



Inside the black box: A comprehensive survey of explainable AI benchmarks, protocols, technical analysis, and open gaps

Habib Khan^a, Faleh Alqahtani^{b,1}, Weiwei Jiang^c, Muhammad Hamza Ali^d, Sareer Ul Amin^e, JaKeoung Koo^{a,*}

^a School of Computing, Gachon University, 1342 Seongnam-daero, Seongnam-si 13120, South Korea

^b School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, Australia

^c School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China

^d Department of Computer Science, Virtual University, Islamabad, Pakistan

^e Department of Applied Art and Technology, Chung-Ang University, Anseong, 17546, South Korea

ARTICLE INFO

Keywords:

Explainable AI
Benchmark analysis
Feature attribution
Performance diagnostics
Robustness assessment
Computational efficiency
Interpretability methods

ABSTRACT

The rapid deployment of artificial intelligence (AI) in real-world pipelines has made explainable decision-making a core requirement for validation, auditing, and responsible use. This survey reviews explainable AI (XAI) methods, with attention to deployable systems, across seven domains: computer vision (CV), natural language processing, tabular data, time series, graphs, reinforcement learning, and recommendation systems. To strengthen practical transparency for advanced deep models for visual intelligence, we synthesize evidence on visual interpretability, characterizing where models attend, how features are integrated across modalities, and how representations evolve across hierarchies. Guided by PRISMA-S search principles, we conducted a documented multi-library search across six source categories and applied explicit inclusion and exclusion criteria to assemble the literature analyzed here. The survey covers 189 works from 2014 to 2026, with primary synthesis drawn from the 2020 to 2026 period and foundational pre-2020 methods retained where they remain standard tools. We organize methods into seven families and, for CV, systematically catalog representative techniques by mechanism, model access, explanation output, and evaluation protocol. Across families, we discuss faithfulness, robustness, and computational cost, emphasizing the balance between efficiency and fidelity that governs practical deployment. To ground the visual interpretability analysis in a concrete system, we apply three established visualization techniques to our lightweight RGB-D saliency detection network (LiSalNet), producing Grad-CAM localization maps, multimodal fusion impact maps, and multiscale hierarchical visualizations with channel activation statistics that reveal how cross-modal evidence and semantic abstraction emerge across network depth. Evidence from human studies is summarized to characterize reported effects on trust, reliance, task accuracy, and cognitive load.

1. Introduction

Explainable Artificial Intelligence (XAI) refers to a set of methods and evaluation practices that make model behaviour more understandable, testable, and actionable for human stakeholders [1]. XAI methods are commonly grouped into feature attribution, saliency based visualization, local surrogate modeling, counterfactual explanation, concept based analysis, and prototype or case based reasoning [2–6]. Feature attribution methods estimate the contribution of input variables to a prediction using gradient based or game theoretic principles, as demonstrated by Integrated Gradients and SHAP [7,8]. Saliency methods

generate visual or token-level importance maps indicating which image regions, words, or features most influence a model decision [9,10]. Local surrogate approaches, such as LIME, approximate the behaviour of a complex model around a specific instance using a simpler interpretable model [11]. Counterfactual explanations identify minimal and feasible input changes that would alter the prediction of the model towards a desired outcome [12,13]. Concept-based methods map internal representations to human-understandable concepts and measure a model's sensitivity to those concepts [14]. Prototype-based models explain predictions by comparing an input with learned representative examples or

* Corresponding author.

Email address: jakeoung@gachon.ac.kr (J. Koo).

¹ Present address: College of Computing and Information Technology, Shaqra University, Saudi Arabia.

<https://doi.org/10.1016/j.cosrev.2026.101037>

Received 21 February 2026; Received in revised form 16 June 2026; Accepted 14 July 2026

Available online 23 July 2026

1574-0137/© 2026 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

parts [15–17]. Although these methods share the general goal of improving transparency, they differ substantially in faithfulness, robustness, human usefulness, and computational cost.

This review covers XAI research published between 2020 and 2026, the period during which the field moved from method proposals toward evaluation, benchmarking, and deployment-oriented analyses. The review spans seven domains: computer vision (CV), natural language processing, tabular data, time series, graph learning, reinforcement learning, and recommendation systems, with a primary technical focus on CV given the maturity and diversity of XAI work in that domain. Multimodal and vision-language models are discussed, which introduce distinct explainability challenges such as cross-modal attribution and attention propagation [18,19]. The survey uses consistent terminology for explanation quality following earlier calls for a more rigorous science of interpretable machine learning [20–23]. Faithfulness refers to the extent to which an explanation reflects the model’s actual decision process, while robustness refers to its stability under input perturbations, random initialization, sampling variation, or distribution shift [21–23].

Despite rapid progress, XAI research faces persistent unresolved challenges. Early surveys mapped the opportunities and open problems that continue to shape the field [24]. Studies often use different definitions, datasets, perturbation rules, and evaluation metrics, making results difficult to compare across methods and domains [3,5,21]. Many methods are still assessed through visual inspection or isolated case studies rather than rigorous quantitative tests, and the lack of agreed evaluation metrics remains a central obstacle [20,22,25]. Gradient-based and perturbation-based methods remain sensitive to baseline selection, masking strategy, and small input changes [26–28]. Prototype-based and concept-based models offer more interpretable structures, but their alignment with human reasoning requires systematic validation [15,17]. High-stakes applications may require models that are interpretable by design rather than relying solely on post hoc explanations [29]. Toolkits such as Quantus and OpenXAI have begun to standardize evaluation, yet they reveal that explanation rankings vary widely across metrics and experimental settings [23,30]. These limitations motivate a review linking XAI method families to evaluation protocols, application domains, human evidence, and practical deployment constraints.

1.1. Motivation and gaps

Practitioners across roles depend on explanations for different purposes. Researchers use them to troubleshoot models and validate behaviour. Engineers identify spurious signals and address biases. Auditors verify compliance and discover risk exposures. Domain experts need to trust predictions that guide medical care, financial decisions, or resource allocation [31]. Despite the breadth of available XAI methods, current literature shows four recurring gaps.

Evaluation inconsistency. Attribution results vary widely with baseline and path sampling choices, and Shapley values are computationally expensive for large models [7,8]. Local surrogates are sensitive to sampling and neighborhood definitions, producing contradictory explanations [11]. Saliency maps can appear convincing while being independent of model weights, as sanity checks confirm [26]. Concept definitions and proxies may not align with user understanding [14], and counterfactual methods must balance feasibility, proximity, and fairness [12]. Surveys consistently highlight a lack of consensus on definitions and metrics, as well as limited adoption of rigorous evaluation protocols [3].

Sparse human evidence. User studies are often small, use narrow tasks, or draw on unrepresentative samples, limiting the generalizability of conclusions about trust, reliance, and task accuracy. Evidence on how people interpret prototype explanations remains particularly limited [15].

Missing cost reporting. Many papers omit computational cost information, leaving practitioners without the information needed to assess

deployment feasibility [32]. Stability under perturbations, reinitialization, and distribution shift also receive limited attention.

Absent visual interpretability protocols. CV pipelines that integrate information across modalities and hierarchical scales are rarely accompanied by consistent visual analysis. Studies display saliency maps without clarifying how multimodal fusion alters feature selectivity, which spatial regions benefit discriminatively after fusion, or how representations evolve from low-level textures to high-level semantics across network depth. Gradient-based visualizations are sensitive to layer choice and noisy gradients, and attention maps in transformer architectures are not always aligned with class evidence [33]. These gaps motivate a unified visual interpretability analysis that measures fusion contributions through gradient-derived impact maps, highlights discriminative regions with Grad-CAM overlays and isolated heatmaps, and exposes scale-specific representation changes with channel activation statistics.

1.2. Contributions of this study

This survey addresses the four gaps mentioned above through the following contributions.

Seven family taxonomy: XAI methods are organized into feature attribution, saliency maps, surrogate and rule-based models, counterfactual and recourse methods, concept-based techniques, prototypes and example-based explanations, and mechanistic analyses. Each family is placed on axes of intrinsic versus post hoc, local versus global, and model-specific versus model-agnostic, with typical roles, evaluation concerns, and documented limitations recorded in a comparative table.

Documented survey of 189 works: Following PRISMA-S search principles, a multi-library search across six source categories with explicit inclusion and exclusion criteria assembles the literature analyzed [34]. The review covers 189 works, with primary synthesis from 2020 to 2026, and foundational pre-2020 methods retained where they remain standard tools. A standardized extraction form captures method details, model families, datasets, licenses, evaluation metrics, reported computational characteristics, sampling budgets, code availability, and human study designs.

Technical analysis with CV emphasis: The survey provides a technical analysis of XAI method families across seven domains, with primary depth in CV. For CV, representative methods are cataloged by mechanism, recording the model access each requires, the explanation it produces, and the evaluation protocol typically applied (Table 6). Cross-domain trends in faithfulness, robustness, and computational cost are characterized, and the balance between efficiency and fidelity that governs practical deployment is discussed.

Visual interpretability analysis: To ground the visual interpretability discussion in a concrete system, three established visualization techniques are applied to LiSalNet, our lightweight RGB-D salient object detection network. The analysis covers (i) multimodal feature fusion impact analysis identifying which spatial regions gain discriminative benefit after cross modal fusion versus those that remain redundant, (ii) Grad-CAM localization with overlay and isolated heatmaps exposing attention concentration and failure modes, and (iii) multiscale feature hierarchy visualization across low, mid, and high-level representations with channel activation statistics that quantify sparsity trends and show how semantic abstraction emerges with network depth.

Human and developer synthesis: Controlled and field studies reporting effects on trust, reliance, task accuracy, and cognitive load are summarized, and findings are translated into a developer workflow covering explanation hooks, logging practices, and guardrails across the model lifecycle.

