stride 4/8
AFM	13.4 px	27.9 px	stride 16	stride 16

matches the physical scale of the underlying structures: cells, then nuclei, then grains. Their absolute values differ by roughly a factor of two, as expected since a component diameter exceeds a correlation radius, but the selected level is unchanged and agrees with the level the descriptor up-weights in all three cases.

C.4 THE ALIGNMENT OBJECTIVE

Why a spatial signal is required. Section 3.2 argues that an objective acting only in representation space cannot constrain a spatial output. Table 9 makes this concrete.

Table 9: **Stage-2 objective** (BBBC039, SegFormer-B5). "final" is the alignment term at the end of training. The two objectives above the rule never observe $p(\mathbf{x})$; the two below use it with different loss geometries. Absolute values come from a separate evaluation path and may differ from Table 1; the comparison of interest is between rows.

Objective	final	S2 IoU	AUC	AP
Cosine to prefix	0.004	0.859	0.472	0.250
Anti-corr. with H_{seg}	-0.929	0.859	0.470	0.235
Pointwise BCE on p	0.148	0.858	0.712	0.454
Rank corr. on p (ours)	0.449	0.853	**0.758**	**0.516**

The cosine objective is driven to 0.004, a cosine similarity of 0.996 between the explanation-weighted summary and the prefix, and the resulting map is no better than the untrained readout. The anti-correlation objective reaches -0.93 and behaves the same way. Both are optimised successfully and both leave the quantity of interest untouched: the normalisation in $\mathbf{v}_{\text{vis}}$ leaves the spatial shape of H^*_{vis} free, so the objective is satisfied by rotating $\widetilde{\mathbf{F}}$ instead.

Rank versus pointwise supervision. The last two rows of Table 9 access the same signal and differ only in the geometry of the loss. Table 10 repeats the comparison on all nine pairs.

Table 10: **Pointwise versus rank objective** on all nine pairs, reported as AUC / AP within Ω_{adm}.

Dataset	Backbone	Pointwise	Rank (ours)
AFM	SegFormer	0.809 / 0.118	**0.824 / 0.196**
AFM	SAM2	0.806 / 0.100	**0.830 / 0.195**
AFM	DINOv2	0.832 / 0.126	**0.837 / 0.204**
LIVECell	SegFormer	0.607 / 0.197	**0.699 / 0.258**
LIVECell	SAM2	0.521 / 0.163	**0.561 / 0.173**
LIVECell	DINOv2	0.626 / 0.212	**0.719 / 0.273**
BBBC039	SegFormer	0.714 / 0.456	**0.758 / 0.516**
BBBC039	SAM2	0.680 / 0.425	**0.712 / 0.466**
BBBC039	DINOv2	0.698 / 0.464	**0.748 / 0.512**

The rank objective is better on all nine pairs, and by more in AP than in AUC, consistent with the reason for preferring it: H^*_{vis} is defined only up to a monotone rescaling, so a pointwise loss spends part of its gradient on differences in value that carry no meaning, and does so most where the positives are sparse.

Auxiliary head and trainable scope. To separate the contribution of the physics descriptor from that of supervision on $p(\mathbf{x})$, we train a matched auxiliary head: a three-layer convolutional head (0.97M parameters) on the frozen Stage-1 features $\widetilde{\mathbf{F}}$, supervised by the same rank objective and cross-validation folds, but with no physics prefix, no scale prior, and no FiLM conditioning. We also vary which modules Stage 2 may update. Table 11 reports both on BBBC039.

On BBBC039 the auxiliary head matches or exceeds PGAD-X on all three backbones, and a Stage 2 that updates only the decoder and pyramid, without the conditioning pathway, does as well as one that includes it. On this dataset the gain over the unaligned readout is therefore attributable to supervision on $p(\mathbf{x})$ rather than to the physics descriptor. This is expected: BBBC039 evaluates against the same rule that supervises alignment, so directly fitting that rule is optimal, and the dataset cannot distinguish the two. The same comparison on AFM, where the evaluation reference is independent of the alignment target, is reported in Table 12.

