Sequential State Estimation for Regime and Transition Detection in Annulus Flow Videos

Alexey Smirnov[1] and Mikhail Voronin[2]

[1]Penza State University, 40 Krasnaya Street, Penza 440026, Russia
[2]Tver State University, 33 Zhelyabova Street, Tver 170100, Russia

Abstract

Flow involving gas and liquid phases inside a vertical annular channel develops into several large-scale flow patterns. These patterns are better understood by observing how they change over time, since their defining visual characteristics emerge through continuous evolution rather than from single, static snapshots. In experimental video footage, factors such as how long certain structures persist, how irregularly they appear or disappear, the way interfaces between phases rearrange, and short-lived mixing events are often just as significant as the momentary visual state captured in any single frame. For this reason, regime recognition based only on single-image classification can miss the temporal structure that makes transitions interpretable and operationally useful. This paper develops a sequential machine-vision framework for regime identification and regime-transition detection in vertical annulus videos. The formulation treats the observed image stream as a noisy projection of an evolving latent flow state and models regime assignment as a temporally coupled inference problem with uncertain boundaries. The proposed treatment integrates frame encoding, sequence representation, transition scoring, soft boundary supervision, and causal filtering so that stable intervals, mixed intervals, and onset events can be handled within one probabilistic pipeline. Particular attention is given to the mismatch between high frame-rate redundancy and comparatively slow regime evolution, the rarity of boundary events relative to stable segments, and the fact that clip-level human interpretation often provides more faithful supervision than isolated frame labels. The paper also specifies evaluation procedures suitable for streaming annulus data, including experiment-disjoint inference, interval-aware transition metrics, calibration near boundaries, and robustness under optical degradation. The resulting methodology is intended to support flow-regime recognition systems that can do more than assign frame labels, namely estimate evolving macrostate trajectories and localize the intervals in which one regime gives way to another.

1 Introduction

Gas–liquid flow in a vertical annulus is often described through a regime vocabulary that compresses complicated interfacial behavior into a manageable set of macrostates [1]. This vocabulary is useful because the pressure gradient, phase distribution, slip behavior, and even the plausibility of certain mechanistic closures depend strongly on how gas and liquid arrange themselves over time. Yet the visual evidence for a regime is seldom confined to a single image. In practice, one recognizes a regime by observing whether gas structures persist, whether dispersed texture remains statistically stable, whether elongated bodies recur with a characteristic cadence, or whether the interface reorganizes into a new pattern over an interval. From this perspective, frame-wise regime classification is a convenient approximation, not a complete statement of the recognition problem.

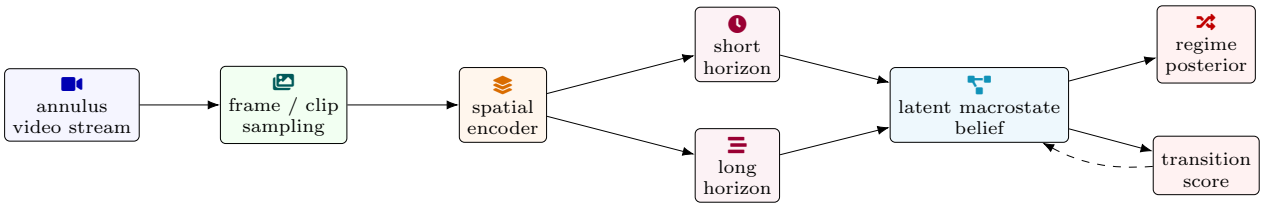

Figure 1: Overall sequential estimation pipeline for inferring the evolving regime state and the probability of transition from annulus flow video observations.

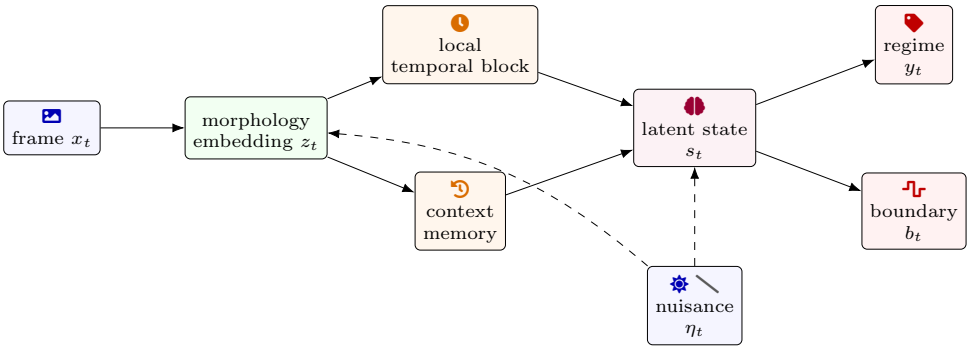

Figure 2: Latent-state view in which each image is encoded into morphology-aware evidence, fused with temporal context, and mapped to both regime identity and transition occupancy.

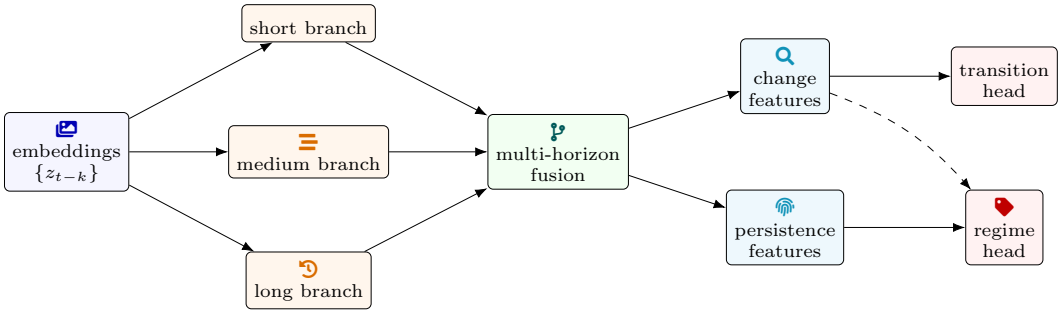

Figure 3: Multi-horizon temporal encoder that combines local continuity, clip-scale persistence, and longer reorganization cues before producing separate regime and change representations.

That distinction matters for two reasons. The first is physical. In annular gas–liquid flow, the same apparent structure may mean different things depending on whether it is newly emerging, steadily persisting, or already fragmenting into another morphology. A single elongated void region in one image can be the center of a stable intermittent pattern or merely the transient trace of a changing interval. The second reason is statistical. High-speed video produces extremely correlated data. Consecutive frames share nearly all low-level image content during stable operation, while the events most important for model improvement and operational interpretation are often the comparatively rare times at which the morphology actually changes. If a learning system treats all frames as independent and equally informative, then the dense stable segments dominate optimization and the transition structure is under-modeled.

The challenge is therefore not simply to improve a classifier but to redefine the inference target. Instead of predicting an isolated label for each frame, the model should estimate a temporally coherent regime process. It should know when it is inside a stable interval, when it is entering a changing interval, and how uncertainty should widen as the boundary is approached. This is a different objective from smoothing noisy frame predictions after training. Post hoc smoothing only regularizes the output sequence. It does not teach the model what a transition looks like, how long transitions tend to last, or which visual cues indicate that the currently dominant macrostate is losing persistence.

A sequence-based formulation also better matches how experts inspect annulus videos [2]. Human reviewers rarely decide regime from a frozen image alone when the sequence is available. They watch short clips, compare successive appearances, and use the progression of interface organization as evidence. High-speed visual observation has long been central to interpreting annular multiphase structures and evolving interfacial behavior in controlled experiments [3]. A machine-vision system should therefore exploit the same dimension of information instead of discarding it for the sake of a simpler benchmark task. The issue is not whether a frame contains useful visual information. It does. The issue is whether the frame alone expresses the semantics of regime persistence and transition timing that engineers actually care about. Often it does not.

