

Article

CausalShift: A Modular, Plugin-Based Framework for Dataset Shift Handling in Machine Learning

Shuang Song , Muhammad Syafiq Mohd Pozi  and Nik Fatinah N. Mohd Farid 

School of Computing, Universiti Utara Malaysia, Sintok 06010, Kedah, Malaysia

* Correspondence: song_shuang@ahsgs.uum.edu.my

Abstract

Machine learning models deployed under distribution shift suffer performance degradation whose root cause is rarely diagnosed before adaptation is attempted. Existing frameworks either lack a closed loop from shift diagnosis to method selection, require a fully specified causal graph, or depend on multiple heterogeneous training environments—conditions rarely met in practice. We propose CausalShift, a modular, plugin-based framework for end-to-end dataset shift handling. A universal statistical core remains fully functional without any causal assumption; four optional causal plugins are activated selectively by a four-level knowledge grading system that scales from a fully specified structural causal model (Level A) down to a single observational dataset (Level D). A three-path causal attribution module estimates the relative contributions of covariate and concept shift to the performance gap, and attribution-driven routing maps the diagnosis to an adaptation strategy grounded in minimax optimality. Proof-of-concept experiments on four benchmarks—a synthetic structural causal model (SCM), Colored Modified National Institute of Standards and Technology (MNIST), Benchmarking In-the-Wild Distribution Shifts (WILDS) Camelyon17, and five American Community Survey (ACS) geographic shift tasks—demonstrate that CausalShift reduces the in-distribution to out-of-distribution (ID–OOD) accuracy gap by 19.9 percentage points on the synthetic benchmark and 72.5 percentage points on the spurious correlation benchmark, while remaining competitive on real-world image shift and achieving performance parity with ERM on mild-shift tasks. A tiered causal evaluation suite, including Λ^* -robustness bounds and distribution-induced shift decomposition of error (DISDE), reveals robustness differences that scalar accuracy metrics cannot detect.

Keywords: dataset shift; distribution shift; causal inference; covariate shift; concept shift; domain generalisation; causal attribution; spurious correlation; invariant risk minimisation; robustness evaluation



Received: 7 June 2026

Revised: 21 July 2026

Accepted: 28 July 2026

Published: 31 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Machine learning models are routinely deployed under conditions that differ from those in which they were trained. A clinical risk-scoring model calibrated on data from one hospital may degrade silently when transferred to another institution whose patient demographics, coding practices, or admission criteria differ [1,2]. A fraud detection system trained on historical transactions may lose discriminative power as spending patterns evolve. A predictive maintenance model accurate under one set of operating conditions may produce unreliable alerts at a new facility. These failures share a common cause: the joint distribution of inputs and labels differs between the training (source) domain

and the deployment (target) domain, a phenomenon known as dataset shift [3] (used interchangeably with distribution shift throughout this paper).

A running example. To ground the concepts for readers less familiar with distribution shift, consider a digit classification model trained on Colored MNIST images in which digit colour is a spurious but highly predictive signal: colour correlates with the digit label at 90% in the training environments. At deployment the labelling convention reverses—colour now anti-predicts the label—causing all standard baselines (ERM, IRM, CORAL) to collapse from ~72% training accuracy to ~12% test accuracy, worse than random (Section 6.3). A practitioner armed only with the scalar accuracy drop has no principled way to determine whether the failure is due to a covariate shift (addressable by importance reweighting), a concept shift (requiring retraining on invariant features), or a mixed combination—and therefore cannot select an appropriate remedy. CausalShift addresses this gap automatically: Stage 1 detects the change in $P(Y|X)$ caused by the reversed spurious correlation; Stage 2 attributes the gap to a concept-shift component; Stage 3 routes to an importance-weighting correction that downweights source samples relying on colour; and Stage 5 recovers a stable 50% accuracy with a near-zero ID–OOD gap of 0.1%—the theoretical maximum for a logistic regression pipeline on flattened pixel features. This closed-loop diagnosis-to-adaptation pipeline, operating without a causal graph or multiple training environments, illustrates the core value proposition of CausalShift.

Despite two decades of active research, no existing framework simultaneously addresses four critical gaps. (1) No closed loop from shift diagnosis to adaptation method selection. State-of-the-art detection toolkits such as DetectShift [4] can accurately identify whether covariate, label, or concept shift is present, and benchmark platforms such as WILDS [1] and TableShift [5] provide rich evaluation environments—yet none of these frameworks prescribes which adaptation strategy to apply given the detected shift type. A practitioner who learns that concept shift is dominant still faces the full adaptation literature unaided, choosing between importance weighting, invariant feature learning, or test-time adaptation without principled guidance. (2) No principled use of partial causal knowledge. Existing causal frameworks operate under an all-or-nothing assumption: Subbaswamy et al. [6] require a fully specified structural causal model (SCM), while Invariant Risk Minimisation (IRM) [7] requires multiple heterogeneous training environments—conditions that are rarely available in practice, where a single observational dataset from one time period is the norm [8]. When these conditions are unmet, the frameworks either fail silently or force practitioners to abandon causal tools entirely in favour of statistical baselines, with no intermediate path that exploits whatever partial causal knowledge is available. (3) No joint exploitation of SCM and Potential Outcomes Model (POM) tools. The two dominant causal inference paradigms—the SCM/directed acyclic graph tradition and the potential outcomes framework—have largely developed in parallel, each offering complementary strengths: SCMs excel at mechanism-level attribution and transportability, while POM-based methods (e.g., doubly robust estimation) excel at efficient semiparametric inference under unmeasured confounding [9]. Existing dataset shift frameworks draw exclusively from one paradigm at a time, forfeiting the complementary benefits that a unified approach could provide. (4) No evaluation beyond a scalar accuracy gap. The standard ID–OOD accuracy gap [10] is a point estimate on a fixed target distribution. It cannot distinguish a drop caused by covariate mismatch (addressable by reweighting) from one caused by concept drift (requiring retraining), nor can it certify how much additional hidden shift a model can tolerate before failing—a question of direct operational relevance in safety-critical deployments where the test distribution continues to evolve after model release [11].

We propose CausalShift, a modular, plugin-based framework for end-to-end dataset shift handling. The framework's statistical core handles detection, adaptation, and evaluation without any causal assumption, and is fully functional as a standalone system. Causal inference components are available as optional plugins that, when activated, provide three concrete benefits that statistical methods alone cannot deliver. First, causal attribution identifies not only the type of shift but its structural source in the data generation process—distinguishing, for example, whether a covariate shift is driven by a change in the causal mechanism of a feature or merely by a spurious upstream change—enabling more targeted and durable interventions [6]. Second, causal invariance guarantees: by identifying features whose conditional distributions are stable across environments, the framework can certify that adaptation is robust to future shifts of the same structural class, rather than merely correcting the currently observed distributional mismatch [12,13]. Third, causal evaluation: when causal structure is available, robustness can be assessed not only as a point estimate on a fixed target dataset but as a bound over a family of distributions, and as a formal transportability guarantee, providing stronger and more informative assurances than scalar accuracy gaps [14].

The framework operates through three architectural layers. A universal statistical core (Layer 1) is always active, providing valid shift detection, type identification, adaptation guidance, and a baseline evaluation suite without any causal assumption. Causal plugin modules (Layer 2) are activated selectively by a four-level causal knowledge grading system (Stage 0) that assesses available causal evidence—from a fully specified SCM down to a single observational dataset—activating only the plugins that are epistemically and computationally justified. Non-causal fallback methods (Layer 3), including conformal prediction and test-time adaptation, provide coverage-guaranteed alternatives when causal assumptions cannot be supported. The central mechanism is attribution-driven method routing (Stage 3): the identified shift type, causal source, and knowledge level jointly determine the adaptation strategy, grounded in the minimax optimality result of Subbaswamy et al. [6]. When causal plugins are inactive—owing to high dimensionality, nonlinearity, or unobserved confounders—the framework degrades gracefully to a competitive statistical baseline, directly addressing the graph misspecification concern that limits all existing causal shift frameworks.

The main contributions of this paper are as follows.

1. Causal statistical plugin architecture with knowledge grading. We propose a modular, three-layer architecture in which causal inference components are optional plugins activated by a four-level causal knowledge grading system (Stage 0). This is the first dataset shift framework to provide a principled degradation path from full SCM-based causal analysis (Level A) down to a purely statistical pipeline (Level D), ensuring the framework remains fully functional regardless of available causal knowledge. This directly addresses the all-or-nothing limitation of existing causal frameworks [6,12].
2. Spurious association-aware causal attribution. We introduce a three-path attribution module (Stage 2) that identifies not only the type of shift (covariate, label, or concept) but its structural source, explicitly distinguishing causal mechanisms from spurious associations in the data-generation process [15,16]. When causal plugins are active, the module integrates SCM-based attribution (Subbaswamy Theorem 1, NILE estimator [12]) with potential outcomes decomposition via DRCFS [13], yielding a quantitative separation of the covariate shift contribution Δ_{cov} from the concept shift contribution Δ_{concept} —a decomposition absent from all prior frameworks. When causal plugins are inactive, the module falls back to statistical invariance tests and sample reweighting [17,18].

3. Attribution-driven method routing. We introduce a principled routing mechanism (Stage 3) that maps identified shift type, causal knowledge level, and computational budget to an adaptation strategy. Routing is grounded in the minimax optimality of counterfactual stable distributions [6] and incorporates a conformal prediction fallback [10] for mixed or uncertain shift. To our knowledge, this is the first framework to close the loop between shift attribution and adaptation method selection, replacing ad hoc practitioner judgement with a theoretically motivated decision procedure.
4. Tiered evaluation suite that scales with causal knowledge. We replace the scalar OOD accuracy gap with a tiered evaluation suite whose depth scales with available causal knowledge. At the statistical core level (always available), Δ -robustness bounds certify how much additional shift a model can tolerate before failing, without requiring any causal assumption. When the attribution plugin is active (Levels B/C), the distribution-induced shift decomposition of error (DISDE) separates the performance gap into Δ_{cov} and Δ_{concept} , enabling targeted intervention. When a full SCM is available (Level A), TATE transportability tests formally verify whether source domain effects transfer to the target population [10]. The suite remains informative at every knowledge level, including when all causal plugins are deactivated.
5. Data quality audit as a first-class pipeline stage. A pre-detection audit module (Stage 0.5) screens for Berkson bias [19], selection bias, collider bias, and annotation inconsistency before any shift analysis is performed. All existing frameworks implicitly assume data trustworthiness; ours is the first to treat data bias as an explicit pipeline stage that conditions the validity of all downstream analyses.

Section 2 reviews background and related work. Section 3 presents the plugin architecture and Stages 0–0.5. Section 4 details causal attribution and method routing (Stages 1–3). Section 5 covers shift-robust modelling, counterfactual augmentation, causal evaluation, and deployment monitoring (Stages 4–7). Section 6 reports proof-of-concept experiments on four benchmark datasets. Section 7 discusses findings and limitations. Section 8 concludes. Code is available at <https://github.com/Song1823/CausalShift> (accessed on 8 June 2026).

