



Article

SEMTRA: Global Semantic Transition and Rough-Set Rules for Auditable Post-Hoc Explainability

Pavlo Radiuk ^{1,*} , Oleksander Barmak ¹ and Iurii Krak ^{2,3} ¹ Department of Computer Science, Khmelnytskyi National University, 11 Instytut's'ka Str., 29016 Khmelnytskyi, Ukraine; barmako@khnmu.edu.ua² Department of Theoretical Cybernetics, Taras Shevchenko National University of Kyiv, 4d Akademika Glushkova Ave, 03680 Kyiv, Ukraine; iurii.krak@knu.ua³ Laboratory of Communicative Information Technologies, V.M. Glushkov Institute of Cybernetics, 40 Akademika Glushkova Ave, 03187 Kyiv, Ukraine

* Correspondence: radiukp@khnmu.edu.ua

Abstract

Deep learning architectures generate highly effective but difficult-to-audit latent representations, creating a practical gap between predictive performance and verifiable explanations. Existing post hoc techniques often produce fragmented local attributions rather than dataset-level rulebooks. In this work, we propose Global SEMantic TRAnsition (SEMTRA), a post hoc framework that maps frozen representation features into semantic attributes, discretizes those attributes, and induces rough-set production rules with explicit coverage, conflict, fidelity, and abstention reporting. Evaluated on the Animals with Attributes 2 (AwA2) Protocol A, the semantic transition achieved a Mean Absolute Error (MAE) of 0.1029 ± 0.0005 . The extracted rulebook covered 84.80% of test instances, yielding a covered accuracy of 39.73% and a covered fidelity to the base predictor of 40.48%. Under the Protocol B split, continuous semantic-prototype transfer reached an unseen-object accuracy of $44.02\% \pm 1.22\%$ as a semantic-transfer validation. Cross-domain validations using SUN and Derm7pt demonstrated that the audit protocol is portable yet strongly dataset-dependent. In the controlled synthetic benchmark, SEMTRA achieved a macro-F1 score of 0.879 at zero semantic noise and degraded to 0.838 at the highest evaluated noise level. Ultimately, SEMTRA serves as a transparent audit layer to expose the verifiable logical subset of a model, rather than replacing the underlying predictor.

Keywords: explainable artificial intelligence (XAI); transition matrix; semantic attributes; rough sets; production rules; discretization; zero-shot learning

Academic Editors: Sheng Du and
Javier Del Ser Lorente

Received: 29 May 2026

Revised: 20 June 2026

Accepted: 24 June 2026

Published: 29 June 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

Deep learning architectures have become strong baselines for visual recognition because they compress high-dimensional observations into effective latent representations. However, this compression mechanism creates a black-box effect. Neural features are optimized primarily for predictive success rather than for human interpretability. As automated systems enter safety-sensitive decision workflows, the demand for transparency has shifted from heuristic approximations toward more reproducible and mathematically specified explanations. The recent manifesto detailing the evolution of Artificial Intelligence (AI) frameworks emphasizes the critical need for verifiable symbolic knowledge [1]. Current approaches addressing this transparency demand fall into diverse methodological categories.

A comprehensive survey of modern explainable frameworks illustrates a prominent divide between local diagnostic techniques and globally interpretable model designs [2].

The problem under consideration is the difficulty of systematically translating compressed latent neural coordinates into a global, conflict-aware rulebook without heavily retraining or modifying the underlying feature extractor. Instance-based post hoc explanations generate input-specific attribution maps, but they usually do not produce a reusable logic policy for dataset-level auditing. Conversely, interpretable-by-design architectures require predefined concept vocabularies and costly structural re-engineering of existing high-performance networks. It is highly important to present novel scientific contributions to address this problem, thereby closing the interpretability gap that complicates routine model verification and clinical decision-support auditing.

To construct this required bridge, the proposed framework implements a continuous semantic transition. Early analytical studies validated the efficacy of transition matrices for interpreting complex networks within structured visual analytics environments [3]. These matrix structures were subsequently adapted and refined to extract understandable features in specialized healthcare informatics contexts [4]. The present manuscript is distinct from the prior Equivariant Transition Matrix (ETM) study. ETM targets symmetry-aware transition operators and Lie-group linearization, with the central objective of preserving geometric consistency under specified transformations [5]. SEMTRA is not an equivariance method and does not claim Lie-group structural consistency. Its contribution is the semantic-to-symbolic audit sequence: semantic reconstruction, discretization, rough-set granulation, conflict-aware rule induction, abstention, and quantitative audit-tax reporting. Table A1 in Appendix A summarizes this separation.

SEMTRA advances this trajectory by converting continuous semantic variables into verifiable logical predicates using established rough-set principles. Initial mathematical formalizations introduced the concept of bounding uncertainty through precise indiscernibility relations [6]. Comprehensive theoretical extensions eventually solidified the applicability of rough sets in structured reasoning about inconsistent data [7].

The goal of this study is to improve the auditability, transparency, and logical verifiability of pre-trained deep learning representations by applying a continuous semantic transition coupled with rigorous rough-set knowledge granulation. The objective of the study explicitly is to formulate and empirically evaluate the SEMTRA framework, ensuring it effectively translates latent neural features into a conflict-aware symbolic rulebook while explicitly identifying ambiguous boundary regions. Accordingly, the research proposes four major contributions:

- First, it establishes a dimensionally explicit transition-matrix formulation that maps compressed deep representations to interpretable semantic attributes without retraining the underlying feature extractor.
- Second, it implements a density-aware discretization module for converting reconstructed continuous semantics into structured symbolic states.
- Third, it introduces a rough-set rule-induction mechanism that constructs a global production rulebook, featuring explicit handling of structural abstentions and boundary regions.
- Finally, it conducts a multi-domain evaluation on benchmark datasets and controlled synthetic data, demonstrating how coverage, fidelity, accuracy, conflict, and sensitivity jointly define the systematic audit tax.

The remainder of this paper is organized as follows. Section 2 details the related works covering explainable methods and symbolic rules. Section 3 outlines the proposed framework methodology, the continuous transition bridge, and the subsequent rough-set extraction mechanisms. Section 4 presents the qualitative and quantitative evaluations.

Section 5 analyzes the implications of the structural explanation tradeoff. Finally, Section 6 concludes the manuscript.

2. Related Works

2.1. Local Post Hoc Explanation

Local post hoc interpretation approaches estimate decision boundaries in the immediate vicinity of individual input samples to generate explanations. Highly ubiquitous frameworks, such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), heavily rely on perturbed neighborhood samples to construct sparse additive feature attributions [8]. In high-stakes medical domains, recent advancements deploy specialized explainable deep learning layers to make complex time-series predictions such as arrhythmia detection transparent [9]. Concurrently, in critical industrial monitoring scenarios, novel fusion methodologies integrate multispectral observation rankings with fuzzy logic to evaluate defect severity [10]. However, even optimized local methodologies fail to yield the exhaustive, dataset-wide conflict resolution protocols essential for comprehensive global diagnostic auditing.

2.2. Concept-Based and Ante Hoc Models

Concept-based techniques aim to directly augment numerical latent features with variables mapped to human-understandable terms. The foundational introduction of concept activation vectors provided a mathematical formulation to determine whether final neural decisions remain sensitive to predefined semantic directions [11]. More sophisticated architectural approaches require networks to accurately predict intermediate semantic variables before executing final classification outputs [12]. Contemporary literature illustrates that these discrete human-aligned concepts can be natively embedded within models during the initial structural training phase [13]. Alternatively, selected optimization algorithms can reliably infer similar aligned concepts entirely through post hoc modifications [14]. The primary limitation of ante-hoc methodologies remains their strict dependence on specialized initial training regimens.

2.3. Formal Knowledge Granulation and Rough Sets

The scientific pursuit of inducing symbolic rule structures predates the massive expansion of deep artificial neural networks. A comprehensive review highlights that classic hierarchical decision trees organically partition observations based on localized entropy metrics to generate completely transparent rule sequences [15]. Concurrently, separate-and-conquer logic algorithms iterate over modular rules to balance overall predictive accuracy with ultimate structural compactness, a process recently enhanced by integrating fuzzy logic components [16]. Modern structured applications successfully deploy these distinct rule-based algorithms to dissect complex unsupervised anomaly detection frameworks [17]. Differentiable, fuzzy, and hybrid rough-set models have also reappeared as neural-network companions, either by smoothing rough membership functions for gradient optimization or by using fuzzy-rough rules to summarize black-box behavior. SEMTRA is positioned differently: it leaves the neural extractor frozen and applies rough-set granulation as a post hoc audit layer whose conflicts and abstentions remain explicit. Nevertheless, most heuristic rule inducers forcefully dictate deterministic boundary classifications. Formulating stable partitions for rough-set structures necessitates robust preprocessing; modern non-monotonic discretization techniques demonstrate substantial utility in stabilizing interval boundaries [18]. Furthermore, novel kernel density estimators effectively map topological probability distributions within large, highly dimensional continuous spaces [19].

2.4. Semantic Transfer as a Validation Proxy

Continuous semantic transfer evaluation serves as a robust mechanism to assess whether a given methodology successfully captures generalized conceptual structures. For instance, the AwA2 benchmark facilitates rigorous testing of continuous semantic projection between seen domains and unseen domains [20]. This foundational repository provides standardized imagery accompanied by precise numeric class-level semantic prototype vectors [21]. Historically, the Direct Attribute Prediction (DAP) mechanisms established baseline predictive benchmarks through probabilistic semantic mappings [22]. Subsequent Indirect Attribute Prediction (IAP) formulations refined structural relationships among intermediate variables to better generalize conceptual knowledge [23].

To continuously elevate Zero-Shot Learning (ZSL) classification performance, researchers have developed various embedding and generative architectures. A Generative Framework for Zero-Shot Learning (GFZSL) improves zero-shot accuracy by synthesizing pseudo-samples derived from target attribute distributions [24]. This feature generation capability has been formally extended to broader any-shot settings [25]. Dense attribute attention mechanisms localize discriminative semantic evidence more explicitly during processing [26]. Attribute prototype networks operate similarly to establish stable classification boundaries [27]. Contrastive learning systems further improve visual-semantic alignment through robust structural objectives [28]. Transformer architectures increasingly guide global representations across predefined semantic constraints [29]. Refined variations of this approach apply cross-attribute awareness to capture complex semantic dependencies [30]. Moreover, modern mechanisms leverage textual document summaries to reliably align conceptual spaces [31].

Diverse analytical methodologies calculate probabilistic Convex combinations of Semantic Embeddings (CONSE) [32]. Cross-Modal Transfer (CMT) architectures project comprehensive textual descriptions directly into visual coordinates [33]. Structured optimization strategies strategically apply Semantic Similarity Embedding (SSE) paradigms [34]. Specialized Attribute Label Embedding (ALE) operations successfully isolate relevant conceptual boundaries [35]. Structured Joint Embedding (SJE) formulations enhance fundamental metric alignment loops [36]. Deep Visual-Semantic Embedding (DEVISE) models explicitly align deep spatial features across distinct domain manifolds [37]. Embarrassingly Simple Zero-Shot Learning (ESZSL) via fast linear regularizations provides mathematically simple yet functionally effective prediction limits [38]. Other frameworks use Synthesized Classifiers (SYNC) with artificial class weights to hallucinate unobserved semantic targets [39]. Semantic Autoencoder (SAE) architectures strictly enforce constraints that preserve inherent topology within projection spaces [40]. Because these projection realms remain highly dimensional, robust implementations rely heavily on efficient randomized Singular-Value Decompositions (SVDs) [41].

The broader scene understanding field expanded evaluation domains through large-scale attribute databases, such as the Scene Understanding (SUN) dataset [42]. The comprehensive structured metadata supporting these visual environments remains actively maintained and accessible online [43]. In the medical context, specialized clinical benchmarks utilize structured dermoscopic assessment criteria [44]. Such repositories supply the standardized diagnostic verification frameworks required for robust testing [45].

Despite noticeable advancements in Explainable Artificial Intelligence (XAI), a scientific gap persists in the systematic bridging of latent features and transparent decision logic without sacrificing auditability. Existing post hoc methods typically generate fragmented visual attributions that fail to establish global rules, while interpretable-by-design architectures demand costly model retraining. Furthermore, traditional heuristic rule inducers forcefully dictate boundary classifications, consistently neglecting the structural indiscerni-

bility inherent in complex deep representations. The proposed research contributions directly address these omissions by decoupling the heavy computational inference of a neural network from a lightweight symbolic extraction layer. By leveraging an explicitly defined continuous transition matrix in conjunction with density-aware discretization and formal rough-set granulation, the presented methodology yields a verified framework capable of extracting honest, conflict-aware diagnostic policies.

3. Materials and Methods

3.1. Framework Architecture Overview

The SEMTRA framework is structurally designed as a modular, post hoc audit pipeline for settings where frozen representation features can be aligned to a semantic attribute dictionary. Operating over a frozen feature extractor, the methodology produces a global symbolic rulebook for the subset of model behavior that can be reconstructed, discretized, and verified. The conceptual architecture, illustrated in Figure 1, explicitly follows a two-phase progression: Phase 1, designated as Neural-to-Semantic Transition, and Phase 2, defined as Semantic-to-Symbolic Granulation.