The case for explicit temporal modeling becomes stronger when one considers ambiguous regions. Near regime boundaries, labels are rarely pointwise truths in a strict sense. A transition occupies a neighborhood of time whose exact extent depends on the operating condition, the optical viewpoint, and the granularity of the chosen taxonomy. Within that neighborhood, neighboring frames may each plausibly support more than one class. A good model should not be punished for preserving uncertainty there. Instead, it should separate two questions that are routinely conflated in static pipelines: which regime is currently most likely, and is the system in a stable interval or a changing one. The first question is categorical [4]. The second is dynamical. Together they define a

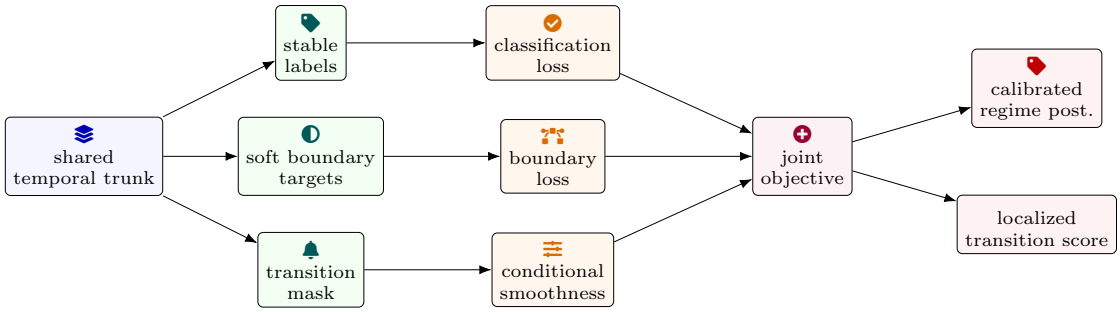

Figure 4: Training design that combines stable supervision, softened boundary targets, and transition-aware regularization to shape posterior behavior in both persistent and mixed intervals.

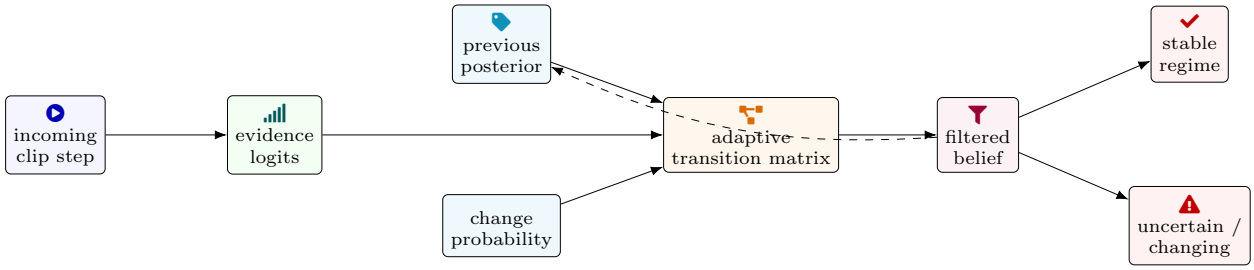

Figure 5: Causal filtering logic in which incoming evidence is combined with persistence-aware dynamics to produce streaming regime estimates and a guarded transition state.

richer and more faithful inference problem.

Temporal modeling is also useful for controlling failure modes. Static models commonly fail by producing jitter, by reacting to momentary artifacts such as glare or blur [5], or by remaining overconfident through boundary regions because the training objective rewards hard classification even where the visual evidence is inherently mixed [6]. A sequence-aware model can address these issues by using persistence priors, boundary-aware objectives, and uncertainty that evolves over time rather than independently at each frame. This does not eliminate all ambiguity, but it does create a more honest relationship between the model output and the physical process being observed [7].

The availability of substantial annulus image corpora makes this direction timely. In particular, Manikonda et al. (2025) describe a vertically oriented annulus image collection whose scale and annotation structure make it especially well suited to the development of temporally informed visual models [8], and the modeling study built around that collection makes clear that the image archive itself furnished the primary empirical basis from which the predictive system was trained, adjusted, and checked. When such data are available, it becomes artificial to insist on a formulation that breaks them into unrelated still images. The temporal continuity of the recording is itself part of the information content.

The present paper develops a different style of treatment from common benchmark-oriented discussions. The emphasis here is on sequential state estimation, transition geometry, and online decision logic. The video stream is regarded as the observation of an evolving latent macrostate. Regime recognition is posed as a structured inference problem. Transition detection is treated not as an afterthought or a threshold on posterior entropy, but as a supervised or weakly supervised objective in its own right. The paper first characterizes the temporal semantics of regime evolution, then introduces a latent-state sequence formulation, then develops temporal encoders and boundary-sensitive training objectives, then addresses supervision and labeling under interval uncertainty, and finally turns to streaming inference, evaluation, and operational design. The goal is to provide a technical framework in which the outputs of a vision system are interpretable as regime trajectories rather than merely as densely sampled isolated classifications.

2 Temporal Semantics of Regime Change

A regime label in annulus flow is often treated as if it were an instantaneous property of the observed image. In reality, it is a temporally aggregated description of morphology. The descriptive words commonly used for these flows already imply time. Terms associated with dispersed structures suggest persistence of a population pattern rather than one frozen bubble arrangement. Terms associated with elongated or intermittent structures suggest recurrence, duration, and continuity [9]. Terms associated with more agitated conditions suggest rapid

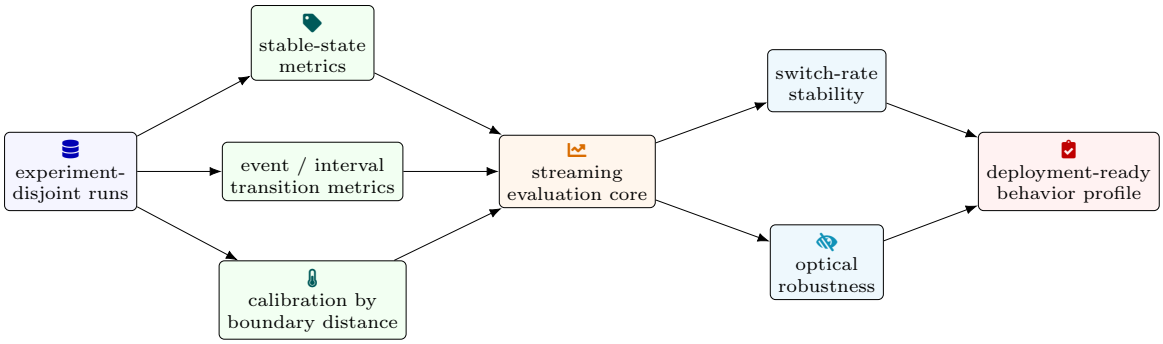

Figure 6: Sequence-aware evaluation protocol emphasizing experiment isolation, transition localization, temporal stability, calibration near boundaries, and resilience under degraded imaging.

Table 1: Problem framing for sequential regime inference in annulus videos

max width=0.82

Aspect	Static view	Sequential view	Main implication
Prediction target	Isolated frame label	Evolving macrostate trajectory	Time-aware inference
Core evidence	Instantaneous appearance	Persistence, recurrence, reorganization	Context becomes essential
Boundary treatment	Hard class switch	Extended mixed interval	Soft supervision needed
Error tendency	Jitter and overconfidence	Delay or misspecification of change	Balance stability and sensitivity
Operational value	Snapshot recognition	Regime and transition localization	Better monitoring utility

reorganization and local breakdown of coherence. The regime, then, is not only what is present in one frame but also how the pattern maintains or abandons itself across neighboring frames.

This observation changes the interpretation of both stable intervals and transitions. A stable interval is not simply a region where the modal class remains constant. It is a time span over which the visual evidence in favor of one macrostate continues to dominate despite local fluctuations. In a high-speed annulus video, even a stable interval can contain large short-time changes at the pixel level. Bubbles move, interfaces stretch, droplets appear or vanish, and specular reflections shift with the flow. Yet at a longer horizon the dominant structure remains recognizable. In other words, temporal stability lives at a coarser level than frame-to-frame similarity. Any method designed for this problem must respect that separation of scales.

Transitions are even more interesting. A transition is not merely a sign change in the identity of the most probable class. It is an interval in which the evidence supporting the old macrostate weakens while evidence for the new one strengthens, often with a period of coexistence or rapid alternation. Some transitions are sharp, with a relatively short mixed segment. Others are extended, with several sub-events before a new stable pattern emerges. If the task is forced into pointwise labels, the transition becomes an arbitrary boundary frame chosen by annotation convention. If the task is treated temporally, the transition becomes a structured object with onset, development, and completion. That interpretation is both more realistic and more useful for later analysis.

One can think of the observed flow as evolving on at least three time scales [10]. The fastest scale is the frame-to-frame visual fluctuation caused by motion and measurement noise. The intermediate scale is the local persistence of morphological motifs such as dispersed bubble texture or large coherent gas structures. The slowest scale is the regime trajectory through operating space over the duration of a run or a substantial segment of a run. The intermediate scale is where most regime evidence lives. The slowest scale is where transitions and operating changes are expressed. A static classifier only sees the fast and local part of this hierarchy. A temporal model can explicitly use the intermediate and slow scales without being distracted by the frame-to-frame redundancy.