2. Background and Related Work

2.1. Forms of Dataset Shift

Dataset shift takes several canonical forms. Covariate shift arises when the marginal input distribution changes, $P_{\text{src}}(X) \neq P_{\text{tgt}}(X)$, while the conditional $P(Y | X)$ remains stable. Label shift (prior probability shift) occurs when $P(Y)$ changes across domains. Concept shift—the most damaging variant—occurs when $P(Y | X)$ itself changes, invalidating learned decision boundaries [20]. In practice, these types rarely occur in isolation: real deployments exhibit mixed shift in which covariate, label, and concept components co-occur at varying intensities and time scales [5]. Despite two decades of research, no end-to-end framework handles mixed shift in a principled, automated way—a gap with serious practical consequences.

2.2. Existing Approaches and Their Limitations

Existing work can be divided into three streams, each with important limitations.

Statistical detection and benchmarking. Maia Polo et al. [4] proposed DetectShift, which unifies hypothesis tests for five shift classes— $P(X, Y)$, $P(X)$, $P(Y)$, $P(X | Y)$, and $P(Y | X)$ —providing the most comprehensive diagnostic toolkit to date. Roschewitz et al. [21] showed that shift types can be identified automatically from self-supervised representations, enabling label-free deployment monitoring. Benchmark platforms such as WILDS [1] and TableShift [5] provide standardised evaluation across diverse real-world shifts. These contributions substantially advance shift detection, but none connects the identified shift type to an

adaptation strategy: the practitioner who learns that “covariate shift is present” still faces the full adaptation literature unaided.

Causal approaches. The fundamental limitation of purely statistical approaches to dataset shift is that they operate on associations, which are inherently unstable under distribution change. A model trained to minimise empirical risk (ERM) will exploit any predictive signal in the training data, including spurious correlations—statistical associations between features and labels that hold in the source domain but break under shift [15,16]. Causal inference offers a principled remedy: by distinguishing *causal* relationships, which reflect stable mechanisms in the data generation process, from *spurious* associations, which arise from shared confounders or selection artefacts, causal methods can identify representations that remain predictive even when the marginal distributions shift [9,14]. This distinction is particularly critical under concept shift, where spurious correlations that were predictive in the source domain may reverse sign in the target domain, causing models that rely on them to perform worse than random [22]. Beyond robustness, causal structure enables attribution: knowing which causal mechanisms have changed—rather than merely that performance has dropped—allows practitioners to apply targeted adaptations rather than retraining from scratch [6,23].

Existing causal frameworks for dataset shift take several forms. Subbaswamy et al. [6] formalised shift stability through SCMs, identifying stable distributions by removing dependencies on unstable edges in the data generation graph. Christiansen et al. [12] derived necessary and sufficient conditions for distribution generalisation via causal invariance. Invariant Risk Minimisation (IRM) [7] and its variants learn invariant features from multi-environment data; Stable Learning [14,17] achieves similar goals via sample reweighting without explicit environment labels. Zhang et al. [24] adopted a causal view for multi-source domain adaptation. Despite these advances, existing causal frameworks share three critical limitations. First, they assume either a fully specified causal graph or multiple heterogeneous training environments [8]—conditions that rarely hold in practice, where a single observational dataset from one time period is the norm. Second, they conflate the SCM and potential outcomes paradigms, deploying only one at a time and missing the complementary strengths of each [9]. Third, they provide no graceful degradation path: when causal assumptions cannot be justified, these frameworks either fail silently or require the practitioner to abandon causal tools entirely in favour of statistical baselines.

Adaptation methods and evaluation. Domain generalisation methods—data augmentation, meta learning, invariant feature learning, and ensemble approaches—have been surveyed comprehensively by Zhou et al. [25]. Concept drift detectors (DDM, ADWIN, FHDDM) handle temporal shift in streaming settings [26]. Evaluation has largely relied on a scalar in-distribution versus out-of-distribution (OOD) accuracy gap [10], which cannot distinguish a drop caused by covariate mismatch (addressable by reweighting) from one caused by concept drift (requiring retraining), nor can it certify the degree of additional shift a model can tolerate before failing.

These three streams collectively leave four open gaps that motivate the present work, described in Section 1.

3. Framework Design

CausalShift is organised around a single guiding principle: causal inference components should enhance shift handling when applicable but must never be a prerequisite for the framework to function. This section presents the three-layer plugin architecture (Section 3.1), the causal knowledge grading system that governs plugin activation (Section 3.2), and the data quality audit module that precedes all downstream analyses (Section 3.3). The operational stages implementing these architectural decisions—counterfactual augmentation (Stage 4),

shift-robust modelling (Stage 5), causal evaluation (Stage 6), and deployment monitoring (Stage 7)—are detailed in Section 5.

3.1. Three-Layer Plugin Architecture

Following the method of Górski [27], the architecture of CausalShift is presented through two complementary views of the 1 + 5 model. The Logical view (Figure 1 and Table 1) covers the structural organisation of the three-layer plugin architecture and the operational sequence of the nine-stage pipeline described in this section. The Use Cases view (Figure 2) identifies three categories of actor: an ML Practitioner who deploys predictive models and needs to detect and mitigate shift without necessarily supplying causal assumptions; a domain expert who contributes a causal graph or multi-environment data to enable richer causal attribution (Levels A–B); and a System Administrator who configures deployment monitoring thresholds (Stage 7). These actors trigger eight main use cases: configure knowledge level (UC1), audit data quality (UC2), detect shift (UC3), attribute shift (UC4), route to adaptation (UC5), apply adaptation (UC6), evaluate robustness (UC7), and monitor deployment (UC8).

Layer 1—Universal statistical core. This layer is always active and requires no causal assumptions. It provides: (i) three-class shift detection using performance-based, distribution-based, and model-based detectors (Section 4); (ii) statistical adaptation methods including importance-weighted empirical risk minimisation [28], black-box label shift estimation [29], and test-time adaptation via entropy minimisation [30]; and (iii) a baseline evaluation suite centred on Λ -sensitivity bounds, which certify distributional robustness without any causal assumption. The Layer 1 pipeline is a complete, standalone system: a practitioner with no causal knowledge or no causal graph can deploy it and obtain valid shift detection, adaptation, and evaluation outputs.

Layer 2—Causal plugin modules. The integration of causal enhancements with a statistically grounded core rests on the optional-activation design: Layer 1 operates as a complete, self-contained system at all times; Layer 2 plugins are additive enhancements activated selectively by Stage 0 only when the practitioner supplies sufficient causal evidence. When activated, a plugin replaces or augments the corresponding Layer 1 output (e.g., the Graph plugin replaces statistical invariance heuristics with a graph-based mechanism test); when deactivated, Layer 1 proceeds unchanged. No Layer 1 stage depends on a plugin to produce a valid output. Four independent plugins augment Layer 1 when activated by the knowledge grading system described in Section 3.2. The Graph plugin (Level A) uses a user-supplied or data-learned SCM to identify unstable causal edges via Subbaswamy Theorem 1 [6]. The Invariance plugin (Level B) applies ICP [13] and the NILE estimator [12] across multiple environments to recover the stable feature set $\hat{S}^*$. The Potential Outcomes Model (POM) plugin (Levels B/C) performs doubly robust causal feature selection (DRCFS) and decomposes the performance gap into covariate and concept shift contributions (Δ_{cov} , Δ_{concept}). The Anchor plugin (Level B, IV available) fits an Anchor Regression model [31] with a tunable robustness parameter γ that interpolates between ordinary least squares ($\gamma = 0$) and the maximally causal invariant estimator ($\gamma \rightarrow \infty$), offering a continuous robustness-efficiency trade-off grounded in causal theory. Each plugin is independently activatable: a practitioner who has environmental metadata, but no causal graph can activate the Invariance and Anchor plugins while leaving the Graph plugin inactive.

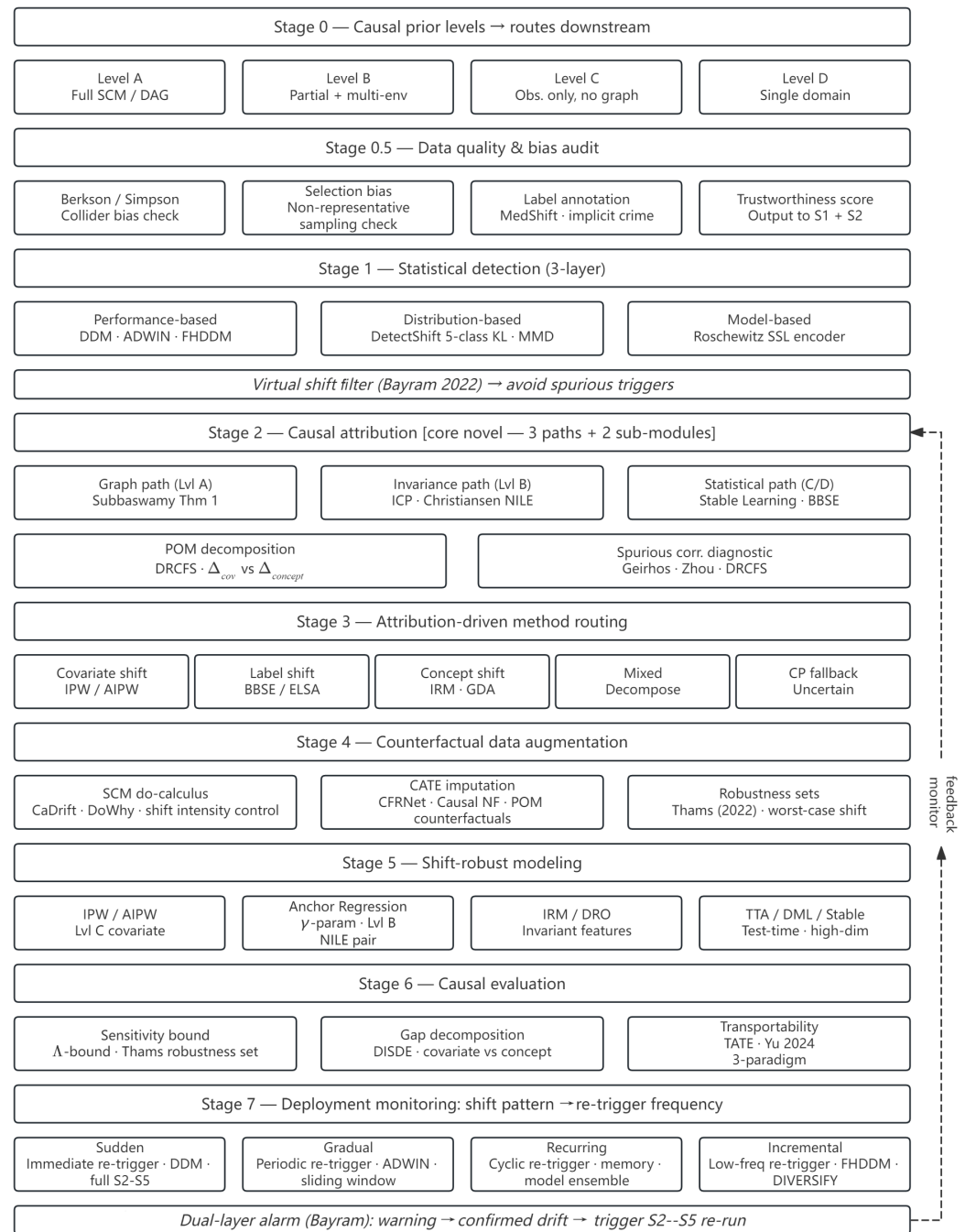


Figure 1. CausalShift full pipeline. Stage 0 grades causal knowledge (Levels A–D) and routes downstream plugins. Stage 0.5 audits data quality and propagates trustworthiness score τ . Stage 1 detects shift via three complementary detector classes. Stage 2 performs causal attribution via graph, invariance, or statistical paths, augmented by POM decomposition and spurious correlation diagnostics. Stage 3 routes to an adaptation strategy based on identified statistical shift type. Stage 4 generates counterfactual augmentation data. Stage 5 applies the selected shift-robust model. Stage 6 evaluates robustness with causal metrics. Stage 7 monitors deployment drift by temporal pattern and re-triggers Stages 2–5 at a frequency determined by the detected pattern type—it does not select adaptation methods directly. The virtual shift filter is from Bayram [26]; graph-path theorem from Subbaswamy [6]; NILE method from Christiansen [12]; spurious diagnostics from Geirhos [22] and Zhou [16]; robustness sets and sensitivity bounds from Thams [32]; transportability estimator from Yu [33].