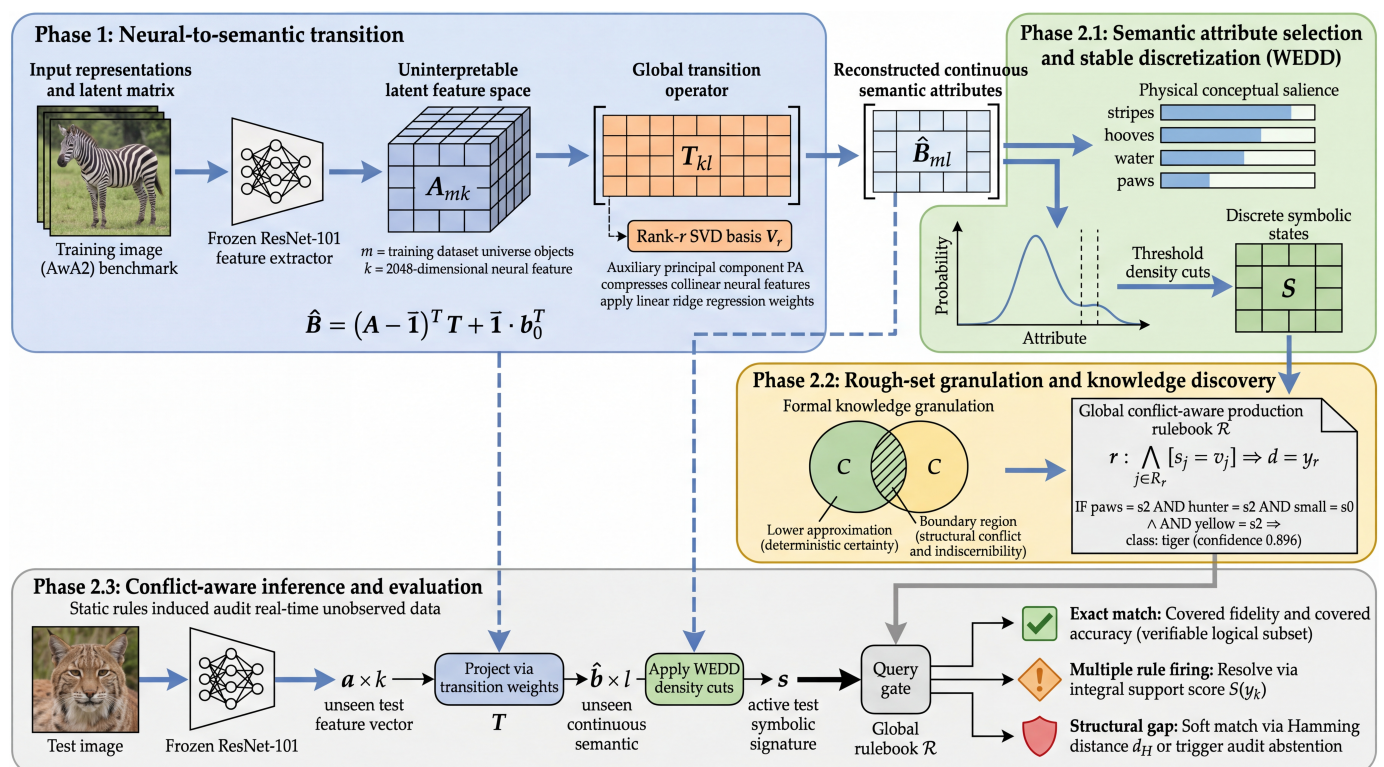


Figure 1. Conceptual architecture of the SEMTRA framework. The arrows denote the processing direction between pipeline stages and are not quantitative length encodings. In Phase 2.1, the blue horizontal bars schematically indicate relative semantic-attribute magnitudes used before WEDD thresholding; they do not represent measured probabilities. The end-to-end pipeline sequentially translates uninterpretable latent neural features into continuous semantic attributes via a global transition matrix and then leverages density-aware discretization and rough-set granulation to induce a transparent, conflict-aware rulebook.

The progression moves from the initial latent space A , mapped through a global transition matrix T , to reconstruct continuous semantics B . Subsequently, these semantics are discretized through the Weighted Entropy-Density Discretization (WEDD) mechanism to induce compact symbolic production rules $\mathcal{R}$. To avoid notational ambiguity, the matrix of all discretized states is denoted as S , with individual symbolic states for object i and

attribute j defined as s_{ij} . The output of the framework follows the standardized rule format r as presented in Equation (1):

$$r : \bigwedge_{j \in R_r} [s_j = v_j] \Rightarrow d = y_r, \quad (1)$$

where R_r is the selected minimal set of semantic antecedents, s_j represents the symbolic variable representing attribute j , and v_j is the corresponding symbolic value bound.

The mathematical formalization of the induced decision-making system S is defined in Equation (2):

$$S = (\mathcal{U}, \mathcal{C}, d, \mathcal{V}, f), \quad (2)$$

where $\mathcal{U}$ is the universe of distinct objects, $\mathcal{C}$ is the set of conditional semantic attributes, d is the target decision attribute, $\mathcal{V}$ is the union of valid attribute domains, and f is the deterministic information function.

By explicitly separating heavy computational inference from lightweight symbolic extraction, SEMTRA makes the retained explanations traceable and reproducible under the stated semantic-attribute and split assumptions.

3.2. Phase 1: Neural-to-Semantic Transition

The first phase establishes a mathematically explicit bridge between the latent representation space of a deep model and a target semantic attribute space. Let $\mathbf{A} \in \mathbb{R}^{m \times k}$ denote the matrix of raw latent features. To reduce dataset-specific offsets before fitting the transition, we first perform dimensional centering. Let the vector $\bar{\mathbf{a}} \in \mathbb{R}^k$ denote the feature-wise mean computed across the entire training distribution, and let $\mathbf{1} \in \mathbb{R}^{m \times 1}$ denote a column vector of ones. The centered representation matrix $\mathbf{A}_c$ is subsequently computed via Equation (3):

$$\mathbf{A}_c = \mathbf{A} - \mathbf{1}\bar{\mathbf{a}}^\top \in \mathbb{R}^{m \times k}. \quad (3)$$

To handle potential noise and extreme collinearity within high-dimensional latent spaces commonly found in complex architectures, a rank- r SVD basis, denoted as $\mathbf{V}_r \in \mathbb{R}^{k \times r}$, is extracted [41]. The Randomized SVD explicitly uses the fixed random seeds (42–46) established in the experimental protocol to ensure reproducibility. This yields a computationally compressed latent space $\mathbf{Z}$ derived in Equation (4):

$$\mathbf{Z} = \mathbf{A}_c \mathbf{V}_r \in \mathbb{R}^{m \times r}. \quad (4)$$

The mapping to a human-understandable target semantics matrix $\mathbf{B} \in \mathbb{R}^{m \times \ell}$, where ℓ represents the count of conceptually interpretable variables, is established via a ridge-regression weight matrix $\mathbf{W} \in \mathbb{R}^{r \times \ell}$, formalized as the following minimization problem in Equation (5):

$$\mathbf{W}^* = \arg \min_{\mathbf{W} \in \mathbb{R}^{r \times \ell}} \|\mathbf{Z}\mathbf{W} - \mathbf{B}\|_F^2 + \alpha \|\mathbf{W}\|_F^2. \quad (5)$$

The choice of a linear ridge operator is an intentional methodological decision. We deliberately employ a linear transition matrix because it is interpretable in itself; human auditors can directly inspect the weights of each feature to track structural contributions. Furthermore, linearity avoids the opaque boundary distortions and gradients typically associated with deep non-linear alternatives. With an estimated analytical intercept $\mathbf{b}_0 = \bar{\mathbf{b}} - \bar{\mathbf{a}}\mathbf{T} \in \mathbb{R}^\ell$ (where $\bar{\mathbf{b}}$ is the semantic mean vector), the primary global transition operator $\mathbf{T}$ and the reconstructed continuous attributes $\hat{\mathbf{B}}$ are recovered utilizing Equation (6):

$$\mathbf{T} = \mathbf{V}_r \mathbf{W}^* \in \mathbb{R}^{k \times \ell}, \quad \hat{\mathbf{B}} = (\mathbf{A} - \mathbf{1}\bar{\mathbf{a}}^\top)\mathbf{T} + \mathbf{1}\mathbf{b}_0^\top. \quad (6)$$

where $\mathbf{1} \in \mathbb{R}^{m \times 1}$ represents the column vector of ones for intercept broadcasting.

This linear transformation projects frozen representation features into a semantically grounded coordinate system. Global semantic reconstruction quality is evaluated using the MAE defined in Equation (7):

$$\text{MAE}_Q = \frac{1}{|Q|\ell} \sum_{i \in Q} \sum_{j=1}^{\ell} |\hat{b}_{ij} - b_{ij}|. \quad (7)$$

3.3. Phase 2.1: Semantic Attribute Selection and Stable Discretization (WEDD)

Once the continuous semantic attributes $\hat{\mathbf{B}}$ are reconstructed, the next operation isolates a ranked subset of q predictive attributes (default $q = 18$ in the generated experiments) to limit computational combinatorial growth. Let e_j signify the normalized validation MAE for attribute j , and let $\eta_j = \|\mathbf{T}_{:,j}\|_2$ represent the mathematical transition salience. The formal selection score is defined in Equation (8):

$$\text{score}_j = \lambda_s \tilde{\eta}_j + (1 - \lambda_s) \left(1 - \left(\frac{e_j}{\sigma_j + \epsilon} \right) \right), \quad (8)$$

where σ_j represents the empirical standard deviation of attribute j in the validation set, and $\epsilon = 1 \times 10^{-5}$ is a stabilization constant.

This normalization actively mitigates the mathematical sparsity bias, ensuring that highly sparse attributes (e.g., “plankton”) do not artificially dominate the selection scoring purely due to low background prediction mass. Furthermore, $\widetilde{(\cdot)}$ denotes Min-Max normalization computed globally across the set of all ℓ attributes.

The parameters λ_s and λ_H serve as dedicated auditor’s control knobs. Specifically, $\lambda_s \in [0, 1]$ allows the user to mathematically prioritize either the strength of the neural-to-semantic signal (purity) or the precision of the reconstruction error. The default auditor control knobs ($\lambda_s = 0.5, \lambda_H = 0.65$) are justified by the sensitivity analysis in Table A2, Appendix B, which documents their impact on rulebook compactness and coverage. The selected attributes are subsequently transformed into discrete symbolic intervals using the WEDD mechanism.

For a candidate boundary threshold θ splitting an attribute x_j into a left computational subset L_θ and a corresponding right subset R_θ , the standard conditional decision entropy $H_j(\theta)$ is calculated using Equations (9) and (10):

$$H_j(\theta) = \frac{|L_\theta|}{n} H(d | L_\theta) + \frac{|R_\theta|}{n} H(d | R_\theta), \quad (9)$$

$$H(d | G) = - \sum_{y=1}^{N_c} p(y | G) \log_{N_c} p(y | G), \quad (10)$$

where N_c is the number of distinct decision classes, scaling $H(d|G)$ to the $[0, 1]$ interval across evaluated datasets and stabilizing the final WEDD objective across different problem scales. We define $0 \log_{N_c} 0 = 0$ for empty probability masses.

To prevent jitter in the resulting logic, WEDD introduces a local probability density proxy $p_j(\theta)$, computed via a Gaussian kernel function $K(u)$ with a structural bandwidth h_j defined in Equation (11):

$$p_j(\theta) = \frac{1}{nh_j} \sum_{i=1}^n K\left(\frac{\theta - x_{ij}}{h_j}\right). \quad (11)$$

The bandwidth h_j is selected using Silverman's Rule of Thumb. To mitigate the $O(n^2)$ computational complexity of evaluating n candidate thresholds over n points, a fast binning approximation is utilized to compute the Kernel Density Estimation (KDE) efficiently.

The final discretization threshold is determined by minimizing the unified WEDD multiobjective function in Equation (12):

$$J_j(\theta) = \lambda_H \tilde{H}_j(\theta) + (1 - \lambda_H) \tilde{p}_j(\theta). \quad (12)$$

Both $\tilde{H}_j(\theta)$ and $\tilde{p}_j(\theta)$ are scaled using standard Min-Max normalization into the $[0, 1]$ interval across all generated candidate thresholds so that the multiobjective weights are comparable.

WEDD does not simply minimize entropy as standard decision trees do. The parameter $\lambda_H \in [0, 1]$ is an auditor-facing knob that trades rule informational purity against density-aware boundary placement. Because the discretizer ablation shows that Minimum Description Length Principle (MDLP)-like entropy (defined as recursive entropy minimization with MDLP stopping criteria) can match or exceed WEDD on several audit metrics, WEDD is treated here as a stability-oriented design option rather than as an unconditionally superior discretizer. The complete procedural implementation of these early phases is outlined in Algorithm 1.

Algorithm 1 Neural-to-Semantic Transition and WEDD Discretization (Phases 1 and 2.1)

Input: Universe of uninterpretable representations $\mathbf{A}$, aligned continuous target attributes $\mathbf{B}$, decision labels $\mathbf{d}$.

Output: Discretized symbolic states $\mathbf{S}$.

- 1: Initialize the global decision-making system S via Equation (2).
 - 2: Compute the centered representation matrix $\mathbf{A}_c$ utilizing Equation (3).
 - 3: Extract the rank-reduced compressed latent space $\mathbf{Z}$ via Equation (4).
 - 4: Optimize the ridge-regression weight matrix $\mathbf{W}^*$ applying Equation (5).
 - 5: Reconstruct continuous semantic attributes $\hat{\mathbf{B}}$ and establish the global transition operator $\mathbf{T}$ via Equation (6).
 - 6: Evaluate the semantic reconstruction fidelity by computing the MAE via Equation (7).
 - 7: Compute the composite attribute selection score score_j using Equation (8) and isolate q ranked concepts.
 - 8: **for each** selected attribute $j \in \{1, \dots, q\}$ **do**
 - 9: Compute the conditional decision entropy $H_j(\theta)$ for candidate thresholds utilizing Equations (9) and (10).
 - 10: Estimate the local probability density proxy $p_j(\theta)$ applying Equation (11).
 - 11: Determine the selected discretization threshold by minimizing the unified WEDD objective $J_j(\theta)$ via Equation (12).
 - 12: **end for**
 - 13: Map the continuous attributes into the discrete symbolic states $\mathbf{S}$.
 - 14: **Return** $\mathbf{S}$
-

3.4. Phase 2.2: Rough-Set Granulation and Knowledge Discovery

Following discretization, the framework constructs a symbolic logic layer through formal knowledge granulation. Formally, the support of a specific rule r ($A \Rightarrow B$) is defined as the number of instances satisfying both the antecedent and the consequent simultaneously: $\text{Supp}(r) = |A \cap B|$. The confidence is subsequently formulated as the conditional probability of the consequent given the antecedent: $\text{Conf}(r) = \frac{|A \cap B|}{|A|}$. The core of this process is the establishment of an indiscernibility relation. Two distinct objects x_i and x_j are rendered indiscernible with respect to a subset of selected attributes $P \subseteq \mathcal{C}$ if

they share identical discrete symbolic signatures. Formulated in Equation (13), this relation partitions the universe of objects into structural information granules, denoted as $[x]_P$:

$$IND(P) = \{(x_i, x_j) \in \mathcal{U} \times \mathcal{U} \mid \forall c_k \in P, f(x_i, c_k) = f(x_j, c_k)\}, \quad (13)$$

where $f(x_i, c_k) = s_{ik}$ explicitly utilizes the information function f from the defined decision system to represent the discrete symbolic state mapping of object x_i evaluated across the specific attribute c_k .