Governance and reproducibility assessment: Drawing on the 189 works, we characterize how governance and reproducibility reporting vary across method families and domains. We observe that documentation artifacts such as datasheets, model cards, and explainability fact

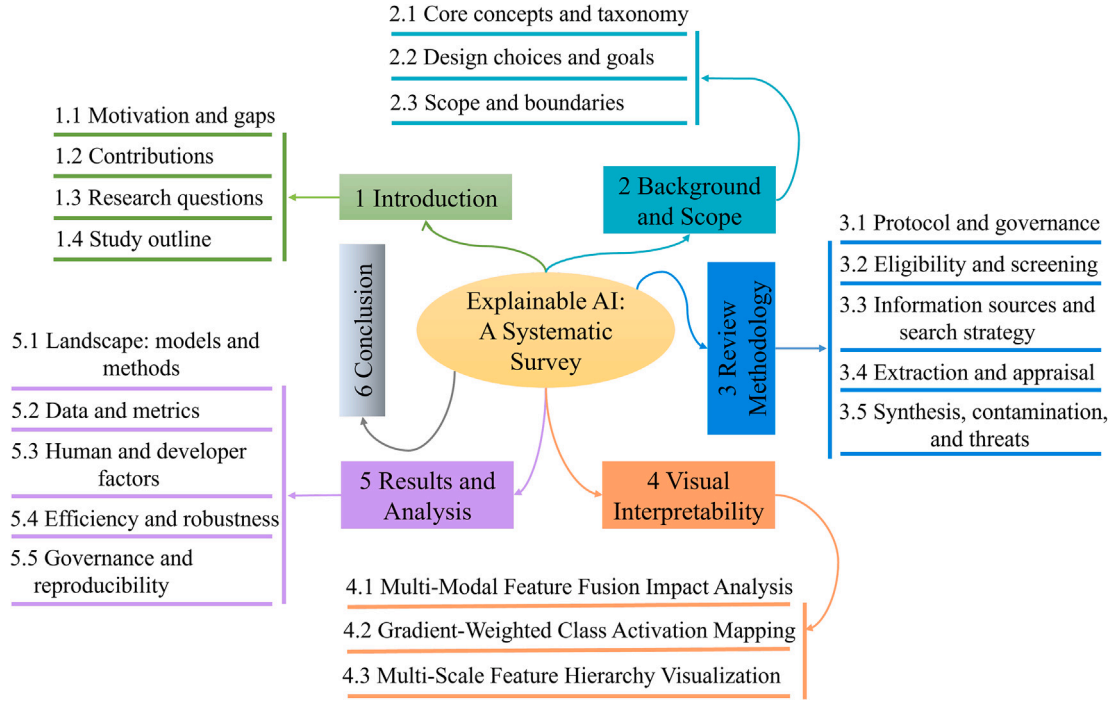


Fig. 1. Survey roadmap includes sections on introduction, background and scope, review protocol, results and analysis, and conclusion.

sheets appear more frequently in vision and NLP work than in graph or reinforcement learning, and that high faithfulness methods rarely disclose deployment costs while efficiency-oriented methods show greater sensitivity to seed variation and distribution shift. We discuss how current regulatory frameworks translate into concrete technical requirements, and identify the reporting gaps that most limit reproducibility across the surveyed literature.

Open challenges: Actionable research directions are identified, covering unified benchmarks, larger-scale human studies, robust evaluation baselines, and governance-integrated deployment pipelines.

1.3. Research questions

Five research questions organize the review and connect each results section to practical decisions.

- **RQ1 (Models and methods).** Which model families and explanation methods appear across the seven surveyed domains, and what patterns emerge in how methods differ by mechanism, model access, explanation output, and evaluation protocol? See [Section 5.1](#).
- **RQ2 (Data and metrics).** Which datasets and evaluation metrics are used to assess XAI methods, and how do protocol choices affect the comparability of reported results? See [Section 5.2](#).
- **RQ3 (Human and developer factors).** What effects do explanations have on trust, reliance, and task accuracy in reported human studies, and how are explanation outputs integrated into developer workflows? See [Section 5.3](#).
- **RQ4 (Efficiency and robustness).** What efficiency strategies are documented for explanation pipelines, and how stable are explanation outputs across input variation, random seeds, and distribution shift? See [Section 5.4](#).
- **RQ5 (Governance and reproducibility).** How do current governance frameworks relate to XAI deployment, and what reproducibility patterns are observed across the included studies? See [Section 5.5](#).

1.4. Study outline

Fig. 1 summarizes the section structure. [Section 2](#) introduces core XAI concepts, defines the seven family taxonomy, and states the scope and boundaries. [Section 3](#) describes the search and selection methodology covering eligibility criteria, search strategy, extraction, quality appraisal, and synthesis. [Section 4](#) presents the visual interpretability analysis applied to LiSalNet, covering fusion impact maps, Grad-CAM localization, and multiscale feature hierarchy visualization. [Section 5](#) reports results across RQ1 through RQ5. [Section 6](#) closes with conclusions and directions for future work.

2. Background and scope

This section defines the core concepts and taxonomic framework used throughout the review. The framework separates intrinsic and post hoc approaches, local and global explanation aims, and model agnostic and model-specific tools, and applies across the seven domains covered: CV, natural language processing (NLP), tabular data, time series, graph learning, reinforcement learning, and recommendation systems. The goal is a stable conceptual base that supports fair comparison, consistent reporting, and practical method selection. CV receives primary technical depth, given the maturity and diversity of XAI work in that domain, though the taxonomy and scope boundaries apply across all seven domains.

2.1. Core concepts and taxonomy

XAI encompasses several families of methods, each with specific roles. Feature attribution assigns importance scores to input features based on their contribution to model output, supporting feature ranking, compliance auditing, and risk assessment across tabular, image, and text data [35,36]. Saliency maps highlight regions or tokens that influence a prediction, supporting visual checks and error analysis in vision and sequence tasks [37,38]. Surrogate and rule-based models approximate a complex predictor with a simple human-readable structure



Fig. 2. End-to-end XAI MIND map organizing the full pipeline from data types and tasks through model access and hooks, explanation method families (attribution and saliency, surrogates and rules, counterfactuals and recourse, concepts, examples and prototypes, mechanistic analyses), evaluation and safety metrics (faithfulness, fidelity, stability, safety probes), reproducibility reporting, serving parameters, and deployment environments. The map serves as a navigation guide connecting method selection to the evaluation protocol and deployment context.

near a point or across a region, aiding debugging and policy review [39,40]. Counterfactual and recourse methods generate nearby inputs that flip a decision under feasibility and fairness constraints, providing actionable guidance and audit trails [41,42]. Concept-based methods link internal representations to human concepts and test sensitivity, enabling domain experts to query models using field-specific concepts [43,44]. Prototype and example-based methods explain a prediction by reference to learned prototypes or selected cases, supporting case-based reasoning and instruction [16,45]. Mechanistic analyses probe internal units, layers, and circuits to reveal computations and failure modes, informing safety reviews and model improvement [10,46,47]. Each XAI family serves a distinct explanation goal. Attribution supports ranked importance, saliency helps triage vision errors, surrogates enable local what-if analysis, counterfactuals guide user action, concept methods assist expert review, prototypes support case-based understanding, and mechanistic analyses audit internal behaviour [48].

Three axes organize these families. Timing separates intrinsic methods built into the model from post hoc methods applied after training. Intrinsic models expose structure and facilitate global audits; post hoc tools preserve model accuracy but require rigorous tests of faithfulness and stability [49,50]. Scope separates local explanations for a single prediction from global summaries across the data. Local tools aid user-facing decisions and debugging; global tools reveal patterns and interactions that guide policy and monitoring [51,52]. Dependency separates model-agnostic tools from model-specific tools. Model agnostic tools are broadly applicable but incur higher computational costs, whereas model-specific tools exploit structural properties to achieve greater efficiency and precision [53,54]. These axes intersect with data

type and risk level. Image classifiers often pair saliency or CAM variants with deletion and insertion tests, while tree ensembles favour model-specific attributions that deliver exact or fast scores.

Fig. 2 sketches the taxonomy by mechanism and aim, placing each family along the three organizational axes of timing, scope, and dependency, and illustrating how attribution, counterfactuals, and mechanistic analyses occupy distinct regions of the design space [5,48,55]. Table 1 complements Fig. 2 by recording, for each family, its explanation form and representative methods, its taxonomic placement, the validation dimensions recommended, and its relevance to this review. The evaluation concerns listed there are operationalized in the review protocol (Section 3) and the results synthesis (Section 5).

2.2. Design choices and goals

Design choices drive faithfulness, stability, and cost across all method families. Attribution depends on the baseline, path schedule, and sampling budget; inappropriate settings distort importance scores and compromise evaluation reliability [61–63]. Perturbation-based saliency depends on the mask design, deletion or insertion protocol, and step size; unstable settings can inflate the results [64]. Surrogates depend on neighborhood definition, kernel, and model class; small changes can flip rules and harm trust [65,66]. Counterfactuals depend on feasibility, proximity, diversity, and fairness constraints; missing any of these constraints yields unsafe recourse advice [59,67]. Concept-based methods depend on the quality of concept sets and validation protocols; proxy concepts can mislead users and mask bias [60,68]. Prototype methods depend on prototype learning and matching quality; weak

Table 1

Validated taxonomy-oriented summary of major XAI method families, synthesized from established XAI taxonomies, survey of surveys evidence, and recent evaluation literature.

