OPEN ACCESS

Flow-Regime Classification Benchmarks for Gas-Liquid Two-Phase Imaging in Vertical Annuli

Bagus Prasetyo,[1] Fajar Mahendra[2]

[1] Universitas Teknologi Sumbawa, Jalan Raya Olat Maras, Batu Alang, Moyo Hulu, Sumbawa, Nusa Tenggara Barat, Indonesia;
[2] Universitas Muhammadiyah Palangkaraya, Jalan RTA Milono Km. 1.5, Palangka Raya, Kalimantan Tengah, Indonesia;

Abstract

Gas–liquid two-phase flow in annular geometries appears in drilling hydraulics, wellbore transients, and multiphase transport where regime identification is coupled to pressure gradient, holdup, and slip. With modern high-speed cameras and inexpensive storage, vision-based regime recognition has become technically feasible, yet published results remain difficult to compare because datasets, sampling policies, and split definitions vary, and because correlated frames can inadvertently leak between training and test. This paper specifies a benchmarking methodology for flow-regime classification in vertical annulus imagery that is designed to separate true model generalization from artifacts of data partitioning and preprocessing. The benchmark definition formalizes the unit of independence at the experiment level, prescribes metadata needed to characterize optical and operating contexts, and provides protocols for constructing frame-level and sequence-level tasks without violating temporal or experimental separation. Evaluation is defined through a metric suite that reports class-balanced discrimination, probability quality, calibration, and robustness under controlled photometric and blur perturbations, together with uncertainty estimates computed by experiment-level resampling rather than frame-level assumptions. A baseline model suite is described to anchor comparisons across classical descriptors, convolutional encoders, and temporal aggregation architectures under standardized optimization and augmentation recipes. Finally, a reproducibility and governance layer is proposed that version-controls labels, documents preprocessing, and enforces evaluator-side metric computation so that reported numbers can be audited and re-scored as the benchmark evolves.

1. Introduction

Flow regimes in gas–liquid annular passages are commonly treated as discrete macrostates that summarize interfacial topology and intermittency over a range of operating conditions [1]. Although regime maps are often drawn in terms of superficial velocities, the regimes themselves are realized through coupled hydrodynamic mechanisms such as bubble coalescence, interfacial shear, wake entrainment, and transient recirculation. In a vertical annulus, confinement and curvature modify buoyancy-driven structures, while the optical accessibility of the test section makes imaging a natural measurement modality. The practical motivation for visual regime recognition is that many quantities of interest are difficult to measure continuously in situ, yet regime changes manifest clearly in spatial patterns of void distribution and interface continuity. A regime label can serve as a compact state variable in supervisory control, in reduced-order modeling, and in the selection or switching of mechanistic correlations that depend on interfacial structure.

Computer vision offers an attractive route to automatic regime recognition because it can exploit the spatial texture and morphology that are evident to a trained observer. However, the transition from qualitative visual assessment to quantitative model comparison has been slowed by the absence of shared benchmarking conventions. In many published pipelines, frames are sampled densely from a small set of videos, and random frame-level splitting is then used for evaluation. This can inflate reported performance because adjacent frames are nearly duplicates in the feature space learned by modern encoders. When the same experiment contributes both training and test frames, the evaluation can become a test of interpolation in time rather than generalization to new runs. Moreover, images acquired from a given optical chain can contain consistent signatures unrelated to the flow itself, including background textures, lens vignetting, overlay graphics, or illumination artifacts. A model can

© Copyright by the Authors.
Licensed as an open access article using a CC BY 4.0 license.

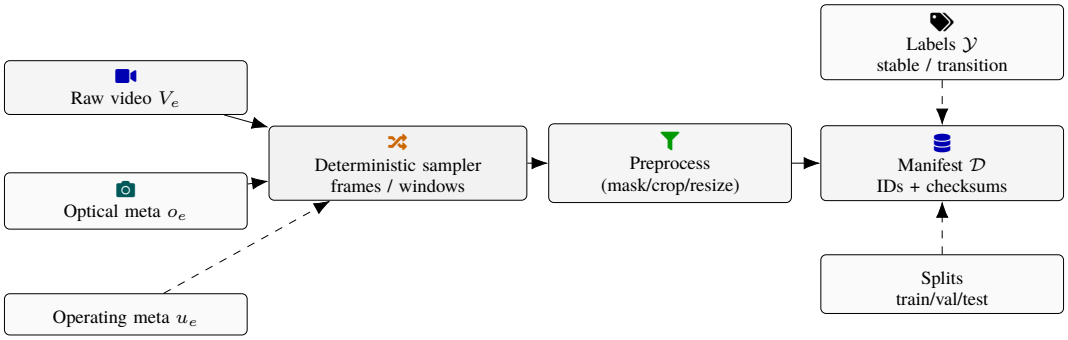


Figure 1: Benchmark data model: each experiment e provides a raw video stream plus optical and operating metadata. A declared deterministic sampler generates frame or window samples, followed by uniform preprocessing. A manifest records sample identifiers, provenance, labels, and fixed split assignments to prevent silent divergence across benchmark versions.

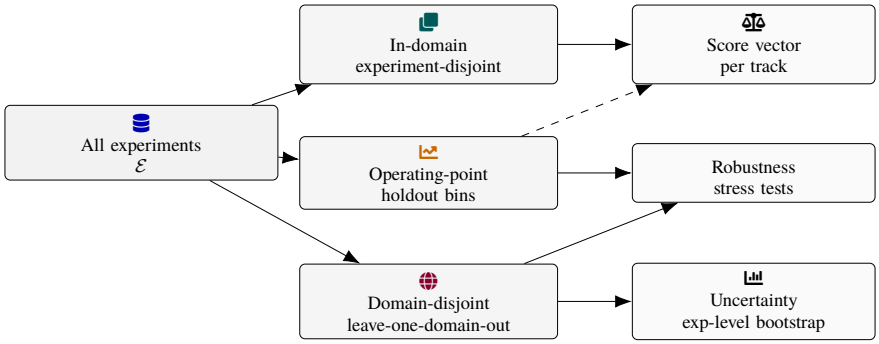


Figure 2: Split tracks define the intended notion of generalization. Experiment-disjoint in-domain evaluation targets new runs under the same facility context; operating-point holdouts test extrapolation across operating space; domain-disjoint evaluation tests portability across facilities or optical chains. Reporting is a track-wise score vector with robustness and experiment-level uncertainty.

inadvertently learn these signatures as shortcuts, producing high in-domain scores that degrade sharply when the optical context changes.

A second obstacle is label semantics. Regime boundaries are not crisp, and the same physical state can appear different under changes in exposure, depth of field, and view window placement. Conversely, different physical states can appear similar in a single frame when viewed through a limited window. Labeling policies thus influence the effective task difficulty [2]. If only highly confident frames are labeled and transitional segments are excluded, the task becomes easier, but the resulting system may be least reliable precisely where operational need is highest, namely near transitions. If, instead, transitional frames are forced into hard labels without expressing uncertainty, the benchmark may penalize models for disagreements that reflect intrinsic ambiguity rather than error. A benchmark intended to guide method development must address these issues explicitly by stating what is labeled, how uncertainty is treated, and what constitutes an independent evaluation example.