Temporal semantics also matter for labeling. When a reviewer tags a clip as belonging to one regime, that label is usually understood as a summary of the clip rather than an assertion that every frame is equally central to the

Table 2: Temporal semantics of stable intervals and regime transitions

max width=0.82

Concept	Temporal meaning	Visual manifestation	Modeling need
Stable interval	Dominant morphology persists over time	Local pixel motion but consistent macro-pattern	Persistence prior
Transition onset	Old regime begins losing support	Early destabilization cues emerge	Change-sensitive features
Mixed interval	Competing evidence coexists	Alternation or coexistence of structures	Soft labels
Transition completion	New regime becomes credible and sustained	Clear new visual organization	Completion logic
Uncertainty	Dynamical rather than frame-local	Confidence widens near change	Temporal calibration

Table 3: Latent-state formulation and inference variables

max width=0.80

Symbol	Interpretation	Role in inference
x_t	Observed video frame at time t	Visual evidence entering the sequence model
s_t	Latent interfacial state	Hidden flow organization behind the image
y_t	Regime macrostate label	Main categorical output over time
b_t	Transition indicator or boundary state	Separates stable and changing intervals
η_t	Imaging nuisance factors	Accounts for blur, glare, and noise
$X_{1:t}$	Observation history up to time t	Supports causal posterior filtering

assigned class. Some frames are prototypical and some are peripheral. If a clip sits near a transition, the correct interpretation may be that the clip contains both a dominant regime and a trend away from it. Therefore, a training pipeline that assumes every clip-level label should be projected unchanged onto every included frame is imposing a precision the review process did not provide. A sequence-aware model can handle this by weighting the most representative frames more heavily or by using soft interval targets that broaden around suspected boundaries.

Another consequence concerns uncertainty. Temporal uncertainty is not just the accumulation of frame-wise uncertainty. A run can contain a sequence of individually high-confidence frames that nonetheless suggest an impending transition when interpreted together. Conversely, a few noisy frames inside a stable region should not force large regime uncertainty if the surrounding context remains consistent. Thus uncertainty itself is a dynamical quantity. It should increase when the latent macrostate is changing or when the evidence supporting persistence is decaying, not simply when one frame is visually awkward in isolation [11]. This makes transition detection and confidence estimation naturally coupled problems.

It is also useful to note that temporal semantics are affected by the chosen video scale. A given flow may appear stable at one frame stride and mixed at another. If one samples extremely sparsely, the model may miss the continuity of a transition. If one samples at the native high speed without any temporal abstraction, the model may overemphasize microscopic changes that are not relevant to regime interpretation. This means that temporal modeling is partly about choosing the right resolution of time for the intended inference. In annulus videos, that resolution is usually coarser than the camera frame interval but finer than the duration of a full operating condition segment. A good sequence model effectively learns or approximates that intermediate view.

The regime-transition problem can therefore be stated more precisely. The system should infer a macrostate trajectory that is temporally persistent, visually justified, and willing to broaden in uncertainty near changing intervals. It should identify not just a dominant class but the geometry of the change: where evidence starts to move, how long ambiguity persists, and when a new stable interpretation becomes credible. This is the sense in which temporal modeling is not a cosmetic extension of image classification. It changes the target object from isolated labels to a structured sequence.

Table 4: Multi-horizon temporal encoder design

max width=0.83

Module	Primary function	Strength in annulus videos	Limitation
Frame encoder ϕ_θ	Extract morphology-aware spatial features	Preserves texture and interface topology	No temporal memory
Short-horizon branch	Capture local continuity and near-term motion	Fast response to local structure	Limited persistence awareness
Medium-horizon branch	Model clip-scale persistence	Supports stable-regime interpretation	May blur sharp onset cues
Long-horizon branch	Track slow reorganization	Detects regime drift and recurring patterns	Higher latency
Fusion layer	Combine scales into h_t	Balances responsiveness and stability	Needs careful weighting

Table 5: Candidate sequence modules for streaming and offline analysis

max width=0.84

Sequence module	Main advantage	Best suited use case	Caution
Dilated temporal convolution	Efficient long receptive field with causality	Online monitoring	Fixed aggregation pattern
Attention-based pooling	Focus on diagnostically important moments	Boundary-rich offline analysis	Overfitting to run signatures
Recurrent unit	Compact state summarization	Resource-limited streaming	Can oversmooth if weakly supervised
3D spatiotemporal encoder	Joint space–time feature learning	Local motion and breakup cues	Sensitive to stride choice
Quality-aware weighting	Downweight degraded frames	Optical disturbance robustness	Requires reliable quality score

3 Latent-State and Sequential Inference Formulation

Let x_t denote the video frame observed at time t and let s_t denote an unobserved latent state summarizing the physically relevant interfacial organization within the visible annulus section. The latent state is not necessarily low dimensional in a literal sense, but it is conceptually useful because it separates what the flow is doing from how the camera renders it. The regime label y_t is a coarse macrostate associated with the recent evolution of s_t rather than only its instantaneous realization. A nuisance variable η_t captures illumination, blur, sensor noise, and other imaging effects. One may write

$$\begin{aligned} x_t &= \mathcal{G}(s_t, \eta_t) \\ y_t &= \mathcal{H}(s_{t-\ell:t+\ell}) \,. \end{aligned} \tag{1}$$

The first line says the image is a projection of the latent flow and nuisance factors [12]. The second says the regime label depends on a local trajectory of latent states. This is a compact way to formalize why sequence information matters: the target is trajectory-dependent.

For streaming inference, the more relevant quantity is the posterior over the current regime given observations up to time t. Let $X_{1:t}$ denote the history of observations. Then the causal objective is to estimate

$$p(y_t \mid X_{1:t}) \,. \tag{2}$$

If a Markov assumption is imposed on a latent regime process, a filtering relation follows:

$$\begin{aligned} p(y_t \mid X_{1:t}) \;&\propto p(x_t \mid y_t, X_{1:t-1}) \\ &\quad \times \sum_{y_{t-1}} p(y_t \mid y_{t-1})\, p(y_{t-1} \mid X_{1:t-1}) \,. \end{aligned} \tag{3}$$

Table 6: Training objectives for stable recognition and transition sensitivity

max width=0.84

Loss term	Purpose	Benefit	Typical scope
$\mathcal{L}_{\text{stab}}$	Classify confidently stable regimes	Preserves discrimination among canonical states	Stable interiors
$\mathcal{L}_{\text{soft}}$	Learn softened regime targets near change	Avoids artificial hard boundaries	Transition neighborhoods
$\mathcal{L}_{\text{trans}}$	Supervise change probability r_t	Improves rare-event detection	Full sequence
$\mathcal{L}_{\text{smooth}}$	Penalize jitter when change is unlikely	Stabilizes posterior trajectory	Non-boundary zones
$\mathcal{L}_{\text{geom}}$	Respect topology of regime confusions	Favors nearby alternatives during change	Ambiguous intervals

Table 7: Supervision sources under boundary uncertainty

max width=0.83

Supervision type	Annotation meaning	Recommended use	Reliability
Stable-interval label	Segment is confidently one macrostate	Frame or clip classification in central region	High
Transition-interval label	Segment contains regime change	Transition head and soft regime targets	Medium
Clip-level dominant label	Clip summarizes prevailing regime	Multiple-instance aggregation	Medium
Approximate boundary marker	Change occurs near reviewed time	Tolerance-aware boundary loss	Moderate
Weak contextual grouping	Neighboring clips share operating context	Persistence regularization	Supportive

This expression is deliberately generic. It does not require a literal hidden Markov model with hand-specified emissions. Instead, it motivates architectures in which learned visual evidence is combined with a temporal prior favoring persistence. The benefit of this viewpoint is that it makes clear how the model should use history: not by memorizing the whole run indiscriminately, but by carrying forward belief about the current macrostate and updating it as new evidence arrives.

Transition detection can be included by augmenting the latent state with a binary or multiclass change variable. Let b_t indicate whether time t lies in a transition neighborhood. Then the inference target becomes

$$p(y_t, b_t \mid X_{1:t}) \,. \tag{4}$$

This factorization is useful because it distinguishes uncertainty due to ambiguity among stable classes from uncertainty due to actual structural change. In practice, a model that predicts both y_t and b_t can keep high transition probability while distributing class mass across neighboring regimes. That is closer to how an expert would describe a mixed interval than forcing a single hard class.

For offline analysis, symmetric context may be used. Let $X_{t-L_1:t+L_2}$ denote a centered window. Then the noncausal posterior

$$p(y_t, b_t \mid X_{t-L_1:t+L_2}) \tag{5}$$

provides an upper bound on what is knowable when future frames are available. This distinction between causal and noncausal inference is important [13]. The former corresponds to online monitoring. The latter corresponds to post hoc regime reconstruction or dataset annotation support. Both are useful, but they answer different questions and should not be conflated in evaluation.

A sequence-energy formulation is another helpful abstraction. Let $Y_{1:T}$ be the regime sequence for a clip or run.