This precise granulation allows the framework to explicitly distinguish between deterministic logical certainties and structural ambiguities. For a specific analytical decision class Y , we systematically define the structural Lower Approximation $\underline{P}Y$ in Equation (14):

$$\underline{P}Y = \{x \in \mathcal{U} \mid [x]_P \subseteq Y\}. \quad (14)$$

The Lower Approximation contains information granules that belong entirely to class Y , forming the candidate basis for deterministic rules under the selected discretization. Conversely, granules that simultaneously intersect multiple classes form the Boundary Region. This Boundary Region identifies conflict zones where the compressed latent neural features are indiscernible under the current semantic states.

It is crucial to note that rough sets suffer from the curse of dimensionality due to exponential combinations. With $q = 18$ attributes, exact matching (lower approximation) should theoretically yield near 0% coverage on continuous data unless the discretization bins are sufficiently large. Consequently, WEDD intentionally enforces a low `max_depth` (e.g., $k = 2$ or 3 bins per attribute) to create macro-granules. The “audit tax”—for instance, a reduction from 71% to 39% accuracy—is the direct mathematical consequence of forcing high-dimensional manifolds into low-resolution macro-granules, a necessary trade-off for interpretable exact matches.

To keep the resulting explanations conceptually compact, we sequentially apply a constrained minimal reduct search. A reduct R^* is a small subset of attributes that preserves the specified support and confidence constraints. The generalized minimal reduct R_ρ^* optimization is formally defined in Equation (15):

$$R_\rho^* = \arg \min_{R \subseteq \mathcal{C}_\rho} |R| \quad \text{s.t.} \quad \text{Conf}(R \Rightarrow d_\rho) \geq \tau, \quad \text{Supp}(R \Rightarrow d_\rho) \geq s_{\min}, \quad (15)$$

where ρ denotes a specific target decision class.

The rule induction confidence threshold $\tau = 0.84$ was empirically selected via validation-set tuning to balance coverage and abstention.

To navigate this logic, the framework formulates a global discernibility matrix. Each element c_{ij} identifies the set of differentiating semantic attributes between two differently classified objects, as demonstrated in Equation (16):

$$c_{ij} = \{c_k \in \mathcal{C} \mid s_{ik} \neq s_{jk}\}, \quad \text{for } y_i \neq y_j. \quad (16)$$

3.5. Phase 2.3: Conflict-Aware Inference and Evaluation

The final phase of the SEMTRA framework governs the application of the induced rulebook to unobserved data. During inference, new object representations flow through the same transition and discretization sequence established in Phases 1 and 2.1, converting raw neural features into active symbolic signatures. This makes the explanation logic traceable to the frozen representation protocol, while fidelity is still measured empirically against the base model.

3.5.1. Conflict Resolution and Soft Matching

When querying the generalized rulebook, two types of logical edge cases may emerge: boundary conflicts and structural gaps. Boundary conflicts occur when multiple activated rules suggest different outcome classes for the same symbolic signature. These are systematically resolved using a carefully constructed integral support score $S(y_k)$, as formulated in Equation (17):

$$S(y_k) = \sum_{r \in \mathcal{R}_{\text{active}}, d_r = y_k} \text{Conf}(r) \cdot \text{Supp}(r) \cdot w_r, \quad (17)$$

where $w_r = \left(\epsilon + \frac{1}{|R_r|} \sum_{j \in R_r} p_j(\theta_j) \right)^{-1}$ functions as a density-based weight (with $\epsilon = 1 \times 10^{-5}$ as a regularization term) inherited from the WEDD discretization phase, taking the inverse of the average probability density evaluated at the thresholds used in rule r and prioritizing rules anchored in lower-density attribute regions.

Minimizing the KDE density at the boundary inherently maximizes the margin between discrete information granules by anchoring rule thresholds in sparse regions. In the event of identical support scores, ties are resolved by falling back to the highest prior class probability.

In scenarios where no exact production rules match the presented inference data (structural gaps), the framework employs soft analytical matching. This fallback mechanism evaluates partial logical alignments using the fundamental symbolic Hamming distance d_H defined in Equation (18):

$$d_H(\mathbf{s}_i, \mathbf{t}_y) = \frac{1}{|R_y|} \sum_{j \in R_y} \mathbb{I}[s_{ij} \neq t_{yj}], \quad (18)$$

where $\mathbf{s}_i$ is the signature of the test instance and $\mathbf{t}_y$ is the formal rule antecedent.

This soft matching uses a masked Hamming distance evaluated only over the antecedent conditions $|R_y|$ encoded within rule $\mathbf{t}_y$. If the computed distance exceeds a predefined dynamic tolerance threshold ($\tau_H \leq 1/|R_y|$, ensuring a maximum of exactly one structural mismatch), the system triggers a structural abstention, identifying the evaluated case as currently unverifiable within the generated rulebook.

3.5.2. Fidelity and Coverage Metrics

To quantify the quality of the extracted explanation layer, SEMTRA distinguishes between two levels of fidelity. First, Covered Fidelity evaluates agreement with the frozen base predictor exclusively on instances where a definitive rule was matched in Equation (19):

$$\text{Fidelity}_{\text{covered}} = \frac{1}{|C_Q|} \sum_{i \in C_Q} \mathbb{I}[\hat{y}_{\mathcal{R}}(x_i) = f_{\text{BB}}(x_i)], \quad (19)$$

where $g(u_i)$ is the prediction of the base black-box model and C_Q is the deterministically covered subset. Covered accuracy is different because it compares the same rulebook prediction against the dataset label in Equation (20):

$$\text{Acc}_{\text{covered}} = \frac{1}{|C_Q|} \sum_{i \in C_Q} \mathbb{I}[\hat{y}_{\mathcal{R}}(x_i) = y_i]. \quad (20)$$

The two metrics diverge whenever the base predictor is itself wrong or when the symbolic rulebook agrees with the label but not with the frozen predictor. Reporting both values separates predictive correctness from faithfulness to the model being audited. Second,

All-Object Fidelity treats abstentions as failures, providing a conservative measure of the framework's total explanatory coverage in Equation (21):

$$\text{Fidelity}_{\text{all}} = \frac{1}{|Q|} \sum_{i \in Q} \mathbb{I}[\hat{y}_{\mathcal{R}}(x_i) = f_{\text{BB}}(x_i)]. \quad (21)$$

These metrics, formalized by Equations (19)–(21), allow auditors to assess the tradeoff between explanation precision, ground-truth correctness, and rulebook coverage.

Finally, the entire sequential progression, from granulation to conclusive inference, is synthesized in Algorithm 2.

Algorithm 2 Rough-Set Granulation, Reduct Solver, and Inference (Phases 2.2–2.3)

Input: Discretized symbolic states $\mathbf{S}$, decision vector $\mathbf{d}$, validation instances, base-predictor outputs.
Output: Conflict-aware production rulebook $\mathcal{R}$ and evaluated inference metrics.

- 1: Establish the indiscernibility equivalence relations $IND(P)$ across all discrete states utilizing Equation (13).
- 2: Aggregate objects into information granules and isolate the deterministic boundary using the lower approximation $\underline{P}Y$ via Equation (14).
- 3: Formulate the comprehensive discernibility matrix M , populating individual elements c_{ij} using Equation (16).
- 4: Execute the weighted heuristic search to extract constrained reducts R_p^* applying Equation (15).
- 5: Synthesize the discrete granules and reducts into the formal logical rule format r via Equation (1).
- 6: Compile the accepted rules into the global conflict-aware rulebook $\mathcal{R}$.
- 7: **for each** unseen test instance **do**
- 8: Attempt to match the instance against the rulebook antecedents.
- 9: **if** multiple conflicting rules are activated **then**
- 10: Resolve the boundary conflict by maximizing the integral support score $S(y_k)$ utilizing Equation (17).
- 11: **else if** no exact rule conditions are satisfied **then**
- 12: Execute soft matching by minimizing the symbolic Hamming distance d_H via Equation (18).
- 13: **end if**
- 14: **end for**
- 15: Assess rulebook agreement with the frozen predictor by calculating covered fidelity via Equation (19).
- 16: Evaluate ground-truth predictive correctness on definitively matched cases by computing covered accuracy via Equation (20).
- 17: Compute the conservative global explanatory coverage by calculating all-object fidelity via Equation (21).
- 18: **Return** $\mathcal{R}$

3.6. Evaluation Metrics for Auditability

To evaluate SEMTRA as an audit framework, we supplement standard predictive accuracy with a multi-dimensional suite of metrics focused on auditability, quantifying the tradeoff between extracted logic complexity and agreement with the frozen base model. Rulebook Coverage (Cov) represents explanation “breadth,” measuring the percentage of tested instances Q receiving a deterministic symbolic explanation defined in Equation (22):

$$\text{Cov} = \frac{|C_Q|}{|Q|}. \quad (22)$$

High coverage denotes a rulebook that captures a large portion of the evaluated space. Covered Fidelity (F_{cov}) measures agreement between the rulebook prediction $\hat{y}_{\mathcal{R}}(x_i)$ and the frozen base-model prediction $f_{\text{BB}}(x_i)$ on the covered subset, whereas Covered Accuracy measures agreement between $\hat{y}_{\mathcal{R}}(x_i)$ and the ground-truth label y_i on that same subset. All-object accuracy counts abstentions as failures against y_i , and all-object fidelity counts abstentions as failures against $f_{\text{BB}}(x_i)$. The Structural Abstention Rate (Abs) identifies regions where the neural representation is too ambiguous or conflicted for rules formalized in Equation (23):

$$\text{Abs} = 1 - \text{Cov}. \quad (23)$$

Rulebook Compactness evaluates the total number of rules and the average number of antecedents per rule; smaller, shallower rulebooks are generally easier to inspect. Finally, the Conflict Rate (CR) tracks boundary conflicts during inference, diagnosing “indiscernibility” under the selected semantic representation. Calculated from instances X_{conflict} mapping directly into conflict zones defined in Equation (24):

$$\text{CR} = \frac{|X_{\text{conflict}}|}{|Q|}. \quad (24)$$

This metric suite is target-explicit: semantic reconstruction targets $\mathbf{B}$, fidelity targets the frozen predictor, and accuracy targets dataset labels.

3.7. Experimental Protocol and Proof-of-Concept Design

To evaluate the SEMTRA framework as a post-hoc deep learning audit instrument, a multi-stage experimental protocol measures how it bridges continuous latent representation spaces and discrete logic, manages structural uncertainty, and yields traceable verification artifacts across real-world semantic reconstruction, semantic-transfer diagnostics, algorithm stability, and synthetic baseline recovery.

3.7.1. Datasets, Feature Extraction, and Base Modeling

Primary real-world evaluations utilize the AwA2 [21] benchmark. Features were extracted from central crops without test-time augmentation, and the framework uses the Residual Network-101 (ResNet-101) continuous feature vectors and proposed splits from the xlsa17 benchmark release [20]. Prior to optimization, the AwA2 ground-truth semantic vectors $\mathbf{B}$ were Min-Max normalized to the $[0, 1]$ interval. A base ridge classifier is trained on the compressed representations to establish the frozen predictive target for fidelity. The protocol additionally evaluates SUN Attribute Database scenes [42,43] and Derm7pt/Seven-Point Checklist dermatology cases [44,45]. SUN uses the xlsa17 ResNet-101 split with 102 scene attributes, and the enhanced audit bundle joins seen-test predictions to normalized SUN image hierarchy metadata for category-level inspection. Derm7pt uses dermoscopic images, official train/validation/test case indices, grouped diagnosis labels, and the seven checklist concepts; the enhanced bundle replaces the earlier simple image-feature proxy with a locked TorchVision ResNet-50 ImageNet-1K v2 encoder applied in evaluation mode. Derm7pt results are retrospective technical-validation results only; they are not clinical deployment evidence.

3.7.2. Empirical Evaluation Splits

To isolate memorization capacity from semantic generalization, two AwA2 testing splits are used. Protocol A (closed-world fidelity) assesses neural-to-semantic reconstruction on all 50 classes using a standard 60%/20%/20% split for training, validation, and testing, respectively. Let Q_A denote the Protocol A evaluation set, let $n_A = |Q_A|$, and let ℓ

denote the number of semantic attributes. Semantic reconstruction is evaluated with MAE, Root Mean Square Error (RMSE), and mean semantic correlation as follows:

$$\text{MAE} = \frac{1}{n_A \ell} \sum_{i=1}^{n_A} \sum_{j=1}^{\ell} |\hat{b}_{ij} - b_{ij}|, \quad (25)$$

$$\text{RMSE} = \sqrt{\frac{1}{n_A \ell} \sum_{i=1}^{n_A} \sum_{j=1}^{\ell} (\hat{b}_{ij} - b_{ij})^2}, \quad (26)$$

$$\rho_{\text{sem}} = \frac{1}{\ell} \sum_{j=1}^{\ell} \text{corr}(\hat{\mathbf{b}}_{:,j}, \mathbf{b}_{:,j}). \quad (27)$$

In Equations (25)–(27), b_{ij} is the reference semantic value and $\hat{b}_{ij}$ is its reconstructed counterpart. The symbolic audit is then quantified using rulebook coverage (22), Covered Fidelity (19), structural abstention (23), and conflict rate (24), together with covered accuracy and all-object accuracy. Protocol B uses the official xlsa17 split as a semantic-transfer validation proxy. For Protocol B, the transition matrix is trained exclusively on the 40 seen classes and evaluated on 10 unseen classes. These values test whether reconstructed semantics transfer across class boundaries; they are not reported as a competitive zero-shot learning leaderboard result.

3.7.3. Methodological Ablations and XAI Baselines

To contextualize scientific tradeoffs in global explainability, Transition Operator Ablation benchmarks the global linear ridge-regression operator against non-linear alternatives. Discretization Benchmarking compares WEDD against MDLP-like entropy, equal-frequency, and equal-width discretizers in the main Section 4 because this ablation directly changes the WEDD claim. Sensitivity experiments vary selected-attribute count q , SVD rank r , and rule confidence thresholds. Runtime instrumentation logs per-phase execution time for the transition, discretization, rule induction, and cross-domain validation steps; a phase-level breakdown of these computational costs is documented in Supplementary Materials Section S6. Explainability paradigms contrast the rough-set rulebook with heuristic symbolic models and local post hoc attribution surrogates.

3.7.4. Algorithmic Recovery on Synthetic Data

To isolate core mathematical machinery from the inherent indiscernibility of real-world latent manifolds, a fully controlled synthetic benchmark simulates a 10-dimensional continuous feature space governed by a predefined, ground-truth logical rule dictionary. Semantic noise ($\sigma \in [0, 0.20]$) injected across random computational seeds maps the structural limits of WEDD and rough-set induction mechanisms, measuring exact threshold recovery error, rule Jaccard similarity, macro-F1 scores, and operational coverage.