Annulus visualization studies have long recognized the importance of careful experimental design and high-speed imaging when interpreting two-phase structures; for example, high-speed visualization of annular flows has been used to resolve interface dynamics and regime evolution under controlled conditions [3]. The present paper focuses on a complementary problem: given that imaging data exist and that labels can be assigned, how should a community benchmark be specified so that model comparisons are fair and informative. The objective is not to define a single universal regime map or to claim that a discrete taxonomy captures all physics. Instead, the objective is to define a measurement instrument for algorithms, analogous to how standardized test sections and reporting practices enable meaningful comparison of pressure-drop correlations. A benchmark for regime recognition should clarify what the model is expected to generalize over, and it should make it difficult for models to exploit leakage or spurious cues.

The paper adopts the term benchmark to mean a specification that includes a dataset manifest, a label tax-

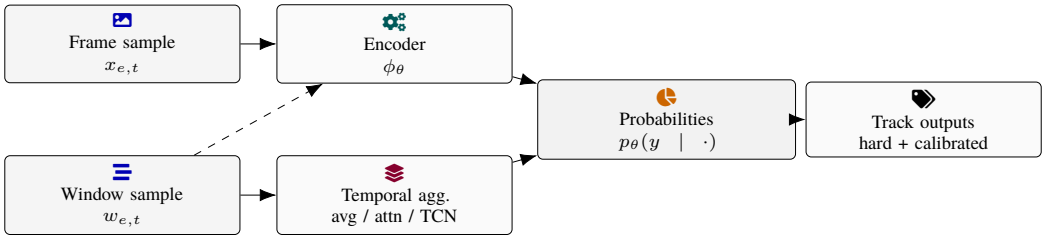


Figure 3: Task formalization for frame-level and sequence-level regime recognition. A shared visual encoder maps images to embeddings, while a temporal aggregator (for windows) summarizes short dynamics before producing a probability vector over regimes. Scoring uses probabilities to enable proper scoring rules and calibration analyses.

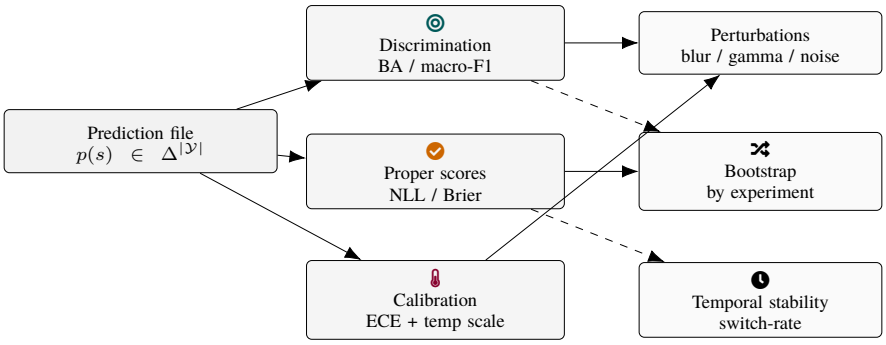


Figure 4: Metric suite emphasizes probability-aware evaluation. Discrimination metrics are class-balanced; proper scoring rules quantify probability quality; calibration is evaluated with fixed binning and optional temperature scaling. Robustness is stress-tested under controlled photometric and blur perturbations, with uncertainty estimated via experiment-level resampling.

onomy, split protocols, metrics, and a reporting template, together with a reference evaluator that computes scores from prediction files keyed by sample identifiers. This specification is designed to support both frame-level classification, where each image is assigned a regime label, and sequence-level classification, where short temporal windows are labeled to reflect the practical use case of streaming inference. A key principle is that the unit of independence is the experiment, not the frame. An experiment is a contiguous acquisition interval under nominally fixed optical configuration and operating conditions. By splitting at the experiment level, the benchmark evaluates whether a model generalizes to new realizations of a regime under similar nominal settings, rather than to nearby frames from the same realization.

The benchmark also recognizes that generalization can be defined at multiple levels. In-domain generalization asks whether a model trained on a subset of experiments from a facility performs well on other experiments from the same facility. Operating-point generalization asks whether a model extrapolates across ranges of superficial velocities or other operating parameters [4]. Cross-domain generalization asks whether a model trained on one facility or optical chain performs well on another. These questions cannot be answered by a single split, so the benchmark is structured as multiple evaluation tracks, each with an explicit scope statement. The result of a model is therefore a vector of scores, not a single number. This design discourages overstating performance and encourages focusing on robustness.

Finally, benchmarking must be feasible. A benchmark that requires extensive new annotation or rare metadata will not be adopted. The specification therefore distinguishes minimal requirements from optional enhancements. Minimal requirements include experiment identifiers, sample identifiers, regime labels for stable segments, and split assignments. Optional enhancements include transition labels, annotator votes, pixel-to-length calibration, and synchronized sensor signals. The benchmark is designed so that it can be instantiated using existing regime-tagged image collections and then progressively extended as new data are added and as governance mechanisms mature. The remainder of the paper formalizes the benchmark definition, describes dataset and label governance, specifies split and statistical protocols, defines metrics and diagnostic evaluations, and proposes baseline suites and reproducibility infrastructure.

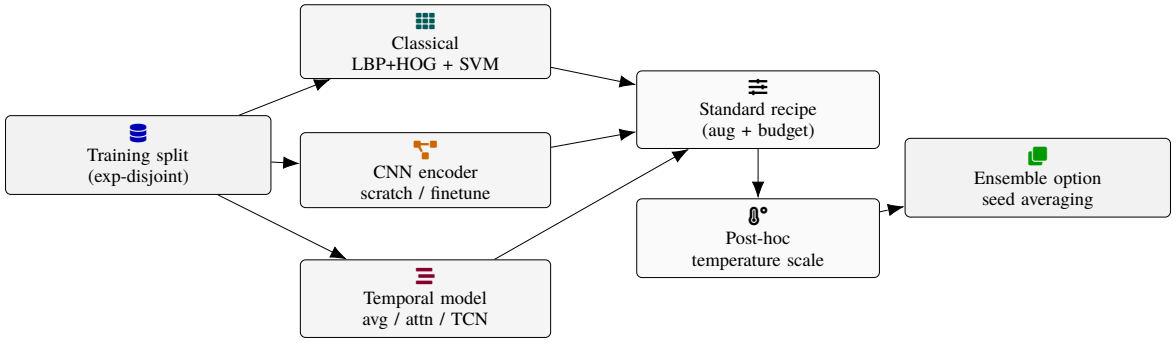


Figure 5: Baseline suite spans classical descriptors, convolutional encoders, and temporal aggregation for windowed inference. A fixed training recipe constrains optimization and augmentation degrees of freedom, while calibration is treated as a pipeline stage and optional ensembles provide a simple reference for improved uncertainty behavior.