Table 11: **Auxiliary head and trainable scope** (BBBC039, AUC / AP within Ω_{adm}). Upper block: matched auxiliary head versus PGAD-X. Lower block: Stage-2 trainable modules, SegFormer-B5 (S2 IoU in parentheses).

Backbone	Auxiliary head	PGAD-X
SegFormer-B5	0.785 / 0.562	0.758 / 0.516
SAM2 Hiera-L	0.784 / 0.558	0.712 / 0.466
DINOv2	0.770 / 0.539	0.748 / 0.512
Stage-2 trainable modules (SegFormer-B5)		
PPE + adapters + FiLM (ours)	0.758 / 0.516	(0.853)
+ decoder	0.777 / 0.542	(0.894)
+ decoder + pyramid	0.780 / 0.555	(0.894)
decoder + pyramid only	0.777 / 0.550	(0.895)

Table 12: **Auxiliary head on AFM** (SegFormer-B5, AUC / AP against the 14 expert-annotated frames). "n" is the number of alignment frames; for $n < 29$ we average three random subsets.

	Auxiliary head	PGAD-X
All alignment frames	0.833 / 0.198	0.824 / 0.196
$n = 4$	0.821	0.794
$n = 8$	0.830	0.804
$n = 16$	0.837	0.820
$n = 29$	0.833	0.828

On the independent AFM reference the two are on par with all alignment frames, and the auxiliary head is ahead by 0.02–0.03 with fewer frames. The descriptor therefore does not improve agreement with the expert reference over direct supervision on $p(\mathbf{x})$. Where the two differ is in what the map carries beyond that reference, which we test next.

Temporal validation on AFM. The 46 AFM frames form one time series of the same field of view (median containment of $\mathbf{m}_t$ in $\mathbf{m}_{t+1}$: 0.975). We therefore ask whether a map computed on frame t predicts the material that transforms between t and $t{+}k$, scoring $y = \mathbf{1}[\mathbf{m}_{t+k} = 1, \mathbf{m}_t = 0]$ within the untransformed region of frame t. No method observes later frames.

Table 13: **Temporal validation** (AFM). AUC over all frame pairs, AUC restricted to the 14 evaluation frames, and AP. The first three rows do not depend on the backbone.

		$k = 1$			$k = 3$		
Map	**Backbone**	AUC	AUC_{eval}	AP	AUC	AUC_{eval}	AP
Random	–	0.496	0.499	0.080	0.500	0.500	0.155
Boundary distance	–	0.939	0.849	0.541	0.919	0.798	0.702
Rule $\Gamma_d(\mathbf{m}_t)$	–	0.778	0.653	0.206	0.763	0.657	0.374
GradCAM	SegFormer	0.863	0.710	0.418	0.780	0.618	0.525
Auxiliary head	SegFormer	0.773	0.692	0.263	0.667	0.619	0.406
PGAD-X	SegFormer	0.865	0.729	0.297	0.782	0.656	0.450
GradCAM	SAM2	0.895	0.814	0.457	0.777	0.680	0.508
Auxiliary head	SAM2	0.882	0.724	0.290	0.803	0.671	0.460
PGAD-X	SAM2	0.898	0.740	0.307	0.895	0.732	0.556
GradCAM	DINOv2	0.854	0.773	0.409	0.758	0.666	0.479
Auxiliary head	DINOv2	0.877	0.719	0.302	0.841	0.672	0.483
PGAD-X	DINOv2	0.867	0.708	0.263	0.822	0.660	0.427