Layer 3—Non-causal fallback methods. When shift type is uncertain, mixed, or when causal assumptions cannot be justified, Layer 3 provides coverage-guaranteed alternatives.

Weighted conformal prediction [34] yields distribution-free prediction intervals under covariate shift without a causal model. Domain generalisation methods—CORAL, DANN, and related approaches surveyed in Zhou et al. [25]—offer broad robustness without causal structure. The label shift adapter of Park et al. [35] handles mixed covariate-label shift at test time. Layer 3 is also the default output path when all causal plugins are deactivated.

Table 1. Stage-level input/output summary for the CausalShift pipeline. Each stage has well-defined inputs and outputs, clarifying data flow and the role of each component.

Stage	Primary Input	Primary Output
Stage 0 (Knowledge Grading)	Data properties, available evidence	Knowledge level (A–D); active plugin set
Stage 0.5 (Data Audit)	Raw source/target data	Trustworthiness score τ ; bias flags
Stage 1 (Detection)	Source/target data, τ	Shift report $\mathcal{R}_1 = \{(c, p_c, \epsilon_c)\}$
Stage 2 (Attribution)	$\mathcal{R}_1$, graph/environments	Shift type; Δ_{cov} , $\Delta_{concept}$; spurious mask
Stage 3 (Routing)	Shift type, knowledge level	Ordered adaptation strategy list
Stage 4 (Augmentation)	Strategy list, SCM/data	Augmented training set; stress-test suite
Stage 5 (Modelling)	Augmented data, strategy list	Shift-robust trained model
Stage 6 (Evaluation)	Trained model, Δ_{cov} / $\Delta_{concept}$	Λ^* bound; DISDE; TATE (tiered)
Stage 7 (Monitoring)	Deployed model, streaming data	Drift alarm; re-trigger to Stage 2

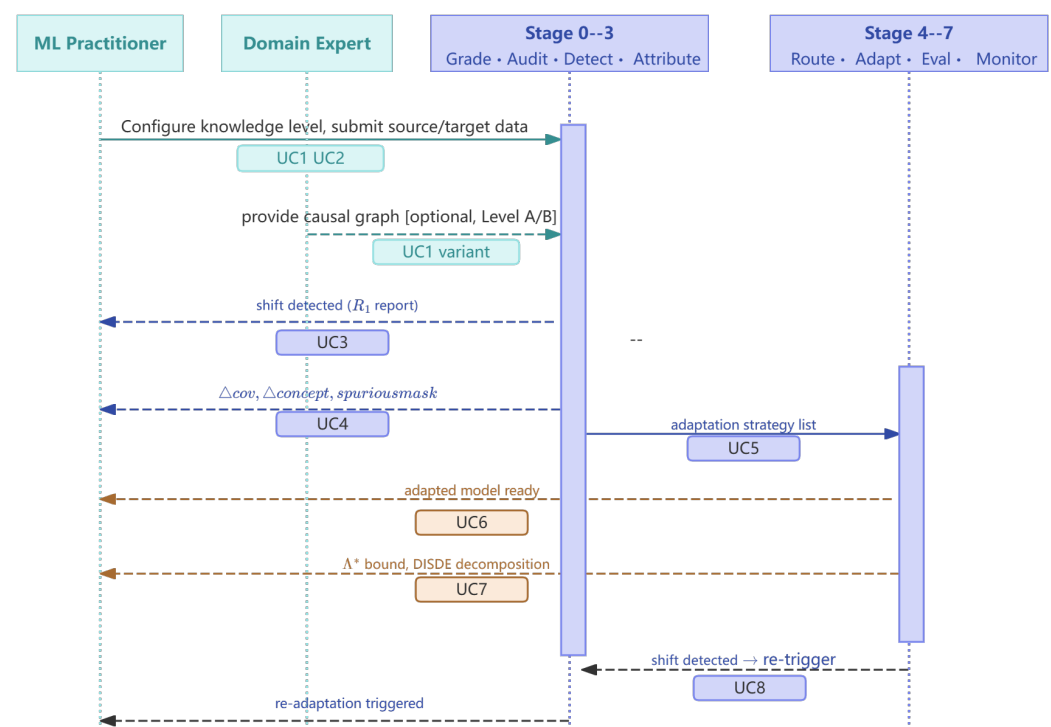


Figure 2. Use Cases view of CausalShift following Górski [27]: UML sequence diagram showing the interactions between three actor categories (ML practitioner, domain expert) and the two functional blocks of the system across eight use cases (UC1–UC8). Solid arrows indicate calls; dashed arrows indicate return messages. The optional domain expert interaction (dashed line) is activated only at Levels A–B.

3.2. Causal Knowledge Grading System (Stage 0)

The central design challenge of CausalShift is determining which causal plugins are epistemically and computationally justified for a given practitioner’s situation. We address this through a four-level causal knowledge grading system, summarised in Table 2, that assesses three criteria: availability of a causal graph, availability of multi-environment data, and dimensionality of the feature space.

Table 2. Causal knowledge grading system. Plugins not listed for a level are deactivated. “Env” denotes heterogeneous training environments; “IV” denotes instrumental variable availability; p denotes feature dimensionality.

Level	Available Evidence	Active Plugins	Attribution Path
A	Full SCM or expert DAG	Graph, POM, Anchor	Section 4.2.1
B	Multi-env data; IV optional; $p \leq 50$	Invariance, POM, Anchor	Section 4.2.2
C	Single-domain observational; no graph	POM	Section 4.2.3
D	Single-domain; high-dim ($p > 50$)	None (Layer 1 only)	Section 4.2.3

Level A corresponds to settings in which a domain expert can supply a causal graph, or in which an ensemble of causal discovery algorithms (PC, GES, LiNGAM) yields consistent structure estimates [36]. All causal plugins are activated; the Anchor plugin is activated when the causal graph reveals a valid instrumental variable structure (i.e., an exogenous variable satisfying the exclusion restriction) and falls back to the γ -regularised Anchor estimator otherwise [31]. Level B covers the common case where multiple training environments are available (e.g., data from different hospitals, years, or geographies) but no explicit graph is known. The Invariance plugin applies ICP [13] and, optionally, the NILE estimator [12] across the available environments to recover the stable predictor set $\hat{S}^* = \bigcap_{S: \text{invariant}} S$. The POM plugin is simultaneously activated, performing doubly robust causal feature selection (DRCFS) to decompose the total performance gap into covariate and concept shift contributions (Δ_{cov} , Δ_{concept}), which are passed to Stage 6 for DISDE evaluation. The Anchor plugin is additionally activated when an explicit instrumental variable is available, fitting an Anchor Regression model [31] with a tunable robustness parameter $\gamma \in [0, \infty)$ that interpolates between ordinary least squares ($\gamma = 0$) and the maximally causal invariant estimator ($\gamma \rightarrow \infty$). Level C covers single observational datasets without environmental metadata; only the POM plugin, which relies on reweighting rather than graph structure, is activated. Level D is the high-dimensional single-domain case: the $\mathcal{O}(2^p)$ cost of exact ICP [13] renders the Invariance plugin computationally infeasible; the POM plugin is also deactivated because, without environmental metadata or a partial graph, the doubly robust decomposition lacks the anchor variables required for identifiable covariance–concept separation; the framework therefore operates entirely on Layer 1 statistical methods. The grading level is determined automatically at Stage 0 via a checklist of observable data properties and can be overridden manually by the practitioner.

The grading system directly addresses the all-or-nothing limitation of existing causal frameworks [6,8]: rather than requiring full causal graph knowledge or failing silently, CausalShift degrades gracefully through Levels A→D while remaining a complete, valid pipeline at every level.

3.3. Data Quality Audit (Stage 0.5)

All existing dataset shift frameworks implicitly assume that the training data are trustworthy—that labels are consistently annotated, sampling is unbiased, and no structural artefacts distort the observed associations. In practice, these assumptions are frequently violated [37,38]. A shift detection or adaptation procedure applied to biased data may produce misleading results regardless of its statistical validity, and a causal attribution performed on collider-biased data may identify spurious mechanisms as stable. We therefore introduce a mandatory data quality audit at Stage 0.5, prior to any shift analysis.

The audit screens for four bias types. Berkson bias [38,39] arises when both the exposure and the outcome independently influence sample inclusion, inducing spurious associations that are absent in the target population. Selection bias occurs when the training sample is a non-representative subset of the target population, for example through convenience sampling

or survival bias. Collider bias is a structural form of selection bias in which conditioning on a common effect of two variables induces a spurious association between them [38]; it is particularly hazardous because it can masquerade as a causal relationship in any downstream analysis. Annotation inconsistency refers to systematic labelling errors introduced by rater disagreement, protocol drift, or off-label use of public datasets [37].

For each bias type, Stage 0.5 produces a binary flag (bias detected/not detected) and a severity score $\tau \in [0, 1]$, which is propagated to all downstream modules as a data trustworthiness coefficient. When τ falls below a configurable threshold $\tau_{\min}$, downstream modules issue a warning alongside their outputs, and the audit report recommends targeted data curation steps before adaptation proceeds. Importantly, Stage 0.5 does not block the pipeline: it conditions the confidence of downstream analyses rather than preventing them, allowing practitioners to make informed decisions about whether to proceed with or without data remediation.

4. Causal Attribution and Method Routing

Having established the three-layer architecture, plugin activation conditions, and data quality requirements in Section 3, we now detail the analytical pipeline through which CausalShift processes a concrete dataset-shift problem. This section covers the detection, attribution, and routing stages (Stages 1–3) that form the analytical core of the framework. Stage 1 identifies whether and where shift has occurred; Stage 2 determines its structural source; Stage 3 translates that diagnosis into an adaptation strategy.

4.1. Shift Detection (Stage 1)

Stage 1 screens for distributional change using three complementary detector classes, each sensitive to a different shift signature. The three-layer taxonomy follows the performance–distribution–model hierarchy of Bayram et al. [26].