Table 8: Causal filtering and decision components

max width=0.82

Component	Function	Effect on output behavior	Deployment role
Evidence vector $e_t(y)$	Encodes current visual support	Updates class belief from new observation	Per-step input
Posterior $\pi_t(y)$	Filtered regime probability	Produces coherent macrostate trajectory	Main state estimate
Transition score ρ_t	Quantifies likelihood of change	Relaxes persistence during boundaries	Change awareness
Adaptive matrix A_t	Modulates stay/switch cost	Prevents abrupt flips without evidence	Temporal regularizer
Hysteresis rule	Requires sustained dominance to switch	Reduces oscillation between neighboring regimes	Decision stability

Table 9: Streaming evaluation metrics for temporal regime modeling

max width=0.84

Metric	What it measures	Why it matters	Evaluation region
Balanced accuracy / macro-F1	Stable-regime discrimination	Handles uneven class frequencies	Stable intervals
Transition event F1	Detection of boundary events with tolerance	Captures event-level usefulness	Reviewed changes
Boundary timing error	Onset or center localization quality	Exposes delay versus false alarm trade-off	Matched events
Switch rate	Frequency of predicted class changes	Quantifies jitter or excessive inertia	Full runs
NLL / Brier near boundaries	Confidence behavior by transition proximity	Tests temporal calibration	Distance-binned frames

One may define

$$E(Y_{1:T}, X_{1:T}) = \sum_{t=1}^{T} U_t(y_t; X_{1:T}) + \sum_{t=2}^{T} V_t(y_{t-1}, y_t; X_{1:T}), \tag{6}$$

where U_t is the unary cost of assigning label y_t at time t and V_t is the transition cost between adjacent labels. Unary terms come from visual evidence. Pairwise terms encode persistence or change sensitivity. If V_t is constant for label changes, the model becomes a standard smooth sequence decoder. If V_t depends on local features, the cost of changing labels can decrease in visually unstable regions and increase in stable ones. This variable transition penalty is much more appropriate in annulus videos, where not all label changes are equally plausible at all times.

The same framework can absorb clip-level supervision. Suppose a clip indexed by j carries a dominant label r_j but frame-level labels inside the clip are not all observed. One can define a bag-level probability through aggregation of frame-wise posteriors:

$$p(r_j \mid X_{I_j}) = \sum_{t \in I_j} \alpha_{j,t}\, p(y_t = r_j \mid X_{I_j}), \tag{7}$$

where $\alpha_{j,t}$ reflects the representativeness of frame t for the clip label. This is a natural way to treat temporally reviewed clips in which only part of the interval is strongly prototypical. It also avoids the unrealistic assumption that every frame inside a labeled interval should inherit the same hard target.

Boundary localization can be expressed either through b_t or through onset and completion times. If a transition interval is represented by (τ^-, τ^+), then the model may predict a start score and an end score rather than one point boundary. This is attractive because real transitions have duration [14]. It also allows the model to separate early

Table 10: Failure modes and practical design tradeoffs

max width=0.84

Issue	Typical symptom	Likely cause	Mitigation
Over-smoothing	Delayed or missed true transitions	Excessive persistence bias	Transition-aware training and calibration
Hyper-reactivity	Spurious boundaries in stable runs	Overweighting local disturbances	Hysteresis and conditional smoothing
Run memorization	Good internal score but poor generalization	Leakage of experiment-specific cues	Experiment-disjoint splits
Noisy boundary learning	Sharp spikes at annotation convention points	Treating approximate labels as exact	Interval-valued supervision
Excessive compute load	High latency or memory demand	Overlong context or dense sampling	Stride control and lighter causal models

destabilization from full adoption of the new regime. In operational terms, this can be useful because different downstream actions may be appropriate at the first sign of change versus once the new regime becomes dominant.

The latent-state viewpoint also clarifies the role of temporal redundancy. Because consecutive frames are highly correlated, the sequence model should not interpret repeated similar images as repeated independent evidence. That would artificially sharpen posterior confidence. Instead, the temporal update should effectively discount redundant information and react most strongly when the visual evidence begins to deviate from the current macrostate expectation. Many practical architectures implement this only implicitly. The conceptual model here makes it explicit: what matters is not frame count but deviation of the incoming observation stream from the pattern expected under persistence.

Thus the sequential inference problem in annulus videos can be summarized as follows. The model observes a stream of images, maintains a belief over a latent macrostate, updates that belief as morphology evolves, and separately tracks whether the process is stable or changing. This formulation is general enough to include convolutional, recurrent, attention-based, and probabilistic decoding architectures, yet specific enough to define what each of them should be trying to infer.

4 Temporal Encoders and Multi-Horizon Representation Design

The sequential formulation requires a representation pipeline that separates spatial morphology extraction from temporal evidence aggregation while still allowing the two to inform each other. A natural decomposition begins with a frame encoder ϕ_θ that maps each frame x_t to a latent vector z_t. The spatial encoder should preserve interface topology, bubble-scale texture, connected gas structures, and coarse void arrangement. In annulus imagery, these cues live at multiple scales, so the frame encoder benefits from hierarchical convolutional design or other multi-scale mechanisms. What the temporal stage receives should not be a generic image embedding but a morphology-aware summary.

Once $\{z_t\}$ are available, the temporal model must answer two different questions at once. It must recognize persistence of the current regime, and it must detect evidence of structural change. These requirements point toward multi-horizon processing. Short windows help identify immediate morphology and local motion coherence [15]. Longer windows help determine whether the current pattern is persistent, decaying, or emerging. A model with only a short receptive field can be reactive but boundary-blind. A model with only a long receptive field can be stable but sluggish. Therefore, a useful sequence encoder should explicitly combine temporal scales.

One practical design is a multi-branch temporal stack. Let g_s, g_m, and g_l denote short-, medium-, and long-horizon temporal modules acting on the frame embeddings. The fused hidden representation at time t may be written as

$$\begin{aligned} h_t \; &= W_s \, g_s(z_{t-\ell_s:t}) \\ &\quad + W_m \, g_m(z_{t-\ell_m:t}) \\ &\quad + W_l \, g_l(z_{t-\ell_l:t}) \, . \end{aligned} \tag{8}$$

The short branch captures immediate morphology continuity. The medium branch captures persistence over a clip-scale interval. The long branch captures slow reorganization or repeated intermittent events. In annulus videos, this division is particularly useful because the regime semantics often depend on all three. A large coherent gas structure can be identified locally, but whether it represents a stable intermittent macrostate or a transient event depends on longer context.

Temporal convolution is a strong candidate for at least one branch. A dilated causal convolution can cover long temporal spans without excessive depth while preserving online usability. If $h_t^{(0)} = z_t$, a layered temporal convolution may take the form

$$h_t^{(\ell+1)} = \sigma\left(\sum_{k=0}^{K-1} W_k^{(\ell)} h_{t-d_\ell k}^{(\ell)} + b^{(\ell)}\right). \tag{9}$$

Dilations d_ℓ allow multiple temporal scales. This architecture is appealing because it is efficient, naturally causal, and well suited to local temporal patterns. However, it does not automatically know which frames are most informative. For transition detection, selective attention can be advantageous because not every frame in a history window contributes equally to the current decision.

Attention-style pooling or attention blocks offer this selectivity [16]. A query at time t can attend over a set of recent embeddings and compute a context vector emphasizing diagnostically important moments. In annulus clips, such moments may correspond to the entry of a large gas structure, the first appearance of interface breakup, or the persistence of a texture pattern that confirms the current stable regime. The benefit of attention is that it can model sparse temporal evidence. The risk is that it may overfit to run-specific patterns if not regularized under experiment-disjoint training.

Recurrent units remain useful as compact state estimators. A gated recurrence can summarize the past with modest memory cost:

$$\begin{aligned} h_t &= \mathcal{R}(z_t, h_{t-1}) \\ p_t &= \mathrm{softmax}(W h_t + b). \end{aligned} \tag{10}$$

Recurrence is especially attractive when the intended use is streaming monitoring and when hardware or latency constraints discourage large attention windows. In this setting, the hidden state functions like an adaptive temporal filter whose memory depends on the input sequence. Yet recurrence by itself does not guarantee good transition sensitivity. If trained only on stable-state classification, it may simply smooth aggressively. It becomes more useful when coupled to explicit transition supervision or auxiliary losses that penalize both jitter and delayed boundary response.

Spatiotemporal clip encoders provide another option. Instead of encoding each frame independently, they operate directly on a short volume of frames and learn joint space–time filters. This can be particularly effective for local motion cues such as interface progression or breakup. In annulus videos, however, such encoders must be used carefully because motion is highly redundant at native frame rate. If the clip is too short or too dense, a 3D encoder can over-focus on trivial local continuity. If the clip stride is chosen sensibly, the same encoder can become a good detector of short-time dynamics that separate stable from changing intervals.