XAI method family	Explanation form and representative methods	Taxonomic placement	Recommended validation dimensions	Relevance to this review
Feature attribution and perturbation-based explanations [2,5]	Feature, token, pixel, or region level importance scores; examples include Integrated Gradients, SHAP, DeepLIFT, occlusion, and removal-based attribution	Mostly post hoc; local explanations that can be aggregated globally; model-specific when gradient-based or model-agnostic when perturbation-based	Faithfulness, sensitivity, baseline dependence, deletion/insertion behaviour, infidelity, and ranking stability [22,23,30]	Central across vision, NLP, tabular, graph, and time series settings; recent work emphasizes scalable attribution, benchmark consistency, and cost-aware evaluation
Visual saliency, activation, and CAM style explanations [9,10,26,37]	Heatmaps or relevance maps over image regions, tokens, layers, or attention pathways; examples include Grad-CAM, Score-CAM, LayerCAM, and transformer relevance propagation	Mostly post hoc and model-specific; usually local; primarily used for visual, multimodal, and transformer-based models	Localization quality, visual plausibility, faithfulness, layer choice sensitivity, sanity checks, perturbation tests, and robustness under input and model randomization [22,23]	The primary focus of the CV technical analysis in this review; recent surveys confirm that visual plausibility alone is insufficient for faithfulness assessment [56–58]
Surrogate and rule-based explanations [2,5,11,39]	Interpretable local or global approximations of black box behaviour; examples include LIME, rule lists, decision trees, anchors, and local symbolic approximations	Usually post hoc and model agnostic; local or global, depending on surrogate design	Local fidelity, stability across sampling neighborhoods, compactness, rule comprehensibility, coverage, and agreement with the target model [22]	Useful in tabular, policy, and audit settings; recent surveys highlight instability and the need to report neighborhood, sampling, and kernel settings
Counterfactual explanations and algorithmic recourse [12,13,41,59]	Alternative input states that would change a prediction; examples include nearest counterfactuals, diverse counterfactuals, and causal recourse, and intervention-based recourse	Mostly post hoc, local, and user-facing; generally model agnostic but often constrained by domain or causal assumptions	Validity, proximity, sparsity, feasibility, diversity, actionability, causal plausibility, robustness, and fairness constraints	Increasingly important for high-stakes and regulated domains; Recent surveys emphasize moving from explanation toward actionable and fair recourse
Concept-based explanations [14,43,44,60]	Explanations grounded in human-understandable concepts or concepts directions; examples include TCAV, completeness-aware concepts, and human-centered concept explanations	Usually model-specific and representation-based; supports both local and global analysis; bridges technical and domain expert interpretation	Concept validity, purity, completeness, and sensitivity to the concept choice, expert alignment, and human interpretability	Relevant where domain concepts are meaningful, such as medicine, science, and safety auditing; recent work stresses the concept validation rather than assuming concept quality
Prototype, example-based, and case-based explanations [15–17,45]	Predictions explained through similar training cases, learned prototypes, image parts, or representative examples; examples include ProtoPNet, ProtoTree, and PIP-Net	Intrinsic or post hoc; usually local but may support a global model inspection through learned prototypes	Prototype purity, coverage, representativeness, diversity, human perceived similarity, class consistency, and case selection bias	Useful for domains where examples are natural evidence; recent prototype models improve interpretability, but still require systematic human and quantitative validation
Mechanistic, representation level, and transformer relevance analysis [10,19,46]	Analysis of neurons, layers, circuits, attention, and relevance flows, internal representations, and learned mechanisms	Model-specific and architecture-dependent; often global or meso level, but supports local explanation through activation or circuit tracing	Internal validity, reproducibility, causal intervention tests, architectural comparability, scalability, and evidence that internal components causally affect outputs	Increasingly relevant for transformers, large language models, and safety-oriented audits, evaluation standards remain less mature than for attribution or perturbation methods
Human-centered and domain-specific evaluation layer [22,31,47,56,58]	Not a separate method family; a validation layer applied to the above families in specific domains, stakeholder contexts, and deployment settings	Cross-cutting across all XAI families; depends on the user group, task risk, data modality, and deployment setting	Task performance, calibrated trust, reliance, cognitive load, clinical or domain relevance, explanation usefulness, reproducibility, and governance readiness	This review treats medical imaging, pathology, and drug discovery, CV and language models as application contexts where method families must be evaluated with domain-specific evidence

training yields brittle cases and poor coverage [69,70]. Across all families, implementation choices affect stability and cost across random seeds, input noise, data augmentations, and model retraining. Common evaluation metrics include deletion and insertion AUC, comprehensiveness and sufficiency, infidelity and sensitivity scores [18,38]. Controls include randomization tests, input invariance checks, and stress tests under distribution shift [71–73].

Practical selection guidelines follow from these design dependencies. For high-risk tasks, combine attribution with deletion and insertion curves, and add randomization checks, prioritizing methods that maintain stable scores across seeds and noise [74,75]. For constrained budgets, prefer model-specific tools that exploit structure for efficiency, or approximate schemes with verified fidelity bounds [61,76]. For user-facing recourse, choose counterfactual methods that are subject to domain constraints and entail trade-offs among feasibility, proximity, and fairness [77]. For domain expert review, use concept-based tests with validated concept definitions and report agreement with attribution or counterfactual signals [78]. For developer workflows, integrate explanations into debugging pipelines and apply sanity checks when

settings change [79,80]. The evolution of these method families and their evaluation protocols is traced in recent taxonomic surveys [5,81].

2.3. Scope and boundaries

The review covers CV, NLP, tabular data, time series, graph learning, reinforcement learning, and recommendation systems in both research and deployment contexts, with primary technical depth in CV. Multimodal coverage appears only where needed to note distinct evaluation requirements [85]. Outside the scope are unverifiable demonstrations, proprietary systems without evaluation details, essays without measurable outcomes, and domain-specific legal analyses without technical validation [86]. Methods include intrinsic and post hoc, local and global, and model-agnostic and model-specific tools. Where reported in included studies, faithfulness, stability, and human study outcomes are discussed; computational cost information is noted where disclosed. Foundational pre-2020 sources are retained where they define methods that remain in standard use; the primary synthesis and characterization draw on works from 2020 to 2026 [87].

Table 2

Comparative positioning of the proposed survey against representative XAI review literature. Each row characterizes a prior study across its scope, evidence base, and treatment of CV, with the final column stating how the proposed survey extends it.

Study	Main scope	Coverage and evidence base	Main XAI focus	Treatment of CV	How proposed survey extends it
Ali et al. [82]	Trustworthy AI and XAI taxonomy	410 articles, mainly 2016 to 2022	Data, model, and post hoc explainability; explanation assessment	Discusses visual explanations, but is not a CV technical analysis	Updates evidence from 2020 to 2026; adds technical analysis of CV methods by mechanism, robustness-aware characterization, and reproducibility-focused discussion
Kalasampath et al. [83]	Application-level XAI review	Literature searched up to 2024	XAI applications across healthcare, finance, agriculture, cybersecurity, law, and education	Covers visualization methods in application contexts without technical model analysis	Adds method family taxonomy, evaluation protocol discussion, and human study synthesis across seven domains
Gipiskis et al. [57]	XAI in semantic image segmentation	Google Scholar with citation chaining; segmentation focused, up to 2024	XAI for dense prediction and segmentation tasks	Strong focus on CV, medical, and industrial segmentation	Extends to seven domains under a unified review framework and adds human study synthesis and governance discussion absent from the segmentation scope
Chaddad et al. [84]	Explainable deep learning for medical image computing	Three medical datasets (brain tumor, skin cancer, chest X-ray) with four CNN models	Grad-CAM++, EigenGradCAM, XGradCAM, AblationCAM, LayerCAM	Strong medical imaging and CAM family focus	Places CAM methods within a wider seven-family taxonomy covering attribution, surrogates, counterfactuals, concepts, prototypes, and mechanistic analysis
Cheng et al. [56]	XAI in CV	534 retrieved; 83 selected after screening; January 2018 to December 2025	Attribution, activation, perturbation, and transformer-based XAI	Strong CV focus across classification, detection, and segmentation tasks	Extends beyond CV to NLP, tabular, time series, and graph learning, RL, and recommendation systems; adds systematic review protocol, human study synthesis, and governance discussion
Ahmed et al. [58]	XAI in medical imaging	PRISMA systematic review; 980 records identified; 133 studies included; 2026	Saliency, attention, gradient-based, rule-based, GNN, and multimodal explanations	Strong radiology and pathology relevance	Generalizes beyond medical imaging to seven domains; adds technical CV analysis through visual interpretability protocols and addresses the standardization gap with a structured review framework
Proposed survey	XAI methods, technical analysis, protocols, and open gaps	452 screened records; 189 works from 2014 to 2026	Seven method families with primary CV technical depth; human studies, governance, and reproducibility	Technical analysis of CV methods by mechanism, model access, and evaluation protocol; visual interpretability applied to LiSalNet covering fusion impact, Grad-CAM localization, and multiscale hierarchy visualization	Provides an updated, protocol-oriented, and deployment-aware survey linking method choice, CV technical analysis, human evidence, and governance discussion across seven domains

Table 2 positions the proposed survey against six representative XAI reviews published between 2023 and 2026. The comparison shows that existing reviews either address individual domains such as medical imaging or image segmentation, or cover application areas without protocol-level evaluation. None simultaneously addresses technical CV analysis by mechanism, human study synthesis, visual interpretability protocols, and governance discussions across seven domains. The proposed survey fills this gap by providing a structured, deployment-aware review that connects method characterization, evaluation evidence, and audit-oriented reporting [88].

3. Search and selection methodology

This section describes the search strategy, source coverage, and selection criteria used to assemble the literature analyzed in this survey. The review window is fixed from 2020 to 2026. Peer-reviewed archival venues are strictly preferred over preprints. Foundational pre-2020 sources are retained where they define methods and reporting standards that remain in standard use. The scope covers seven domains: vision, natural language processing, tabular data, time series, graphs, reinforcement learning, and recommendation systems. Method families include feature attribution, saliency maps, surrogate models, counterfactuals, concept methods, prototypes, and mechanistic analyses. The section details the search sources, eligibility criteria, and selection steps. Search queries, source lists, and selection decisions are documented for reproducibility.

3.1. Protocol and governance

The search objectives, eligibility criteria, and query families were defined before conducting searches to ensure consistent and reproducible selection. The scope, inclusion boundaries, screening decision logic, and data extraction schema were documented at the outset and applied uniformly throughout the selection process. Source lists, query strings, run dates, and platform versions were recorded to enable the search to be reproduced. Version control procedures were applied to research scope, eligibility criteria, search strings, and analysis. Any amendments to the original scope or criteria were logged, along with their rationale and the affected steps. Records were screened against a standardized checklist to maintain consistency in decisions, and a brief rationale was documented for each exclusion to build an audit trail. Data handling procedures complied with all applicable license terms, and restricted corpora remained subject to those terms. The final documentation package contained the search queries, the selection flow, screening decisions, and the extraction records used to produce the synthesis in this survey [87].

3.2. Eligibility and screening

Study selection follows written criteria agreed upon before conducting the search. The goals are topical relevance, research merit, and results that others can verify. The scope encompasses XAI for CV, natural language processing, tabular data, time series, graph learning,

reinforcement learning, and recommendation systems. Records must be in English with full text available. A standardized checklist is applied to all records to maintain consistency in decisions. Reviewers document a brief rationale for each inclusion or exclusion decision to build an audit trail. Reference lists and citation chains are reviewed under the same inclusion criteria. Records from multiple sources are deduplicated before screening to prevent double-counting. When a preprint has a corresponding peer-reviewed version, the archival publication is retained, and the preprint is linked. The criteria below govern inclusion and exclusion.