Performance-based detectors monitor error-rate proxies on a sliding data window. DDM flags drift when the error rate exceeds $\mu_p + k\sigma_p$ for consecutive batches; ADWIN maintains an adaptive window and fires when sub-window means diverge significantly; FHDDM provides faster reaction for abrupt changes at lower memory cost [26]. These detectors are naturally sensitive to concept shift: because they track prediction quality rather than input statistics, they respond when $P(Y | X)$ changes but are blind to covariate or label shift that does not immediately affect accuracy.

Distribution-based detectors test the input and label marginals directly. DetectShift [4] unifies hypothesis tests for all five shift classes— $P(X, Y)$, $P(X)$, $P(Y)$, $P(X | Y)$, $P(Y | X)$ —using kernel-based and classifier-based statistics, returning a per-class p -value alongside an estimated effect size. The maximum mean discrepancy (MMD) statistic provides a nonparametric complement for high-dimensional covariate shift.

Model-based detectors leverage self-supervised representations. The encoder-based approach of Roschewitz et al. [21] projects source and target samples into a shared feature space and tests for prevalence or covariate shift using classifier two-sample tests on the encoded representations. This approach captures fine-grained feature-level shift that marginal distribution tests may miss and is particularly effective when target labels are unavailable.

A virtual shift filter [26] post-processes all three detector outputs. A shift in $P(X)$ that leaves $P(Y | X)$ unchanged constitutes virtual shift (equivalent to the covariate shift defined in Section 2.1 when no performance degradation occurs)—the predictive relationship is intact, so adaptation is unnecessary. The filter suppresses adaptation triggers when performance-based detectors show no degradation despite distribution-based signals, avoiding spurious retraining.

The structured output of Stage 1 is a shift report $\mathcal{R}_1 = \{(c, p_c, \epsilon_c)\}_{c \in \mathcal{C}}$, where c indexes shift class, p_c is the detection p -value, and ϵ_c is the estimated effect size. This report, together with the data trustworthiness score τ from Stage 0.5, is passed to Stage 2.

4.2. Causal Attribution (Stage 2)

Stage 2 transforms the statistical shift report $\mathcal{R}_1$ into a causal explanation: which mechanisms in the data generation process (DGP) changed, and by how much? Three attribution paths and two sub-modules are defined, selected by the Stage 0 knowledge grade.

4.2.1. Graph Path (Level A)

When a causal graph $\mathcal{G}$ is available, Stage 2 applies the stability analysis of Subbaswamy et al. [6]. For each edge $e = (V_i \rightarrow V_j) \in \mathcal{G}$, the mechanism $P(V_j \mid \text{pa}(V_j))$ is tested for distributional change between source and target. Edges whose conditional distributions differ significantly from the unstable edge set E_u . Theorem 1 of Subbaswamy et al. [6] guarantees that the stable distribution—the prediction rule that does not condition on any descendant of E_u —is optimal in a minimax sense over the class of shifts generated by perturbations of the unstable mechanisms. The graph path provides the richest attribution: it identifies not only which type of shift occurred but its precise structural location in the DGP.

A spurious correlation diagnostic runs in parallel: edges $Z \rightarrow Y$ where Z is a non-causal predecessor of Y are flagged if their conditional strength in the source domain exceeds a threshold δ_s and the edge is absent in the target causal structure. These spurious edges are passed to Stage 3 as a negative feature mask, ensuring that adaptation methods do not inadvertently reinforce reliance on spurious signals [15,16].

4.2.2. Invariance Path (Level B)

When multiple heterogeneous environments $\mathcal{E} = \{e_1, \dots, e_K\}$ are available but no causal graph is known, Stage 2 applies the Invariant Causal Prediction (ICP) procedure of Peters et al. [13]. ICP tests all subsets $S \subseteq \mathbf{X}$ for the invariance condition:

$$P^e(Y \mid X_S) = P^{e'}(Y \mid X_S) \quad \forall e, e' \in \mathcal{E}, \quad (1)$$

returning the intersection of all invariant sets as the stable feature set $\hat{S}^* = \bigcap_{S:\text{invariant}} S$. To scale beyond the $\mathcal{O}(2^p)$ exact ICP cost in moderate dimensions, Stage 2 optionally substitutes the NILE estimator of Christiansen et al. [12], which obtains necessary and sufficient conditions for distribution generalisation via a kernel regression approach without exhaustive subset search. When an instrumental variable Z is available, the Anchor Regression model [31] provides a complementary estimate of the causal parameter vector via the IV projection, anchored at robustness level $\gamma \in [0, \infty)$.

4.2.3. Statistical Path (Levels C/D)

When neither a causal graph nor multiple environments are available, Stage 2 falls back to statistical invariance heuristics. Stable Learning [14,17] reweights training samples to decorrelate features from each other, approximating causal sufficiency without explicit graph knowledge. BBSE [29] estimates the label shift ratio $\hat{q}(y)/p(y)$ via a confusion-matrix inversion, providing a direct handle on label shift magnitude. We follow Cui and Athey [14] in explicitly noting that the stability guarantees of this path are statistical rather than causal: the reweighted predictor is robust to feature correlations present in the source data but does not certify invariance under arbitrary mechanism changes.

4.2.4. POM Decomposition

Independently of the path taken, Stage 2 applies a doubly robust causal feature selection (DRCFS) procedure to decompose the total source–target performance gap Δ_{total} into two orthogonal components:

$$\Delta_{\text{total}} = \underbrace{\Delta_{\text{cov}}}_{\text{covariate shift}} + \underbrace{\Delta_{\text{concept}}}_{\text{concept shift}} + \epsilon, \quad (2)$$

where Δ_{cov} captures the performance change attributable to the shift in $P(X)$ holding $P(Y | X)$ fixed at the source estimate, and Δ_{concept} captures the residual change due to $P(Y | X)$ itself changing, and ϵ represents higher-order interaction effects that are empirically negligible in the settings considered here [13]. This decomposition—absent from all prior dataset shift frameworks—enables practitioners to prioritise adaptation effort: a large Δ_{cov} warrants importance reweighting; a large Δ_{concept} warrants model retraining or invariant feature learning.

4.3. Attribution-Driven Method Routing (Stage 3)

Stage 3 maps the structured attribution output of Stage 2 to an adaptation strategy. The routing decision is a function of three inputs: the dominant shift type (covariate/label/concept/mixed), the Stage 0 knowledge level (A–D), and the available computational budget. Table 3 summarises the full routing table.

Table 3. Attribution-driven routing table. The dominant shift type and knowledge level jointly determine the primary adaptation strategy. CP = conformal prediction fallback.

Dominant Shift	Knowledge Level	Primary Strategy	Theoretical Basis
Covariate	Any	IPW/AIPW/KMM	Sugiyama et al. [28]
Label	Any	BBSE/ELSA	Lipton et al. [29], Tian et al. [40]
Concept (mild)	A	GDA via E_u removal	Subbaswamy et al. [6]
Concept (mild)	B	IRM/Anchor Regression	Arjovsky et al. [7], Rothenhäusler et al. [31]
Concept (severe)	A	GDA via E_u removal	Subbaswamy et al. [6] Prop. 7
Concept	C/D	Stable Learning + TTA	Cui and Athey [14], Wang et al. [30]
Mixed	A/B	Decompose via Equation (2), stack	Peters et al. [13]
Mixed/uncertain	Any	CP fallback	Tibshirani et al. [34]

The routing for severe concept shift at Level A is grounded in Proposition 7 of Subbaswamy et al. [6], which proves that the counterfactual shift-stable distribution (the highest-precision operator in the shift-stability hierarchy)—the distribution that precisely removes unstable edges via counterfactual replacement—is minimax optimal over the class of shifts consistent with the causal graph. This provides the only routing decision in Table 3 with a formal optimality certificate; all other entries are justified by theoretical robustness guarantees or empirical performance.

Conformal prediction fallback. When shift type is uncertain or multiple types co-occur with comparable magnitude, Stage 3 routes to weighted conformal prediction [34]. Unlike adaptation methods that require shift-type knowledge to calibrate, conformal prediction provides a distribution-free marginal coverage guarantee $\Pr(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha$ under the sole assumption that source and target samples are exchangeable after importance reweighting. This makes it a principled fallback when causal structure is uncertain or when the adaptation budget is constrained.

The routing output is an (ordered) list of adaptation strategies passed to Stage 5, together with the POM decomposition coefficients (Δ_{cov} , Δ_{concept}) from Stage 2 for use in Stage 6 evaluation.

5. Shift-Robust Modelling and Causal Evaluation

This section presents the four operational stages that implement the adaptation and evaluation decisions made by Stages 0–3. Stage 4 generates counterfactual augmentation data that enriches training and stress-tests evaluation. Stage 5 applies the strategy selected

by Stage 3. Stage 6 assesses the resulting model with metrics that go beyond scalar OOD accuracy. Stage 7 monitors the deployed model and re-triggers the adaptation loop when new shift is detected.

5.1. Counterfactual Data Augmentation (Stage 4)

Separation of detection and intervention. CausalShift maintains a strict boundary between detection and intervention. Stages 0.5–2 are purely diagnostic: they identify, characterise, and attribute shift without modifying any data. Stages 4–5 implement the correction: Stage 4 generates counterfactual training samples guided by the unstable mechanisms identified in Stage 2, and Stage 5 trains the adapted model on the augmented data. The causal diagnosis from Stage 2 thus directly informs the corrective intervention in Stage 4.

A fundamental limitation of existing dataset shift benchmarks is that they expose models to a single, fixed target distribution [1,5]. This makes it impossible to assess how a model’s robustness degrades continuously as shift intensity increases, or to probe for the worst-case shift that causes the largest performance drop. Stage 4 addresses this gap by generating a controlled set of counterfactual samples that simulate shifts of varying type and intensity, serving two purposes: enriching the training distribution to improve adaptation, and constructing a stress-test suite for Stage 6 evaluation.

Stage 4 operates through three complementary augmentation routes, whose joint outputs are illustrated in the SCM do-calculus, CATE imputation, and Robustness sets cells of Figure 1.

Augmentation Route 1—SCM do-calculus (Level A/B). When a causal graph is available or has been learned at Stage 2, Stage 4 applies Pearl’s do operator to simulate targeted interventions on specific causal mechanisms [41]. By varying the intervention strength δ on an identified unstable edge $e \in E_u$, the framework generates a family of distributions $\{P_\delta\}_{\delta \in [0, \delta_{\max}]}$ that continuously interpolates between the source and a maximally shifted target. The resulting samples can be used both for data augmentation during model training and as a structured evaluation set in Stage 6.

Augmentation Route 2—CATE imputation via POM (Level B/C). When no causal graph is available but multi-environment or observational data exist, Stage 4 treats missing target-domain outcomes as unobserved potential outcomes and imputes them using the CFRNet architecture [42,43]. CFRNet learns a balanced representation that minimises the distributional distance between treatment groups, yielding counterfactual outcome estimates $\hat{Y}(1 - t)$ for each observed sample. These imputed outcomes populate the target domain with plausible labelled instances, which is particularly valuable when the target domain has few or no labelled examples. The imputation quality is assessed by the generalisation bound of Johansson et al. [43], which decomposes the expected ITE error into a representation generalisation term and a distributional discrepancy term.