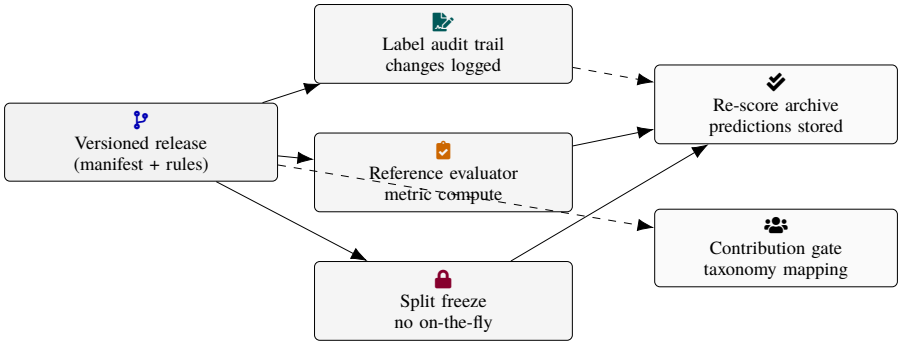


Figure 6: Reproducibility and governance layer: releases are versioned and manifest-driven, label edits are tracked with an audit trail, and split assignments are fixed. Centralized evaluator-side scoring produces consistent reports and enables re-scoring of archived predictions when metrics or benchmark diagnostics evolve.

2. Benchmark Definition and Task Formalization

A benchmark begins with a formal description of the data objects, the learning tasks, and the intended notion of independence. Let $\mathcal{E}$ denote the set of experiments. Each experiment $e \in \mathcal{E}$ is associated with a raw video stream V_e and a metadata record m_e. The metadata record includes an optical descriptor o_e and an operating descriptor u_e, where o_e summarizes camera, lens, illumination, and view-window geometry, and u_e summarizes the operating point, such as superficial velocities, pressure level, and temperature when available. The benchmark defines a deterministic sampling operator that maps each V_e to a set of samples. A sample can be a frame or a fixed-length window of frames [5]. The sampling operator must be declared and versioned because sampling density controls redundancy and, therefore, the effective task difficulty.

For a frame-level task, the sampled set for experiment e is $\{x_{e,t}\}_{t\in\mathcal{T}_e}$, where t indexes time or frame number and $\mathcal{T}_e$ is the chosen sampling index set. Labels are assigned as $y_{e,t} \in \mathcal{Y}$, where $\mathcal{Y}$ is a finite set of regime classes. For a sequence-level task, the sampled object is a window $w_{e,t} = (x_{e,t}, x_{e,t+\Delta}, \ldots, x_{e,t+(L-1)\Delta})$ with stride Δ and length L, and the label is $y^{(L)}_{e,t}$. The benchmark permits $y^{(L)}_{e,t}$ to be defined either as the modal regime within the window or as a distribution over regimes derived from sub-annotations, depending on whether a hard-label or soft-label sequence task is used.

The dataset is represented as a manifest of tuples keyed by unique identifiers. Each sample has an identifier $s = (e, t)$ for frame-level samples or $s = (e, t, L, \Delta)$ for sequence-level samples, enabling reconstruction of its provenance. The benchmark dataset is then

$$\begin{aligned} \mathcal{D} &= \{(s_i,\, x_i,\, y_i,\, e_i,\, m_{e_i})\}_{i=1}^{N} \\ \text{with} \quad s_i &= (e_i,\, t_i) \;\text{ or }\; (e_i,\, t_i,\, L,\, \Delta)\,. \end{aligned} \tag{1}$$

The explicit inclusion of e_i is not redundant; it is the key object for defining split policies and for statistical inference. In the benchmark, independence is assumed across experiments but not across samples within an experiment.

Table 1: Benchmark data objects and provenance keys

Object	Key fields	Identifier example	Notes
Experiment e	optical + operating context	`E_042`	Unit of independence
Raw stream V_e	video path, fps, resolution	`E_042.mp4`	Source for sampling
Metadata m_e	o_e, u_e descriptors	`meta/E_042.json`	Machine-readable, versioned
Frame sample $x_{e,t}$	time/frame index t	(`E_042`, t=18750)	Used in frame task
Window $w_{e,t}$	(L, Δ) + start t	(`E_042`, t=9000, L=16, Δ=2)	Used in sequence task

Table 2: Minimal metadata fields for benchmark releases

Category	Field	Type/Units	Purpose
Identity	experiment_id	string	split + resampling
Acquisition	date, fps	ISO date, Hz	timing provenance
Optics	camera_id, lens	strings	domain definition
Geometry	view region	normalized ROI	comparable crops
Imaging	resolution	pixels	input standardization
Operations	constancy flag	boolean/text	stability assumption

This does not claim that experiments are fully independent in a physical sense, since the same facility can reproduce similar behavior, but it provides a defensible statistical unit that avoids the most severe leakage.

A model is a parameterized mapping from samples to probabilities over regimes. For frame-level prediction, the model produces $p_\theta(y \mid x)$; for sequence-level prediction, it produces $p_\theta(y \mid w)$. In both cases, a benchmark submission is a prediction file that provides, for each sample identifier s, a probability vector $p(s) \in \Delta^{|\mathcal{Y}|}$ where $\Delta^{|\mathcal{Y}|}$ is the simplex. Hard predictions are derived by $\hat{y}(s) = \arg\max_k p_k(s)$, but benchmark scoring relies on the probabilities as well, so that calibration and proper scoring rules can be computed.

The benchmark distinguishes between stable-regime samples and transition samples. Stable-regime samples are those drawn from time intervals annotated as having a single dominant regime under the adopted taxonomy. Transition samples are those drawn from intervals where the regime changes or where multiple morphologies coexist. The benchmark defines a primary track that uses stable-regime samples with hard labels, and an auxiliary track that includes transition samples with either soft labels or an additional label that indicates uncertainty. This structure reduces pressure to exclude difficult data while keeping the primary task well-defined.

The label taxonomy is defined for annulus imagery in a manner intended to be visually discriminable and operationally consistent [6]. The taxonomy must be specified in the benchmark release, and changes to it constitute a major version increment. In an annulus, common macrostates include dispersed-bubble-dominated patterns, intermittent large-structure patterns, and highly agitated churn-like patterns. Because different laboratories use slightly different naming conventions, the benchmark defines each label by qualitative visual criteria rather than by laboratory-specific nomenclature. The criteria emphasize interface continuity, the presence and coherence length of gas structures, the degree of void coalescence, and the spatiotemporal variability of the projected void field. The benchmark does not require that labels correspond one-to-one with a specific mechanistic correlation regime, but it encourages recording operating descriptors so that downstream analyses can relate predicted regimes to operating space.

A key requirement is that all preprocessing that could alter the information content must be declared. Preprocessing includes cropping, resizing, color conversion, histogram equalization, masking overlays, and compression. Because optical artifacts can be correlated with labels, preprocessing choices can change the shortcut opportunities available to models. The benchmark therefore specifies a default preprocessing pipeline intended to re-

Table 3: Evaluation tracks and split strategies

Track	Split unit	Holdout criterion	Tests
In-domain	experiment	disjoint experiments	new runs, same setup
Operating-point	experiment bins	held-out operating ranges	extrapolation in u_e
Domain-disjoint	domain (facility/optics)	leave-one-domain-out	domain shift robustness
Stable-only	intervals/frames	stable segments only	crisp discrimination
Transition-aware	intervals/frames	include transitions/soft labels	ambiguity handling

Table 4: Sampling modes for frame-level and sequence-level tasks

Mode	Sample type	Key parameters	Guardrail
Frame-sparse	frame	stride in frames/time	reduces redundancy
Frame-dense	frame	high-rate extraction	requires exp-level splits
Seq-dense	window	small stride, overlap	exp-level aggregation
Seq-sparse	window	non-overlap/large stride	closer to independence
Manifest-fixed	both	declared operator + version	prevents silent drift

move obvious non-flow cues, such as timestamp overlays, while preserving flow structure. If overlays cannot be removed without affecting the image, a consistent mask is applied to all samples. The benchmark also records the spatial region of interest in normalized coordinates so that different crops can be compared and re-applied.