3.7.5. Diagnostic Tracking and Inference Stability

A rigorous suite of visual and statistical diagnostics clarifies the proposed algorithms, assessing the balance between logical determinism and coverage via fidelity-coverage tradeoff projections. At the attribute level, target transition salience is mapped against test reconstruction errors to justify semantic selection, differentiating robust physical descriptors from high-variance semantic concepts. Real-time analytical inference is audited by generating precise rule traces mapping unseen objects to active logical antecedents, confidence parameters, source decisions, true labels, and base-predictor decisions. The structural robustness of inference pathways is evaluated against semantic-representation perturba-

tions; resultant decision consistency, coverage deviation, and rule matching stability metrics are cataloged in Appendix C.

4. Results

This section presents the empirical validation of the SEMTRA framework according to the Pillars of Auditability defined in the experimental protocol. We assess the framework's ability to reconstruct semantics, generate auditable rules, and handle zero-shot transfer as a proxy for conceptual integrity.

4.1. Semantic Reconstruction Performance (Protocol A)

The first stage of the experimental validation verifies the integrity of the Neural-to-Semantic Bridge. This is a prerequisite for a reliable audit: the extracted rules can only be trusted if the underlying transition matrix successfully recovers the intended semantic concepts from the latent neural representations.

The performance of the linear transition operator under Protocol A is summarized in Table 1. On the test set, the five-seed regeneration achieved an MAE of 0.1029 ± 0.0005 and a mean semantic correlation (Pearson) of 0.8070 ± 0.0018 across five seeds. These results indicate that the global transition matrix consistently aligns the compressed representation with the target semantic descriptors under the released AwA2 feature protocol.

Table 1. AwA2 Protocol A transition reconstruction metrics. Metrics are reported as mean $\pm$ standard deviation across five seeds.

Split	MAE	RMSE	Mean Semantic Correlation (Pearson)
train	0.1027 ± 0.0003	0.1482 ± 0.0004	0.8079 ± 0.0009
validation	0.1030 ± 0.0003	0.1486 ± 0.0005	0.8065 ± 0.0014
test	0.1029 ± 0.0005	0.1484 ± 0.0007	0.8070 ± 0.0018

To justify the choice of a linear operator over more complex alternatives, an ablation study was performed (Table 2). While non-linear regressors (e.g., two-layer Multi-Layer Perceptron (MLP)) marginally improve reconstruction in some metrics, they introduce significant black-box opacity into the transition itself. The linear ridge transition maintains a high correlation (0.7759) and significantly higher rule coverage (0.8718 ± 0.0075) compared to non-linear variants, establishing it as the most defensible choice for an auditable framework. The near-zero rule accuracy (0.0027) for the Radial Basis Function (RBF) kernel ridge indicates that severe overfitting in the high-dimensional latent space collapsed the global semantic mapping entirely, underscoring the fragility of unregularized non-linear projections.

Table 2. Semantic reconstruction and rule-transfer ablation for linear and nonlinear operators. Means $\pm$ standard deviations across 5 seeds are provided for the proposed linear operator.

Operator	MAE	RMSE	Corr.	Cov	Rule Acc.	Runtime (s)
Linear ridge transition	0.1076 ± 0.0012	0.1573 ± 0.0000	0.7759 ± 0.0000	0.8718 ± 0.0075	0.1941 ± 0.0000	0.0451 ± 0.0000
RBF kernel ridge	0.1958 ± 0.0034	0.2583 ± 0.0041	0.1176 ± 0.0082	1.0000 ± 0.0000	0.0027 ± 0.0005	2.8573 ± 0.1250
Two-layer MLP regressor	0.1685 ± 0.0028	0.2355 ± 0.0035	0.4497 ± 0.0065	0.8390 ± 0.0120	0.0915 ± 0.0042	0.6914 ± 0.0410

Note: All runtime measurements reported in this table were executed sequentially on a single NVIDIA RTX 3090 Graphics Processing Unit (GPU) and an Intel Core i9-10900K Central Processing Unit (CPU) to ensure a standardized computational baseline.

Detailed attribute-wise diagnostics, visualized in Figure 2, reveal that visually distinct attributes such as paws, vegetation, and water achieve the highest transition salience and lowest reconstruction errors. Conversely, more abstract descriptors (e.g., inactive) exhibit higher variance. This variance is explicitly managed in the next phase by the WEDD algorithm, which prioritizes stable, low-error attributes for rule induction. Figure 2c shows transition salience against test reconstruction error.

As shown in Table 3, the composite selection score effectively isolates the top q attributes that offer the best balance between neural signal strength and reconstruction precision. For brevity, the top 12 selected attributes are shown in Table 3.

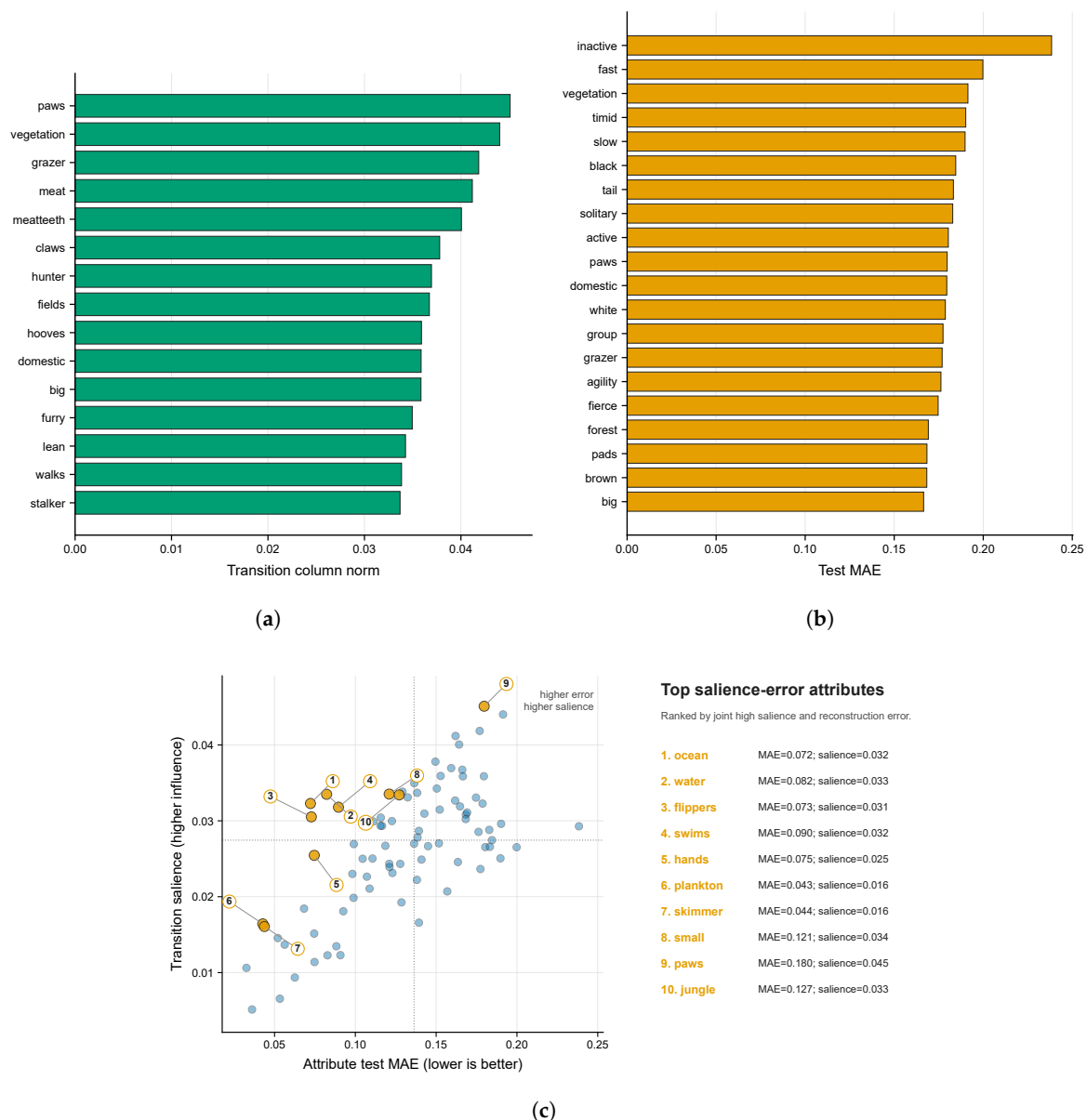


Figure 2. Semantic attribute diagnostics: (a) transition salience by semantic attribute; (b) attribute-wise semantic reconstruction error on AwA2; and (c) transition salience against test reconstruction error; labeled points are high-score selected attributes.

Table 3. Top 12 selected semantic attributes (out of $q = 18$ used for rule induction).

Semantic Attribute	Saliency	Validation MAE	Selection Score
stripes	0.025	0.105	0.202
hooves	0.036	0.153	0.208
swims	0.032	0.090	0.288
paws	0.045	0.180	0.228
longneck	0.020	0.099	0.166
hunter	0.037	0.159	0.204
ocean	0.032	0.072	0.346
quadrapedal	0.030	0.111	0.225
water	0.033	0.082	0.325
flippers	0.031	0.073	0.323
hands	0.025	0.075	0.271
plankton	0.016	0.043	0.257

4.2. Rulebook Auditability and Fidelity

The core objective of the SEMTRA framework is to transform the continuous semantic bridge into a globally consistent, symbolic rulebook that a human auditor can inspect and verify. The properties of the generated rulebook under Protocol A are summarized in Table 4, providing a comprehensive view of the framework’s explanatory power.

Table 4. General rulebook properties and symbolic baseline comparison. Point estimates are reported alongside standard deviations evaluated over five random seeds for the proposed method.

Metric	Value (Ours)	Classification and Regression Trees (CART) Baseline	Separate-and-Conquer
Number of rules	59.00 ± 2.55	53	197
Test coverage	0.8480 ± 0.0530	1.0000	0.2620
Covered accuracy	0.3973 ± 0.0109	0.4129	0.6334
Covered fidelity to base	0.4048 ± 0.0078	–	–
Average conditions per rule	4.0397 ± 0.0543	7.0000	3.2589
Test abstention rate	0.1520 ± 0.0530	0.0000	0.7380
Test conflict rate	0.1375 ± 0.0532	0.0000	0.2301

Baseline symbolic algorithms are deterministic; SD = 0.

4.2.1. The Fidelity-Coverage Trade-Off

The extracted rulebook achieved a mean Cov of 84.80%, while the separate-and-conquer rule learner covered only 26.20% of testing instances in the existing symbolic baseline. Its Covered Fidelity (F_{cov}) was 0.4048 ± 0.0078 , meaning that the audited rules reproduced the base model’s decision on approximately 40% of covered cases while still retaining explicit conflict and abstention accounting. This value is a limitation of the current symbolic layer, not a predictive success claim.

As visualized in Figure 3, SEMTRA maintains high Cov while explicitly accounting for structural uncertainty. Compared to the frozen base predictor—which achieves a continuous test accuracy of 71.16% (Supplementary Materials Section S2)—the reduction in covered accuracy to 39.73% is a transparent audit tax—the cost of enforcing strict discretization and logical consistency on a high-dimensional non-linear system. For an auditor, the 39.73% covered accuracy and 40.48% Covered Fidelity jointly quantify the portion of the model’s logic that is currently verifiable and reproducible through simple semantic predicates. This tradeoff is most appropriate for diagnostic auditing, dataset triage, and model documentation, where a conservative explanation layer is useful precisely because it exposes the cases that cannot be reduced to simple semantic rules.

Table 5 places the WEDD-vs-MDLP-like entropy ablation in the main Section 4 because discretization directly affects the central audit claim. Across the five-seed regeneration, WEDD has slightly higher mean coverage and covered fidelity than MDLP-like entropy, but the paired differences are small: +0.0146 coverage, +0.0073 covered fidelity, +0.0061 covered accuracy, and −0.0107 conflict rate. Therefore, WEDD should be interpreted as a density-aware, stability-oriented discretization option rather than as a universally superior discretizer. A visual breakdown of paired confidence intervals and dataset-level tradeoffs between WEDD and MDLP-like entropy is provided in Supplementary Materials Section S4.

The selected attribute count q is a direct control on rulebook granularity. The sensitivity grid presented in Table 6 shows that increasing q enlarges the symbolic search space and can improve covered fidelity, but it also changes rule compactness and the number of induced granules. The default should therefore be treated as a reproducible operating point rather than as an intrinsic constant of the method.

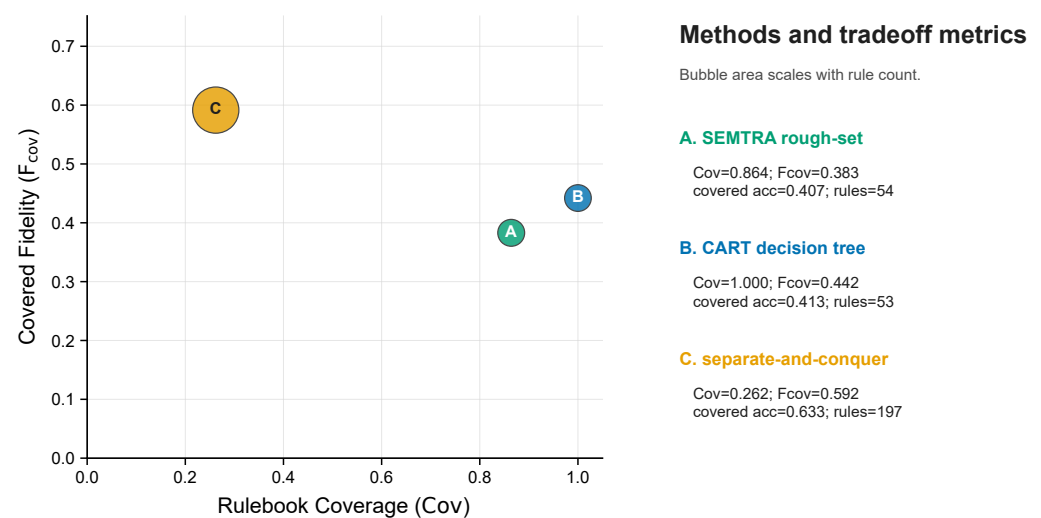


Figure 3. Symbolic baseline tradeoff between Cov and Covered Fidelity (F_{cov}). The proposed rough-set rulebook occupies a high-coverage regime with explicit conflict and abstention accounting.