Inclusion

- **Topic fit:** Work on XAI with quantifiable evaluation metrics. Includes methods and evaluations for feature attribution, saliency, rule and surrogate models, counterfactuals and recourse, concept-based techniques, prototype or example-based explanations, mechanistic analyses, human studies, benchmarks, and toolkits with reported metrics.
- **Time and language:** Published in English, with the full text available.
- **Sources and venues:** ACM Digital Library, IEEE Xplore, NeurIPS and ICML proceedings (Curran Associates and PMLR), ACL Anthology for NLP venues, AAAI and IJCAI conference proceedings, arXiv preprints when no archival version exists, and selected standards and policy documents from NIST, ISO, and IEC, the EU AI Act, the IEEE Standards Association, ACM policy, and OpenRAIL. OpenReview is used only for ICLR final versions. Archival venues are preferred over preprints.
- **Contribution type:** Method, evaluation, dataset, or benchmark, toolkit, or system report with quantitative results.
- **Reporting:** Clear description of setup, datasets, protocols, and metrics. Studies must report at least one quantitative evaluation metric

and specify the dataset used. Records missing both a dataset name and a performance metric are excluded.

Exclusion

- **Out of scope:** Unverifiable demonstrations. Proprietary systems without evaluation details. Essays without measurable outcomes. Domain-specific legal analyses lacking technical evaluation components. Long multi-modal deep dives outside the primary domains.
- **Non-research items:** Editorials, opinion pieces, tutorials, conference posters, or abstracts without accompanying full papers.
- **Coverage:** Non-English items. Work published outside the 2020 to 2026 timeframe. Records with missing full text.
- **Insufficient detail:** Records with missing or unclear methodological descriptions, datasets, evaluation metrics, or experimental configurations.
- **Access and licensing:** Benchmarks without public splits or terms that permit academic use.
- **Duplicates:** Duplicate records or preprints superseded by a peer-reviewed version.

Records are screened first by title and abstract, then by full text, against the inclusion and exclusion criteria listed above, with a brief rationale documented for each exclusion decision.

3.3. Information sources and search strategy

The search strategy employs multiple digital libraries, conference archives, journal platforms, preprint services, and policy repositories. This approach captures XAI research across CV, natural language processing, tabular data, time series, graph analysis, reinforcement learning, and recommender systems. Fig. 3 illustrates the main digital libraries and conference series that form the core search sources, showing the

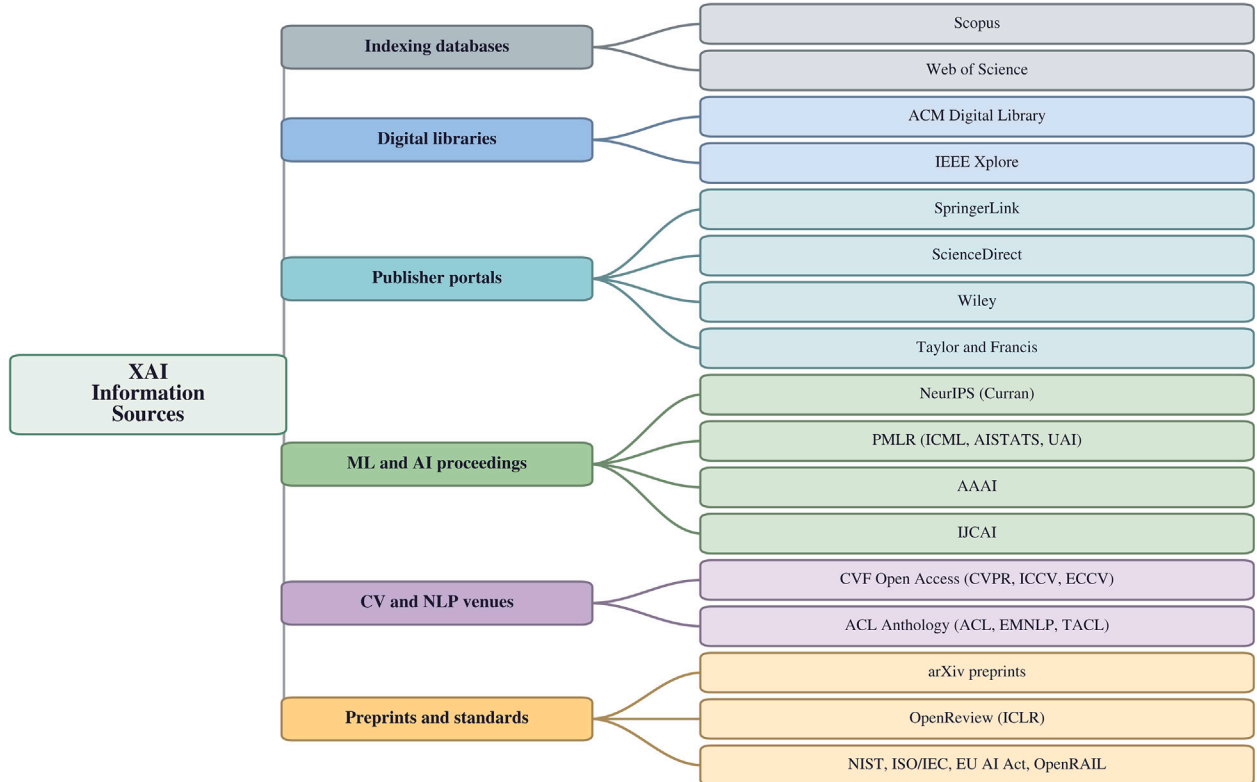


Fig. 3. Hierarchy of digital libraries, proceedings, and policy sources used for the XAI search, organized from broad indexing databases through publisher portals to domain-specific archival venues.

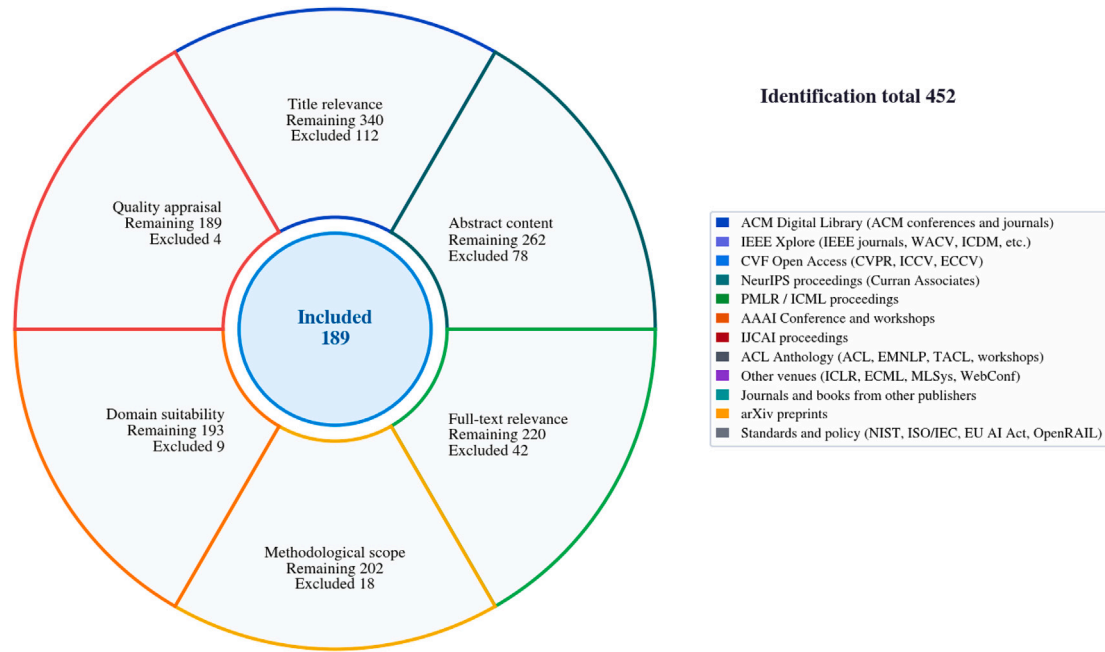


Fig. 4. Selection flow summarizing the identification and screening of 452 XAI records across six stages, yielding 189 works analyzed in this survey.

Table 3

Sources, venue mapping, fields searched, filters, and archival preference, grouped by the six source categories in Fig. 3.

Source category	Venue mapping	Fields searched	Filters	Archival rule
Indexing databases	Scopus, Web of Science	Title, abstract, author keywords	English; 2020 to 2026	Prefer archival
Digital libraries	ACM Digital Library, IEEE Xplore	Title, abstract, author keywords	English; 2020 to 2026	Prefer archival
Publisher portals	SpringerLink, ScienceDirect, Wiley, Taylor and Francis	Title, abstract, author keywords	English; 2020 to 2026	Prefer archival
ML and AI proceedings	NeurIPS (Curran), PMLR (ICML, AISTATS, UAI), AAAI, IJCAI	Title, abstract, author keywords	English; 2020 to 2026	Archival
CV and NLP venues	CVF Open Access (CVPR, ICCV, ECCV), ACL Anthology (ACL, EMNLP, TACL)	Title, abstract, author keywords	English; 2020 to 2026	Archival
Preprints and standards	arXiv, OpenReview (ICLR), NIST, ISO/IEC, EU AI Act, OpenRAIL	Title, abstract, author keywords	English; 2020 to 2026	Preprint linked to archival; standards reference only

hierarchical relationship between broad indexing databases and domain-specific archival venues used to ensure comprehensive coverage. Core sources include the ACM Digital Library, IEEE Xplore, which covers major CV venues (CVPR, ICCV, WACV, ICDM), NeurIPS proceedings archived by Curran Associates, and the Proceedings of Machine Learning Research for ICML and related conferences. Additional archival sources include the AAAI Digital Library, IJCAI proceedings, and the ACL Anthology, which covers ACL, EMNLP, TACL, and associated workshops. Other conference sites include ICLR, ECCV, ECML, MLSys, and the Web Conference. Journals and books from other publishers are queried through their own platforms. Preprints are taken from arXiv only when no archival or camera-ready version is available. Standards and policy documents are collected from NIST, ISO, IEC, the EU AI Act, the IEEE Standards Association, ACM, and the OpenRAIL license family. Searches are limited to English-language records from 2020 to 2026. A small curated set of foundational XAI papers from 2014 to 2019 is added by backward citation chaining from key methods such as LIME, SHAP, Integrated Gradients, and Grad-CAM. Boolean query families target seven method categories: feature attribution, saliency maps, surrogate and rule-based models, counterfactual and recourse approaches, concept-based techniques, prototype and example-based explanations, and mechanistic circuit-level analyses. Additional queries target evaluation metrics, faithfulness criteria, human-subject studies, decision-support applications, and computational efficiency measure-

ments. Fielded searches span titles, abstracts, and author keywords for each source. Backward and forward citation chaining is applied to seed papers in top venues. Recent XAI tracks in flagship conferences are hand-searched. Deduplication proceeds by DOI match, then exact title, then normalized title with an author-year tiebreaker. The retained record keeps the most complete metadata available. All merges are logged for reproducibility. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow documents the steps for identification, screening, eligibility, and inclusion. Fig. 4 summarizes these stages using a radial PRISMA diagram, showing the record counts at each of the six screening stages and the reasons for exclusion that reduced the initial pool of 452 records to the 189 works analyzed in this survey.