A useful addition is explicit quality-aware weighting. Not all frames in an interval are equally informative [17]. Blur, glare, occlusion, and partial field-of-view obstruction can make a frame less reliable, even if nearby frames remain useful. Let a scalar quality score q_t be estimated from the frame or its embedding. Then a weighted context summary can be formed as

$$\begin{aligned} \bar{z}_t &= \sum_{u=t-L+1}^{t} \alpha_{t,u}\, z_u \\ \alpha_{t,u} &= \frac{\exp(a_{t,u} + \lambda q_u)}{\sum_{v=t-L+1}^{t} \exp(a_{t,v} + \lambda q_v)}. \end{aligned} \tag{11}$$

This formulation lets the model prefer frames that are both temporally relevant and visually reliable. Such weighting can materially improve transition detection, because boundary intervals often coincide with visually difficult segments in which a few frames carry disproportionate information.

Sequence encoders should also expose both a regime representation and a change representation. A shared trunk can feed two heads: one for the current macrostate and one for transition likelihood. This separation is useful because the features needed to say "still in the same regime" are not always identical to the features needed to say "evidence of change is emerging." In practice, the change head can be driven by differences between short-

and long-horizon summaries, by the derivative of hidden state, or by a dedicated boundary representation learned jointly with the regime head. This architecture makes transition detection an explicit modeling objective instead of a heuristic based on posterior instability.

Finally, representation design must guard against one pervasive risk: learning run identity rather than regime dynamics. Since temporal context includes repeated lighting, camera angle, and local field-of-view cues, a powerful sequence model can memorize experiment-specific patterns more easily than a frame-only model. This is why experiment-disjoint evaluation is indispensable and why augmentations that alter nuisance appearance while preserving morphology are helpful even in temporal learning. A good temporal encoder should capture how morphology evolves, not merely how one particular run looks while it evolves.

5 Boundary Geometry and Transition-Sensitive Objectives

Once the temporal representation is defined, the learning objective must express what counts as correct behavior in stable segments and near transitions. A purely frame-wise cross-entropy objective with hard labels at every time point is not well suited to this problem. It rewards certainty even where the annotation is interval-valued and even where the physical process is mixed. It also gives overwhelming weight to stable intervals simply because they contain more frames. Transition-sensitive learning therefore requires both different targets and different regularization.

For stable frames or clip centers far from reviewed boundaries, an ordinary classification term is still appropriate [18]. If $p_t(y)$ denotes the predicted regime posterior at time t and the stable label is $y_t^\star$, one has

$$\mathcal{L}_{\mathrm{stab}} \;=\; -\sum_{t\in\mathcal{S}} \log p_t(y_t^\star)\,, \tag{12}$$

where $\mathcal{S}$ indexes stable supervision points. This term encourages discrimination among canonical macrostates. However, it should not dominate the full objective to the extent that the model learns to ignore change points.

For boundary neighborhoods, a soft target is more reasonable. Suppose an annotated transition connects stable regimes r^- and r^+ over an interval centered near τ. Then instead of a hard label switch, one can define a time-varying target distribution

$$\begin{aligned} q_t(y) \;&=(1-\gamma_t)\,\delta_{y=r^-} \;+\; \gamma_t\,\delta_{y=r^+} \\ \gamma_t \;&=\sigma\!\left(\frac{t-\tau}{\kappa}\right). \end{aligned} \tag{13}$$

The scale parameter κ encodes boundary width. This target does not imply that the physical regime linearly blends between two states. It encodes the annotation fact that the change is not pointwise precise. Training against q_t preserves uncertainty where hard labels would be artificial.

A dedicated transition head can be trained in parallel. Let r_t denote the predicted probability that time t belongs to a changing interval. If $b_t^\star \in \{0,1\}$ is a boundary indicator or dilated boundary mask, then

$$\begin{aligned} \mathcal{L}_{\mathrm{trans}} \;=\; -\sum_t \Big[&\beta\, b_t^\star \log r_t \\ &+ (1-b_t^\star)\log(1-r_t)\Big]. \end{aligned} \tag{14}$$

The positive weight β is needed because boundary times are far rarer than stable times. Without it, the model can minimize loss by predicting no transitions at all. A more refined variant predicts onset and completion separately, but even a single change score is valuable because it disentangles boundary evidence from regime identity.

Temporal smoothness should be conditional rather than global. A model should be smooth inside stable intervals and flexible near transitions. This can be expressed by penalizing successive posterior changes only when the transition head is low:

$$\mathcal{L}_{\mathrm{smooth}} \;=\; \sum_t (1-r_t)\,\|p_{t+1}-p_t\|_2^2\,. \tag{15}$$

This term discourages jitter in stable regions without forcing the model to smear genuine change [19]. It is more faithful than applying a fixed low-pass filter to the output sequence, because the amount of smoothing now becomes part of the learned decision process.

Another useful objective component is boundary ranking. Suppose the model outputs a scalar change score c_t. Around a real transition center, one would like scores nearer the boundary to exceed scores farther away. This can be encoded through pairwise ranking constraints or through a distance-weighted regression target. Such

objectives help the model learn boundary geometry rather than only boundary presence. In practice, this improves localization because the transition head then represents not just whether change exists but how strongly the current instant resembles the center of a changing interval.

Regime geometry can also be exploited. In many taxonomies, some confusions are more reasonable than others. A model that spreads probability between two neighboring regimes during a boundary interval is behaving differently from a model that abruptly jumps to a distant class. A regime-distance regularizer can reflect this:

$$\mathcal{L}_{\text{geom}} = \sum_{t} \sum_{i,j} q_t(i)\, p_t(j)\, D_{ij}\,, \tag{16}$$

where D_{ij} measures the dissimilarity between regimes i and j. Such a term is especially sensible during transitions, where the model should retain probability mass near adjacent macrostates rather than collapse into implausible alternatives.

The full objective can then be written as

$$\begin{aligned}\mathcal{L} = {} & \lambda_1 \mathcal{L}_{\text{stab}} + \lambda_2 \mathcal{L}_{\text{soft}} \\ & + \lambda_3 \mathcal{L}_{\text{trans}} + \lambda_4 \mathcal{L}_{\text{smooth}} \\ & + \lambda_5 \mathcal{L}_{\text{geom}}\,.\end{aligned} \tag{17}$$

This expression makes the design philosophy explicit. Stable classification, boundary uncertainty, change detection, jitter control, and regime topology are not competing hacks. They are distinct aspects of the same sequence-learning problem. Their relative importance can vary with the dataset and deployment mode, but omitting any of them tends to reintroduce familiar failure modes.

A final point concerns annotation imprecision. If boundary annotations are approximate, then the model should not be judged or trained as though a single ground-truth frame were exact [20]. Temporal tolerance should be built into both training and evaluation. This can be done by dilating boundary targets, by using Gaussian kernels around reviewed transition centers, or by predicting boundary intervals instead of points. In annulus videos, where change often occupies a visually extended segment, such tolerance is not merely convenient. It is the correct representation of the uncertainty inherent in the task.

6 Supervision Design and Data Use Under Boundary Uncertainty

Temporal regime modeling is constrained not only by architecture but by the type of supervision available. In annulus video datasets, labels are often attached to intervals, clips, or selected key frames rather than to every frame with precise boundary timing. This is not a weakness of the data collection process. It reflects the fact that experts naturally reason over short clips and that exact frame-level boundaries are costly and, in many cases, not meaningfully unique. A sound training design should therefore treat the supervision as interval-valued and potentially heterogeneous rather than forcing it into a fully pointwise schema.

A helpful distinction is between three supervision types. The first is stable-interval labeling, in which a reviewed segment is judged to represent one macrostate with high confidence. The second is transition-interval labeling, in which a segment is known to contain change, possibly with an approximate onset and completion. The third is weak contextual supervision, in which clips can be grouped by operating condition or by neighborhood within a run even if no exact label is assigned to every time point. A sequence model can use all three if its losses and batch construction are designed accordingly.

Stable-interval labels are relatively straightforward. The central portion of such an interval can supervise frame-wise or clip-wise classification, while its edges may be downweighted unless the reviewer explicitly confirmed that the interval is uniformly stable. Transition-interval labels are richer. They can supervise the transition head directly and can provide soft targets for the regime head. Weak contextual supervision can be used to regularize persistence or to prevent the model from assigning implausible rapid alternations inside apparently homogeneous run segments [21].

Manikonda et al. (2025) present a vertical-annulus image dataset whose organization makes it especially effective for constructing temporally aware recognition pipelines [8], and the modeling study built on that archive demonstrates that the visual collection itself carried most of the evidentiary burden for fitting, adjusting, and checking the learned regime predictors. This kind of resource is valuable not merely because it contains many images, but because it contains them as parts of coherent experimental sequences. For transition-sensitive modeling, that continuity is as important as the label set itself.