With SegFormer and SAM2, PGAD-X, aligned to the same rule as the auxiliary head, predicts future transformation better than the head, and the margin widens with the horizon: from $+0.092$ and $+0.016$ in AUC at k=1 to $+0.115$ and $+0.092$ at k=3 (results at k=2 lie between), with the

same ordering on the evaluation frames alone. With DINOv2 the auxiliary head is ahead at every horizon (0.841 versus 0.822 at k=3). The descriptor-conditioned map thus carries information about the ongoing transformation beyond what direct supervision on $p(\mathbf{x})$ recovers on two of the three backbones, but not on all. Two further qualifications apply. GradCAM is competitive over short horizons and exceeds PGAD-X in AP at k=1 on every backbone, and distance to the crystal boundary is better than every learned map: over short horizons the film grows along all fronts, whereas $p(\mathbf{x})$ targets the impingement corridors that are consumed last. Temporal prediction is therefore a partial validation of the map's physical content rather than evidence that it localizes the fastest-changing material.

Supervision weight. The weight λ sets how strongly the alignment term competes with segmentation. Explanation quality rises steeply between λ=0 and λ=0.1 and is nearly flat thereafter, rising by 0.019 across a fiftyfold further increase. At λ=0 the alignment term is absent and Stage 2 reduces to further segmentation training, which alone lifts the readout to 0.562 and segmentation IoU to 0.861 (versus 0.853 at λ=1), so the IoU change after Stage 2 is due to this additional training rather than to alignment. The Stage-1 BBBC039 model was trained on 16-bit images linearly rescaled to 8 bits, whereas Stage 2 and all explanation scores use per-image percentile scaling. Evaluated with its training-time scaling, Stage 1 reaches IoU 0.859, 0.879 and 0.684 (SegFormer, SAM2, DINOv2), which Table 1 reports; under percentile scaling it drops to 0.638, 0.314 and 0.516. Stage 2 absorbs this shift through its conditioning modules, which is why λ=0 recovers the same IoU. We use λ=1, where the explanation is within 0.002 of its best value over the sweep.

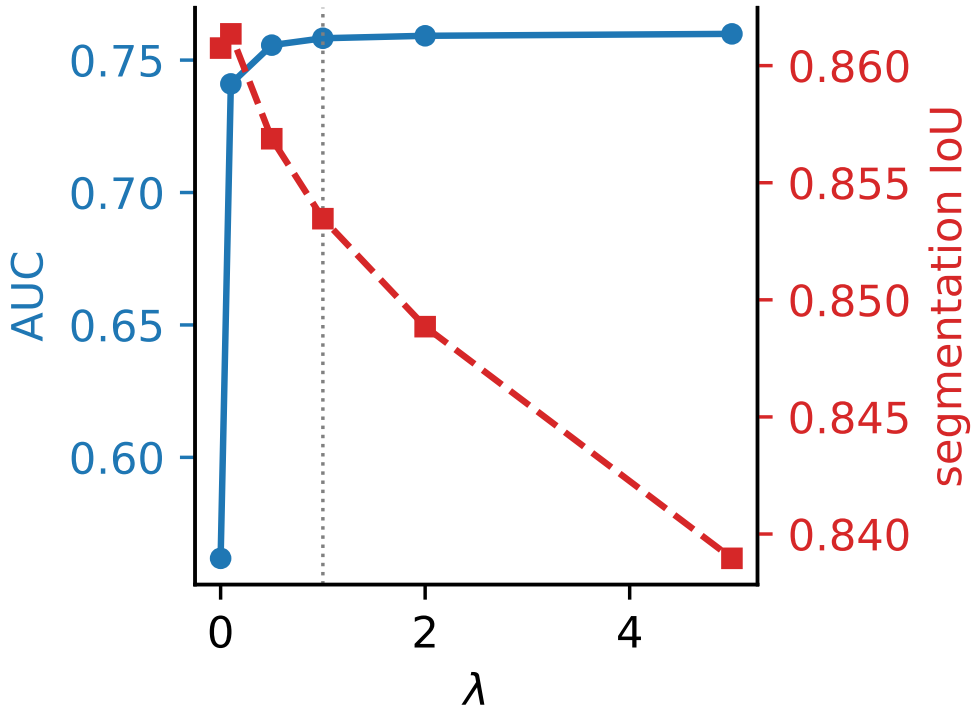

Figure 5: **Supervision weight** λ (BBBC039, SegFormer-B5). Explanation quality rises steeply between λ=0 and λ=0.1 and is nearly flat thereafter. We use λ=1, marked, near the knee of the curve.