Table 3 summarizes data sources, venue mappings, search fields, temporal filters, and archival preferences, and follows the same six source categories shown in Fig. 3. It documents query procedures for each source and export processes for automated deduplication and logging. Fig. 5 reports the yearly selection profile after deduplication and screening, revealing the growth in XAI publications after 2019 and confirming that the 2020 to 2026 window captures the period of most active methodological development. It shows the number of XAI papers per publication year from 2014 to 2026, with publications from 2014 to 2019 classified as foundational literature and 2020 to 2026 forming the main review window.

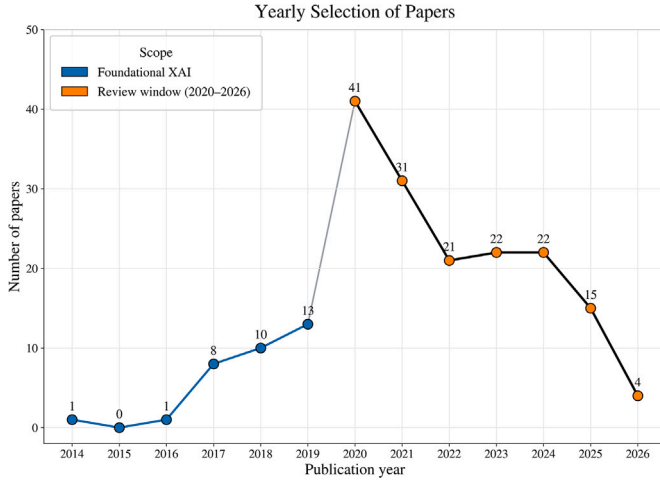


Fig. 5. Year-wise distribution of the 189 works analyzed in this survey from 2014 to 2026. Foundational methods from 2014 to 2019 are retained as standard tools; the primary synthesis window covers 2020 to 2026.

Table 4

Data items captured for each included study.

Field block	Details captured
Bibliography	Title, authors, venue, year, DOI, stable identifiers, links to code or data if disclosed
Method family and scope	Method family; intrinsic or post hoc; local or global; model-agnostic or model specific
Model and training	Model family, architecture tags, training objective, datasets used for training or fine-tuning if relevant
Datasets and splits	Dataset names, size, domain, license, version pins, splits, leakage notes
Metrics and definitions	Primary and secondary metrics, exact definitions or references, graders if any
Protocol settings	Seeds, shots, baselines or background sets, path or step counts, sampling budgets, prompt patterns if used
Efficiency and hardware	Devices, precision, batch size, peak memory, throughput, end-to-end latency, energy or carbon if reported
Human studies	Design, tasks, sample sizes, blinding, outcomes, effect sizes where available
Controls and stress tests	Sanity checks, ablations, robustness tests, adversarial or noise tests, retraining-based tests
Governance and risk	Bias or fairness analysis, privacy notes, license clarity, policy or audit notes
Reproducibility assets	Released code, models, configs, commit hash, data availability, run logs
Reporting gaps	Missing fields, unclear settings, absent seeds, absent energy, other gaps with flags

3.4. Extraction and appraisal

Each included study is converted into a uniform record using a locked extraction form. Extraction fields capture seven categories of information: method characteristics, experimental details, evaluation protocols, computational metrics, reproducibility information, human study designs, and governance considerations. Units are normalized where safe. Original units are preserved and flagged when normalization is not possible. Records receive gap flags when items are missing or unclear. Quality appraisal scores each record on five pillars: methodological clarity, evaluation protocol transparency, presence of validation controls or sanity checks, reproducibility indicators, and governance and risk assessment reporting. A three-level scale is used for each pillar, with marks of adequate, partial, or absent, scored as 2, 1, and 0, respectively. See Table 4 for the data items captured.

3.5. Synthesis, contamination, and threats

Results are synthesized by domain and method under matched settings. Matched settings are operationally defined as studies that share all of the following conditions: the same baseline input or background set, the same sampling budget, the same model architecture family, the same evaluation dataset and split, the same random seed specification, and the same hardware profile or a documented equivalent. For each method family and task type, performance metrics (deletion/insertion AUC, comprehensiveness, sufficiency) and computational metrics (latency, memory, energy) are compared only under these identical experimental conditions. Where studies report them, sensitivity to baselines, background sets, and sampling budgets is noted as a factor affecting stability. Duplicate and near-duplicate records are removed during deduplication, and contamination concerns such as train-test overlap and leakage are noted where the source studies discuss them. Where reported, prediction accuracy is summarized alongside computational cost. Threats to validity include selection bias from database coverage, publication bias from positive reporting, metric mismatch across studies, benchmark drift over time, confounding by compute budget, code or data unavailability, and reviewer disagreement. To mitigate these threats, the review applies multi-source database searches, systematic citation chaining, PRISMA-style reporting, documented searches, and explicit recording of negative findings. The National Institute of Standards and Technology (NIST)'s four principles of explainable AI guide the controls. These principles state that explanations must be evidence-based, meaningful to users, accurate in reflecting the system's process, and bounded by known knowledge limits [89]. They inform risk reporting and highlight where current studies fall short. By transparently documenting synthesis and threats, the review supports reproducible and trustworthy conclusions. In summary, the pipeline proceeds as follows. Search objectives, eligibility criteria, and queries are defined before searching. Multi-library searches across six source categories retrieve an initial pool of 452 records, which are deduplicated by DOI, title, and author-year tiebreaker. Records are screened by title and abstract, then by full text, against the inclusion and exclusion criteria, yielding 189 works analyzed in this survey. Each work is recorded in a uniform extraction form, and results are synthesized by domain and method family, with threats to validity documented and mitigated through PRISMA-style reporting and explicit recording of negative findings [87].

4. Visual interpretability techniques

This section presents our LiSalNet framework for visual interpretability as a complement to the synthesis in Sections 3 and 5. It is neither a review of visual interpretability papers nor an empirical comparison of competing methods. Rather, it is an illustrative and methodological contribution that shows how three established visualization techniques can be applied together in a unified pipeline to expose model reasoning in a concrete and auditable way. The goal is to address the gap identified in Section 1.1, where reviewed studies lacked consistent visual protocols that simultaneously show the impact of fusion, attention localization, and multiscale feature behaviour. The visualizations use techniques documented in the reviewed literature, namely gradient-weighted class activation mapping [9], fusion impact analysis using gradient signals [85], and hierarchical feature visualization [10], which are applied to representative examples to illustrate their complementary roles. They are not new empirical results and should be read as a demonstration of best practices rather than as primary evidence. The three components are organized as follows. Section 4.1 analyzes multimodal feature fusion and its interpretable impact through gradient weighted visualizations (Fig. 6). Section 4.2 details Grad-CAM for localizing discriminative regions (Fig. 7). Section 4.3 examines multiscale feature hierarchies from low-level to high-level representations (Fig. 8).

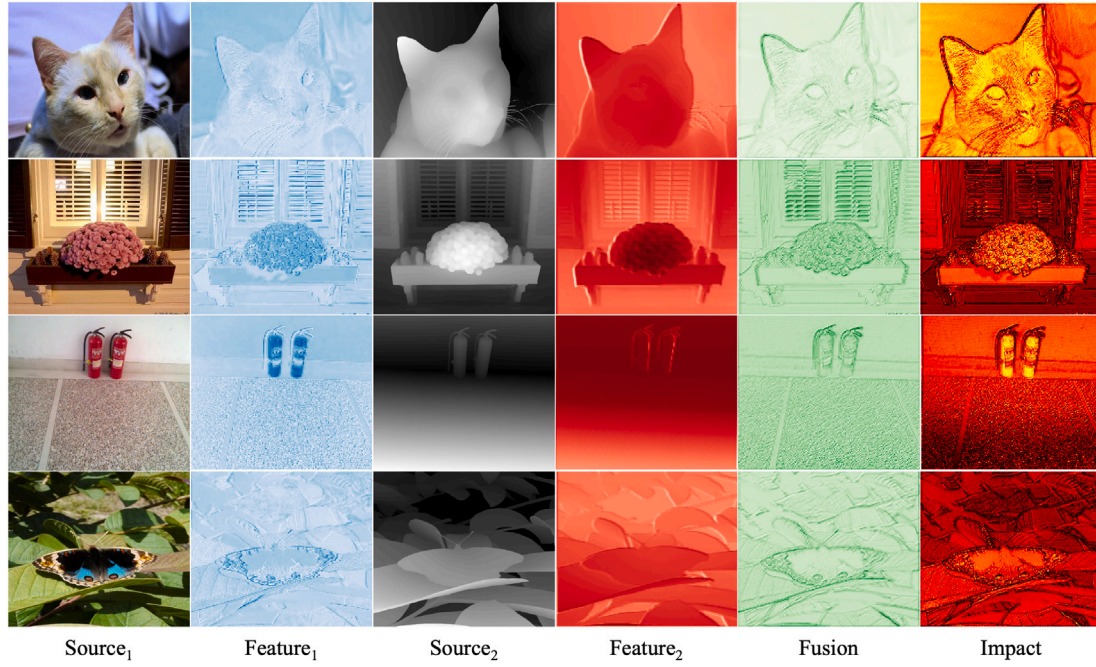


Fig. 6. Multimodal feature fusion impact analysis. The pipeline shows sensor inputs, extracted features, fused representation, and impact map. High-impact regions indicate maximal discriminative benefit, while low-impact areas suggest redundancy.