To link the benchmark to statistical learning objectives, the benchmark can be interpreted as estimating a generalization risk under a split-induced distribution. Let P_{tr} and P_{te} denote the training and test distributions over experiments and samples induced by the split protocol. A learned predictor $\hat{p}$ is evaluated by an expected loss under P_{te}. Proper scoring rules such as log loss are preferred for probability outputs. When hard labels are used, a canonical loss is cross-entropy,

$$\mathcal{L}_{\text{CE}}(\hat{p}) = \mathbb{E}_{(x,y)\sim P_{\text{te}}}\big[-\log \hat{p}(y \mid x)\big]. \tag{2}$$

When soft labels are available, the loss generalizes to the cross-entropy between the target distribution $q(y \mid x)$ and the prediction $\hat{p}(y \mid x)$. The benchmark does not mandate a training objective, but by scoring probabilities it encourages methods that produce meaningful confidence estimates rather than only hard decisions [7].

Finally, the benchmark defines what constitutes a valid submission. A valid submission must provide predictions for all samples in the test manifest for a given track, must not use the test labels, and must declare whether any external data were used. External data are permitted in an open track but not in a closed track. Because regime recognition in annulus imagery is often sensitive to domain properties, allowing external pretraining can confound comparisons. The benchmark therefore separates closed and open tracks explicitly, enabling fair comparison within each constraint set.

3. Dataset Standardization and Label Governance

The utility of a benchmark depends strongly on how the dataset is organized and governed. In annulus imagery, the same raw video can generate many benchmark variants depending on frame extraction, cropping, and label selection. To prevent silent divergence, the benchmark standardizes the data model through a manifest and requires versioned releases. Each release contains a list of experiments, a list of samples derived from those experiments, labels for each sample, and metadata pointers. The release is accompanied by a deterministic loader that repro-

Table 5: Core metric suite for regime recognition benchmarking

Metric	Category	Uses probabilities	Notes
Accuracy	discrimination	optional	reported, not primary
Balanced accuracy	discrimination	no	mean per-class recall
Macro-F1	discrimination	no	sensitive to FP/FN per class
NLL (log loss)	proper scoring	yes	penalizes overconfidence
Brier score	proper scoring	yes	probability MSE vs target
ECE	calibration	yes	fixed binning rule
Coverage–accuracy	abstention	yes	confidence threshold sweep

Table 6: Controlled robustness perturbations for test-time stress testing

Perturbation	Levels (example)	Deterministic	Motivation
Brightness/contrast	mild $\rightarrow$ strong	yes (seeded)	illumination drift
Gamma transform	$\gamma < 1$ and $\gamma > 1$	yes	exposure nonlinearity
Gaussian blur	increasing σ	yes	defocus/motion blur
Additive noise	increasing std	yes	sensor/compression noise
ROI shift/crop	small translations	yes	view-window variation
Border/background mask	fixed mask region	yes	shortcut sensitivity

duces the mapping from raw data to benchmark samples. If raw videos cannot be distributed, the release can distribute derived images with checksums and still preserve provenance through experiment and timestamp identifiers.

Metadata fields are stratified into required and optional categories. Required fields include an experiment identifier, acquisition date, camera identifier, nominal frame rate, image resolution, and a description of the view region relative to the annulus. Required operating fields include at least a qualitative descriptor indicating whether operating conditions are constant within the experiment segment. Optional fields include superficial velocities, pressure, temperature, and additional sensor streams [8]. The benchmark encourages storing operating descriptors in machine-readable form with units, but it does not require them for participation in the primary visual classification track. This choice reflects the reality that many historical image collections have incomplete sensor logs, while still enabling richer tracks when the data exist.

Label governance begins with an annotation protocol. Annotators label contiguous time intervals rather than isolated frames whenever possible. Interval labeling supports temporal consistency and reduces the risk that rare frames are mislabeled due to momentary occlusions or lighting fluctuations. Each interval is assigned a primary regime label if it is judged to be stable under the taxonomy, or it is marked as transitional if a change is occurring or if multiple morphologies co-exist. Frames sampled from stable intervals inherit the interval label. Frames sampled from transitional intervals are either withheld from the primary track or included in an auxiliary track with soft labels. Soft labels can be generated by having multiple annotators vote on the frame or by representing a mixture between the labels of adjacent stable intervals. The benchmark does not require a single method for soft labeling, but it requires that the chosen method be documented and that any soft labels be stored explicitly in the manifest.

The benchmark also addresses annotator disagreement. Disagreement is expected near boundaries and in highly agitated patterns. Rather than forcing a single ground truth, the benchmark encourages storing both a consensus label and an agreement score for a subset of samples. Agreement scores can be derived from vote entropy or from pairwise agreement rates. These scores enable stratified evaluation where performance is reported separately on high-agreement and low-agreement subsets [9]. Such stratification helps interpret whether a model is failing on intrinsically ambiguous data or on clearly la-

Table 7: Reference baseline model families and outputs

Family	Example method	Output	Calibration
Classical	LBP + HOG + linear model	class probs	Platt scaling
CNN (scratch)	modest ResNet-style encoder	class probs	temperature scaling
CNN (finetune)	generic pretraining + head	class probs	temperature scaling
Temporal (avg)	per-frame encoder + mean pool	window probs	temperature scaling
Temporal (attn)	encoder + attention pooling	window probs	temperature scaling
Ensemble	mean of K seeds	averaged probs	implicit + optional temp

Table 8: Standardized training and evaluation recipe elements

Component	Default choice	Rationale	Notes
Early stopping	val NLL	probability quality	avoids overconfidence
Imbalance handling	weights or balanced sampling	stable comparisons	report which used
Augmentations	mild shift/photometric/blur	label-preserving	avoid large rotations
Validation usage	model select + temp scaling	no test peeking	exp-disjoint
Sequence scoring	exp-level aggregation	overlap-safe	dense vs sparse modes
Uncertainty intervals	exp-level bootstrap	correlation-aware	percentile CI
Reporting	per-track score vector	scope clarity	no single-number claims

beled data, and it discourages cherry-picking only easy subsets.

A frequent pitfall in imaging benchmarks is the presence of non-flow cues correlated with labels. In annulus experiments, such cues can include differences in exposure settings across operating points, changes in background texture due to instrumentation changes, or overlays that encode run identifiers. The benchmark mitigates this by defining an overlay-handling policy. If overlays exist and can be removed, they are removed uniformly. If not, a fixed mask is applied to a region that covers overlays across all samples. The mask shape and location are part of the benchmark definition. This does not eliminate all shortcuts, but it prevents the most direct ones and increases the likelihood that models attend to flow structure.