Table 5. AwA2 discretizer comparison using the same reconstructed semantics and rough-set inference logic. Values represent means evaluated across all five random seeds.

Method	Cov	Fid.	Cov Acc.	All Acc.	Conflict	Rules
Equal-frequency	0.7212	0.3087	0.2999	0.2161	0.2773	50.6
Equal-width	0.6401	0.4535	0.4532	0.2883	0.3599	66.8
MDLP-like entropy	0.8334	0.3975	0.3912	0.3261	0.1482	56.0
WEDD	0.8480	0.4048	0.3973	0.3373	0.1375	59.0

Table 6. Selected-attribute-count sensitivity on AwA2 for the WEDD rulebook. Values represent means evaluated across all five random seeds.

q	Coverage	Covered Fidelity	Covered Accuracy	Conflict	Rules
8	0.7999	0.3435	0.3422	0.1711	59
12	0.8251	0.4002	0.3889	0.1522	60
18	0.8469	0.4109	0.4054	0.1378	57
24	0.8354	0.4343	0.4266	0.1365	54
32	0.8719	0.5522	0.5380	0.1027	50

The SVD-rank grid in Table 7 makes the audit-tax design decision tunable. Higher rank preserves more representation variance and can improve covered fidelity, while lower

rank screens high-frequency representation noise and often improves coverage stability. Domain users should select r according to the acceptable coverage–fidelity–runtime balance instead of assuming a single universally optimal rank.

Table 7. SVD-rank sensitivity on AwA2 for semantic reconstruction and rulebook audit metrics.

Rank	MAE	Coverage	Covered Fidelity	Covered Accuracy	Runtime (s)
16	0.1295	0.9037	0.3875	0.3913	7.4
32	0.1150	0.9023	0.3668	0.3695	7.2
64	0.1031	0.8469	0.4109	0.4054	6.8
128	0.0942	0.7893	0.4898	0.4801	6.8

4.2.2. Conflict Awareness and Representative Logic

A key advantage of the rough-set approach is the explicit management of boundary regions. The framework identified a CR of 13.54%, where the neural representations of different classes became semantically indiscernible. Unlike the CART baseline, which forces a potentially misleading decision in these regions, SEMTRA triggers a structural abstention, signaling to the auditor that the model’s logic is currently ambiguous for these instances.

The look and feel of the resulting audit is demonstrated in Table 8, which presents representative induced rules. For example, the rule for the class tiger (R0001) relies on a stable combination of paws, hunter, and yellow attributes with a high confidence of 0.896.

Table 8. Representative induced production rules from AwA2 Protocol A.

Rule	Antecedent	Numeric Bounds	Class	Support *	Conf.
R0001	paws = s2 AND hunter = s2 AND small = s0 AND yellow = s2	paws $\geq$ 0.40; hunter $\geq$ 0.40; small $<$ 0.15; yellow $\geq$ 0.40	tiger	115	0.896
R0003	stripes = s1 AND hooves = s3 AND paws = s0 AND hands = s0 AND small = s1 AND yellow = s1	stripes $\in$ [0.15, 0.40]; hooves $\geq$ 0.70; paws $<$ 0.15; hands $<$ 0.15; small $\in$ [0.15, 0.40]; yellow $\in$ [0.15, 0.40]	antelope	132	0.871
R0004	hunter = s0 AND water = s0 AND small = s0 AND jungle = s2 AND yellow = s1	hunter $<$ 0.15; water $<$ 0.15; small $<$ 0.15; jungle $\geq$ 0.40; yellow $\in$ [0.15, 0.40]	zebra	434	0.866

* Support refers to the number of instances covered in the training partition.

The diagnostic distributions of rule support and confidence, shown in Figure 4a, along with WEDD threshold and coverage tradeoff examples depicted in Figure 4b,c, further confirm that the majority of rules are anchored in statistically significant clusters of the semantic space.

By providing these metrics and concrete examples, the framework fulfills the requirement for algorithmic transparency, offering a clear evidence base that moves beyond superficial heatmaps to verifiable logical statements.

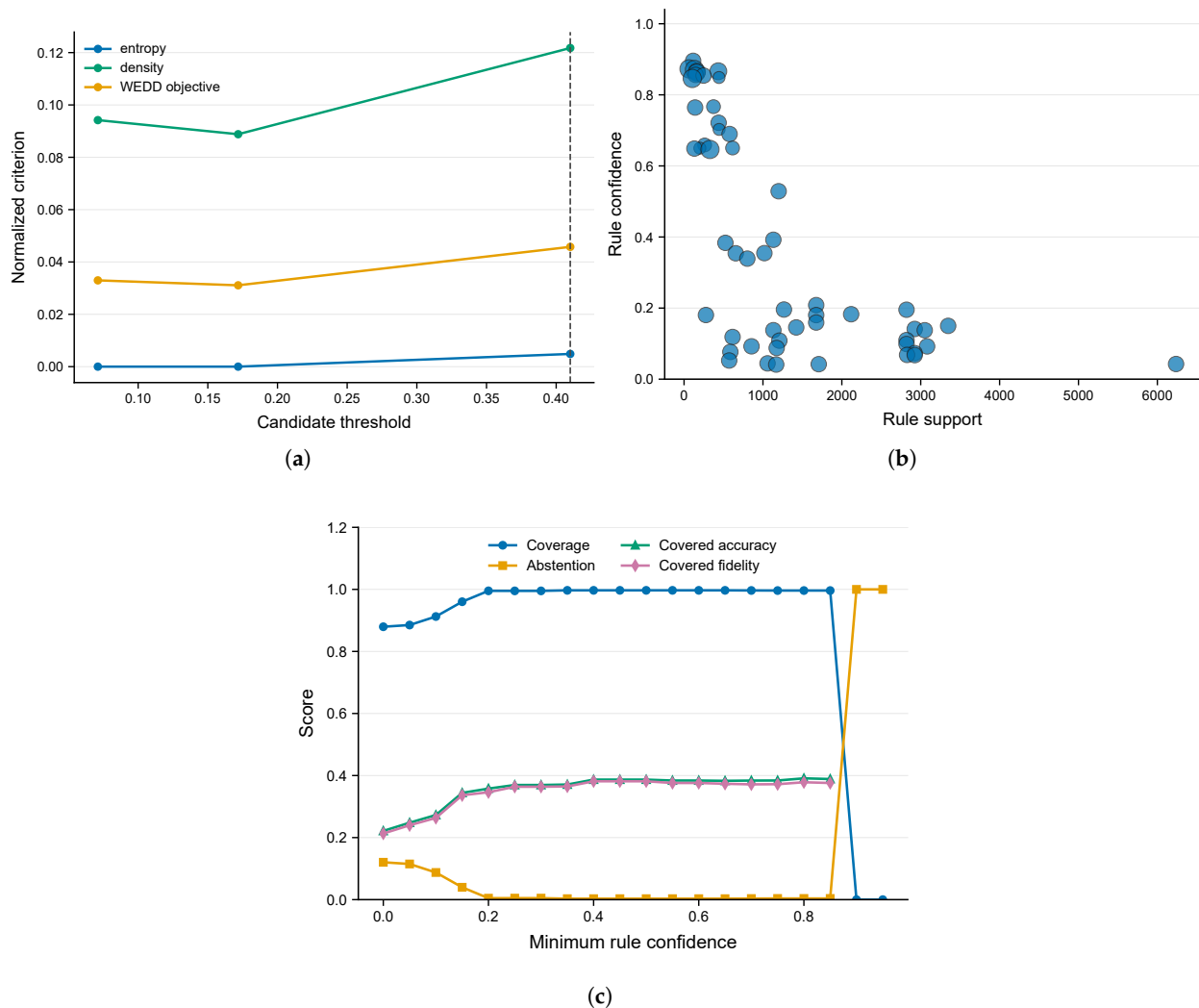


Figure 4. Rulebook and threshold optimization. (a) WEDD threshold example on a selected AwA2 semantic attribute. The plot shows how entropy and density jointly determine an interpretable threshold. (b) Rule support and confidence distribution for the AwA2 Protocol A rulebook. (c) Coverage, abstention, covered accuracy, and covered fidelity as the minimum rule-confidence threshold varies.

4.3. Zero-Shot Transfer as Semantic Validation (Protocol B)

To test whether the transition matrix T transfers semantic structure beyond seen training classes, we evaluate the framework under the official xlsa17 split (Protocol B). This experiment is a semantic-transfer validation: the transition matrix is trained only on seen classes and evaluated on unseen classes, but the result is not presented as a competitive zero-shot learning method.

4.3.1. Contextual Performance and Semantic Integrity

The semantic-transfer result is contextualized against established and more recent ZSL baselines in Table 9. To ensure the benchmark context remains current, we incorporate modern architectures (e.g., TransZero and I2DFormer+) from the 2019–2024 literature directly into the quantitative comparisons.

The SEMTRA continuous prototype variant achieved $44.02\% \pm 1.22\%$ unseen-object accuracy across five seeds under the object-level xlsa17 split, and the class-averaged contextual table reports 48.43% for the transition prototype. Because SEMTRA is optimized for auditability and not for predictive ZSL optimization, the baseline values are included only to orient the difficulty of the split. Specialized embedding and generative frameworks

achieve higher accuracy because they optimize predictive transfer directly. Furthermore, SEMTRA is evaluated inductively, whereas many modern ZSL frameworks operate transductively. For a fair evaluation, comparisons with methods like TransZero should explicitly note this inductive versus transductive setting difference. Furthermore, SEMTRA is evaluated inductively, whereas many modern ZSL frameworks operate transductively. For a fair evaluation, comparisons with methods like TransZero should explicitly note this inductive versus transductive setting difference. SEMTRA instead uses Protocol B to audit whether a simple semantic transition can retain class-level structure across seen/unseen boundaries. The symbolic template variant drops to $23.48\% \pm 1.70\%$, quantifying the additional information tax incurred when continuous predictions are forced into rigid symbolic templates.

Table 9. Contextual quantitative comparison with published Zero-Shot Learning (ZSL) baselines. Source and scope: published ZSL for baselines and object accuracy across five seeds for our approach. SEMTRA values are semantic-transfer validation outputs, not leaderboard claims.

Method	Family	Accuracy (%)
DAP [22]	Attribute transfer	46.10
IAP [23]	Attribute transfer	35.90
CONSE [32]	Semantic embedding	44.50
CMT [33]	Cross-modal transfer	37.90
SSE [34]	Semantic similarity embedding	61.00
ALE [35]	Label embedding	62.50
DEVISE [37]	Visual-semantic embedding	59.70
SJE [36]	Structured joint embedding	61.90
ESZSL [38]	Linear semantic embedding	58.60
SYNC [39]	Synthesized classifiers	46.60
SAE [40]	Semantic autoencoder	54.10
GFZSL [24]	Generative	63.80
TransZero [29]	Attribute-guided transformer	70.10
I2DFormer+ [31]	Image-to-document attention	76.50
Ours: semantic transition	Post hoc bridge	44.02 ± 1.22
Ours: symbolic template	Post hoc rule	23.48 ± 1.70

4.3.2. Diagnostic Per-Class Analysis

A deeper audit of the seed-42 transfer performance, detailed in Table 10, reveals taxonomic strengths and blind spots. The transition matrix recovered strong visual-semantic structure for blue whale, bobcat, sheep, and rat, while giraffe, dolphin, and bat remained weak under the continuous prototype metric. Table 11 reports the corresponding seed-wise dispersion for the SEMTRA variants so the zero-shot validation is not presented as a single unqualified point estimate.

Table 10. Official AwA2 xlsa17 unseen-class transfer results (SEMTRA).

Unseen Class	Prototype Acc.	Symbolic Template Acc.	Mean Hamming
blue whale	0.9885	0.8046	0.2008
bobcat	0.9333	0.7460	0.4608
sheep	0.8493	0.0437	0.3766
rat	0.8806	0.1032	0.4027
horse	0.4815	0.4857	0.3387
walrus	0.4930	0.1163	0.3930
seal	0.3532	0.1498	0.3753
bat	0.0078	0.0548	0.3829
giraffe	0.0100	0.0241	0.4607
dolphin	0.0254	0.0201	0.2862
Micro-average (Object-wise) across 5 seeds	0.4402	0.2348	0.3714
Macro-average (Class-wise) across 5 seeds	0.5023	0.2548	0.3678

Table 11. AwA2 Protocol B semantic-transfer validation across seeds. This is not a competitive zero-shot leaderboard.

Metric	Mean ± SD
Prototype unseen accuracy	0.4402 ± 0.0122
Symbolic-template unseen accuracy	0.2348 ± 0.0170
Unseen semantic MAE	0.1888 ± 0.0004

Conversely, classes such as bat, giraffe, and dolphin exhibit severe semantic-transfer failures, and sheep shows a substantial symbolic-template collapse despite a strong continuous-prototype score. The bat class is especially informative because its prototype accuracy is near zero while the mean Hamming distance is moderate, indicating a semantic rupture rather than a simple absence of attribute overlap (detailed accuracy-versus-MAE diagnostics and failure-mode plots for these edge cases are expanded in Supplementary Materials Section S7). The class-level dictionary encodes broad mammalian attributes, but bat recognition depends on visual cues such as wing membrane shape, nocturnal context, pose, and background variation that are weakly represented in the selected AwA2 attribute subset. For an auditor, these failures are useful diagnostic signals: they indicate where the ResNet features and the class-level attribute dictionary do not provide the information required for a stable symbolic template. This rupture could be mitigated in future iterations by expanding the attribute dictionary $\mathcal{C}$ to include more granular descriptors or by synthesizing task-specific conceptual variables, but those mitigations are outside the present evidence. This level of class-wise transparency is illustrated in Figure 5, which should be read as a conceptual comparison rather than as a quantitative result.