Clip-level supervision can be expressed as multiple-instance learning. Let C_j denote a clip with reviewed dominant label r_j. Instead of forcing every frame in C_j to match r_j, one may define a clip posterior through

attention over frame predictions:

$$p(r_j \mid C_j) = \sum_{t \in C_j} \alpha_{j,t}\, p_t(r_j)$$
$$\alpha_{j,t} = \frac{\exp(u_{j,t})}{\sum_{s \in C_j} \exp(u_{j,s})}. \tag{18}$$

This allows the model to learn which frames are most representative of the clip label. It is particularly appropriate when the clip contains a few highly diagnostic frames surrounded by more ambiguous ones. It also respects the annotation process more closely than naive label replication.

Boundary supervision is often weaker. If an expert marks that a transition occurs within a clip but does not specify exact frame times, the model can still benefit. One can train the transition head with interval targets and use latent alignment within the interval. For example, the loss may require high boundary mass somewhere inside the reviewed region and low boundary mass far outside it. This is a form of weak sequence supervision that fits the practical realities of regime review. It recognizes that the expert knows a change occurred without claiming a frame-perfect event time.

Batch construction should reflect the rarity and value of transition clips. If clips are sampled uniformly by frame count, stable segments dominate because they are long and dense [22]. Training then emphasizes persistence at the expense of detection. A better strategy samples by segment type or by reviewed interval, ensuring that transition-containing clips appear sufficiently often. The ratio should not be so extreme that stable-state modeling becomes poor, but it should counteract the natural imbalance of the archive. In temporal annulus modeling, this rebalancing is essential because the boundary events are where the sequence formulation earns its keep.

Another design choice is whether to use pseudo-labels on unlabeled portions of the runs. In principle, a temporally aware model can propagate stable labels through nearby unlabeled frames or infer likely boundary neighborhoods through self-consistency. This can be helpful if done conservatively, but it carries risk. A model that is under-trained on boundary geometry can generate over-sharp pseudo-labels that later reinforce its own mistakes. Therefore, pseudo-labeling is most defensible in clearly stable interiors and least defensible in uncertain transition corridors. The sequence setting makes this distinction easier to operationalize because transition head output and temporal disagreement provide cues about where pseudo-labels should be trusted or withheld.

Calibration of supervision confidence is also important. Not all annotations are equally certain. Stable-interval labels reviewed by agreement among multiple experts may deserve near-unit weight. Boundary intervals flagged as approximate should receive softer treatment. Interface quality, optical clarity, and reviewer notes can all be used to weight losses or target entropies. This weighting is not an admission of weakness. It is a way of keeping the model from learning false certainty where the supervision itself is imprecise.

A further benefit of sequence-aware supervision is that it naturally supports mixed annotation granularity inside one archive. One run may have only dominant clip labels [23]. Another may have approximate transition markers. A short evaluation subset may have detailed frame-wise review. Instead of discarding the weaker portions or coercing everything to the weakest common denominator, the model can absorb all of them with task-appropriate losses. This is especially valuable in annulus research, where labeled video resources are costly and heterogeneous.

In short, supervision design for temporal regime learning should reflect what the labels actually mean. Stable segments provide class certainty. Transition segments provide temporal structure and uncertainty zones. Clip labels provide bag-level evidence rather than exact frame truth. Visual archives are most useful when that structure is preserved. A temporal model trained under these principles is better aligned with both the data and the phenomenon than a model trained on artificially homogenized frame labels.

7 Online Filtering, Calibration, and Decision Logic

A sequence model becomes especially useful when it is embedded in a causal decision process. In streaming operation, the system receives one new frame or clip step at a time and must update its belief about the current regime, the possibility of transition, and the confidence appropriate to that judgment. This is not the same as running a noncausal sequence encoder and reading off one label. It requires a filtering logic that balances persistence against incoming evidence and that can delay commitment when change is suspected but not yet resolved.

Let $\pi_t(y) = p(y_t = y \mid X_{1:t})$ be the causal regime posterior and let $\rho_t = p(b_t = 1 \mid X_{1:t})$ be the transition probability. A practical filtering recursion can be built on the output of the temporal encoder. Suppose the encoder produces an evidence vector $e_t(y)$ and a transition score ρ_t. Then a persistence-aware update can be

written as

$$\begin{aligned}\tilde{\pi}_t(y) &= e_t(y) \sum_{y'} A_t(y', y)\, \pi_{t-1}(y') \\ \pi_t(y) &= \frac{\tilde{\pi}_t(y)}{\sum_k \tilde{\pi}_t(k)}\,, \end{aligned} \tag{19}$$

where A_t is a transition matrix whose diagonal dominance may depend on ρ_t. If ρ_t is low, the matrix favors staying in the same regime [24]. If ρ_t is high, off-diagonal movement becomes cheaper. This approach has two advantages. It improves interpretability, and it prevents sudden class switching in the absence of transition evidence.

A simple transition-conditioned parameterization is

$$\begin{aligned} A_t(y', y) &= (1 - \rho_t)\, \mathbf{1}[y' = y] \\ &\quad + \ \rho_t\, B(y', y)\,, \end{aligned} \tag{20}$$

where B is a normalized matrix encoding plausible regime changes. This makes the role of the transition head explicit. The regime posterior need not jump merely because one incoming frame is ambiguous. It moves when the model believes a changing interval is underway and when the new evidence is strong enough to outweigh persistence.

Calibration is crucial in this filtering stage. If the evidence vector e_t is overconfident, the filter will switch too abruptly. If it is underconfident, the filter will be sluggish. Post hoc temperature scaling can help at the regime-head level, but temporal calibration should be assessed after filtering as well because the persistence update changes the effective confidence. One can temperature-scale the prefilter logits or calibrate the combined posterior on a held-out validation stream. In either case, the goal is not merely to improve average log loss but to make the posterior trajectory behave plausibly through stable and changing intervals.

Decision logic can then operate on π_t and ρ_t. A straightforward policy declares the current regime as $\arg\max_y \pi_t(y)$ when ρ_t is low. When ρ_t exceeds a threshold or when the regime posterior remains diffuse, the system may instead output a transition or uncertain state. This is often preferable to forcing a hard class label because the operational question is frequently whether the flow is changing rather than which endpoint class currently has the largest posterior mass. The threshold itself should be calibrated on validation data and may depend on the relative cost of missed transitions versus false alarms [25].

A hysteresis mechanism is often beneficial. Without it, the model may oscillate when two neighboring regimes have similar posterior mass over several frames. Hysteresis can be implemented by requiring stronger evidence to leave the current stable regime than to remain in it, or by requiring that a candidate new regime dominate for several steps before commitment. This is not ad hoc if it is written into the filtering logic and tuned under streaming evaluation. It is simply a recognition that regime interpretation has inertia. The key is to prevent hysteresis from masking genuine rapid changes, which is why its interaction with ρ_t should be explicit.

Another useful quantity is boundary imminence. Instead of treating transition detection as binary, the system may track whether the current state appears to be approaching a change. This can be estimated from sustained growth in transition probability or from monotonic drift in regime posterior away from the current dominant class. Such a signal is valuable because it offers early warning without pretending that the new regime is already established. In many annulus applications, early warning of loss of stability can be more useful than precise labeling of the eventual successor regime at the earliest possible moment.

Online decision logic should also include out-of-pattern handling. If the current clip is visually unlike the training support, the model may become confidently wrong unless uncertainty is monitored. Ensemble disagreement, unusually high transition probability without a plausible boundary shape, or deviation of intermediate embeddings from the training distribution can all signal that the stream has moved into unfamiliar territory. In such cases, the system can fall back to an uncertain or review-needed state. This is especially relevant in annulus experiments because optical drift or unanticipated operating behavior can create sequences not represented in the labeled archive.

Finally, the filter logic makes the temporal model more actionable. It turns a sequence of probabilities into a state-estimation process with interpretable persistence, change, and uncertainty. This is valuable not only for deployment but for scientific analysis [26]. A filtered posterior can be inspected as a regime trajectory, compared with operating events, and used to identify intervals worth later human review. In this sense, causal temporal modeling does more than classify images. It produces a structured interpretation of the evolving flow.

8 Experimental Design and Streaming Performance Measurement

Temporal regime models should be evaluated under protocols that respect the sequential nature of the data and the intended use of the predictions. Standard frame-wise accuracy on a random split is particularly misleading in this domain because it ignores both temporal redundancy and the role of transitions. A better evaluation design uses

experiment-disjoint splits, streaming inference, and metrics that separately probe stable classification, boundary detection, and confidence behavior.

Let the data be partitioned by experiment into training, validation, and test sets with no experiment overlap. Within each experiment, the model should be run on full sequences or on long contiguous segments rather than on shuffled windows. This matters because the filtering dynamics, boundary logic, and calibration are sequence-dependent. If evaluation is performed only on isolated windows, the model is not being tested in the mode for which it was designed.