Dataset balance is another governance issue. Some regimes occur rarely in typical operating campaigns, and naive sampling can yield severe imbalance. A benchmark that reflects the natural frequency is useful for deployment-like evaluation, but a benchmark that is too imbalanced can produce unstable metrics and can encourage trivial classifiers. The benchmark therefore defines two sampling views: a natural-frequency view and a balanced-evaluation view. In the natural-frequency view, the test set reflects the available data, and metrics such as balanced accuracy and macro-F1 are used to counteract imbalance. In the balanced-evaluation view, the test set is subsampled to equalize classes at the experiment level when feasible, and performance is reported on that view as well. The balanced-evaluation view is not intended to replace the natural view but to provide an additional lens for comparing methods without dominance by frequent regimes.

A benchmark also benefits from a clear relationship between the dataset and the label taxonomy. When new data are added, labels must be assigned consistently with prior releases. The benchmark therefore adopts a governance model where label guidelines are version-controlled and where label changes are logged. If a label is corrected,

Table 9: Reproducibility and governance policies for benchmark evolution

Item	Policy	Trigger	Artifact
Manifest	authoritative sample list	any release	versioned manifest
Splits	fixed, published assignments	new split set	named split files
Taxonomy	qualitative visual criteria	label definition change	major version bump
Label fixes	audit trail + rationale	correction/review	change log entry
Evaluator	centralized metric code	metric update	pinned evaluator tag
Provenance	checksums + extraction params	raw/derived updates	reproducible loader

the reason is recorded, such as resolution of annotator disagreement or discovery of an annotation shift caused by a guideline misinterpretation. This audit trail makes it possible to understand why a benchmark score changes across versions and to re-evaluate older models if desired.

The framework is particularly well suited to modern vertical annulus image collections that provide regime-tagged samples at scale. Manikonda et al. (2025) describe an annulus image corpus with regime annotations that functions as a practical training and validation substrate for vision models, and their modeling pipeline demonstrates that the learned regime recognizers were constructed and judged primarily on the basis of those images rather than on a separate surrogate signal [10]. This observation motivates an emphasis on benchmark discipline: when images are the dominant evidence source for training and evaluation, the benchmark protocol largely determines whether reported gains reflect genuine generalization or exploitation of correlation structure in the data.

A final aspect of standardization concerns file formats and color representations. Some annulus imaging setups use monochrome cameras; others use color cameras where illumination spectrum can vary. The benchmark defines a canonical input representation for evaluation, such as grayscale conversion with a specified formula or retention of RGB with a specified normalization. The chosen representation must be consistent across all samples to prevent models from using color differences that reflect illumination changes rather than flow morphology. If both monochrome and color data exist, the benchmark can define separate tracks or can convert all to a common grayscale representation. The appropriate choice depends on the dataset composition and must be documented.

4. Split Strategies and Statistical Inference Protocols

Evaluation splits are the central design choice that determines whether a benchmark measures true generalization [11]. In annulus imagery, a random frame-level split is inadequate because it allows leakage through temporal correlation and repeated operating sequences. The benchmark therefore defines split strategies at the experiment level and provides statistical inference procedures that respect that structure.

The default split is experiment-disjoint within a domain. A domain is defined as a set of experiments that share a facility and optical configuration, or more generally a set expected to have low domain shift. The training, validation, and test sets are formed by partitioning experiments, not frames, so that no experiment contributes samples to more than one set. This split measures generalization to new runs under similar nominal conditions. It answers the question of whether a model can recognize regimes across repeated realizations without memorizing run-specific textures. The split also supports a realistic deployment scenario where a model is trained on historical runs and then applied to new runs in the same facility.

A second split strategy is operating-point-disjoint evaluation. Here, experiments are grouped into bins based on operating descriptors, such as ranges of superficial velocities. One or more bins are held out for testing. This strategy measures extrapolation across operating space, which is important if a model is expected to function outside the range of operating points seen during training. Because operating descriptors can be noisy or incomplete, the benchmark requires that binning rules be specified and that the number of experiments per bin be reported. If operating metadata are missing, this track is not used, but

when available it provides a stronger test than purely random experiment splits, because it reduces the chance that the test set contains near-identical operating conditions.

A third strategy is domain-disjoint evaluation, where an entire domain is held out [12]. Domains can correspond to facilities, camera systems, illumination setups, or viewwindow geometries. In a leave-one-domain-out setting, a model is trained on experiments from multiple domains and tested on a new domain. This evaluates robustness to domain shift and is often the most informative indicator of practical portability. However, it can be harsh if domains differ in label semantics or if some regimes appear only in the held-out domain. The benchmark therefore requires that domain-disjoint tracks report which regimes are shared and which are absent, and it recommends reporting both overall performance on shared regimes and performance on the full label set where defined.

In addition to defining splits, the benchmark defines how statistical uncertainty is computed. Reporting a single number without uncertainty is misleading when the number of experiments is limited. Moreover, treating frames as independent samples underestimates uncertainty because within-experiment correlation reduces the effective sample size. The benchmark therefore computes confidence intervals by resampling experiments rather than frames. Let $\mathcal{E}_{\text{te}}$ be the set of test experiments. An experiment-level bootstrap draws $|\mathcal{E}_{\text{te}}|$ experiments with replacement, includes all their samples, and computes the metric on the aggregated sample set. Repeating this process yields a bootstrap distribution for the metric, from which percentile intervals can be computed. This procedure reflects variability across runs and yields more realistic uncertainty intervals.

The benchmark also defines how hyperparameter tuning interacts with splits. A common leakage pathway is tuning based on test performance, even indirectly. The benchmark therefore fixes a validation set at the experiment level and requires that model selection, early stopping, and calibration be performed using only the training and validation experiments. The test set is used only once per submission for final scoring [13]. If participants wish to perform extensive hyperparameter searches, they should do so via nested cross-validation at the experiment level, but the benchmark recognizes that this can be computationally burdensome. As a practical compromise, the benchmark defines a standard training budget and prescribes baseline hyperparameters for reference models. Submissions can deviate from these, but they must report their tuning strategy.

Sequence-level evaluation introduces additional split nuances. If windows overlap heavily, then even an experiment-disjoint split can produce inflated scores if the evaluation treats overlapping windows as independent and if the model exploits repeated structures within an experiment. The benchmark addresses this by defining two sequence sampling modes. In a dense mode, windows are extracted with small stride to mimic streaming inference, but metrics are aggregated per experiment before being averaged across experiments, so that each experiment contributes equally. In a sparse mode, windows are extracted with larger stride or with non-overlapping segmentation, yielding a set closer to independent samples. Both modes can be reported, with dense-mode results emphasizing operational performance and sparse-mode results emphasizing statistical stability.

The benchmark also defines aggregation strategies for converting frame-level predictions into experiment-level summaries, which can be useful for monitoring. For an experiment e with samples $s \in \mathcal{S}_e$, an experiment-level predicted regime distribution can be computed by averaging probabilities,

$$\bar{p}_e(k) = \frac{1}{|\mathcal{S}_e|} \sum_{s \in \mathcal{S}_e} p_s(k) \,. \tag{3}$$

An experiment-level hard label can then be derived by $\arg\max_k \bar{p}_e(k)$. The benchmark can score both sample-level and experiment-level predictions. Experiment-level scoring is useful when the objective is to identify the dominant regime over a run, while sample-level scoring is useful when the objective is to detect rapid changes. Reporting both clarifies whether a model is accurate only when averaged over time or also accurate at fine temporal resolution.