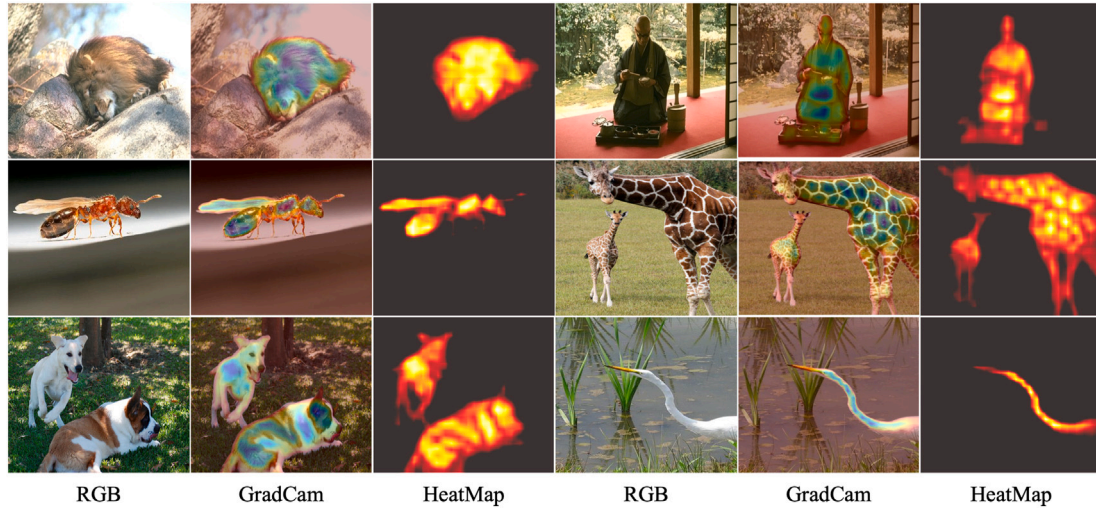


Fig. 7. Gradient weighted class activation mapping for visual attention localization. Each row shows the input, a Grad-CAM overlay with warm colours for high attention and cool colours for low attention, and an isolated heatmap showing raw activation intensity. (For interpretation of the references to colour in this Figure legend, the reader is referred to the web version of this article.)

4.1. Multimodal feature fusion impact analysis

Multimodal fusion combines complementary information from heterogeneous sensors to improve robustness and discriminative power. Fig. 6 traces the progression from raw sensor inputs through intermediate feature representations to fusion outcomes and their interpretable impact. High-impact regions concentrate on semantically meaningful areas such as object boundaries and foreground regions, while low-impact areas correspond to redundant or background information. The first column (Source₁) presents RGB images capturing photometric appearance and texture. The second column (Feature₁) shows feature maps from the RGB encoder, downsampled by 8× to 16× with 256 to 512 channels. The third column (Source₂) displays the complementary modality, depth for RGB-D fusion or thermal for RGB-T fusion. The fourth column (Feature₂)

provides the corresponding representations from the secondary encoder with matched spatial and channel dimensions.

The fifth column (Fusion) visualizes the integrated feature space after the fusion operator, which must preserve spatial correspondence while maximizing information gain. The sixth column (Impact) quantifies which spatial regions contribute most to the final prediction, with high-impact regions marking where fusion provides the greatest benefit and low-impact regions indicating redundancy. The benefits differ by scenario: in the cat image (row 1), RGB provides texture, while depth disambiguates spatial geometry and sharpens the facial region; in the fire extinguisher (row 3), fusion sharpens boundaries and suppresses the background.

Fusion can be performed early (raw sensor concatenation), late (score level combination), or at the intermediate feature level, and

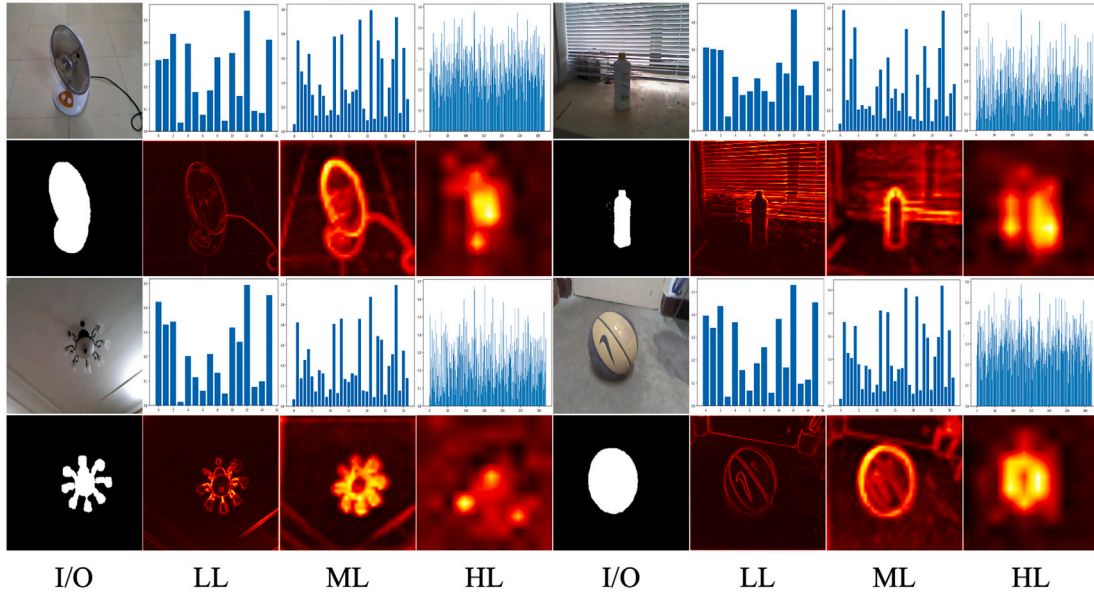


Fig. 8. Multiscale feature hierarchy visualization from low-level to high-level representations. Each row displays input and output (I/O), low-level features (LL) that capture edges and textures, mid-level features (ML) that encode part-based representations, and high-level features (HL) that abstract semantic class information. Bar charts quantify channel activation distributions.

cross-modal attention modules compute query, key, and value projections across modalities to weight evidence adaptively by scene context. The fusion process is written as

$$\mathbf{F}_{\text{fused}} = \phi(\mathbf{F}_1, \mathbf{F}_2; \theta), \quad (1)$$

where $\mathbf{F}_1 \in \mathbb{R}^{H \times W \times C_1}$ and $\mathbf{F}_2 \in \mathbb{R}^{H \times W \times C_2}$ are the features from the two modalities, ϕ is the fusion operator parameterized by θ , and $\mathbf{F}_{\text{fused}} \in \mathbb{R}^{H \times W \times C_{\text{out}}}$ is the integrated representation. The impact map $\mathbf{I} \in \mathbb{R}^{H \times W}$ is

$$\mathbf{I} = \left\| \frac{\partial y}{\partial \mathbf{F}_{\text{fused}}} \right\|, \quad (2)$$

which highlights the regions where fusion contributes most to the prediction.

4.2. Gradient weighted class activation mapping

Grad-CAM produces coarse localization maps from the gradient of the target class score with respect to the final convolutional feature map, followed by global average pooling to obtain channel-wise importance weights. Fig. 7 applies Grad-CAM across object categories and localizes discriminative regions even for small objects such as ants and for multi-instance scenes with more than one subject. The first column (RGB) shows the input resized to the model resolution. The second column (Grad-CAM) overlays the gradient-weighted activation map, with discriminative regions in warm colours and less relevant areas in cool colours. The third column (HeatMap) shows the raw activation intensity without the overlay, giving a clearer view of the spatial attention distribution.

Grad-CAM extracts feature maps $\mathbf{A}^k \in \mathbb{R}^{H \times W}$ from the last convolutional layer, where k indexes channels and H and W are the spatial height and width. For a target class c , the importance weight α_k^c for channel k is

$$\alpha_k^c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \frac{\partial y^c}{\partial A_{i,j}^k}, \quad (3)$$

where y^c is the class score before softmax. The Grad-CAM heatmap $I_{\text{Grad-CAM}}^c \in \mathbb{R}^{H \times W}$ is

$$I_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c \mathbf{A}^k \right), \quad (4)$$

where ReLU keeps only positive influences. The map is upsampled to the input resolution by bilinear interpolation and normalized to $[0, 1]$. The examples in Fig. 7 show precise localization across scenarios: row 1 activates strongly on the animal body and head, the ant in row 2 concentrates on the thorax and legs despite its small size, and the dog scene in row 3 highlights both animals, indicating multi-instance attention.

Beyond gradient-based maps, attention heatmaps offer complementary visualization by reading learned attention weights directly from transformer architectures. They visualize the attention scores $\mathbf{A}_{\text{att}} \in \mathbb{R}^{N \times N}$ from the self attention operation, where N is the number of tokens or spatial patches. For a query token q_i , the attention weight to key token k_j is

$$\mathbf{A}_{\text{att}}(i, j) = \text{softmax} \left(\frac{q_i \cdot k_j}{\sqrt{d_k}} \right), \quad (5)$$

where d_k is the key dimension. These weights reveal which spatial regions or semantic concepts the model associates during prediction. Unlike Grad-CAM, which needs a backward pass, attention heatmaps are available from forward pass activations, which makes them efficient for real-time interpretation, and multi-head attention yields several maps per layer that expose diverse feature relationships across representation subspaces.

4.3. Multiscale feature hierarchy visualization

Deep networks build hierarchical representations across multiple scales of abstraction. Fig. 8 shows this hierarchy across input and output (I/O), low-level (LL), mid-level (ML), and high-level (HL) representations, where channel activation distributions become progressively sparser with depth as the network shifts from dense texture detection to selective semantic encoding. The first column (I/O) displays the input image with the binary ground truth mask to establish the task goal,

and the bar charts above each feature panel summarize the distribution of activation magnitudes across channels. The second column (LL) shows low-level features from early layers, typically at 1/4 or 1/2 of the input resolution with 64 to 128 channels, capturing edges, corners, textures, and colour gradients with high spatial fidelity but limited semantic content.