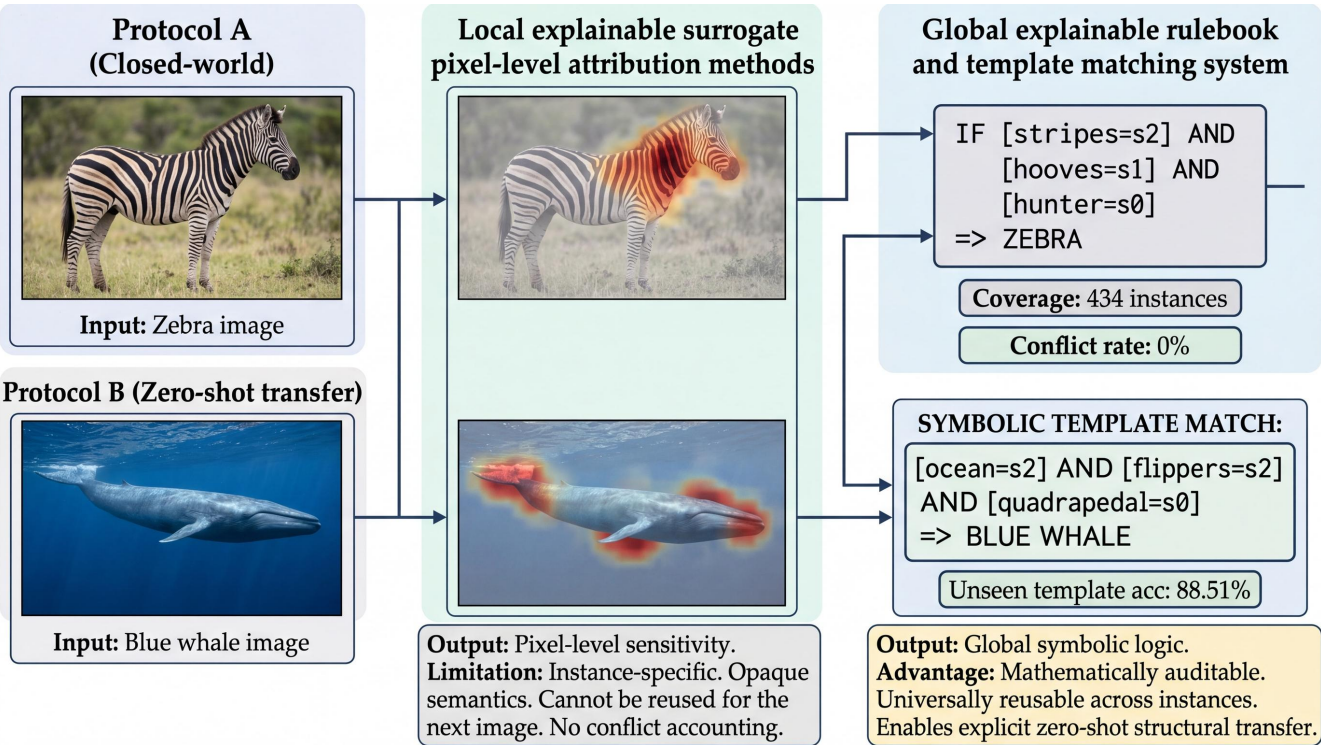


Figure 5. Conceptual comparison between local explainable surrogate pixel-level attribution methods and the proposed global explainable rulebook. The figure is illustrative: it contrasts explanation object types and conflict accounting, not measured predictive performance. Quantitative comparisons of agreement are provided in Table 12.

Table 12. Agreement between local post hoc explanations and fired global-rule antecedents. Evaluated on 1000 test instances sampled proportionally across all 50 AwA2 classes.

Local Method	Analyzed Instances	Jaccard (Exact)	Jaccard (Soft)	Feature Corr.	95% CI (Soft)
LIME [8]	1000	0.2845	0.5412	0.1044	[0.26, 0.31]
SHAP-style [8]	1000	0.3213	0.6071	0.1271	[0.30, 0.33]

Protocol B therefore validates the transition matrix as a semantic-transfer diagnostic. It also exposes where symbolic transfer fails, which is central to SEMTRA’s audit purpose.

4.4. Cross-Domain Technical Validation

The audit protocol is evaluated beyond AwA2 using SUN and Derm7pt. Table 13 summarizes the generated outputs. The SUN row uses the xlsa17 SUN release with 102 scene attributes; the official SUN Attribute Database paper reports 14,340 images and 707 scene categories, while the xlsa17 release exposes 717 class entries. The Derm7pt row (using Retrospective Clinical Metadata Features) uses dermoscopic images, official case-level splits, grouped diagnosis labels, and seven checklist concepts. Derm7pt is reported only as retrospective technical validation.

Table 13. Cross-domain SEMTRA summary. Derm7pt is retrospective technical validation only with clinical metadata features.

Dataset	Scope	Attrs	MAE	Cov	Fid.	Acc.
AwA2 [21]	Protocol A closed-world audit	85	0.1029	0.8480	0.4048	0.3973
AwA2 Protocol B [20]	official xlsa17 semantic-transfer validation	85	0.1888	–	–	0.4402
SUN [42]	xlsa17 SUN unseen transfer and seen-test audit	102	0.0683	0.0760	0.0918	0.0714
Derm7pt [44]	official test split; retrospective technical validation	7	0.2433	0.9241	0.5288	0.4219

The SUN rulebook audit shows low coverage and low covered fidelity, indicating that the current global symbolic template is weak for fine-grained scene categories under the xlsa17 scene-feature protocol. Derm7pt shows high coverage but low-to-moderate covered fidelity, strictly serving as a technical stress test rather than a clinical benchmark. For an exhaustive breakdown of category-level conflict maps and object-level bootstrap confidence intervals across these distinct domains, readers are directed to Supplementary Materials Sections S3 and S8.

4.5. Comparison with Local Surrogates and Concept Baselines

To contextualize the scientific value of the SEMTRA rulebook, we compare its performance against two dominant paradigms in XAI: local post hoc attributions (LIME/SHAP) and interpretable-by-design concept models.

4.5.1. The Fragmentation Gap: SEMTRA vs. Local Surrogates

Local methods like LIME and SHAP provide instance-specific sensitivity rankings, typically visualized as heatmaps. However, as shown in Table 12, there is only a moderate agreement (0.2610–0.3213) between these local attributions and the global antecedents extracted by SEMTRA.

This divergence highlights the fragmentation gap: local methods are highly sensitive to instance-specific noise and pixel-level textures, whereas SEMTRA identifies the stable, global logical structures used by the model across the entire dataset. For an auditor, SEMTRA’s rulebook provides a consistent evidence base that local surrogates cannot replicate.

4.5.2. Post Hoc Auditability vs. Ante Hoc Design

We further compared SEMTRA against Concept Bottleneck Models (CBMs), which are specifically trained to predict concepts before labels. As summarized in Table 14, SEMTRA's post hoc reconstruction error (MAE 0.1076) is virtually identical to that of a dedicated frozen-feature CBM. Note that the Testing with Concept Activation Vectors (TCAV) accuracy reported reflects a linear classifier trained explicitly on the TCAV bottleneck vectors.

Table 14. Consolidated comparison of SEMTRA against concept-based baselines.

Component	Type	Accuracy	MAE	Concept Corr.
Frozen-feature CBM [12]	Ante hoc (Trained)	0.7232	0.1076	0.7759
SEMTRA (Ours)	Post hoc (Audit)	0.4073	0.1076	0.7759
TCAV [11]	Concept direction	0.4850	–	–

While the CBM achieves higher predictive accuracy because it is structurally optimized for a specific concept vocabulary, SEMTRA provides a comparable level of semantic reconstruction integrity without requiring any retraining of the underlying black-box. The result supports SEMTRA as a practical post-hoc audit layer for frozen representations when an appropriate concept dictionary is available, rather than as a universal substitute for interpretable-by-design architectures.

Unlike traditional symbolic learners that provide a single path to a decision, SEMTRA's use of rough-set granulation allows for a diagnostic audit. While a CART decision tree forces a prediction (leading to zero abstentions but potentially high error in boundary regions), SEMTRA identifies where the model's logic becomes indiscernible. This supports SEMTRA as a diagnostic audit layer for identifying the limits of a neural network's conceptual reconstruction, not as a classifier replacement.

4.6. Algorithmic Recovery (Synthetic Benchmark)

To isolate the framework's internal logic from the noise and ambiguity of real-world neural representations, we conducted a controlled synthetic benchmark. This stage serves as an algorithmic sanity check, testing whether the combination of WEDD discretization and rough-set induction can recover a known ground-truth logic when the semantic signal is clearly defined.

4.6.1. Exactness of Logic Extraction

The synthetic environment consisted of a 10-dimensional feature space with a rule dictionary of known logical structures (Table 15).

Table 15. Synthetic benchmark ground-truth rule dictionary.

Class	Ground-Truth Rule	Numeric Cut Summary
1	b_1 high AND b_3 low	$b_1 > 0.70; b_3 < 0.35$
2	b_2 medium AND b_5 high	$0.35 < b_2 \leq 0.65; b_5 > 0.70$
3	b_4 low, b_6 high, AND b_8 medium	$b_4 < 0.35; b_6 > 0.70; 0.35 < b_8 \leq 0.65$

As summarized in Table 16, at zero noise ($\sigma = 0.000$), the framework achieved a macro-F1 score of 0.879 and a rule Jaccard similarity of 0.700, with a near-perfect operational coverage of 97.10%. The minor baseline threshold error (0.014) observed at zero noise stems from the Gaussian KDE bandwidth smoothing used in the WEDD algorithm, represents the systematic bias introduced by the KDE bandwidth smoothing, not an optimization failure.

Table 16. Synthetic stress-test results averaged over random seeds.

Noise (σ)	MAE	Threshold Error	Rule Jaccard	Macro-F1	Coverage
0.000	0.000	0.014	0.700	0.879	0.971
0.100	0.021	0.018	0.763	0.881	0.962
0.200	0.041	0.017	0.743	0.838	0.952

4.6.2. Robustness Under Controlled Noise

The stability of the framework was tested by injecting increasing levels of semantic noise ($\sigma \in [0, 0.20]$). As visualized in Figure 6a, the operational coverage remained relatively stable as noise increased, indicating that the WEDD thresholds were not dominated by small perturbations in this synthetic setting. While the exact threshold recovery error inevitably fluctuates (Figure 6b), the framework maintained a macro-F1 score above 0.838 throughout the tested noise range.

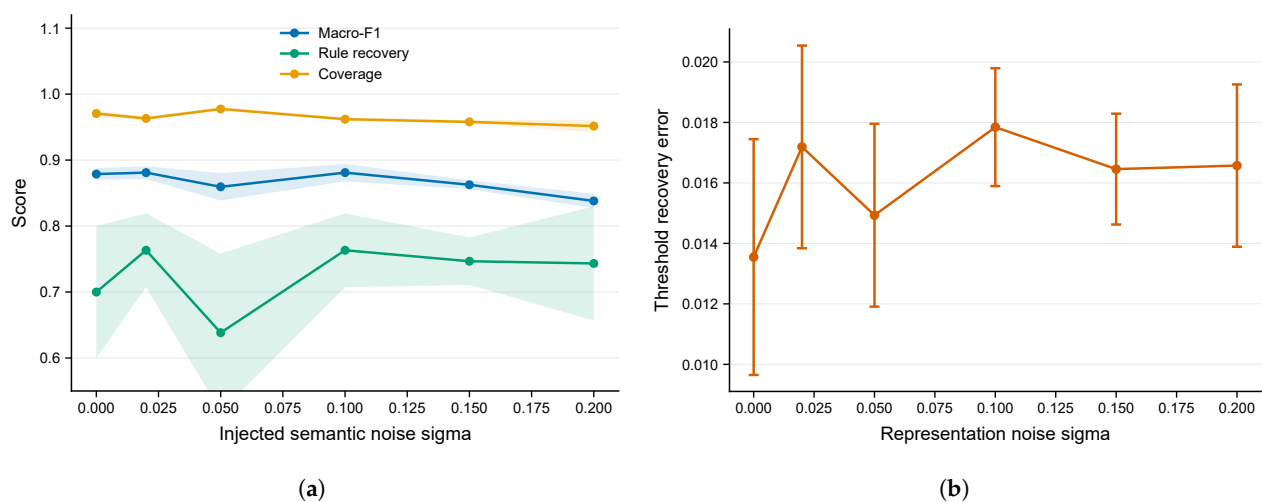


Figure 6. (a) Synthetic benchmark with 95% confidence bands across seeds. Macro-F1, rule-recovery Jaccard, and coverage remain high across the tested semantic-noise range. (b) Synthetic threshold recovery error across noise levels.

The synthetic results demonstrate that the core mathematical machinery of SEMTRA, specifically the ability to translate continuous variables into discrete granules and extract minimal reducts, behaves predictably when the semantic signal is controlled. For an auditor, this is a useful finding: it suggests that the lower fidelity observed in the real-world AwA2 experiments is driven at least partly by semantic indiscernibility in the frozen features and class-level attributes, rather than by an isolated failure of the rule-induction routine. Consequently, SEMTRA is best viewed as a diagnostic audit engine that reports the verifiable portion of a model's semantic logic, including the portions where that logic fragments or becomes noisy.

5. Discussion

5.1. The Scientific Value of Global Auditability

The overarching objective of this manuscript is to establish a reproducible, post hoc analytical pathway that maps semantically opaque neural representations into globally auditable symbolic knowledge. The current trajectory of AI inherently demands verifiable symbolic knowledge in high-stakes decision-making workflows, as emphasized by the XAI 2.0 manifesto [1]. However, the modern interpretability landscape is characterized by a methodological schism [2]. On the one hand, local post hoc explanations, such as LIME and SHAP [8], dominate the field but routinely produce fragmented, instance-specific visual

heatmaps. On the other side, interpretable-by-design architectures, such as CBMs [12], provide structural transparency but require costly retraining and structural re-engineering of existing high-performance networks.

The proposed SEMTRA framework effectively bridges this fragmentation gap by operating over frozen representations to generate a globally applicable rulebook. As empirically observed in Section 4.5 and Table 12, local surrogate attributions demonstrate only moderate agreement with our globally extracted logical antecedents. This divergence underscores a critical advantage of our approach: rather than offering an unstable, locally perturbed approximation of importance, SEMTRA explicitly extracts the stable, dataset-wide logical policies that the network utilizes. By structuring the output as formal production rules (1), the framework shifts the explanation paradigm from localized diagnostic hints to a systematic, reusable audit artifact.

5.2. Interpreting the Audit Tax: Accuracy vs. Verifiability

A core conceptual contribution of this work is the explicit quantification of the “audit tax,” which we define as the inevitable performance cost of enforcing mathematical discretization (12) and logical transparency on a highly dimensional, non-linear latent system. During the AwA2 Protocol A evaluation, the underlying neural predictor achieved a competitive accuracy of 71.16%. Conversely, the extracted symbolic rulebook yielded a covered accuracy of 39.73% and a covered fidelity to the base model of 40.48% (Table 4).