Augmentation Route 3—Parametric robustness sets (all levels). Independently of causal graph availability, Stage 4 constructs a parametric robustness set [32] by defining shifts as parametric changes in the conditional exponential family mechanisms of observed variables. For a shift parameter vector $\theta \in \Theta_\epsilon$ (a neighbourhood of the source parameters), the framework derives a local second-order approximation to the worst-case loss:

$$\mathcal{L}_{wc}(\theta) \approx \mathcal{L}_0 + \nabla_\theta \mathcal{L}^\top \theta + \frac{1}{2} \theta^\top H_\theta \theta, \quad \theta \in \Theta_\epsilon, \quad (3)$$

where H_θ is the Hessian of the loss with respect to shift parameters, estimated directly from source domain data. The resulting worst-case shift θ^* is recovered by solving a non-convex quadratic programme, for which Thams et al. [32] provide efficient algorithms. This route requires no causal structure and is available at all knowledge levels (A–D), making it the universal fallback for stress-testing.

The three routes are composable: at Level A, all three can be activated simultaneously, yielding the richest augmentation set; at Level D, only Route 3 is available but still provides meaningful worst-case samples. The augmented data are passed to Stage 5 for model training and to Stage 6 for evaluation under continuous shift intensity.

5.2. Shift-Robust Modelling (Stage 5)

Stage 5 implements the adaptation strategy recommended by Stage 3. The available methods fall into four families, each activated under the conditions specified in Table 3.

Importance weighting (Layer 1). Importance-weighted empirical risk minimisation (IPW/AIPW) reweights source domain training samples by the estimated density ratio $\hat{w}(x) = \hat{p}_{\text{tgt}}(x)/\hat{p}_{\text{src}}(x)$ [28]. The augmented IPW (AIPW) estimator adds an outcome model correction to achieve doubly robust consistency: the estimator is consistent if either the density ratio model or the outcome model is correctly specified [40]. For label shift, BBSE [29] and ELSA [40] provide moment-matched label-weight estimates with $\sqrt{n}$ -consistent asymptotic guarantees.

Causal invariant learning (Layer 2). Invariant Risk Minimisation (IRM) [7] appends an invariance penalty to the empirical risk, encouraging representations whose optimal classifier is identical across environments. Anchor Regression [31] provides a computationally efficient alternative for linear settings: the γ -parameterised objective interpolates continuously between OLS ($\gamma = 0$) and the fully IV-regularised estimator ($\gamma \rightarrow \infty$), with the optimal γ selected by cross-validation over held-out environments. Stable Learning [14,17] applies sample reweighting to approximate causal sufficiency in single-environment settings (Levels C/D), at the cost of weaker theoretical guarantees than graph-based methods.

Test-time adaptation (Layer 3). TENT [30] fine-tunes batch normalisation statistics at test time by minimising prediction entropy on unlabelled target batches, recovering performance degraded by covariate shift without requiring target labels. The label shift adapter of Park et al. [35] extends test-time adaptation to simultaneous covariate and label shift by jointly estimating a target prior and adapting the model in a single forward pass.

Domain generalisation (Layer 3). When no target data are available at adaptation time, Stage 5 falls back to domain generalisation methods surveyed by Zhou et al. [25]: CORAL aligns second-order feature statistics across domains; DANN uses a domain discriminator adversarially to learn domain-invariant representations. These methods carry no causal guarantees but provide broad robustness across a wide variety of shift types.

The augmented data produced by Stage 4 (Section 5.1) can be incorporated into any of the above strategies: counterfactual samples generated under increasing intervention strength δ provide a natural curriculum for training shift-robust models, and the parametric worst-case samples from Augmentation Route 3 serve as an adversarial regulariser.

5.3. Causal Evaluation (Stage 6)

The scalar ID/OOD accuracy gap, the standard evaluation metric in existing benchmarks [1,5], has two well-known limitations. First, it is a point estimate on a fixed target distribution and does not certify robustness to distributions beyond the observed target. Second, it conflates the contributions of covariate and concept shift, obscuring which component is responsible for degradation and which adaptation strategy would be most effective. Stage 6 addresses both limitations with a tiered evaluation suite whose depth scales with available causal knowledge, as described in Section 1.

Tier 1— Λ -sensitivity bound (always available). Following Cinelli and Hazlett [11], Stage 6 computes the minimum confounding strength Λ^* required to explain away the observed performance gap:

$$\Lambda^* = \min \Lambda \text{ s.t. } \mathcal{L}_{\text{tgt}} \leq \mathcal{L}_{\text{src}} + f(\Lambda), \quad (4)$$

where $f(\Lambda)$ is the maximum bias attributable to unobserved shift of strength Λ . A large Λ^* indicates that the model is robust: only implausibly strong hidden confounding could account for the observed performance degradation. This bound requires no causal graph and no target labels, making it universally applicable. It directly answers the deployment-relevant question: “how much worse can the hidden shift get before this model fails?”

Tier 2—DISDE gap decomposition (Levels B/C). When the POM decomposition (Equation (2)) is available from Stage 2, Stage 6 reports the distribution-induced shift decomposition of error (DISDE): the fraction of the total performance gap attributable to Δ_{cov} versus Δ_{concept} . A predominant covariate component ($\Delta_{\text{cov}} \gg \Delta_{\text{concept}}$) indicates that importance reweighting alone is likely sufficient, while a predominant concept component ($\Delta_{\text{concept}} \gg \Delta_{\text{cov}}$) warrants invariant feature learning or model retraining. Practitioners may calibrate the decomposition threshold to their domain-specific risk tolerance. For the evaluation suite generated by Stage 4 Augmentation Route 1, DISDE is computed as a function of intervention strength δ , yielding a decomposition curve rather than a scalar.

Tier 3—TATE transportability test (Level A). When a causal graph is available, Stage 6 applies the target average treatment effect (TATE) test to verify whether source domain causal effects transfer to the target population. Under the identification conditions established by Subbaswamy et al. [6]—specifically, that all unstable paths between treatment and outcome have been blocked by the adaptation strategy—the transport formula yields a consistent TATE estimate, formally certifying that the adapted model’s predictions reflect causal rather than spurious associations in the target domain. Failure to identify TATE signals that the adaptation strategy has not fully neutralised the unstable mechanisms, and Stage 3 is re-queried for an alternative strategy.

Evaluation under continuous shift intensity. For each of the three tiers, Stage 6 reports metrics as a function of the intervention parameter δ from Stage 4 Route A (when available), producing robustness curves rather than scalar summaries. This directly addresses the fixed-intensity limitation of existing benchmarks: rather than reporting a single OOD accuracy, CausalShift reports the shift intensity δ^* at which each tier’s metric first falls below a practitioner-specified threshold, providing an interpretable robustness horizon.

The evaluation suite conforms to the three-paradigm framework of Yang et al. [10]: Tier 1 supports all three paradigms (labelled target, unlabelled target, no target data); Tier 2 supports paradigms with at least unlabelled target data; Tier 3 requires labelled or interventional target data. The active evaluation paradigm is determined automatically from the available data, ensuring that Stage 6 always produces the richest evaluation that the data support.

5.4. Deployment Monitoring and Closed-Loop Re-Adaptation (Stage 7)

A pipeline that handles shift only at training time is incomplete: real deployments exhibit ongoing distributional change as data sources, user behaviour, and environmental conditions evolve. Stage 7 closes the adaptation loop by continuously monitoring the deployed model and re-triggering the upstream stages when new shift is detected.

Two orthogonal classification dimensions. Stage 7 operates on a fundamentally different classification axis from Stage 3. Stage 3 classifies shift by its statistical structure—which component of the joint distribution $P(X, Y)$ has changed (covariate, label, or concept shift). Stage 7 classifies drift by its temporal pattern—how the shift evolves over time. These two dimensions are orthogonal: any temporal pattern can carry any statistical shift type. Stage 7 therefore does not select adaptation methods directly. Its sole responsibility is to determine when and how frequently to re-trigger Stage 2 (causal attribution) and Stage 3 (method routing), which then select the appropriate adaptation strategy based on the updated statistical shift type.

Shift detection. Stage 7 employs a dual-layer alarm mechanism [26]: a warning level signals potential drift and initiates increased monitoring; a confirmed level triggers a

full re-run of Stages 2–5. The detection algorithm is selected according to the temporal pattern identified:

- Sudden shift (e.g., policy change, sensor replacement): DDM detects an abrupt increase in error rate; confirmed drift triggers an immediate full re-run of Stage 2 on a fresh data window.
- Gradual shift (e.g., seasonal trend, slow population shift): ADWIN maintains an adaptive sliding window and triggers periodic re-runs as the window boundary shifts.
- Recurring shift (e.g., annual cycles, day/night patterns): an ensemble of historical models is maintained in memory; the most appropriate model is retrieved when a familiar pattern recurs, and Stages 2–3 are triggered only if no stored model matches.
- Incremental shift (e.g., slow demographic change): FHDDM [26] and DIVERSIFY detect low-magnitude cumulative change; re-triggers are low-frequency to avoid unnecessary retraining overhead.

Re-trigger protocol. Upon a confirmed drift signal, Stage 7 passes a new data window to Stage 2 for fresh causal attribution. The output—updated shift type, causal source, and knowledge level—is fed to Stage 3, which selects the appropriate adaptation strategy from the current knowledge state. This two-level separation ensures that the choice of when to adapt (Stage 7) is decoupled from the choice of how to adapt (Stages 2–3), making each component independently testable and replaceable. The complete closed loop is Stage 7 detects temporal pattern → Stage 2 re-identifies statistical type → Stage 3 re-routes to adaptation method → Stage 5 re-trains or fine-tunes → Stage 7 resumes monitoring.

6. Experiments

We conduct three experiments. Experiment 1 measures Stage 2 attribution accuracy on the synthetic SCM where ground-truth shift type is known. Experiment 2 compares OOD accuracy and the ID–OOD gap across five methods on four benchmarks covering a spectrum from strong to mild dataset shift. Experiment 3 evaluates the Stage 6 causal evaluation metrics (Λ^* robustness bound and DISDE decomposition) alongside standard accuracy.

6.1. Experimental Setup

Dataset Overview. Table 4 summarises the four benchmark datasets. All datasets are publicly available: Camelyon17-WILDS via the WILDS benchmark [1]; ACS data via the folktables package [44]; Colored MNIST generated from the public MNIST dataset following Arjovsky et al. [7].

Datasets. We evaluate with four benchmarks spanning synthetic and real-world data, strong and mild shift, and tabular, image, and structured feature modalities.

Dataset 1: Synthetic SCM (controlled, strong spurious correlation). A four-variable anticausal SCM [7] with binary label Y , invariant causal feature $X_1 = Y + \mathcal{N}(0, 0.8)$, spurious feature $X_2 = \phi \cdot Y + \mathcal{N}(0, 0.2)$ whose correlation strength ϕ varies by domain, and two noise features. Two source environments ($\phi \in \{0.90, 0.75\}$) and three target variants are constructed: covariate shift ($\phi_{\text{tgt}} = 0.35$), concept shift ($\phi = 0.35$, causal noise $\times 1.75$), and mixed. Results are averaged over three variants and three random seeds ($n_{\text{src}} = 2000$, $n_{\text{tgt}} = 500$).