The third column (ML) shows mid-level features from intermediate layers, downsampled to 1/8 or 1/16 resolution, with 256 to 512 channels, which encode part-based representations, spatial relationships, and moderate semantics, balancing spatial precision and abstraction. The fourth column (HL) shows high-level features from deep layers, downsampled to 1/16 or 1/32 resolution with 512 to 2048 channels, which abstract away spatial detail and encode semantic class information, global context, and task-specific discriminative cues. The activation bar charts quantify feature diversity and sparsity: low-level features are dense and nearly uniform with many active channels, mid-level features are moderately sparse with selective activation, and high-level features are highly sparse with a few dominant channels encoding class-specific patterns. Sparsity is measured by the Gini coefficient or the L_1/L_2 norm ratio, with values closer to 1 indicating higher sparsity. The Gini coefficient is

$$G = \frac{\sum_{i=1}^C (2i - C - 1) \cdot a_i}{C \sum_{i=1}^C a_i}, \quad (6)$$

where C is the number of channels, a_i is the activation magnitude of channel i sorted in ascending order, and $G \in [0, 1]$. The L_1/L_2 ratio is

$$R = \frac{\|\mathbf{a}\|_1}{\sqrt{C} \cdot \|\mathbf{a}\|_2}, \quad (7)$$

where $\mathbf{a} \in \mathbb{R}^C$ is the channel activation vector, with $R \approx 1$ for uniform activations and $R \ll 1$ for sparse ones. This analysis informs architectural design by identifying which layers contribute most to performance, guiding pruning, and exposing redundancy. Removing low-level features in dense prediction causes boundary localization errors, while removing high-level features degrades semantic classification, which is why multi-scale architectures add lateral connections that fuse high-level semantic and low-level spatial information.

5. Results and analysis

This section reports XAI results from 2020 to 2026 in a single, reproducible format. Section 5.1 defines the base predictors and their hooks in Table 5. Section 5.2 fixes datasets, protocols, and metrics in Table 7 so scores are comparable. Section 5.3 summarizes reported effects on trust, reliance, accuracy, and time in Table 8 and provides a developer workflow with hooks, logs, and guardrails in Table 9. Section 5.4 presents an adaptation menu of efficiency levers with their supported efficiency and fidelity effects and the validation each requires in Table 10. Section 5.5 aligns governance and reproducibility with current standards and delivers a concise checklist.

5.1. Landscape: models and methods

This section charts the landscape of XAI models from 2020 to 2026 in Table 5, where each row supports benchmarking. Domain sets the area and Task the goal, such as classification, detection, segmentation, forecasting, node or edge prediction, control, click-through rate (CTR), or sequence modeling. Family names the model line, and Input sets the default image size or token budget used for timing and floating-point operations per second (FLOPs) for frameworks. Prec lists numeric precision, fp16 for most vision and natural language processing (NLP) encoders, and fp32 for forecasting, graphs, tabular, and reinforcement learning (RL). XAI hooks specify the accessible model components, for example, gradients, attention, intermediate layers, masks, feature importance, or subgraph views. Typical data lists representative bench-

marking datasets, and Env records the runtime for frameworks such as PyTorch, JAX, or TensorFlow. Vision classification uses convolutional networks and Vision Transformers at 224 with fp16 [90–92]. Object detection covers architectures such as DETR and YOLO [93,94]. Segmentation includes SegFormer and Mask2Former [95,96]. NLP spans masked encoders and encoder–decoder sequence-to-sequence models [97,98], together with causal and mixture of experts language models [99,100]. Time series covers Temporal Fusion Transformers and Informer [101,102] as well as TimesNet and PatchTST [103,104]. Graph learning includes Graphormer, spectral attention, and hybrid GPS [105–107]. RL includes MuZero and DreamerV3 [108,109]. Recommender systems include DCN v2 for CTR and the sequence models S3-Rec and DuoRec [110–112].

Explanation methods pair with these model families across the seven domains. For CV, Table 6 catalogs representative techniques by mechanism, recording the model access each requires, the explanation it produces, the evaluation typically applied, and its publication venue. Gradient and activation saliency maps include Grad-CAM, Grad-CAM++, and Score-CAM [9,37,113,114], complemented by perturbation methods such as RISE, occlusion, and meaningful perturbations [27,115,116]. For NLP, common choices include Integrated Gradients and DeepLIFT [7,117], the local surrogates LIME and SHAP [8,11], and concept testing with TCAV [14]. Graph explainers return subgraph masks through GNNExplainer or PGExplainer [118,119], and reinforcement learning relies on agent saliency and rollout perturbations [27,120]. Across all domains, validation uses checks such as sanity tests and ROAR [26,74], following the input specifications and numerical precision in Table 5 and logging fixed random seeds, preprocessing pipelines, batch sizes, peak memory, and median latency.

Several cross-domain trends emerge from the synthesized evidence. Gradient-based methods such as Grad-CAM and Integrated Gradients dominate vision and NLP for their low overhead but trade localization fidelity for speed, especially under distribution shift. Perturbation-based methods such as RISE and SHAP, and exact Shapley estimators, score higher on deletion and insertion benchmarks but cost far more, making them impractical for edge deployment. Surrogate methods such as LIME are the least stable, adequate in tabular and recommender settings but contradictory in high-dimensional vision and NLP. Graph learning and reinforcement learning show the largest evaluation gaps, with fewer standardized benchmarks, smaller human study samples, and less computational profiling than vision and NLP. Three patterns cut across all domains: high-faithfulness methods trade off against efficiency and energy transparency while low-overhead methods trade stability for speed; human-subject evidence is the weakest part of the evidence base everywhere, with small samples and inconsistent effect-size reporting; and the absence of energy reporting recurs throughout, leaving practitioners unable to make informed deployment decisions under resource constraints.

5.2. Data and metrics

Table 7 summarizes the task benchmarks and task metrics used across the model families cataloged in Table 5. For vision classification, ImageNet Real is evaluated with Top-1 accuracy [157]. Object detection uses Open Images V6 with mean average precision (mAP) [158], and segmentation uses Waymo Open with mean intersection over union (mIoU) [159]. For natural language understanding, SuperGLUE provides accuracy and F1 [160]. Time series forecasting on ETT and ECL is evaluated with mean absolute error (MAE) and mean squared error (MSE) [102]. Graph learning follows the Open Graph Benchmark (OGB) using accuracy and ROC AUC [161]. For reinforcement learning, D4RL reports normalized return for offline evaluation, and Progen reports average return across procedurally generated environments [162, 163]. Recommender systems use MIND for news recommendation and KuaiRec for short-video recommendation, both evaluated using AUC, NDCG, and HR@K [164,165]. Tabular learning follows TabPFN style

Table 5

Base predictor catalog by domain, task, model family, input format, precision, XAI hooks, typical datasets, environment, and references.

Domain	Task	Family/model	Input	Prec.	XAI hooks	Typical data	Env.	Ref.
Vision	Cls	ConvNeXt	224×224	fp16	grad, layers	ImageNet-1K	PyTorch	[90]
	Cls	Vision Transformer	224×224	fp16	grad, attn	JFT-300M, ImageNet-21K, ImageNet-1K	JAX or PyTorch	[91]
	Det	DETR	800×1333	fp16	grad, attn	COCO	PyTorch	[93]
	Det	YOLOX	640×640	fp16	grad, layers	COCO	PyTorch	[94]
	Seg	SegFormer	512×512	fp16	grad, attn	ADE20K, Cityscapes	PyTorch	[95]
	Seg	Mask2Former	512×512	fp16	grad, attn	COCO, ADE20K, Cityscapes, Mapillary	PyTorch	[96]
	Seg	Segment Anything Model	prompt + image	fp16	masks, prompts, attn	SA-1B, zero-shot segmentation	PyTorch	[121]
	Cls/Feat	DINOv2	224×224 or variable	fp16	layers, attn	LVD-142M, ImageNet, dense tasks	PyTorch	[122]
	Cls/Backb.	MambaVision	224×224	fp16	grad, layers	ImageNet-1K, COCO, ADE20K	PyTorch	[123]
NLP	Det	YOLOv12	640×640	fp16	grad, layers, attn	MSCOCO 2017	PyTorch	[124]
	NLU/Cls	ELECTRA pre-trained encoder	512 tok	fp16	grad, attn	BookCorpus, Wikipedia, GLUE	TF or PyTorch	[125]
	Long seq	Longformer	4096 tok	fp16	grad, attn	Long-document tasks	PyTorch or TF	[126]
	Cls	DeBERTa	512 tok	fp16	grad, attn	GLUE, SuperGLUE	PyTorch	[97]
	Gen	T5 / Seq2Seq Transformer	512 tok	fp16	grad, attn	C4, supervised text-to-text tasks	JAX or PyTorch	[98]
	Gen	LLaMA 2 causal LM	4k tok	fp16	grad, attn	Web text and code	PyTorch	[99]
	Gen	Mixtral MoE LM	32k tok	fp16	grad, attn	Web text and code	PyTorch	[100]
	Seq	Mamba SSM	long tokens	fp16	states, activations	language, audio, genomics	PyTorch	[127]
	Gen	Gemma open LM	8k tok	fp16/bf16	grad, attn	web text, code, instruction data	JAX or PyTorch	[128]
Tabular	Gen/Reason	DeepSeek-R1	long context	fp16/bf16	grad, attn, traces	reasoning, math, code, STEM	PyTorch	[129]
	Gen/MoE	Qwen3	long context	fp16/bf16	grad, attn	multilingual web, code, math	PyTorch	[130]
	Cls	TabNet	p features	fp32	masks	UCI, credit, census	PyTorch	[131]
	Cls	FT-Transformer	p features	fp32	grad, attn	Tabular benchmarks	PyTorch	[132]
	Cls	TabPFN / prior-fitted Transformer	p features	fp32	post-hoc feat. import.	Small tabular classification	PyTorch	[133]
	Cls/Reg	TabR	p features	fp32	retrieval, feat. contrib.	public tabular benchmarks	PyTorch	[134]
	Cls/Reg	TabM	p features	fp32	ensemble contrib.	public tabular benchmarks	PyTorch	[135]
	Cls/Reg	TabPFN v2	p features	fp32	post-hoc feat. import.	small/medium tabular data	PyTorch	[136]
Time series	Fcst	Temporal Fusion Transformer	$L \times D$	fp32	attn, var-select	Multi-horizon forecasting data	PyTorch	[101]
	Fcst	TimesNet / TimesBlock	$L \times D$	fp32	grad	Forecasting benchmarks	PyTorch	[103]
	Fcst	PatchTST / patch Transformer	$L \times D$	fp32	attn	Long-horizon forecasting data	PyTorch	[104]
	Fcst	TimesFM	context window	fp32/bf16	attn, residuals	large-scale time-series corpus	JAX	[137]
	Fcst	Moirai	$L \times D$	fp32/bf16	attn, patches	LOTSa, Monash, zero-shot data	PyTorch	[138]
	Fcst	Chronos	tokenized series	fp32/bf16	token scores, attn	public and synthetic time-series data	PyTorch	[139]
	Fcst	Time-MoE	variable context	fp16/bf16	attn, experts	Time-300B, multi-domain data	PyTorch	[140]
Graph	GNN	Graphormer	n nodes	fp32	subgraph, attn	OGB, OGB-LSC	PyTorch	[105]
	GNN	Spectral Graph Transformer	n nodes	fp32	attn	Graph benchmarks	PyTorch	[106]
	GNN	GraphGPS / hybrid Transformer	n nodes	fp32	attn	OGB, graph benchmarks	PyTorch	[107]
	GNN	Exphormer	n nodes	fp32	sparse attn	graph benchmarks, OGB	PyTorch	[141]
	GNN	SGFormer	n nodes	fp32	global attn	large-graph benchmarks	PyTorch	[142]
	GNN	Graph Mamba Networks	n nodes	fp32	states, layers	long-range graph benchmarks	PyTorch	[143]
	GNN	Graph-Mamba	n nodes	fp32	states, node selection	long-range graph sequence data	PyTorch	[144]
RL	Control	MuZero planning model	96×96 or state	fp32	saliency	Atari, Go, chess, shogi	JAX or reimpl.	[108]
	Control	DreamerV3 world model	64×64 or state	fp32	saliency, latent states	Atari, DMControl, Minecraft	JAX	[109]
	Control	EfficientZero V2	image or state	fp32	latent states, rollout	Atari100K, Proprio Control, Vision Control	PyTorch	[145]
	Control	Genie world model	video/image	fp16/bf16	latent actions, rollouts	unlabelled video, interactive worlds	JAX	[146]
	Control	TD-MPC2	prompt state or image	fp32	latent states, planning	DMControl, Meta-World, ManiSkill2	PyTorch	[147]
	Control	Genie 3 world model	text/image prompt	fp16/bf16	latent rollouts	interactive generated environments	internal/research preview	[148]