Finally, the benchmark defines procedures for handling missing classes in certain splits. In domain-disjoint tracks, it is possible that a regime class does not appear in the held-out domain [14]. Metrics such as balanced accuracy and macro-F1 should then be computed over the set of classes present in the test set, but the benchmark also reports the class set explicitly so that comparisons are fair. If a class appears in the test set but not in training, the task becomes a form of open-set recognition. The benchmark treats this as a separate challenge and discourages conflating it with standard closed-set classification. When open-set conditions occur, the benchmark recommends reporting additional metrics that quantify how often the model assigns low confidence to unseen classes, but it does not mandate a specific approach.

5. Metrics, Calibration, and Robustness Diagnostics

A benchmark for annulus regime recognition should evaluate more than top-1 accuracy. The probability outputs of modern models carry information that can be crucial in deployment, especially near transitions. Moreover, imbalance and ambiguity necessitate class-aware metrics and proper scoring rules. The benchmark therefore defines a metric suite with discrimination, probability quality, calibration, and robustness components, and it specifies how each component is computed from prediction files.

Discrimination metrics begin with accuracy but emphasize class-balanced measures. Balanced accuracy is computed as the mean of per-class recall. For class set $\mathcal{Y}$ with $C = |\mathcal{Y}|$ and counts n_c,

$$\mathrm{BA} = \frac{1}{C} \sum_{c \in \mathcal{Y}} \frac{1}{n_c} \sum_{i:\, y_i = c} \mathbf{1}[\hat{y}_i = c] \,. \tag{4}$$

Macro-F1 is also reported because it is sensitive to both false positives and false negatives for each class. Because macro-F1 can be unstable when a class is rare, the benchmark requires reporting class counts and suggests computing intervals by experiment-level bootstrapping.

Probability quality is assessed using proper scoring rules. Negative log-likelihood (log loss) is computed on predicted probabilities and penalizes overconfident errors. The Brier score is computed as the mean squared error between the probability vector and the one-hot target [15]. These metrics are particularly informative when comparing methods that trade accuracy for better-calibrated probabilities. For sample index i with predicted probability p_i and one-hot target q_i,

$$\begin{aligned} \mathrm{NLL} &= -\frac{1}{N} \sum_{i=1}^{N} \log p_i(y_i) \\ \mathrm{Brier} &= \frac{1}{N} \sum_{i=1}^{N} \|p_i - q_i\|_2^2 \,. \end{aligned} \tag{5}$$

When soft labels are used in a transition track, q_i is not one-hot and the same definitions apply.

Calibration is evaluated using expected calibration error and reliability diagrams. Let $\hat{c}_i = \max_k p_i(k)$ and $\hat{y}_i = \arg\max_k p_i(k)$. Samples are partitioned into bins by $\hat{c}_i$. The expected calibration error is computed by comparing average confidence to empirical accuracy in each bin. Because ECE depends on binning, the benchmark fixes the number of bins and the binning rule. In addition, the benchmark reports calibration both before and after a standard post-hoc temperature scaling procedure fitted on the validation experiments. This is important because temperature scaling can correct systematic overconfidence without changing classification decisions, and comparing calibrated outputs clarifies whether differences in probability quality are intrinsic or correctable.

A principal concern in annulus imaging is robustness to photometric and optical variations. The benchmark includes a controlled perturbation suite applied at test time. Perturbations include brightness and contrast adjustments, gamma transformations, Gaussian blur, and additive noise. Each perturbation is applied at multiple severities. For each severity, the full metric suite is computed, and robustness is summarized as the mean balanced accuracy across severities or as the area under a degradation curve. The objective is not to simulate every real artifact but to provide a standardized stress test that detects models that rely on brittle cues. The benchmark specifies that perturbations be applied deterministically given a seed so that results can be reproduced exactly [16].

In addition to perturbations, the benchmark includes diagnostics for shortcut learning. One diagnostic is to mask non-flow regions such as borders and fixed backgrounds, then re-score. If performance drops sharply when borders are masked, the model may have relied on border artifacts. Another diagnostic is to evaluate on cropped regions that shift the field of view slightly. Robust models should tolerate small shifts without large degradation, provided the shift does not remove critical flow content. These diagnostics are optional but recommended because they provide interpretable evidence about what cues the model uses.

Transition sensitivity is evaluated in the auxiliary track. If transition intervals are annotated, the benchmark reports performance as a function of distance to the nearest transition boundary. For example, samples can be grouped by whether they are within a short time of a transition and scored separately. Models that perform well on stable segments but fail near boundaries can then be identified. If soft labels are available, the benchmark also reports the divergence between predicted probabilities and the soft targets, which can reveal whether the model captures uncertainty in mixed patterns.

Temporal consistency is relevant even in the frame-level track because deployments often apply frame-level models in streaming mode. The benchmark defines a temporal smoothness diagnostic computed per experiment by counting the number of predicted label changes per unit time and comparing it to a reference derived from annota-

tions. Excessive switching suggests instability, while too little switching can indicate over-smoothing if a temporal filter is applied. The benchmark does not mandate a particular smoothness target, but it requires reporting the switch rate so that temporal behavior is visible [17].

To support decision-making under uncertainty, the benchmark includes an abstention analysis when models provide meaningful confidence estimates. An abstention rule rejects samples with confidence below a threshold. The benchmark then reports accuracy on the retained samples as a function of coverage. This produces a coverage–accuracy curve that indicates whether a model can trade coverage for reliability. Because not all models support abstention in a calibrated way, this analysis is optional, but it is valuable for operational contexts where low-confidence frames can be flagged for additional sensing or for human review.

Finally, the benchmark prescribes that all metrics be computed by a reference evaluator that reads prediction files. This prevents subtle inconsistencies across implementations. The evaluator also enforces that predictions sum to one and that all samples are present. When a submission is missing predictions, the evaluator rejects it rather than silently imputing. This strictness is important because missing predictions can bias metrics, particularly if missingness correlates with difficult cases.

6. Baseline Models and Standardized Training Recipes

Baselines anchor a benchmark by establishing expected performance ranges and by clarifying how much improvement is meaningful. In annulus imagery, baselines should cover classical feature-based models, modern convolutional models, and temporal aggregation models for sequence tasks. The baseline suite described here is intended to be reproducible under standardized recipes rather than to exhaustively optimize performance.

A classical baseline begins with deterministic preprocessing, including resizing, grayscale conversion if required, and normalization. Feature extraction uses multiscale texture and edge descriptors [18]. In annulus images, dispersed bubbles generate fine-grained texture, while large coherent gas structures generate long edges and regions of uniform intensity. A classical descriptor suite can capture both by combining local binary patterns for microtexture with histogram-of-oriented-gradients for edge orientation statistics. The resulting feature vector is standardized and reduced via principal component analysis to control dimensionality. A linear SVM or logistic regression classifier then produces class scores, which can be calibrated via Platt scaling on the validation set. This baseline is computationally inexpensive and provides a useful reference that is less prone to overfitting through high-capacity memorization, although it can still exploit shortcuts if preprocessing leaves non-flow cues.