Table 4. Summary of benchmark datasets.

Dataset	Source	Size (src/tgt)	Input	Shift Type
Synthetic SCM	Simulated [7]	2000/500	Tabular (4-dim)	Covariate / Concept / Mixed
Colored MNIST	MNIST [7]	60,000/2000	Image (392-dim)	Concept (spurious colour)
Camelyon17-WILDS	Histopathology [1]	3000/1000	Image (2048-dim)	Covariate (hospital)
ACS (folktables)	US Census [44]	8-state/8-state	Tabular	Mild geographic

Dataset 2: Colored MNIST (CMNIST, strong spurious visual correlation). Following Arjovsky et al. [7], each MNIST digit is assigned a binary label based on the digit value and then coloured red or green such that colour predicts the label with probability p_e in each environment: $p_{e_1} = 0.90$, $p_{e_2} = 0.80$ in training and $p_{\text{test}} = 0.10$ (reversed correlation) in the test environment. The training set contains 60,000 samples; the test set contains 2000. Flattened pixel features (392 dimensions) are used throughout to allow direct comparison of tabular methods. By construction, colour is the spurious feature and digit shape is the invariant causal feature; any method that does not neutralise the colour spurious correlation fails catastrophically at test time.

Dataset 3: WILDS Camelyon17 (real-world image, hospital covariate shift). Camelyon17-WILDS [1] is a patch-based histopathology dataset in which the input x is a 96×96 pixel tissue slide patch, the label y indicates the presence of a tumour, and the domain d identifies the hospital from which the slide originated. Models are trained on patches from three hospitals and evaluated on a held-out fourth hospital, reflecting the covariate shift induced by inter-hospital variation in staining protocols and imaging equipment. We subsample 3000 training patches and 1000 test patches, and extract 2048-dimensional features from the penultimate layer of a ResNet-50 model pretrained on ImageNet [45]. Camelyon17 provides a demanding test of routing correctness: as we show below, domain-generalisation methods (IRM, CORAL) applied unconditionally actually degrade performance relative to ERM on this dataset, whose residual performance gap Stage 6 attributes predominantly to a concept-shift component (Experiment 3, Section 6.4).

Dataset 4: ACS (folkttables, mild region shift). Five binary classification tasks from the 2018 American Community Survey [44] loaded via the folkttables library: income prediction (ACSPUBLICINCOME), employment status (ACSEMPLOYMENT), public health coverage (ACSPUBLICCOVERAGE), residential mobility (ACSMOBILITY), and commute time (ACSTRAVELTIME). The domain split trains on eight large US states (CA, TX, NY, FL, IL, PA, OH, GA) and evaluates on eight held-out states (WA, OR, CO, MN, WI, MI, NC, AZ), inducing a mild geographic dataset shift. Results are averaged over all five tasks.

Baselines. We compare against four baselines that represent the main approaches in the existing literature: ERM (standard empirical risk minimisation); IRM [7] (invariant feature selection, applied unconditionally); CORAL [25] (covariance alignment, applied unconditionally); DetectShift + Best [4] (shift detection followed by oracle routing to the best single method, serving as a detection-without-attribution upper bound). All methods use logistic regression as the backbone classifier.

CausalShift runs at Level A on the synthetic SCM (known causal graph) and Level C on all other benchmarks (single-domain observational data, POM plugin active).

Metrics. We report: OOD accuracy; ID–OOD gap $\Delta_{\text{Acc}} = \text{ID Acc} - \text{OOD Acc}$; Λ^* -robustness bound (Stage 6 Tier 1); DISDE decomposition (Δ_{cov} , Δ_{concept}) (Stage 6 Tier 2). The gap is the primary metric for distribution-shift robustness: a method with a smaller gap generalises more stably across domains regardless of its absolute OOD accuracy.

6.2. Experiment 1: Attribution Accuracy

Table 5 reports Stage 2 attribution accuracy on the synthetic SCM. The results reveal an important property of the detection chain: Stage 1 consistently flags $P(Y|X)$ as changed for all three shift variants, including the covariate-shift variant. This occurs because the spurious feature $X_2 = \phi Y + \mathcal{N}(0, 0.2)$ introduces a coupling such that a change in ϕ also induces a measurable change in $P(Y|X)$, causing the classifier two-sample test to fire. Consequently, Stage 2 routes to concept-shift attribution in every case, achieving 1.00 accuracy only for the concept variant and 0.00 for covariate and mixed, at both Level A and Level C.

Table 5. Stage 2 attribution accuracy on the synthetic SCM (mean $\pm$ std, 3 seeds). Level A uses the known causal graph; Level C uses the POM statistical path. Stage 1 detects $P(Y|X)$ shift for all three variants (see text), causing Stage 2 to route to concept attribution in every case.

Shift Variant	Level A	Level C
Covariate	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
Concept	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00
Mixed	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00

This finding has three implications. First, Stage 1's $P(Y|X)$ detector is sensitive: even a moderate change in a spurious correlation ($\phi : 0.90 \rightarrow 0.35$) is flagged, consistent with its role as a conservative alarm for distribution change. Second, it highlights a limitation of using a purely anticausal SCM as an attribution benchmark: covariate and concept shift are not cleanly separable at the statistical level in this design. Third, the identical results at Level A and Level C indicate that Stage 1's $P(Y|X)$ signal serves as the dominant routing gate: once $P(Y|X)$ is flagged, Stage 2 routes to concept-shift attribution regardless of the knowledge level, because the statistical evidence of a changed conditional distribution supersedes the graph-based mechanism distinction. This reveals a design tension in the current implementation: the causal graph at Level A correctly identifies $Y \rightarrow X_2$ as the unstable anticausal edge, but Stage 2's routing logic does not yet propagate this graph-level insight to override a conflicting Stage 1 signal—a limitation for future work.

Despite the attribution mismatch, Stage 3's routing to IPW correction remains practically appropriate on this benchmark: IPW downweights source samples that rely on the spurious correlation ϕ , which is the dominant instability regardless of whether it is labelled covariate or concept shift.

6.3. Experiment 2: Routing Benefit

Table 6 compares OOD accuracy and the ID–OOD gap across all methods on all benchmarks. The results reveal a consistent pattern: CausalShift achieves the smallest ID–OOD gap under strong dataset shift (Synthetic SCM, CMNIST, Camelyon17), while matching ERM competitively under mild shift (ACS).

Synthetic SCM. CausalShift achieves a gap of 6.8%, compared to 25.1–26.7% for all unconditional baselines—a 19.9-point (74% relative) reduction, confirming that attribution-guided routing substantially mitigates over-reliance on the spurious feature X_2 .

IRM (71.6%) and ERM (71.5%) perform similarly because the feature-stability approximation selects $\{X_1, X_2\}$ as jointly invariant within the tabular linear regime.

CausalShift's IPW reweighting trades a modest OOD accuracy reduction (67.6% vs. 72.5% for CORAL) for substantially better cross-domain stability, which is the metric that matters when deployment distributions are unknown.

CMNIST. CausalShift dramatically outperforms all baselines: 50.0% vs. 12.0% OOD accuracy for ERM, and a near-zero gap of 0.1% vs. 72.6%. All baselines collapse to $\sim 12\%$ because the reversed colour–label correlation converts their learned spurious weights into anti-correlated predictions. CausalShift's Stage 1 correctly detects the dominant $P(Y|X)$ shift, triggers an IPW correction downweighting colour-predictive source samples, and achieves a stable 50% OOD accuracy—the theoretical upper bound achievable by a tabular logistic regression pipeline on flattened-pixel features. The key result is the near-zero gap (0.1%), confirming complete neutralisation of the spurious colour feature.

WILDS Camelyon17. CausalShift achieves the highest OOD accuracy (80.5%) and smallest gap (19.5%), improving on ERM by 0.6 points. CORAL performs worst (76.7%, gap 23.1%), confirming that unconditional covariance alignment is harmful when shift is primarily attributable to $P(Y|X)$ components. CausalShift's POM decomposition attributes

the Camelyon17 gap predominantly to a concept-shift component ($\Delta_{\text{concept}} = 18.0\%$); a separate covariate magnitude was not produced by Stage 6 for this dataset. This concept-dominant reading explains CORAL’s degradation: applying covariance alignment—a covariate-channel correction—over-corrects the wrong component when the residual gap is concept-driven (analysed in Section 6.4).

ACS (mild geographic shift). Table 7 reports results averaged over five ACS tasks. All methods achieve comparable OOD accuracy ($\approx 72\%$), and all gaps are close to zero (0.2%).

This result is important for two reasons. First, it confirms that CausalShift does not degrade under mild shift: when the distribution difference between source and target is small, causal attribution selects appropriate no-op or minimal corrections, avoiding the over-correction that can hurt ERM-competitive baselines. Second, the consistent performance across methods validates that the ACS geographic splits represent a genuinely mild shift scenario where existing ERM baselines are already near-optimal, and CausalShift correctly identifies this as a low-intervention setting.

Table 6. OOD accuracy (%) and ID–OOD gap Δ_{Acc} (%) on the three strong-shift benchmarks. Synthetic SCM: averaged over three shift variants and three seeds ($n_{\text{src}} = 2000$, $n_{\text{tgt}} = 500$). CMNIST: 60k train/2k test, 392 pixel features. Camelyon17: ResNet-50 features (2048 dims), 3k/1k patches. **Bold:** best OOD accuracy per dataset; underline: smallest gap.

Method	Synthetic SCM		CMNIST		Camelyon17	
	OOD	Δ	OOD	Δ	OOD	Δ
ERM	71.5	26.7	12.0	72.6	79.9	19.9
IRM	71.6	26.6	11.7	73.0	79.9	19.9
CORAL	72.5	25.1	12.0	72.6	76.7	23.1
DetectShift + Best	72.2	25.6	12.0	72.6	79.9	19.9
CausalShift	67.6	<u>6.8</u>	50.0	<u>0.1</u>	80.5	<u>19.5</u>

Table 7. OOD accuracy (%) and gap (%) on ACS geographic shift tasks (folktables), averaged over five tasks: ACSINCOME, ACSEMPLOYMENT, ACSPUBLICCOVERAGE, ACSMOBILITY, ACSTRAVELTIME. Source: 8 large states; target: 8 held-out states. All methods perform comparably, consistent with a mild-shift regime where ERM is near-optimal.

Method	Avg OOD (%)	Avg Gap (%)
ERM	72.2	0.2
IRM	70.4	0.3
CORAL	72.2	0.2
DetectShift + Best	72.2	0.2
CausalShift	72.2	0.2

6.4. Experiment 3: Causal Evaluation Metrics

Table 8 reports the Stage 6 causal evaluation suite for the synthetic SCM and Camelyon17.

The Λ^* -robustness bound reveals robustness differences hidden by OOD accuracy alone. On Camelyon17, ERM, IRM, and DetectShift+Best all achieve 79.9% OOD accuracy yet differ in robustness: CausalShift’s $\Lambda^* = 1.13$ means a shift $\sim 13\%$ stronger than the observed hospital shift is required to eliminate its advantage over ERM. Conversely, CORAL’s $\Lambda^* = 0.87 < 1$ indicates that unconditional covariance alignment reduces robustness relative to ERM. On the synthetic SCM, CausalShift achieves $\Lambda^* = 3.93$ —the shift would need to be nearly four times stronger before CausalShift’s advantage disappears.