Stable-state performance should be reported first, but only on intervals judged to be confidently stable. Balanced accuracy and macro-averaged F1 remain useful here because regime frequencies are uneven. Yet these metrics alone are insufficient. They tell how well the model recognizes canonical macrostates once ambiguity is mostly removed. They do not reveal whether the model behaves sensibly near change points or whether its temporal smoothing is excessive.

Transition performance therefore needs its own metrics. If reference transitions are represented by centers or intervals, event-level precision and recall can be computed using a tolerance window. Predicted events are matched to reference events if they lie within an allowable temporal distance. This yields a transition F1 score that is more meaningful than per-frame boundary classification because it reflects whether the model detected the event rather than whether it guessed the exact center frame [27]. When transition intervals are annotated, interval overlap metrics such as temporal intersection-over-union or onset and completion error provide more detail.

Boundary timing error is another important measure. For matched events, one may compute mean absolute onset error, mean absolute completion error, or simply mean absolute boundary-center error depending on the annotation format. These metrics expose the trade-off between early detection and false alarms. A model with strong smoothing may achieve good stable classification yet show delayed transitions. A highly reactive model may detect early but produce many spurious alarms. Reporting both event detection and timing error makes this trade-off visible.

Sequence stability should be quantified explicitly. Let $\hat{y}_t$ be the predicted dominant class after filtering. The switch rate is the number of times $\hat{y}_t$ changes, normalized by run duration. This metric should be compared with the switch rate implied by the annotations or by expert-reviewed intervals. Too high a rate indicates jitter. Too low a rate may indicate excessive inertia. For temporal models, sequence stability is not a secondary property. It is a central performance dimension because the value of temporal reasoning lies partly in stabilizing the inferred macrostate trajectory.

Calibration should be measured both globally and by transition proximity. Let d_t be the distance from time t to the nearest reference transition interval. Then negative log-likelihood, Brier score, and calibration error can be computed within bins of d_t. This reveals whether confidence falls appropriately near changing intervals while remaining sharp in stable regions. A model that stays highly confident deep into the transition bin is likely misrepresenting uncertainty even if its average accuracy is strong.

Causal and noncausal results should be reported separately [28]. The noncausal model provides an upper bound on what can be achieved when future context is allowed, which is valuable for offline analysis and annotation support. The causal model reflects the online monitoring problem. The gap between them is itself informative, because it measures how much future context matters for transition timing in the given dataset. A small gap suggests that causal cues suffice. A large gap suggests that boundary evidence is distributed on both sides of the event and that online detection will inevitably involve latency or greater uncertainty.

Robustness testing should also be sequence-aware. Applying independent random perturbations to frames does not adequately simulate the degradations encountered in real annulus recordings. More meaningful tests include coherent episodes of blur, brightness drift, sustained glare, or reduced frame rate. The model should be evaluated under these conditions for stable-state accuracy, transition detection, and calibration. Temporal models may partially absorb a few corrupted frames by relying on surrounding context, but they can also fail systematically when an entire episode is degraded. Reporting only clean-data performance would conceal this.

Another useful design is stratified evaluation by run difficulty. Some experiments contain long stable segments and few ambiguous events. Others contain many short transitions or optically difficult conditions. Average metrics can be dominated by the easier runs. Therefore, per-run distributions should be reported or at least summarized. Experiment-level bootstrap confidence intervals are essential because frames within a run are not independent and because the number of independent sequences is usually far smaller than the number of evaluated time points.

Finally, evaluation should include simple baselines that clarify what temporal modeling contributes. A strong frame encoder with no temporal context, the same encoder with post hoc smoothing only, and a persistence-only filter using static frame scores are all useful comparators [29]. These baselines help determine whether the gains of the full model come from genuine transition modeling, from generic smoothing, or simply from additional context length. Without them, one cannot tell whether the temporal system has actually learned the dynamics of regime change or merely regularized a static classifier.

9 Failure Modes, Computational Tradeoffs, and Design Governance

Temporal modeling introduces capabilities not available to static classification, but it also introduces new failure modes. One major risk is over-smoothing. If persistence penalties or recurrent memory dominate the evidence, the model can delay or suppress genuine transitions. This may look appealing in aggregate metrics because jitter decreases and stable-state accuracy may rise, yet the output becomes less faithful to the actual process. The opposite risk is hyper-reactivity, where the model interprets short-lived visual disturbances as boundaries and fragments stable runs into many short pseudo-regimes. Good temporal design is therefore not about maximizing smoothness. It is about learning when smoothness is justified.

Another failure mode is sequence memorization. Because runs have characteristic optics, camera placement, and operating trajectories, a powerful temporal model can learn the signature of particular runs or domains rather than abstract regime evolution. This is especially easy if training and test windows from the same run leak across the split. Experiment-disjoint protocols mitigate this, but even then the model may still exploit domain-specific temporal patterns if domains are not diverse. Regularization, nuisance-aware augmentation, and domain-aware validation remain important in temporal as well as static learning.

Boundary supervision itself can create artifacts. If approximate boundary labels are treated as exact, the transition head can become trained on annotation noise rather than on visual change. This often manifests as a model that produces narrow, high-amplitude transition spikes aligned with the annotation convention rather than with the visually mixed interval. Such a model can appear precise under point-based metrics and yet be less useful for interpretation. Interval-valued supervision and tolerance-aware evaluation are necessary to prevent this failure [30].

There is also a computational trade-off between long context and practical deployment. Attention over long histories can improve transition interpretation, but it increases latency and memory cost. Heavy 3D clip encoders can learn rich motion cues, but they may be too expensive for real-time use or for large hyperparameter sweeps on long videos. In many annulus applications, a hybrid approach with moderate frame encoding cost and lightweight causal filtering may offer the best balance. The appropriate architecture is therefore partly determined by whether the goal is offline annotation support, online monitoring, or general research benchmarking.

Sampling design matters here as well. Native high-speed video rates often exceed what is necessary for regime inference. Processing every frame can be computationally wasteful and may even hurt learning by flooding the model with trivial continuity. Temporal downsampling or adaptive stride selection can reduce this burden, but the stride should be chosen in physical time, not just in frame count. Otherwise a model trained on one camera rate may not transfer sensibly to another. This is a common but underappreciated issue in sequence models for fluid videos.

Governance of model outputs is another concern. A temporal system that estimates regime trajectories can be tempting to use as a definitive automated annotator. That would be premature unless calibration, domain robustness, and uncertainty have been validated under the actual target conditions. The right stance is more measured. The model should be treated as a structured estimator whose outputs support human review, downstream control logic, or mechanistic analysis, but not as an oracle immune to optical shift or taxonomy ambiguity. This is especially true near transitions, where the model may appropriately output wide uncertainty rather than a clean answer.

A related governance issue is dataset feedback. Once a temporal model exists, it can guide the curation of future annulus video datasets by identifying underrepresented transition types, runs with persistent uncertainty, or domains where boundary behavior is poorly learned [31]. In that sense, sequence modeling helps define what kinds of additional labels are valuable. However, if future curation is driven entirely by the current model's uncertainty, the archive can become over-specialized to that model family. A balanced strategy preserves both actively enriched difficult intervals and passively sampled baseline intervals so that future models are not trapped by the inductive biases of the current one.

Interpretability should also be governed carefully. Temporal attention weights, hidden-state changes, or filtered posterior trajectories can be informative, but they do not automatically reveal physical causation. A weight peak on a particular frame shows that the model found that frame useful, not that the frame alone physically caused the regime change. Such outputs are best used as diagnostic aids rather than as proofs of mechanistic insight. Still, they are often more meaningful than frame-level saliency maps because they at least respect the sequential nature of the task.

Finally, temporal models change what counts as a scientifically useful result. A modest gain in stable-state accuracy accompanied by a substantial improvement in transition timing and calibration may be more valuable than a larger gain in isolated-frame accuracy. Conversely, a sequence model that slightly improves average score while losing boundary fidelity may be less useful than a simpler baseline. The design and governance of temporal annulus recognition should therefore focus on the full behavior profile of the model, not on one compressed leaderboard number.

10 Conclusion

This paper presented a sequence-centered framework for regime recognition and transition detection in vertical annulus gas–liquid videos. The central argument was that flow regimes in these recordings are temporally organized macrostates and that the operationally important phenomenon is often the trajectory of those macrostates, not merely the class of one frame. On that basis, the paper described a latent-state formulation in which video frames are observations of evolving interfacial structure, regime identity is inferred from temporally organized evidence, and changing intervals are represented explicitly rather than being reduced to abrupt label flips. Multi-horizon temporal encoders, boundary-sensitive objectives, clip- and interval-aware supervision, and causal filtering logic were then developed as components of one coherent inference pipeline. The resulting system is designed to recognize stable patterns, preserve uncertainty near ambiguous intervals, and localize the onset and completion of transitions without relying on post hoc smoothing alone. Evaluation was framed in streaming terms, with experiment-disjoint sequences, transition event metrics, boundary timing error, sequence stability, and calibration by transition proximity, thereby aligning the measurement of model quality with the intended use of the outputs. The broader implication is that annulus image archives contain information not only in their visual content but in their continuity through time, and that exploiting this continuity makes regime recognition more faithful to both the physical evolution of the flow and the way experts actually interpret video records [32].