C.5 Robustness of the Reported Margins

Variance across seeds. Stage 2 is stochastic through the ordering of training images and, for the rank objective, through the random subsample of locations drawn at each step.

Table 14: **Variance across three seeds** (SegFormer-B5, mean and standard deviation over seeds).

Dataset	AUC	AP
AFM	0.823 ± 0.004	0.197 ± 0.001
LIVECell	0.697 ± 0.003	0.259 ± 0.002
BBBC039	0.761 ± 0.003	0.517 ± 0.003

The seed-to-seed spread is at most 0.004 in AUC and 0.003 in AP, an order of magnitude below the margins reported in Table 1, so those margins are not attributable to the choice of seed.

Construction of the AFM prior. On AFM the Stage-2 target and the evaluation reference are produced by different procedures. The target $p(\mathbf{x})$ is a morphological closing of the segmentation mask, which fills the narrow corridors between approaching crystallites; the reference $p^\star(\mathbf{x})$ is a set of expert annotations on 14 frames, drawn independently of the masks. Table 15 compares the two, together with a watershed reconstruction that partitions the transformed phase into individual grains and takes the resulting ridges.

Table 15: **Agreement with expert annotation** on the 14 annotated AFM frames. Precision and recall use a 5 px tolerance.

Construction	P@5px	R@5px	F_1
Morphological closing (r=11)	0.556	0.981	0.707
Watershed ridges	0.084	0.144	0.100

The closing recovers 98% of the expert-drawn corridors within 5 px. Its precision is lower because it returns every narrow gap in the mask, whereas the expert marked only those where two crystallites were judged to be actively approaching. The watershed reconstruction, which uses only the geometry of the transformed phase, recovers almost none of the expert annotation, so $p^\star$ is not a deterministic function of the mask and the target and the reference are not interchangeable. The mean per-image correlation between $\mathbf{m}$ and $p^\star(\mathbf{x})$ is -0.10.

Sensitivity to the closing radius. The radius is the one free parameter of this construction. Table 16 rebuilds the target at five radii, retrains Stage 2 on each, and evaluates against the same expert annotations.

Table 16: **Closing radius** (AFM, SegFormer-B5).

r	AUC	AP
7	0.828	0.191
9	0.825	0.191
11	0.825	0.192
13	0.821	0.187
15	0.814	0.184

The result varies by 0.014 in AUC across a radius range of more than a factor of two, comparable to the seed-to-seed spread of Table 14, and the value we use is not the best of the five. The construction is therefore insensitive to this choice within the range where the closing still corresponds to a narrow corridor.

D Qualitative Comparison across Datasets

Fig. 6 extends the case study of Section 4 to all three datasets, two frames each, chosen as the highest-scoring held-out images. On LIVECell and BBBC039 the frames come from the fold the displayed checkpoint did not train on.