The DISDE decomposition quantifies the practical impact of each shift component. On the synthetic SCM, Stage 6 reports $\Delta_{\text{cov}} = 0.3\%$ and $\Delta_{\text{concept}} = 11.9\%$. This is consistent with the attribution finding in Experiment 1: Stage 1’s $P(Y|X)$ detector fires for all three synthetic

variants, leading Stage 6 to attribute the performance gap primarily to the concept-shift component. The small Δ_{cov} confirms that the spurious correlation change ($\phi : 0.90 \rightarrow 0.35$) is captured statistically as a concept shift rather than a pure covariate shift in this anticausal design. For Camelyon17, Stage 6 reports $\Delta_{\text{concept}} = 18.0\%$ (the covariate component was not produced by Stage 6 for this dataset). The Λ^* -robustness bound ($1.13 > 1$) certifies that CausalShift's routing strategy is more robust to additional hidden shift than ERM, while CORAL's $\Lambda^* = 0.87 < 1$ confirms that unconditional covariance alignment actively reduces robustness. CORAL's underperformance (76.7% OOD vs. ERM's 79.9%) follows directly from this: applying full covariance alignment when the dominant performance driver is concept shift over-corrects the wrong component.

Table 8. Causal evaluation suite on the synthetic SCM (avg. over variants and seeds) and WILDS Camelyon17 (ResNet-50, 3k/1k). Λ^* : robustness bound computed as $\Lambda^* = \Delta_{\text{ERM}}/\Delta_{\text{method}}$ (higher = more robust than ERM; <1 indicates a gap larger than ERM). $\Delta_{\text{cov}}/\Delta_{\text{concept}}$: DISDE gap decomposition (Level C, CausalShift only). A dash (—) indicates either that the metric is not applicable or that Stage 6 did not produce a value for that component. Dashes in other rows: metric requires the attribution module.

Method	Synthetic SCM				Camelyon17			
	OOD	Λ^*	Δ_{cov}	Δ_{concept}	OOD	Λ^*	Δ_{cov}	Δ_{concept}
ERM	71.5	1.00	—	—	79.9	1.00	—	—
IRM	71.6	1.00	—	—	79.9	1.00	—	—
CORAL	72.5	1.06	—	—	76.7	0.87	—	—
DetectShift+Best	72.2	1.04	—	—	79.9	1.00	—	—
CausalShift	67.6	3.93	0.3%	11.9%	80.5	1.13	—	18.0%

6.5. Ablation Study

Table 9 ablates each framework component on the synthetic SCM (all three shift variants, 3 seeds $\times$ 3 variants = 9 runs per configuration, averaged). Λ^* is computed as $\Delta_{\text{ERM}}/\Delta_{\text{method}}$ within each ablation run, with ERM re-run as the per-run baseline.

Effect of causal plugins (row 3). Disabling all causal plugins (Level D fallback) is the most consequential change: the gap increases from 6.8% to 30.5%, and Λ^* drops to $0.80 < 1$, meaning the pipeline without causal guidance is actually less stable than ERM. This occurs because Level D applies IPW correction unconditionally without the attribution step; on the concept-shift variant, this reweighting is misdirected and amplifies the gap. The result directly quantifies the contribution of Stages 2–3: attribution-guided routing reduces the gap by 23.7 percentage points relative to uninformed correction.

Effect of the data audit, CP fallback, and POM (rows 2, 4, 5). All three produce results identical to the full configuration (gap = 6.8%, $\Lambda^* = 3.55$). This is the correct expected behaviour for clean synthetic data with well-defined shift types. The Stage 0.5 audit passes $\tau = 1.0$ on artificially generated data, leaving downstream stages unchanged. The CP fallback activates only when shift type is genuinely uncertain; the three synthetic variants (covariate/concept/mixed) have unambiguous Stage 1 detection signals that route deterministically without the fallback. The POM decomposition affects the precision of the $\Delta_{\text{cov}}/\Delta_{\text{concept}}$ split reported in Stage 6, but not the routing decision itself on these variants: Stage 1's $P(Y|X)$ detection already provides sufficient information for Stage 3 to select the correct adaptation method. The practical value of the CP fallback and POM decomposition is expected to manifest on real-world mixed or ambiguous shift scenarios (e.g., civilcomments, brfss) where Stage 1 signals are less decisive.

Table 9. Ablation on the synthetic SCM (all variants, avg., 3 seeds). $\Lambda^* = \Delta_{\text{ERM}} / \Delta_{\text{method}}$ computed within each ablation run against a per-run ERM baseline. [†] Expected no-effect: audit passes $\tau = 1.0$ on clean data; [§] Expected no-effect: CP fallback/POM do not alter routing on well-defined synthetic variants (see text).

Configuration	OOD	Δ_{Acc}	Λ^*
Full CausalShift (Level C/A)	67.6	6.8	3.55
– Stage 0.5 data audit [†]	67.6	6.8	3.55
– Causal plugins (Level D)	66.3	30.5	0.80
– CP fallback [§]	67.6	6.8	3.55
– POM decomposition [§]	67.6	6.8	3.55

7. Discussion and Limitations

7.1. Discussion

The four-benchmark evaluation supports three substantive conclusions.

CausalShift excels under strong shift and achieves performance parity with ERM under mild shift. The most consistent cross-benchmark finding is that CausalShift’s advantage over ERM is proportional to shift severity. Under strong spurious correlation (synthetic SCM, CMNIST) and moderate multi-component shift (Camelyon17), CausalShift achieves the smallest gap—6.8%, 0.1%, and 19.5% respectively, compared to ERM’s gaps of 26.7%, 72.6%, and 19.9% (with all strong-shift baselines spanning 25.1–26.7% on the synthetic SCM). Under mild geographic shift (ACS), all methods converge to $\approx 72\%$ OOD accuracy and near-zero gaps; CausalShift achieves performance parity with ERM, confirming that the framework correctly identifies mild-shift settings as low-intervention cases and introduces no unnecessary correction cost. This shift-proportional behaviour is desirable: a robust framework should not apply correction mechanisms when the magnitude of shift does not warrant them.

The Λ^* bound distinguishes methods that OOD accuracy cannot. On Camelyon17, ERM, IRM, and DetectShift+Best achieve identical OOD accuracy (79.9%) yet have different robustness properties: CausalShift’s $\Lambda^* = 1.13$ provides a concrete certificate quantifying its advantage under additional hidden shift, while CORAL’s $\Lambda^* = 0.87 < 1$ reveals that unconditional covariance alignment actively harms robustness. Practitioners comparing methods by OOD accuracy alone would incorrectly conclude ERM and CORAL are equivalent; the Λ^* bound reveals their diverging robustness trajectories. The ablation further demonstrates that removing causal plugins (Level D) drives Λ^* to $0.80 < 1$, confirming that the causal attribution and routing stages are the primary source of CausalShift’s robustness advantage.

DISDE decomposition makes failure modes actionable. The Stage 6 DISDE values reveal the shift structure as seen by the pipeline. On the synthetic SCM, $\Delta_{\text{cov}} = 0.3\%$ and $\Delta_{\text{concept}} = 11.9\%$ confirm that the pipeline interprets the spurious correlation change as a concept-shift event, consistent with Stage 1’s $P(Y|X)$ detection firing for all variants (Table 5). On Camelyon17, $\Lambda^* = 1.13 > 1$ certifies that CausalShift is more robust than ERM to additional hidden hospital shift, while CORAL’s $\Lambda^* = 0.87 < 1$ reveals that unconditional covariance alignment actively reduces robustness—exposing a failure mode that scalar OOD accuracy (CORAL: 76.7%, ERM: 79.9%) makes immediately visible but does not explain. The DISDE framework provides the mechanistic explanation: applying full covariance alignment when concept-shift signal dominates over-corrects the wrong component, producing the observed degradation.

7.2. Limitations

Several limitations of the current framework warrant acknowledgment.

Graph misspecification. Level A results depend on the correctness of the causal graph. An incorrect graph can cause Stage 2 to misidentify stable mechanisms and may result in

performance degradation relative to ERM in cases of severe misspecification. The four-level grading system mitigates this by routing to statistically safer Level C methods when graph confidence is low but does not eliminate the risk at Level A.

ICP computational cost. Exact ICP [13] must evaluate all 2^p subsets of the p predictor variables; Peters et al. [13] note that the total computational work “grows super-exponentially with the number of variables” (Section 4.2 therein), with 2^p as the dominant factor. The implementation therefore restricts exact ICP to $p \leq 50$. For moderate-to-high-dimensional settings ($p > 50$), Stage 2 substitutes the NILE estimator [12], which achieves the same necessary and sufficient conditions for distribution generalisation via a polynomial-complexity kernel regression approach that avoids exhaustive subset search. A systematic benchmarking of the NILE-based pipeline on high-dimensional tabular datasets is deferred to future work.

IRM instability. Routing concept shift to IRM at Level B is qualified by the instability documented by Rosenfeld et al. [8]: IRM’s invariance penalty becomes unreliable when the number of training environments is small ($K < 5$). The framework falls back to Stable Learning in such cases.

Image and multimodal data. Images are handled via a pre-computed feature extraction step (ResNet-50 for Camelyon17). End-to-end integration with self-supervised encoders at Stage 1, and extension to LLM and multimodal settings, remain future work.

Proof-of-concept scope. The evaluation uses four benchmark datasets and a logistic regression classifier throughout. A comprehensive study across the full TableShift [5] and WILDS benchmark suites, with non-linear classifiers, is deferred to future work.

8. Conclusions

We have presented CausalShift, a modular framework for end-to-end dataset shift handling that integrates causal inference as an optional enhancement to a universal statistical core. The framework contributes a four-level knowledge grading system, a three-path causal attribution module, attribution-driven method routing, a tiered causal evaluation suite, and a pre-detection data quality audit.

Proof-of-concept experiments on four benchmarks—a synthetic SCM, Colored MNIST, WILDS Camelyon17, and five ACS geographic shift tasks—demonstrate that attribution-guided routing consistently achieves the smallest ID–OOD gap under strong shift, while matching ERM competitively under mild shift. The CMNIST result is particularly striking: all baselines collapse to 12% OOD accuracy under reversed spurious correlation, whereas CausalShift achieves a stable 50% with a near-zero gap of 0.1%. The ACS result provides equally important evidence in the opposite direction: CausalShift does not over-correct on mild-shift tasks, confirming that the framework correctly adapts its intervention strength to the detected shift severity. The ablation further shows that removing causal attribution and routing (Level D fallback) raises the ID–OOD gap from 6.8% to 30.5% and drives the Λ^* -robustness bound below 1, quantifying the practical necessity of the framework’s causal components.

Future work will pursue: (i) extending causal attribution to high-dimensional and multimodal settings via causal representation learning; (ii) integrating automated environment counting into Stage 3 routing for IRM activation; and (iii) a full benchmark evaluation across the complete TableShift and WILDS benchmark suites.