A convolutional baseline uses a residual-style encoder with a modest parameter count. The benchmark includes two variants: training from scratch and fine-tuning from generic pretraining. Training from scratch provides a clear measure of what the annulus dataset alone supports. Fine-tuning provides a practical reference because many workflows begin with pretrained encoders. The benchmark fixes the input resolution, optimizer type, learning rate schedule, batch size, and augmentation policy for these baselines to reduce degrees of freedom. Augmentations are limited to those that plausibly preserve regime semantics, such as small translations, mild brightness jitter, and mild blur. Large rotations are avoided because gravitational direction and illumination asymmetry can make rotation non-invariant. Horizontal flips are treated cautiously; if the imaging setup is symmetric and the view window is centered, flips may be permissible, but if there are fixed asymmetries, flips can introduce unrealistic patterns. The benchmark therefore defines flip augmentation only in a track where symmetry has been verified or where flips are disabled by default.

Temporal baselines for sequence-level tasks begin with a per-frame encoder that produces embeddings $z_t = \phi_\theta(x_t)$ [19]. A simple temporal aggregator averages embeddings across the window and applies a linear head. A more expressive aggregator uses attention weights that depend on embeddings, enabling the model to focus on frames that contain clear interface structures. Another aggregator uses a temporal convolution over embeddings, capturing short-range dynamics without requiring recurrent state. The benchmark provides at least one temporal baseline and requires that temporal models be evaluated with experiment-disjoint splits and with an explicit window sampling policy to avoid leakage through overlapping windows.

Training recipes emphasize probability quality. Early stopping is based on validation negative log-likelihood rather than validation accuracy so that models are discouraged from becoming overconfident. Class imbalance is handled by either class-weighted loss or balanced sampling; the baseline suite includes both strategies in separate variants because each can behave differently under domain shift. Class weighting can cause instability when rare classes have noisy labels, while balanced sampling

can increase variance because it oversamples rare cases. By including both, the benchmark reveals how sensitive performance is to imbalance handling.

Calibration is treated as part of the baseline pipeline. After training, temperature scaling is fitted on the validation set to minimize log loss, and calibrated probabilities are then used for scoring. The benchmark reports both calibrated and uncalibrated metrics. This is important because some architectures are systematically miscalibrated, and calibration can change ranking under proper scoring rules even when accuracy is unchanged.

The benchmark also defines a minimal ensemble baseline because ensembles often improve both accuracy and calibration. The ensemble baseline trains multiple instances of the same architecture with different random seeds and averages their predicted probabilities [20]. This provides a reference for uncertainty-aware methods without requiring specialized Bayesian inference. The benchmark requires reporting the number of ensemble members and the computational cost, because ensemble improvements must be weighed against inference constraints in deployment.

To reduce the risk that baselines become moving targets, the benchmark pins baseline code and hyperparameters to the benchmark version. When a new dataset version is released, baselines are re-run under the same recipes, and results are updated. This provides continuity over time and allows researchers to interpret whether performance gains are due to better data or better methods.

A practical concern is that different participants may have different compute budgets. The benchmark addresses this by defining baseline tiers. A small tier uses a lightweight encoder suitable for real-time inference on modest hardware. A medium tier uses a standard residual encoder. A large tier uses a higher-capacity model. Each tier has fixed training recipes. Participants can choose a tier or report multiple tiers. This encourages fair comparison among methods with similar capacity and avoids conflating algorithmic improvements with brute-force scaling.

Baselines also support diagnostic reporting. For each baseline, the benchmark encourages reporting per-class confusion patterns and per-experiment performance variability [21]. In annulus imagery, it is common for a model to perform well on certain experiments and poorly on others due to optical differences. Reporting distributions over experiments makes this variability visible and helps identify whether a model improvement is broad or concentrated. A method that improves the lower tail of per-experiment performance can be operationally meaningful even if the mean improvement is small, because failures are often unacceptable in monitoring contexts.

7. Reproducibility Infrastructure and Benchmark Governance

A benchmark becomes durable when it is reproducible and governed. Reproducibility requires that data provenance, preprocessing, splitting, and metric computation be controlled and auditable. Governance requires that changes to labels and dataset composition be tracked and that community contributions be integrated without breaking comparability.

The benchmark adopts a manifest-driven design. The manifest lists all experiments, all samples, their labels, and their split assignments. Each sample entry includes a checksum of the image file or a pointer to the raw video location and extraction parameters. This enables verification that two users are evaluating on the same data. If derived images are distributed, the manifest includes checksums to detect corruption or accidental modification. If raw videos are distributed, the manifest includes frame indices and extraction scripts so that derived images can be regenerated.

Splits are defined in the manifest, not generated on the fly. This prevents accidental changes in split assignment due to differences in random seeds or library versions. The benchmark can provide multiple predefined splits for statistical robustness, but each split is still fixed and named [22]. Participants report which split they used. If cross-validation is used, folds are defined at the experiment level and published as part of the benchmark. This ensures that results can be compared across papers without ambiguity about how folds were constructed.

Metric computation is centralized in a reference evaluator. Participants submit prediction files keyed by sample identifiers. The evaluator computes metrics and diagnostics and generates a standardized report. Centralized evaluation reduces opportunities for inconsistent implementation choices, such as differences in macro averaging or binning for calibration metrics. It also enables the benchmark maintainers to update metrics or diagnostics and to re-score older submissions if desired, provided that prediction files are archived.

Benchmark governance includes versioning policies. A major version change occurs when the label taxonomy changes or when the split protocol changes in a way that affects comparability. A minor version change occurs when new experiments are added or when labels are corrected in a way that may affect scores. A patch change occurs

when metadata are corrected without affecting labels. Every reported result must specify the benchmark version. This requirement enables readers to interpret whether two results are directly comparable.

Label corrections and disagreements are handled through an audit trail [23]. When a label is corrected, the benchmark records the original label, the new label, the reason, and the evidence used. Evidence can include annotator re-review, consensus meetings, or additional sensor data when available. The audit trail is important because regime labeling is partly subjective. The benchmark does not assume that labels are perfect; it assumes that labels are governed transparently so that the community can understand changes.

A key governance issue is how to incorporate new data from different facilities. Adding a new domain can strengthen the benchmark by enabling domain-disjoint tracks, but it can also introduce taxonomy mismatch if regimes are labeled differently. The benchmark therefore requires that new data contributions include a mapping to the benchmark taxonomy and a description of how labeling guidelines were applied. If a new facility uses a different regime naming convention, the mapping must be justified. If a regime is not observable in a given view, this must be noted. The benchmark can then decide whether to include the new data in existing tracks, create a new track, or hold it as an external validation set.

Another governance issue is protecting against overfitting to the test set. If the test set is public and repeatedly used, the community can inadvertently tune to it. The benchmark mitigates this by encouraging the use of multiple splits and by optionally maintaining a hidden test set for periodic evaluations. Even without a hidden set, maintaining a time-based holdout where newer experiments are added and held out can provide a moving target that reduces overfitting. The benchmark also encourages reporting performance on all predefined splits rather than reporting only the best split [24].