References

[1] S. B. Olawale, P. O. Longe, and S. F. Ofesi, "Evaluating the effect of drill string rotation and change in drilling fluid viscosity on hole cleaning," *Journal of Petroleum Exploration and Production Technology*, vol. 11, no. 7, pp. 2981–2989, Jun. 22, 2021. DOI: 10.1007/s13202-021-01209-y

[2] D. Chavis, *Savannah river site interim waste management program plan fy 1991–1992*, May 1, 1992. DOI: 10.2172/5092395

[3] A. A. Zahid, S. R. Ur Rehman, S. Rushd, A. Hasan, and M. A. Rahman, "Experimental investigation of multiphase flow behavior in drilling annuli using high speed visualization technique," *Frontiers in Energy*, vol. 14, no. 3, pp. 635–643, 2020.

[4] R. O. Prum and R. H. Torres, "Structural colouration of mammalian skin: Convergent evolution of coherently scattering dermal collagen arrays," *The Journal of experimental biology*, vol. 207, no. 12, pp. 2157–2172, May 15, 2004. DOI: 10.1242/jeb.00989

[5] J. McDonald, *Valve stem freeze seal for high-temperature sodium*, Jul. 30, 1960. DOI: 10.2172/4181737

[6] D. Anderson, "Well integrity, plugging and abandonment: Abrasive cutting applications for well severance," in *SPE Symposium: Decommissioning and Abandonment*, SPE, Dec. 3, 2018. DOI: 10.2118/193958-ms

[7] R. Reiner, A. Zedrosser, H. Zeiler, K. Hackländer, and L. Corlatti, "Population reconstruction as an informative tool for monitoring chamois populations," *Wildlife Biology*, vol. 2020, no. 4, pp. 1–13, Dec. 7, 2020. DOI: 10.2981/wlb.00757

[8] K. Manikonda, C. Obi, A. A. Brahmane, M. A. Rahman, and A. R. Hasan, "Vertical two-phase flow regimes in an annulus image dataset-texas a&m university," *Data in Brief*, vol. 58, p. 111 245, 2025.

[9] C. Varadharajan, *Summary report on co2 geologic sequestration & water resources workshop*, Feb. 15, 2012. DOI: 10.2172/1062103

[10] T. Hemphill, "Hole-cleaning model evaluates fluid performance in extended-reach wells," *Oil & Gas Journal*, vol. 95, no. 28, pp. 56–64, Jul. 14, 1997.

[11] E. Waters and M. Shockley, *Results of thermal-hydraulic experiments with kvns self-supported fuel in a zircaloy tube – k reactors*, Mar. 2, 1964. DOI: 10.2172/10152252

[12] B. Demirdal and J. Cunha, "Investigation of effect of equivalent diameter definitions on determination of pressure losses of non-newtonian fluids in annuli," in *Canadian International Petroleum Conference*, PETSOC, Jun. 12, 2007. DOI: 10.2118/2007-146-ea

[13] N. Available, *Preliminary design of a special casing joint for a well equipped twin horizontal drainholes in the oxnard field*, Dec. 31, 1993. DOI: 10.2172/10140206

[14] M. Farahat, "Regression approach to calculated the effective mud annular viscosity during drilling oil wells.(dept.m)," *MEJ. Mansoura Engineering Journal*, vol. 21, no. 4, pp. 74–91, Mar. 30, 2021. DOI: 10.21608/bfemu.2021.159941

[15] J. Zhang, B. Li, and Y. Liu, "Theory and application for helical flow of drilling fluid in the annuli of directional wells," *SPE Advanced Technology Series*, vol. 5, no. 01, pp. 146–155, May 1, 1997. DOI: 10.2118/30824-pa

[16] T. Reed and A. Pilehvari, "A new model for laminar, transitional, and turbulent flow of drilling muds," in *SPE Production Operations Symposium*, SPE, Mar. 21, 1993. DOI: 10.2118/25456-ms

[17] S. M. Khan et al., "Rural communities experience higher radon exposure versus urban areas, potentially due to drilled groundwater well annuli acting as unintended radon gas migration conduits.," *Scientific reports*, vol. 14, no. 1, pp. 3640–3640, Feb. 26, 2024. DOI: 10.1038/s41598-024-53458-6

[18] S. Sukumar, R. Weijermars, I. N. Alves, and S. F. Noynaert, "Analysis of pressure communication between the austin chalk and eagle ford reservoirs during a zipper fracturing operation," *Energies*, vol. 12, no. 8, pp. 1469–, Apr. 18, 2019. DOI: 10.3390/en12081469

[19] S. M. Willson, "A wellbore stability approach for self-killing blowout assessment," in *All Days*, SPE, Jun. 20, 2012. DOI: 10.2118/156330-ms

[20] D. Walker, R. Noland, F. McCusig, and C. Stone, *Borax-iv reactor: Manufacture of fuel and blanket elements*, Mar. 1, 1958. DOI: 10.2172/4349104

[21] A. M. Sharf, H. A. Jawan, and F. A. Almabsout, "The influence of the tangential velocity of inner rotating wall on axial velocity profile of flow through vertical annular pipe with rotating inner surface," *EPJ Web of Conferences*, vol. 67, pp. 02 105–, Mar. 25, 2014. DOI: 10.1051/epjconf/20146702105

[22] M. Mj, "Cannon–thurston maps for kleinian groups," *Forum of Mathematics, Pi*, vol. 5, May 22, 2017. DOI: 10.1017/fmp.2017.2

[23] W. C. Chin and X. Zhuang, "Exact non-newtonian flow analysis of yield stress fluids in highly eccentric borehole annuli with pipe or casing translation and rotation," in *International Oil and Gas Conference and Exhibition in China*, SPE, Jun. 8, 2010. DOI: 10.2118/131234-ms

[24] M. Stocking, "Almost normal surfaces in 3-manifolds," *Transactions of the American Mathematical Society*, vol. 352, no. 1, pp. 171–207, Sep. 21, 1999. DOI: 10.1090/s0002-9947-99-02296-5

[25] P. Oudeman and L. Bacarreza, "Field trial results of annular pressure behavior in a high-pressure/high-temperature well," *SPE Drilling & Completion*, vol. 10, no. 02, pp. 84–88, Jun. 1, 1995. DOI: 10.2118/26738-pa

[26] M. Sorgun and M. E. Ozbayoglu, "Predicting frictional pressure loss during horizontal drilling for non-newtonian fluids," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 33, no. 7, pp. 631–640, Jan. 31, 2011. DOI: 10.1080/15567030903226264

[27] I. Azouz and S. A. Shirazi, "Numerical simulation of drag reducing turbulent flow in annular conduits," *Journal of Fluids Engineering*, vol. 119, no. 4, pp. 838–846, Dec. 1, 1997. DOI: 10.1115/1.2819506

[28] C. Elendu et al., "The diagnostics and recompletion strategy of a well with sustained casing pressure," in *SPE Nigeria Annual International Conference and Exhibition*, SPE, Aug. 1, 2022. DOI: 10.2118/211908-ms

[29] P. Oudeman and M. Kerem, "Transient behavior of annular pressure build-up in hp/ht wells," *SPE Drilling & Completion*, vol. 21, no. 04, pp. 234–241, Dec. 20, 2006. DOI: 10.2118/88735-pa

[30] A. Nowamooz, F.-A. Comeau, and J.-M. Lemieux, "Evaluation of the potential for gas leakage along wellbores in the st. lawrence lowlands basin, quebec, canada," *Environmental Earth Sciences*, vol. 77, no. 8, pp. 1–17, Apr. 16, 2018. DOI: 10.1007/s12665-018-7483-6

[31] T. W. Marriott, S. Chase, I. Khallad, R. Bolt, and P. Whelan, "Reverse-circulation cementing to seal a tight liner lap," in *Offshore Technology Conference*, OTC, Apr. 30, 2007. DOI: 10.4043/18839-ms

[32] V. Vanita and A. Kumar, "Effect of radial magnetic field on free convective flow over ramped velocity moving vertical cylinder with ramped type temperature and concentration," *Journal of Applied Fluid Mechanics*, vol. 9, no. 6, pp. 2855–2864, Nov. 1, 2016. DOI: 10.29252/jafm.09.06.26060