While this numerical reduction might initially appear as a predictive detriment, it accurately maps the verifiable boundaries of the model’s logic. The fidelity-coverage tradeoff, visualized in Figure 3, demonstrates that SEMTRA intentionally occupies a high-coverage operational regime while explicitly documenting structural uncertainty. The synthetic benchmark evaluations presented in Section 4.6 further validate this hypothesis; SEMTRA successfully achieved exact rule recovery at zero semantic noise. Therefore, the audit tax observed in real-world scenarios primarily stems from the inherent semantic indiscernibility and feature entanglement in the ResNet latent space, rather than from a mechanical failure of the rough-set induction engine.

5.3. The Power of Honest Silence and Indiscernibility

Traditional heuristic symbolic learners, such as standard decision trees [15], are structurally compelled to force a deterministic classification even when the underlying data manifold is densely entangled. SEMTRA fundamentally diverges from this paradigm by integrating formal rough-set theory [6,7]. This mathematical foundation empowers the framework to embrace “honest silence”. By computing strict lower approximations (14), the framework acknowledges structural ambiguities, yielding an explicit structural abstention rate (23) of 15.20% and a conflict rate (24) of 13.75% on the AwA2 dataset.

When visually similar but taxonomically distinct classes map into overlapping semantic states, the framework refuses to hallucinate a false boundary. Instead, the resulting boundary region acts as a high-fidelity diagnostic warning for auditors. This structural design transforms inference abstentions from algorithmic failures into critical indicators of where the neural network’s conceptual understanding structurally breaks down.

5.4. Linear Transition as a Defensive Choice

The strategic implementation of a globally linear ridge-regression operator (5) serves as a deliberate, defensive safeguard optimized for maximum algebraic transparency. Previous analytical investigations have validated the efficacy of continuous transition matrices in structured visual analytics [3] and complex healthcare diagnostics [4]. Furthermore, unlike the ETM which strictly prioritizes geometric consistency under predefined Lie-group trans-

formations [5], SEMTRA is exclusively tailored for semantic-to-symbolic rule extraction and granular logic reporting (see the distinction summarized in the Supplementary Material).

The ablation studies detailed in Table 2 empirically justify this linear constraint. While complex non-linear regressors marginally minimize continuous reconstruction errors, they simultaneously introduce severe black-box opacity into the transition bridge, thereby degrading downstream rule coverage and interpretability. By maintaining linearity, the neural-to-semantic transition weights remain mathematically transparent, allowing human auditors to directly trace and stress-test the origin of semantic derivations.

5.5. Limitations and Open Challenges

Despite providing a robust mechanism for global auditability, the SEMTRA framework exhibits several profound limitations, omissions, and methodological shortcomings that present immediate challenges for future research.

First and foremost is the strict dependency on predefined, human-annotated concept dictionaries. The framework assumes that the provided semantic attributes ($\mathcal{C}$) sufficiently encapsulate all predictive signals utilized by the neural network. In reality, deep models routinely exploit confounding variables, background context, image artifacts, and intra-class variances (such as lighting or object pose) that are absent from the dictionary. Consequently, when the framework encounters decisions based on these unmapped latent cues, it suffers from a semantic bottleneck, inevitably leading to structural abstentions or reduced covered fidelity.

Second, the framework suffers from sequential error propagation. Because the pipeline is disconnected, operating sequentially through latent projection, WEDD discretization, and rough-set granulation, information is irretrievably lost at each stage. Errors originating in the continuous semantic reconstruction phase inevitably distort the subsequent density-aware thresholds, which in turn degrade the logical purity of the final rough-set reducts.

Third, algorithmic scalability remains a critical bottleneck. The combinatorial nature of minimal reduct search (Equation (15)) is computationally expensive. As the dimensionality of the semantic dictionary increases, the search space expands exponentially. This limits the framework's current applicability to datasets with hundreds of granular semantic features unless aggressive pruning or dimensionality reduction (q -sensitivity bounding) is strictly enforced.

Finally, the cross-domain technical evaluations exposed severe dataset dependence. While SEMTRA performed adequately on the physical object attributes of AwA2, its application to the abstract, globally distributed scene features of the SUN dataset resulted in extremely low rulebook coverage and fidelity. Similarly, while Derm7pt technical validations showed promise, relying on general ImageNet-pretrained encoders for highly specialized clinical concepts remains diagnostically insufficient.

These limitations define a clear trajectory for future research. Open questions remain about how to implement end-to-end differentiable rough-set algorithms that mitigate sequential error propagation without sacrificing post hoc transparency. Additionally, exploring automated concept discovery, potentially via integration with Large Language Models, could alleviate the rigid dependency on static human-annotated dictionaries.

6. Conclusions

In this work, we proposed the SEMTRA framework to translate uninterpretable neural representations into auditable, rough-set production rules. It effectively bridges the transparency gap in deep learning by establishing explicit metrics for coverage, fidelity, and conflict reporting. Evaluated on the AwA2 Protocol A, the semantic transition achieved a test MAE of 0.1029 ± 0.0005 . The induced rulebook covered 84.80% of instances, successfully

reproducing base predictor decisions with a covered fidelity of 40.48% and a covered accuracy of 39.73%. Under the Protocol B split, continuous semantic transfer achieved 44.02% unseen-object accuracy, serving as a robust diagnostic of conceptual transfer. On controlled synthetic data, the framework reached a macro-F1 score of 0.879 at zero noise and 0.838 under maximum noise, directly validating the rule extraction engine. Despite these achievements, the framework remains highly dependent on predefined concept dictionaries, and sequential error propagation limits overall performance in complex cross-domain applications like SUN and Derm7pt.

Future research should address these challenges by investigating instance-level semantic supervision, developing search-efficient rough-set combinatorial algorithms, and establishing domain-specific validations to safely deploy transparent audit layers in critical environments.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/make8070181/s1>, File S1: “Supplementary Materials for ‘SEMTRA: Global Semantic Transition and Rough-Set Rules for Auditable Post-hoc Explainability’.” Completed supplementary captions and legends are as follows: Table S1, Feature extractor, base predictor, and experiment hyperparameters; Table S2, Base predictor performance on AwA2 Protocol A; Table S3, SUN xlsa17 semantic-transition and rulebook audit summary; Figure S1, Object-level audit outcomes computed from enhanced prediction exports. (a) Object-level audit outcomes. (b) Conflict and abstention limits. (c) Object-level bootstrap intervals. The panels separate coverage, conflict, abstention, covered accuracy, covered fidelity, and bootstrap uncertainty so that predictive correctness and faithfulness to the frozen base model are not conflated; Figure S2, SUN category-level coverage and conflict map. The plot highlights the hierarchy-dependent failure modes of the current scene-audit configuration; Table S4, Derm7pt retrospective technical-validation summary using official case-level splits; Table S5, Derm7pt retrospective technical-validation summary with locked ResNet-50 features; Figure S3, Derm7pt grouped-diagnosis diagnostics under the locked ResNet-50 technical-validation protocol. The figure should not be interpreted as clinical validation; Table S6, Selected-attribute-count sensitivity on AwA2 for the WEDD rulebook; Figure S4, Discretizer evidence across coverage, accuracy, fidelity, conflict, and paired WEDD-minus-MDLP-like entropy intervals. (a) Coverage by discretizer. (b) Accuracy by discretizer. (c) Fidelity by discretizer. (d) Conflict by discretizer. (e) Paired WEDD-MDLP-like entropy intervals. The panels support cautious framing: the observed WEDD advantages are modest and should be treated as operating-point evidence rather than universal dominance; Table S7, SVD-rank sensitivity on AwA2 for semantic reconstruction and rulebook audit metrics; Figure S5, Sensitivity and runtime controls for the SEMTRA audit. (a) Selected-attribute count sensitivity. (b) Retained SVD rank sensitivity. (c) Rule-confidence threshold behavior. (d) Phase-level runtime summary; Table S8, Per-phase Runtime summary from JSONL instrumentation; Table S9, AwA2 Protocol B semantic-transfer validation across seeds. This is not a competitive zero-shot leaderboard; Figure S6, Protocol B semantic-transfer diagnostics. The panels summarize seed-level variation, accuracy-versus-MAE behavior, and per-class failure modes for bat, giraffe, dolphin, and sheep; Table S10, Cross-domain SEMTRA summary. Derm7pt is retrospective technical validation only; Figure S7, Cross-domain portability viewed through coverage and fidelity. The figure shows that SEMTRA’s audit behavior is strongly dataset-dependent; Table S11, Object-level bootstrap intervals from enhanced prediction exports.

Author Contributions: Conceptualization, P.R., O.B. and I.K.; methodology, P.R.; software, P.R.; validation, P.R. and O.B.; formal analysis, P.R., O.B. and I.K.; investigation, P.R.; resources, O.B. and I.K.; data curation, P.R.; writing—original draft preparation, P.R.; writing—review and editing, O.B. and I.K.; visualization, P.R.; supervision, I.K.; project administration, O.B. 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 for the computational experiments reported here. AwA2 and SUN are public benchmark datasets; Derm7pt is a public, de-identified retrospective dermatology dataset used only for technical validation.

Informed Consent Statement: Not applicable for the analysis of public benchmark data. No human participants were recruited and no clinical data were collected.

Data Availability Statement: AwA2 is available from the official Animals with Attributes 2 dataset page [21]; xlsa17 AwA2/SUN features are available from the MPI zero-shot benchmark resources [20]; SUN Attribute Database metadata are available from the official SUN Attribute Database page [43]; Derm7pt metadata and official split files are available from the authors' public resources [45]. The GitHub repository <https://github.com/radiukpavlo/transition-matrix-dss> (accessed on 21 June 2026) includes scripts and generated artifacts; the bundle contains prediction exports, SUN category metadata, locked Derm7pt encoder manifests, and object-level bootstrap intervals.

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

Abbreviations

The following abbreviations are used in this manuscript:

Abs	Structural Abstention Rate
AI	Artificial Intelligence
AUROC	Area Under the Receiver Operating Characteristic
AwA2	Animals with Attributes 2
CART	Classification and Regression Trees
Cov	Rulebook Coverage
CR	Conflict Rate
ECE	Expected Calibration Error
ETM	Equivariant Transition Matrix
KDE	Kernel Density Estimation
LIME	Local Interpretable Model-agnostic Explanations
MAE	Mean Absolute Error
MDLP	Minimum Description Length Principle
MLP	Multi-Layer Perceptron
PCA	Principal Component Analysis
RBF	Radial Basis Function
ResNet	Residual Network
RMSE	Root Mean Square Error
SEMTRA	Global Semantic Transition
SHAP	Shapley Additive Explanations
SJE	Structured Joint Embedding
SSE	Semantic Similarity Embedding
SUN	Scene Understanding
SVD	Singular-Value Decomposition
TCAV	Testing with Concept Activation Vectors
WEDD	Weighted Entropy-Density Discretization
XAI	Explainable Artificial Intelligence
ZSL	Zero-Shot Learning

Appendix A. ETM vs. SEMTRA Comparison

Table A1. Comparison of the prior ETM framework and the present SEMTRA framework.

Dimension	ETM [5]	SEMTRA	Relationship
Core problem	Symmetry-consistent explanation mapping	Semantic auditability of frozen representations	Complementary post-hoc transition goals
Mathematical operator	Lie-group-constrained transition matrix	Ridge semantic transition plus rough-set rules	Shared transition family, different constraints
Constraint	Equivariance and Lie-algebra intertwining	Discretization stability, support, confidence, and abstention	Non-overlapping objective functions
Semantic layer	Geometric mental-model coordinates	Human-interpretable semantic attributes	SEMTRA makes semantics the rule substrate
Discretization	Not central	WEDD, entropy, quantile, and width comparisons	SEMTRA adds symbolic state construction
Symbolic rule induction	Not included	Rough-set granules, reducts, and production rules	SEMTRA adds auditable rulebooks
Conflict/abstention	Not primary output	Explicit conflict and structural abstention metrics	SEMTRA exposes unverifiable regions
Primary output	Symmetry-aware transition explainer	Coverage-aware semantic rulebook	Different audit artifacts
Evaluation focus	Geometric consistency and reconstruction	Fidelity, coverage, accuracy, conflict, runtime, and sensitivity	Distinct empirical questions

Appendix B. Experimental Protocol and Hyperparameters

Table A2 reports the one-factor sensitivity grid for the two auditor-facing control knobs. The results show that increasing λ_s changes the selected semantic basis and modifies rulebook compactness, while increasing λ_H shifts WEDD toward entropy-dominant thresholds and affects the coverage–fidelity balance.

Table A2. Sensitivity of the auditor control knobs on rulebook compactness, coverage, fidelity, and covered accuracy. The λ_s sweep varies semantic-attribute selection while holding $\lambda_H = 0.65$; the λ_H sweep varies WEDD thresholding using the released default attribute set corresponding to $\lambda_s = 0.50$.

Control Knob	Value	Rules	Rulebook Coverage	Covered Fidelity	Covered Accuracy
λ_s	0.25	50	0.7687	0.4796	0.4414
λ_s	0.50	54	0.8640	0.3829	0.4073
λ_s	0.75	57	0.7847	0.3856	0.3935
λ_H	0.25	58	0.7136	0.3514	0.3535
λ_H	0.65	54	0.8640	0.3829	0.4073
λ_H	0.90	53	0.8977	0.4349	0.4205

The feature extractor utilizes the released ResNet-101 representation natively provided with the Awa2 dataset. The base predictor is systematically trained on the representation matrix to provide reproducible probabilities, robust class predictions, and explicit fidelity targets. Table A3 outlines the detailed feature and classifier configurations, while Table A4 reports the corresponding validation and test performance metrics, including the Area Under the Receiver Operating Characteristic (AUROC) and Expected Calibration Error (ECE).

Table A3. Feature extractor, base predictor, and experiment hyperparameters.

Component	Value	Implementation Detail
Feature extractor	ResNet-101 features	ImageNet Large Scale Visual Recognition Challenge (ILSVRC)-pretrained representation layer; no image-level augmentation in this package
Feature dimension	2048	Global average/penultimate representation coordinates
Auxiliary compression	64	Principal Component Analysis (PCA) reduced to 64 components retaining 95% variance
Base predictor	Ridge classifier	Trained on variance-screened representation coordinates
Random seeds	42, 43, 44, 45, 46	Used for AwA2 seed-wise regeneration
Semantic bridge	Ridge regression	Grid alpha in {0.01, 0.1, 1, 10, 100}
Rule thresholds	WEDD	alpha = 0.65, max_depth = 2, min_bin_size = 30, min_gain = 0.002
Rule induction	Greedy reducts	tau = 0.84, s_min = 18 for AwA2 Protocol A
Soft matching threshold	$\tau_H = 0.25$	Maximum allowed masked Hamming distance

Table A4. Base predictor performance on AwA2 Protocol A.