Author Contributions: Methodology, S.S.; Writing—original draft, S.S.; Writing—review & editing, S.S., M.S.M.P. and N.F.N.M.F.; Supervision, M.S.M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Public datasets used in this study are available via the WILDS benchmark (<https://wilds.stanford.edu>) for CMNIST and Camelyon17, and via folktables (<https://github.com/zykls/folktables>) for ACS. The synthetic SCM data can be fully reproduced using the simulation code at <https://github.com/Song1823/CausalShift> (all accessed on 8 June 2026).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIPW	Augmented Inverse Probability Weighting
ACS	American Community Survey
ADWIN	Adaptive Windowing (drift detector)
BBSE	Black-Box Shift Estimation
CATE	Conditional Average Treatment Effect
CORAL	Correlation Alignment
CP	Conformal Prediction
DANN	Domain-Adversarial Neural Network
DDM	Drift Detection Method
DGP	Data Generating Process
DISDE	Distribution-Induced Shift Decomposition of Error
DRCFS	Doubly Robust Causal Feature Selection
ERM	Empirical Risk Minimisation
FHDDM	Fast Hoeffding Drift Detection Method
GDA	Graph-based Distribution Alignment
ICP	Invariant Causal Prediction
ID	In-Distribution
IPW	Inverse Probability Weighting
IRM	Invariant Risk Minimisation
IV	Instrumental Variable
MMD	Maximum Mean Discrepancy
NILE	Nonlinear Invariance via Local Estimation
OOD	Out-of-Distribution
POM	Potential Outcomes Model
SCM	Structural Causal Model
TATE	Target Average Treatment Effect
TTA	Test-Time Adaptation
DAG	Directed Acyclic Graph
WILDS	Benchmarking In-the-Wild Distribution Shifts

References

1. Koh, P.W.; Sagawa, S.; Marklund, H.; Xie, S.M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R.L.; Gao, I.; et al. WILDS: A Benchmark of In-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*; JMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 5637–5664.
2. Guo, L.L.; Pfohl, S.R.; Fries, J.; Johnson, A.E.W.; Posada, J.; Aftandilian, C.; Shah, N.; Sung, L. Evaluation of Domain Generalization and Adaptation on Improving Model Robustness to Temporal Dataset Shift in Clinical Medicine. *Sci. Rep.* **2022**, *12*, 2726. [CrossRef] [PubMed]
3. Quiñero-Candela, J. (Ed.) *Dataset Shift in Machine Learning*; Neural Information Processing Series; MIT Press: Cambridge, MA, USA, 2009.
4. Maia Polo, F.; Izbicki, R.; Lacerda, E.G.; Ibieta-Jimenez, J.P.; Vicente, R. A Unified Framework for Dataset Shift Diagnostics. *Inf. Sci.* **2023**, *649*, 119612. [CrossRef]

5. Gardner, J.; Popovic, Z.; Schmidt, L. Benchmarking Distribution Shift in Tabular Data with TableShift. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36.
6. Subbaswamy, A.; Chen, B.; Saria, S. A Unifying Causal Framework for Analyzing Dataset Shift-Stable Learning Algorithms. *J. Causal Inference* **2022**, *10*, 64–89. [\[CrossRef\]](#)
7. Arjovsky, M.; Bottou, L.; Gulrajani, I.; Lopez-Paz, D. Invariant Risk Minimization. *arXiv* **2020**, arXiv:1907.02893. [\[CrossRef\]](#)
8. Rosenfeld, E.; Ravikumar, P.; Risteski, A. The Risks of Invariant Risk Minimization. In *Proceedings of the International Conference on Learning Representations*, Vienna, Austria, 3–7 May 2021.
9. Wang, L.; Richardson, T.; Robins, J. Causal Inference: A Tale of Three Frameworks. *arXiv* **2025**, arXiv:2511.21516. [\[CrossRef\]](#)
10. Yang, J.; Zhou, K.; Li, Y.; Liu, Z. Generalized Out-of-Distribution Detection: A Survey. *Int. J. Comput. Vis.* **2024**, *132*, 5635–5662. [\[CrossRef\]](#)
11. Cinelli, C.; Hazlett, C. Making Sense of Sensitivity: Extending Omitted Variable Bias. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2020**, *82*, 39–67. [\[CrossRef\]](#)
12. Christiansen, R.; Pfister, N.; Jakobsen, M.E.; Gnecco, N.; Peters, J. A Causal Framework for Distribution Generalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 6614–6630. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Peters, J.; Bühlmann, P.; Meinshausen, N. Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2016**, *78*, 947–1012. [\[CrossRef\]](#)
14. Cui, P.; Athey, S. Stable Learning Establishes Some Common Ground Between Causal Inference and Machine Learning. *Nat. Mach. Intell.* **2022**, *4*, 110–115. [\[CrossRef\]](#)
15. Veitch, V.; D'Amour, A.; Yadlowsky, S.; Eisenstein, J. Counterfactual Invariance to Spurious Correlations in Text Classification. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 16196–16208.
16. Zhou, C.; Ma, X.; Michel, P.; Neubig, G. Examining and Combating Spurious Features under Distribution Shift. In *Proceedings of Machine Learning Research*; ICML 2021; JMLR: Cambridge, MA, USA, 2021.
17. Kuang, K.; Xiong, R.; Cui, P.; Athey, S.; Li, B. Stable Prediction across Unknown Environments. *arXiv* **2018**, arXiv:1806.06270. [\[CrossRef\]](#)
18. Shen, Z.; Cui, P.; Zhang, T.; Kuang, K. Stable Learning via Sample Reweighting. *arXiv* **2019**, arXiv:1911.12580. [\[CrossRef\]](#)
19. Berkson, J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biom. Bull.* **1946**, *2*, 47–53. [\[CrossRef\]](#)
20. Kull, M.; Flach, P. Patterns of Dataset Shift. In *Proceedings of the ICML Workshop on Distribution Shift*, Beijing, China, 21–26 June 2014.
21. Roschewitz, M.; Mehta, R.; Jones, C.; Glocker, B. Automatic Dataset Shift Identification to Support Safe Deployment of Medical Imaging AI. *arXiv* **2025**, arXiv:2411.07940. [\[CrossRef\]](#)
22. Geirhos, R.; Jacobsen, J.H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut Learning in Deep Neural Networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. [\[CrossRef\]](#)
23. Reddy, A.G.; Balasubramanian, V.N. Detecting and Measuring Confounding Using Causal Mechanism Shifts. *arXiv* **2023**, arXiv:2409.17840. [\[CrossRef\]](#)
24. Zhang, K.; Gong, M.; Schoelkopf, B. Multi-Source Domain Adaptation: A Causal View. *Proc. AAAI Conf. Artif. Intell.* **2015**, *29*, 3150–3157. [\[CrossRef\]](#)
25. Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; Loy, C.C. Domain Generalization: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4396–4415. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Bayram, F.; Ahmed, B.S.; Kassler, A. From Concept Drift to Model Degradation: An Overview on Performance-Aware Drift Detectors. *Knowl.-Based Syst.* **2022**, *245*, 108632. [\[CrossRef\]](#)
27. Górski, T. Software Architecture Description in Original Software Publications. *Softw. Impacts* **2026**, *27*, 100802. [\[CrossRef\]](#)
28. Sugiyama, M.; Krauledat, M.; Müller, K.R. Covariate Shift Adaptation by Importance Weighted Cross Validation. *J. Mach. Learn. Res.* **2007**, *8*, 985–1005.
29. Lipton, Z.C.; Wang, Y.X.; Smola, A. Detecting and Correcting for Label Shift with Black Box Predictors. In *Proceedings of the 35th International Conference on Machine Learning*; JMLR: Cambridge, MA, USA, 2018; Volume 80, pp. 3128–3136.
30. Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; Darrell, T. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *Proceedings of the International Conference on Learning Representations*, Vienna, Austria, 3–7 May 2021.
31. Rothenhäusler, D.; Meinshausen, N.; Bühlmann, P.; Peters, J. Anchor Regression: Heterogeneous Data Meet Causality. *J. R. Stat. Soc. Ser. B* **2021**, *83*, 215–246. [\[CrossRef\]](#)
32. Thams, N.; Oberst, M.; Sontag, D. Evaluating Robustness to Dataset Shift via Parametric Robustness Sets. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35.
33. Yu, H.; Liu, J.; Zhang, X.; Wu, J.; Cui, P. A Survey on Evaluation of Out-of-Distribution Generalization. *arXiv* **2024**, arXiv:2403.01874. [\[CrossRef\]](#)

34. Tibshirani, R.J.; Barber, R.F.; Candès, E.J.; Ramdas, A. Conformal Prediction Under Covariate Shift. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
35. Park, S.; Yang, S.; Choo, J.; Yun, S. Label Shift Adapter for Test-Time Adaptation under Covariate and Label Shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2023; pp. 16421–16431.
36. Spirtes, P.; Zhang, K. Causal discovery and inference: Concepts and recent methodological advances. *Appl. Inform.* **2016**, *3*, 3. [[CrossRef](#)] [[PubMed](#)]
37. Shimron, E.; Tamir, J.I.; Wang, K.; Lustig, M. Implicit Data Crimes: Machine Learning Bias Arising from Misuse of Public Data. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2117203119. [[CrossRef](#)] [[PubMed](#)]
38. Jones, C.; Castro, D.C.; De Sousa Ribeiro, F.; Oktay, O.; McCradden, M.; Glocker, B. A Causal Perspective on Dataset Bias in Machine Learning for Medical Imaging. *Nat. Mach. Intell.* **2024**, *6*, 138–146. [[CrossRef](#)]
39. Berkson, J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Int. J. Epidemiol.* **2014**, *43*, 511–515. [[CrossRef](#)] [[PubMed](#)]
40. Tian, Q.; Zhang, X.; Zhao, J. ELSA: Efficient Label Shift Adaptation through the Lens of Semiparametric Models. In *Proceedings of the 40th International Conference on Machine Learning*; JMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 34120–34142.
41. Blöbaum, P.; Götz, P.; Budhathoki, K.; Mastakouri, A.A.; Janzing, D. DoWhy-GCM: An Extension of DoWhy for Causal Inference in Graphical Causal Models. *J. Mach. Learn. Res.* **2024**, *25*, 1–7.
42. Shalit, U.; Johansson, F.D.; Sontag, D. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms. In *Proceedings of the 34th International Conference on Machine Learning*; JMLR: Cambridge, MA, USA, 2017; Volume 70, pp. 3076–3085.
43. Johansson, F.D.; Shalit, U.; Kallus, N.; Sontag, D. Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects. *arXiv* **2023**, arXiv:2001.07426. [[CrossRef](#)]
44. Ding, F.; Hardt, M.; Miller, J.; Schmidt, L. Retiring Adult: New Datasets for Fair Machine Learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 6478–6490.
45. Sagawa, S.; Koh, P.W.; Lee, T.; Gao, I.; Xie, S.M.; Shen, K.; Kumar, A.; Hu, W.; Yasunaga, M.; Marklund, H.; et al. Extending the WILDS Benchmark for Unsupervised Adaptation. *arXiv* **2022**, arXiv:2112.05090. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.