The benchmark infrastructure supports richer analyses by preserving metadata and enabling experiment-level aggregation. For example, if superficial velocities are available, benchmark reports can include performance stratified by velocity ranges. This can reveal whether a model fails systematically at certain operating extremes. If optical metadata are available, reports can include performance stratified by camera or illumination configuration. Such stratification can guide domain adaptation research and can help determine whether additional data collection is needed in specific contexts.

Finally, governance includes ethical and safety considerations specific to the user context and to engineering practice. While regime recognition benchmarks do not involve personal data, they can still influence operational decisions. The benchmark therefore encourages reporting model limitations and failure modes, and it discourages interpreting benchmark scores as guarantees of safe deployment. The benchmark is intended to support research and development; operational use requires additional validation under the target environment and appropriate engineering controls.

8. Conclusion

This paper specified a benchmarking methodology for flow-regime classification from vertical annulus imagery, with the objective of making computer vision results comparable across studies and facilities. The benchmark definition formalized experiments as the unit of independence, defined manifest-driven dataset standardization, and separated stable-regime evaluation from transition-aware auxiliary evaluation. Split strategies were prescribed to address in-domain, operating-point, and domain-disjoint generalization, and statistical uncertainty was defined through experiment-level resampling rather than frame-level assumptions. A metric suite was proposed that combines class-balanced discrimination, proper scoring rules for probability quality, calibration evaluation, and robustness stress tests under controlled perturbations, together with diagnostics for temporal consistency and shortcut sensitivity. Baseline models and standardized training recipes were described to anchor comparisons across classical descriptors, convolutional encoders, and temporal aggregation models, with calibration treated as a first-class component. Finally, reproducibility infrastructure and governance policies were outlined to ensure that preprocessing, labels, splits, and metric computation remain auditable and version-controlled as the benchmark evolves. Under these conventions, future papers can report regime-recognition performance in a way that clarifies the scope of generalization, exposes trade-offs between accuracy and reliability, and reduces the likelihood that improvements are driven by leakage or non-flow artifacts rather than by genuinely improved modeling of annular two-phase visual structure [25].

References

[1] S. Kalita, E. M. Teschner, P. M. Sander, and D. Konietzko-Meier, “To be or not to be heavier:

The role of dermal bones in the buoyancy of the late triassic temnospondyl amphibian metoposaurus krasiejowensis.," *Journal of anatomy*, vol. 241, pp. 1459–1476, 9 2022.

[2] T. N. Ofei, S. Irawan, and W. Pao, *Flow Profile Distribution in Narrow Annuli Using Modified Yield Power-Law Fluid Model*, pp. 209–217. Springer Singapore, 3 2015.

[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] J. Brock and K. Bromberg, *Kleinian Groups and Hyperbolic 3-Manifolds: Cone-manifolds and the density conjecture*, pp. 75–94. Cambridge University Press, 11 2003.

[5] M. tao Fan, J. Li, G. hui Liu, H. li Zhang, G. Tang, Z. Lang, X. hui Hu, and J. hua Guo, "The new method to analysis the influence of trapped pressure on casing stress in multi-stage fracturing of shale gas," *DEStech Transactions on Engineering and Technology Research*, 10 2017.

[6] N. E. Karraker, M. D. Kusrini, J. R. Atutubo, R. M. Healey, and A. Yusratul, "Non-marine turtle plays important functional roles in indonesian ecosystems.," *Ecology and evolution*, vol. 10, pp. 9613–9623, 9 2020.

[7] G. T. Llewellyn, F. L. Dorman, J. L. Westland, D. Yoxtheimer, P. L. Grieve, T. Sowers, E. Humston-Fulmer, and S. L. Brantley, "Evaluating a groundwater supply contamination incident attributed to marcellus shale gas development," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. 6325–6330, 5 2015.

[8] B. R. Hammond, B. R. Wooten, S. E. Saint, and L. M. Renzi-Hammond, "The effects of a blue-light filtering versus clear intraocular implant on color appearance.," *Translational vision science & technology*, vol. 10, pp. 25–25, 10 2021.

[9] S. E. Salem and M. N. El-Din, "Drillpipe eccentricity prediction during drilling directional wells," in *Canadian International Petroleum Conference*, PETSOC, 6 2006.

[10] 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. 111245, 2025.

[11] G. Waters and B. Wray, "Narrow annulus cementing in the deep anadarko basin," *SPE Drilling & Completion*, vol. 10, pp. 53–60, 3 1995.

[12] J. A. Byrd and H. E. Helms, "Ceramic applications in turbine engines," in *18th Joint Propulsion Conference*, American Institute of Aeronautics and Astronautics, 6 1982.

[13] T. Li, "Taut foliations of 3-manifolds with heegaard genus two," 2 2022.

[14] Y. Liu, E. R. Upchurch, E. M. Ozbayoglu, S. Baldino, J. Wang, and D. Zheng, "Design and calculation of process parameters in bullheading and pressurized mud cap drilling," in *SPE/IADC International Drilling Conference and Exhibition*, SPE, 3 2023.

[15] N. Somaratne and G. Hallas, "Review of risk status of groundwater supply wells by tracing the source of coliform contamination," *Water*, vol. 7, pp. 3878–3905, 7 2015.

[16] M. M. Huque, S. Imtiaz, A. Rahman, and M. E. Hossain, "Kick detection and remedial action in managed pressure drilling: a review," *SN Applied Sciences*, vol. 2, pp. 1–29, 6 2020.

[17] K. Finlinson and J. S. Purcell, "Volumes of montesinos links," *Pacific Journal of Mathematics*, vol. 282, pp. 63–105, 2 2016.

[18] Z. Liu, R. Samuel, A. Gonzales, and Y. Kang, "Modeling and simulation of annular pressure buildup (apb) mitigation using rupture disk," *IADC/SPE Drilling Conference and Exhibition*, 3 2016.

[19] D. K. Jafar, "Slot flow approximation for herschel-bulkley drilling fluid through eccentric annuli," 5 2015.

[20] 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, pp. 635–643, 9 2018.

[21] F. R. Straehl, T. M. Scheyer, A. M. Forasiepi, R. D. E. MacPhee, and M. R. Sánchez-Villagra, "Evolutionary patterns of bone histology and bone compactness

in xenarthran mammal long bones.," *PloS one*, vol. 8, pp. e69275–, 7 2013.

[22] S. D. Newsome, M. A. Etnier, C. M. Kurle, J. R. Waldbauer, C. P. Chamberlain, and P. L. Koch, "Historic decline in primary productivity in western gulf of alaska and eastern bering sea: isotopic analysis of northern fur seal teeth," *Marine Ecology Progress Series*, vol. 332, pp. 211–224, 3 2007.

[23] Z. Chen, S. Chaudhary, and J. Shine, "Intermixing of cementing fluids: Understanding mud displacement and cement placement," *IADC/SPE Drilling Conference and Exhibition*, 3 2014.

[24] W. B. Daniel and M. Radonjic, "Application of geomimicry to engineering problems:gilsonite as a solution to leaky wellbores," in *Proceedings of The International conference on innovative research in Science Engineering & Technology*, Acavent, 12 2018.

[25] G. M. Holm, J. A. Owens, G. Page, and A. Ritchie, "The hod geohazard : a unique overpressured interval. cooperation in it's recognition, evaluation and risk mitigation.," *SPE Offshore Europe Oil and Gas Exhibition and Conference*, 9 2005.