Split	Top-1	Top-5	Macro-F1	Weighted-F1	AUROC	ECE
Validation	0.7093	0.9306	0.5399	0.6639	0.9832	0.6622
Test	0.7116	0.9291	0.5434	0.6681	0.9836	0.6645

It is noted that the selected α hit the lower bound of the grid search space (0.01); thus, exploring smaller regularization values could potentially yield marginal improvements. Furthermore, the disparity between the Top-1 Accuracy (70.93%) and the Macro-F1 score (53.99%) indicates that the base Ridge classifier experiences partial majority-class collapse, favoring well-represented classes over minority taxonomies due to class imbalance.

Appendix C. Rule Inference Traces and Structural Perturbation Stability

Detailed rule inference traces, structural stability analyses, and per-phase runtime logs are retained as essential programmatic audit artifacts. Table A5 illustrates representative inference traces successfully mapping individual objects to explicit rule antecedents, while Supplementary Materials Section S4 and S5 report the full diagnostic frontiers for q-sensitivity, rank-sensitivity, rule-confidence thresholds, and detailed reduct-search runtime components.

Table A5. Representative rule inference traces with object ID, prediction, matched rule evidence, and antecedent states.

Case	Object	True	Pred.	Mode	Rule	Support	Conf.	Antecedent States
correct exact	5083	collie	collie	exact	R0027	1675	0.208	stripes = s0; hooves = s0; swims = s1; paws = s3; longneck = s0; hunter = s1; ocean = s1; quadrapedal = s3
abstention	24,024	ox	abstain	exact	R0029	2820	0.195	stripes = s0; hooves = s2; swims = s1; paws = s0; longneck = s1; hunter = s0; ocean = s1; quadrapedal = s2
fallback error or boundary	16,781	hamster	mole	fallback	–	–	–	stripes = s0; hooves = s1; swims = s1; paws = s2; longneck = s0; hunter = s0; ocean = s1; quadrapedal = s2

Trace Walkthrough for a Zebra Instance

For test object 36914 (zebra), the 2048-dimensional ResNet representation $\mathbf{a}_{36914}$ is mapped through the transition matrix $\mathbf{T}$ to obtain the reconstructed continuous semantic vector $\hat{\mathbf{b}}_{36914}$. The relevant WEDD-discretized components are $\hat{b}_{\text{hunter}} = 0.0000$, $\hat{b}_{\text{water}} = 0.0501$, $\hat{b}_{\text{small}} = 0.0000$, $\hat{b}_{\text{jungle}} = 0.2632$, and $\hat{b}_{\text{yellow}} = 0.0879$. Using the learned WEDD thresholds, these values become hunter = s_0 , water = s_0 , small = s_0 , jungle = s_2 , and yellow = s_1 . This state vector exactly satisfies Rule R0004, whose antecedent is hunter = s_0 AND water = s_0 AND small = s_0 AND jungle = s_2 AND yellow = s_1 , yielding a verified zebra prediction with support 434 and confidence 0.866.

Furthermore, Table A6 rigorously reports the operational rule consistency and prediction stability under varying degrees of semantic-representation perturbations. Reconstructed semantic representations were perturbed using additive Gaussian noise, as formalized in Equation (A1):

$$\hat{\mathbf{b}}_{i,\text{perturbed}} = \hat{\mathbf{b}}_i + \epsilon_i, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2 \hat{s}_j^2), \quad (\text{A1})$$

where $\hat{s}_j$ is the empirical standard deviation of reconstructed semantic attribute j on the test split.

Because WEDD anchors thresholds in low-density valleys rather than dense class-overlap regions, rules exhibit high robustness against micro-perturbations up to $\sigma = 0.05$. The degradation curve is therefore gradual: Rule consistency remains 0.8667 and coverage remains essentially unchanged at 0.8816 at $\sigma = 0.05$, while the larger $\sigma = 0.10$ setting produces the expected decline in rule and decision consistency.

Table A6. Rule consistency under semantic-representation perturbations.

Sigma	Rule Consistency	Decision Consistency	Cov	Cov Change	Abstention	Conflict
0.0000	1.0000	1.0000	0.8796	0.0000	0.1204	0.1180
0.0100	0.9745	0.9663	0.8792	−0.0004	0.1208	0.1183
0.0250	0.9326	0.9176	0.8808	0.0012	0.1192	0.1168
0.0500	0.8667	0.8401	0.8816	0.0020	0.1184	0.1160
0.1000	0.7307	0.7112	0.8940	0.0145	0.1060	0.1026

References

1. Longo, L.; Brcic, M.; Cabitza, F.; Choi, J.; Confalonieri, R.; Del Ser, J.; Guidotti, R.; Hayashi, Y.; Herrera, F.; Holzinger, A.; et al. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Inf. Fusion* **2024**, *106*, 102301. [\[CrossRef\]](#)
2. Mersha, M.; Lam, K.N.; Wood, J.; AlShami, A.K.; Kalita, J. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing* **2024**, *599*, 128111. [\[CrossRef\]](#)
3. Radiuk, P.; Barmak, O.; Manziuk, E.; Krak, I. Explainable deep learning: A visual analytics approach with transition matrices. *Mathematics* **2024**, *12*, 1024. [\[CrossRef\]](#)
4. Barmak, O.; Krak, I.; Yakovlev, S.; Manziuk, E.; Radiuk, P.; Kuznetsov, V. Toward explainable deep learning in healthcare through transition matrix and user-friendly features. *Front. Artif. Intell.* **2024**, *7*, 1482141. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Radiuk, P.; Barmak, O.; Bedratyuk, L.; Krak, I. Equivariant transition matrices for explainable deep learning: A Lie group linearization approach. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 92. [\[CrossRef\]](#)
6. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [\[CrossRef\]](#)
7. Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning About Data*; Theory and Decision Library D; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1991; Volume 9. [\[CrossRef\]](#)
8. Salih, A.M.; Raisi-Estabragh, Z.; Boscolo Galazzo, I.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv. Intell. Syst.* **2025**, *7*, 2400304. [\[CrossRef\]](#)
9. Kovalchuk, O.; Barmak, O.; Radiuk, P.; Klymenko, L.; Krak, I. Towards transparent AI in medicine: ECG-based arrhythmia detection with explainable deep learning. *Technologies* **2025**, *13*, 34. [\[CrossRef\]](#)

10. Radiuk, P.; Rusyn, B.; Melnychenko, O.; Perzynski, T.; Sachenko, A.; Svystun, S.; Savenko, O. Criticality assessment of wind turbine defects via multispectral UAV fusion and fuzzy logic. *Energies* **2025**, *18*, 4523. [\[CrossRef\]](#)
11. Nicolson, A.; Schut, L.; Noble, J.A.; Gal, Y. Explaining explainability: Recommendations for effective use of concept activation vectors. *Trans. Mach. Learn. Res.* **2025**, *2025*, 1–21. [\[CrossRef\]](#)
12. Srivastava, D.; Yan, G.; Weng, T.W. VLG-CBM: Training concept bottleneck models with vision-language guidance. In *Proceedings of the Advances in Neural Information Processing Systems 37*; Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C., Eds.; ACM: Red Hook, NY, USA, 2024; Volume 37, pp. 79057–79094. [\[CrossRef\]](#)
13. Espinosa Zarlenga, M.; Barbiero, P.; Ciravegna, G.; Marra, G.; Giannini, F.; Diligenti, M.; Shams, Z.; Precioso, F.; Melacci, S.; Weller, A.; et al. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Proceedings of the Advances in Neural Information Processing Systems 35*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; ACM: Red Hook, NY, USA, 2022; Volume 35, pp. 21400–21413.
14. Yükseköğül, M.; Wang, M.; Zou, J. Post-hoc concept bottleneck models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda, 1–5 May 2023.
15. Mienye, I.D.; Jere, N.R. A survey of decision trees: Concepts, algorithms, and applications. *IEEE Access* **2024**, *12*, 86716–86727. [\[CrossRef\]](#)
16. Bollaert, H.; Palangetić, M.; Cornelis, C.; Greco, S.; Słowiński, R. FRRI: A novel algorithm for fuzzy-rough rule induction. *Inf. Sci.* **2025**, *686*, 121362. [\[CrossRef\]](#)
17. Zhang, Y.; Li, R.; Wu, N.; Li, Q.; Lin, X.; Hu, Y.; Li, T.; Jiang, Y. Dissect black box: Interpreting for rule-based explanations in unsupervised anomaly detection. In *Proceedings of the Advances in Neural Information Processing Systems 37*; Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C., Eds.; ACM: Red Hook, NY, USA, 2024; Volume 37, pp. 84169–84196. [\[CrossRef\]](#)
18. Şenozan, H.; Soylu, B. A flexible non-monotonic discretization method for pre-processing in supervised learning. *Pattern Recognit. Lett.* **2024**, *181*, 77–85. [\[CrossRef\]](#)
19. Biroli, G.; Mézard, M. Kernel density estimators in large dimensions. *SIAM J. Math. Data Sci.* **2026**, *8*, 46–76. [\[CrossRef\]](#)
20. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-shot learning: A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2251–2265. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Lampert, C.H.; Pucher, D.; Dostal, J. Animals with Attributes 2. Dataset Website. 2017. Available online: <https://cvml.ista.ac.at/AwA2/> (accessed on 22 May 2026).
22. Lampert, C.H.; Nickisch, H.; Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami, FL, USA, 20–25 June 2009; pp. 951–958. [\[CrossRef\]](#)
23. Lampert, C.H.; Nickisch, H.; Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 453–465. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Verma, V.K.; Rai, P. A simple exponential family framework for zero-shot learning. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II*; Ceci, M., Hollmén, J., Todorovski, L., Vens, C., Džeroski, S., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; Volume 10535, pp. 792–808. [\[CrossRef\]](#)
25. Xian, Y.; Sharma, S.; Schiele, B.; Akata, Z. f-VAEGAN-D2: A feature generating framework for any-shot learning. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, CA, USA, 16–20 June 2019; pp. 10275–10284. [\[CrossRef\]](#)
26. Huynh, D.; Elhamifar, E. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, Seattle, WA, USA, 14–19 June 2020; pp. 4483–4493. [\[CrossRef\]](#)
27. Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; Akata, Z. Attribute prototype network for zero-shot learning. In *Proceedings of the Advances in Neural Information Processing Systems 33*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H.T., Eds.; ACM: Red Hook, NY, USA, 2020; Volume 33, pp. 21969–21980.
28. Han, Z.; Fu, Z.; Chen, S.; Yang, J. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, Nashville, TN, USA, 19–25 June 2021; pp. 2371–2381.
29. Chen, S.; Hong, Z.; Liu, Y.; Xie, G.S.; Sun, B.; Li, H.; Peng, Q.; Lu, K.; You, X. TransZero: Attribute-guided transformer for zero-shot learning. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 330–338. [\[CrossRef\]](#)
30. Chen, S.; Hong, Z.; Hou, W.; Xie, G.S.; Song, Y.; Zhao, J.; You, X.; Yan, S.; Shao, L. TransZero++: Cross attribute-guided transformer for zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12805–12821. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Naeem, M.F.; Xian, Y.; Van Gool, L.; Tombari, F. I2DFormer+: Learning image to document summary attention for zero-shot image classification. *Int. J. Comput. Vis.* **2024**, *132*, 3806–3822. [\[CrossRef\]](#)

32. Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G.S.; Dean, J. Zero-shot learning by convex combination of semantic embeddings. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014)*, Banff, AB, Canada, 14–16 April 2014.
33. Socher, R.; Ganjoo, M.; Manning, C.D.; Ng, A.Y. Zero-shot learning through cross-modal transfer. In *Proceedings of the Advances in Neural Information Processing Systems 26*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; ACM: Red Hook, NY, USA, 2013; Volume 26, pp. 935–943.
34. Zhang, Z.; Saligrama, V. Zero-shot learning via semantic similarity embedding. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV 2015)*, Santiago, Chile, 7–13 December 2015; pp. 4166–4174. [[CrossRef](#)]
35. Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1425–1438. [[CrossRef](#)] [[PubMed](#)]
36. Akata, Z.; Reed, S.; Walter, D.; Lee, H.; Schiele, B. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Boston, MA, USA, 7–12 June 2015; pp. 2927–2936. [[CrossRef](#)]
37. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. DeViSE: A deep visual-semantic embedding model. In *Proceedings of the Advances in Neural Information Processing Systems 26*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; ACM: Red Hook, NY, USA, 2013; Volume 26, pp. 2121–2129.
38. Romera-Paredes, B.; Torr, P.H.S. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*; *Proceedings of Machine Learning Research*; Bach, F., Blei, D., Eds.; Springer: Lille, France, 2015; Volume 37, pp. 2152–2161.
39. Changpinyo, S.; Chao, W.L.; Gong, B.; Sha, F. Synthesized classifiers for zero-shot learning. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, 27–30 June 2016; pp. 5327–5336. [[CrossRef](#)]
40. Kodirov, E.; Xiang, T.; Gong, S. Semantic autoencoder for zero-shot learning. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, Honolulu, HI, USA, 21–26 July 2017; pp. 3174–3183. [[CrossRef](#)]
41. Halko, N.; Martinsson, P.G.; Tropp, J.A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **2011**, *53*, 217–288. [[CrossRef](#)]
42. Patterson, G.; Xu, C.; Su, H.; Hays, J. The SUN attribute database: Beyond categories for deeper scene understanding. *Int. J. Comput. Vis.* **2014**, *108*, 59–81. [[CrossRef](#)]
43. Patterson, G.; Hays, J. SUN Attribute Database. Dataset Website. 2012. Available online: <https://cs.brown.edu/people/gmpatter/sunattributes.html> (accessed on 15 June 2026).
44. Kawahara, J.; Daneshvar, S.; Argenziano, G.; Hamarneh, G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 538–546. [[CrossRef](#)] [[PubMed](#)]
45. Kawahara, J.; Daneshvar, S.; Argenziano, G.; Hamarneh, G. Seven-Point Checklist Dermatology Dataset. Dataset Website. 2019. Available online: <https://derm.cs.sfu.ca/Welcome.html> (accessed on 15 June 2026).

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.