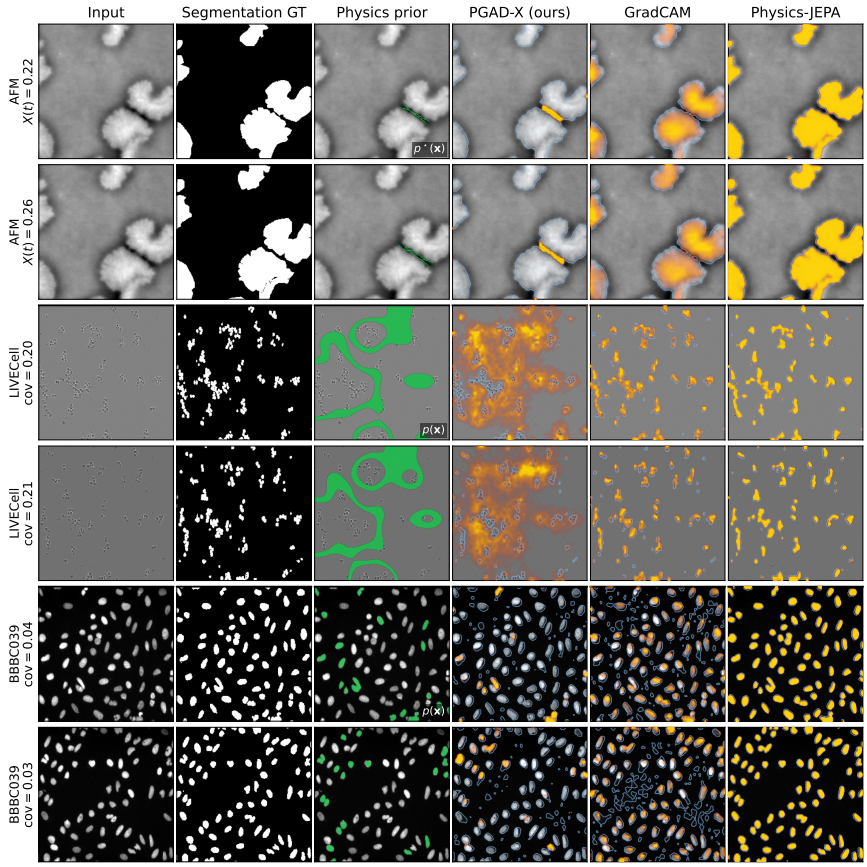

Figure 6: **Qualitative comparison across all three datasets** (SegFormer-B5). Columns: the input; the segmentation target $\mathbf{m}$; the physics prior in green, $p^\star(\mathbf{x})$ on AFM and $p(\mathbf{x})$ elsewhere; and the explanation produced by PGAD-X, by a GradCAM readout and by Physics-JEPA. In the last three columns light blue is that method's own predicted segmentation and red-to-yellow is its explanation, with locations below the display threshold left transparent. Physics-JEPA has no explanation head, so its map is its segmentation output. Rows are labelled by transformed fraction on AFM and by prior coverage elsewhere.

AFM: which interfaces are still advancing. The prior is a set of thin corridors, and PGAD-X responds along them, most strongly where two crystallites have closed to a narrow gap. What this offers a practitioner is a spatial reading of the impingement term $(1 - X)$ in the Avrami law, which assumes the untransformed fraction is available for growth everywhere. The map shows which part of it actually is: interfaces still facing open matrix, as against interfaces that have already met and stopped. Comparing the two rows, at $X(t) = 0.22$ most crystallites are isolated and the corridors are few, while at $X(t) = 0.26$ several have begun to converge. A rate constant fitted under the well-mixed assumption becomes progressively less trustworthy as that transition proceeds, and the map indicates when. GradCAM instead covers the crystallite interiors, which is where $\mathrm{d}X/\mathrm{d}t$ is exactly zero, and so reports the opposite of what is being asked.

LIVECell: where proliferation is not yet inhibited. Here the prior is a region rather than a curve, marking where the local density sits near half saturation. PGAD-X responds over that region and stays out of both the empty field and the densest clusters, the two regimes in which $\dot{N} = \rho N(1 - N/N_\infty)$

vanishes. The practical reading is a map of where a culture is still in its exponential phase. A biologist estimating a doubling time from a single field has to assume that field is representative; the map shows which parts of it are, and which have already saturated and would bias the estimate downward. The distinction is invisible in the mask, since a saturated cluster and a growing colony are both simply cells, and GradCAM accordingly responds to cell bodies wherever they appear, irrespective of their local density.

BBBC039: which nuclei are dividing. The prior is neither a curve nor a region but a subset of the segmented objects, so the question is not where in the image to look but which of the nuclei are condensing. PGAD-X responds on individual nuclei and leaves their neighbours dark, which is what a mitotic index requires: the fraction of nuclei in division, not the number of nuclei present. The mask already gives the latter. GradCAM covers essentially every nucleus, and this is visible in Table 1 as an AP of 0.305–0.367 against a positive rate of 0.253, only modestly above assigning the base rate to every nucleus in the field. It recovers the mask rather than the subset of it that is dividing.

Physics-JEPA returns the segmented phase entire on all three datasets, since it has no spatial readout to return anything else, and its two stages therefore produce identical maps.

BBBC039 also illustrates why the admissible domain matters. Scored on the full frame, a map that simply highlights the nuclei separates them from a background occupying four fifths of the image and attains an AUC above 0.88 (Table 3); scored within the nuclei, where the question is which of them are condensing, the same map is constant and scores exactly at chance. The visual difference between the GradCAM and PGAD-X columns on those rows is precisely the difference the restriction is designed to measure.

E Related Work

We provide a more detailed comparison of related work.

Scientific image segmentation and visual attribution. General - purpose vision backbones, including SegFormer Xie et al. (2021), SAM2 Ravi et al. (2025), and DINOv2 Oquab et al. (2023), are increasingly adapted to scientific images, often through parameter-efficient fine-tuning Hu et al. (2022). These models produce accurate segmentation masks, but the masks do not indicate whether the learned representation captures the physical or biological mechanism of interest.

Visual explanation methods typically attribute a prediction to influential pixels, features, or learned concepts. Representative approaches include gradient-based attribution Selvaraju et al. (2017); Sundararajan et al. (2017), concept or bottleneck explanations Losch et al. (2019), prototype-based reasoning Chen et al. (2019), and game-theoretic attribution Lundberg & Lee (2017).

Existing work on segmentation explanation similarly focuses primarily on evidence supporting the predicted mask Gipiškis et al. (2024). Our objective is different: we seek to localize the *active dynamics* specified by a governing law. Because these regions may not coincide with the segmentation target, an attribution map can be faithful to the prediction while remaining misaligned with the scientific mechanism. PGAD-X therefore derives both outputs from shared visual features, but generates the explanation independently of the segmentation gradient.

Physics-informed and physics-conditioned learning. Physics-informed neural networks incorporate differentiable equation residuals into model training Raissi et al. (2019), while neural operators learn mappings between fields or solution functions Li et al. (2020). Related methods use physical structure to improve cell representations Kang et al. (2023), control and risk estimation Wang et al. (2024), and dynamical-system modeling Kapoor et al. (2024). These approaches primarily constrain or improve the model's predictions; they do not address how to produce a spatial explanation of where the governing dynamics are active.

Our setting may involve empirical laws with unknown parameters, for which direct equation-residual supervision is unavailable. Rather than solving or enforcing the complete governing equation, PGAD-X encodes it as a compact descriptor and extracts a measurable characteristic length that connects the mechanism to an appropriate visual feature scale. We use FiLM Perez et al. (2018) and lightweight adapters to condition the shared representation, followed by a cross-attention branch that produces an explicit spatial explanation.

Spatial explanation and physics alignment. Spatial explanations are commonly evaluated as localization outputs using pointing games or threshold-based overlap measures Selvaraju et al. (2017). Such evaluation can be misleading in our setting: large parts of an image may be physically incapable of exhibiting the mechanism, and reference maps may be partly approximated from image intensity, object boundaries, or the segmentation mask. We therefore evaluate within a physics-admissible domain using threshold-free ranking metrics, and compare against explicit image- and mask-derived reference floors.

Joint-embedding methods such as I-JEPA Assran et al. (2023) align representations by predicting target embeddings in latent space. Although such alignment can encode physics-related information, latent agreement alone does not identify where that information is expressed spatially. We adapt this design as a Physics-JEPA baseline to distinguish latent physics alignment from active-dynamics localization. In contrast, PGAD-X injects the physics descriptor into spatial backbone features and directly produces a physics-aligned explanation map.