(continued on next page)

Table 5 (continued)

Domain	Task	Family/model	Input	Prec.	XAI hooks	Typical data	Env.	Ref.
Recsys	CTR	DCN-V2 hybrid model	id + cont.	fp32	feat. contrib.	Criteo, web-scale ranking data	TF or PyTorch	[110]
	Seq	S3-Rec self-supervised model	seq ids	fp32	attention	Amazon sequential data	PyTorch	[111]
	Seq	Contrastive sequential model	seq ids	fp32	attention	Amazon sequential data	PyTorch	[112]
	Graph/LLM	LLMRec	graph + text	fp16	graph aug., token scores	recommendation graphs	PyTorch	[149]
	Seq/Gen	TIGER generative retrieval	semantic id seq	fp16	token scores	Amazon, sequential data	PyTorch	[150]
	Seq	Mamba4Rec	seq ids	fp32	states, sequence saliency	Amazon, MovieLens	PyTorch	[151]
	Seq	Selective Gated Mamba	seq ids	fp32	states, gates	sequential recommendation data	PyTorch	[152]

Table 6

Representative explainable AI methods for CV, grouped by mechanism. For each method we list the model access it requires, the explanation it produces, the evaluation typically applied, and its publication venue.

Method	Family	Model access required	Explanation output	Typical evaluation	Venue, year
Grad-CAM [9]	Gradient-based CAM	Gradients and last convolutional feature maps	Class-discriminative heatmap	Weakly supervised localization, pointing game, class discrimination study (AMT)	ICCV, 2017
Grad-CAM++ [114]	Gradient-based CAM	Gradients and feature maps	Heatmap with improved multi-instance localization	Average drop, confidence increase, win percentage	WACV, 2018
XGrad-CAM [153]	Axiom-based gradient CAM	Gradients and feature maps	Heatmap satisfying sensitivity and conservation axioms	Class discrimination study, confidence drop (perturbation analysis), axiom analysis (sensitivity and conservation)	BMVC, 2020
LayerCAM [154]	Hierarchical gradient CAM	Gradients at multiple layers	Multi-layer fused heatmap	Weakly supervised localization and segmentation	IEEE TIP, 2021
Score-CAM [37]	Gradient-free CAM	Forward passes with masked activations	Class heatmap without gradients	Average drop, confidence increase, deletion/insertion AUC, energy-based pointing game	CVPR Workshops, 2020
Ablation-CAM [155]	Gradient-free CAM	Forward passes with channel ablation	Class heatmap	Average drop and percent increase in confidence and activation score, win percentage	WACV, 2020
Eigen-CAM [156]	Gradient-free CAM	Activations of the convolutional layer	Heatmap from principal components of feature maps	Localization assessment	IJCNN, 2020
Occlusion sensitivity [116]	Perturbation	Forward passes with sliding occlusion	Occlusion saliency map	Classification accuracy, occlusion, feature visualization	ECCV, 2014
Meaningful perturbations [27]	Perturbation	Gradients and learned mask optimization	Optimized saliency mask	mask representativeness across constant/noise/blur perturbations, minimality of deletion region	ICCV, 2017
RISE [115]	Perturbation, black box	Forward passes with random masks	Saliency map for any black-box model	Deletion game (softmax suppression), mask representativeness across constant/noise/blur perturbations, minimality of deletion region	BMVC, 2018
Transformer attribution [10]	Relevance propagation	Attention and gradient-based relevance	Relevance map for vision transformers	Positive and negative perturbation AUC, segmentation pixel-accuracy/mIoU/mAP, token-F1 on Movie Reviews (ERASER)	CVPR, 2021
ProtoNet [15]	Prototype, intrinsic	Learned prototype layer	Prototype activations and similarity scores	Classification accuracy on CUB-200-2011 and Stanford Cars, qualitative prototype visualization	NeurIPS, 2019
ProtoTree [16]	Prototype tree, intrinsic	Soft decision tree over prototypes	Prototype decision path	Top-1 accuracy, number of prototypes as interpretability measure, ensemble accuracy	CVPR, 2021
PIP-Net [17]	Prototype, intrinsic	Patch-based prototypes	Sparse prototype scores	Classification accuracy, prototype correlation with ground-truth object parts, sparsity ratio, out-of-distribution detection	CVPR, 2023
TCAV [14]	Concept-based, post hoc	Activations and concept activation vectors	Concept sensitivity scores	Statistical significance testing, sorting and activation maximization, ground truth tracking on synthetic dataset, bias analysis on standard networks	ICML, 2018
Sanity checks [26]	Evaluation protocol	Model and label randomization	Pass or fail diagnostic	Randomization test, data label permutation test, similarity measured by Spearman rank correlation, Gradients and Grad-CAM pass, Guided BackProp and Guided Grad-CAM fail	NeurIPS, 2018
ROAR [74]	Evaluation protocol	Iterative remove and retrain	Accuracy-degradation curve	Retraining benchmark	NeurIPS, 2019

OpenML evaluations using accuracy, AUC, and log loss [133]. Regardless of the benchmark, the evaluation protocol, including preprocessing, splits, random seeds, and any task-specific evaluator, should be reported explicitly so that scores remain reproducible and comparable across studies.

Beyond these task benchmarks, XAI methods are assessed with explanation-specific metrics that quantify how well an explanation reflects the model rather than how well the model performs its task. Faithfulness is commonly measured by deletion and insertion AUC and by comprehensiveness and sufficiency [38,115], reliability by infidelity

Table 7

Task benchmarks and task metrics by domain for the surveyed model families. These are the datasets and task-level metrics on which the base predictors and their explanations are evaluated; explanation-specific evaluation metrics are summarized in Section 2.2 and recorded per method in Table 6.

Domain	Task	Dataset	Size	Metric	Ref
Vision	Cls	ImageNet ReaL	50k	Top-1 acc	[157]
Vision	Det	Open Images V6	1.9M	mAP	[158]
Vision	Seg	Waymo Open	large	mIoU	[159]
NLP	NLU	SuperGLUE	8 tasks	Acc, F1	[160]
Time series	Fcst	ETT, ECL	1.3M	MAE, MSE	[102]
Graph	Node/Link	OGB	varies	Acc, ROC AUC	[161]
RL	Offline	D4RL	many	Norm. return	[162]
RL	Progen	Progen	16 games	Avg. return	[163]
Recsys	News	MIND	160k users	AUC, NDCG	[164]
Recsys	Video	KuaiRec	22M inter.	AUC, NDCG, HR@K	[165]
Tabular	Cls	TabPFN suite	small tasks	Acc, AUC, logloss	[133]

and sensitivity [75], and trustworthiness by randomization-based sanity checks [26]. These metrics, their typical use, and representative methods are described in Section 2.2 and recorded per method in Table 6. Protocol choices, including the baseline or background set, masking strategy, sampling budget, and layer selection, strongly affect these scores, so comparability across studies requires reporting these settings alongside the metric.

5.3. Human and developer factors

Human and developer factors define whether explanations help real work. Human factors encompass how users interpret, trust, rely upon, and act upon AI explanations, including cognitive load and time. Developer factors encompass how teams design, integrate, log, and govern explanation capabilities in tools and services. This section reads Tables 8 and 9 together with the model catalog in Table 5. Table 8 quantifies effects in controlled and field studies from 2020 to 2026. Domain and Task set the context. Explainer names the method in plain terms, for example, gradient class activation mapping, heatmaps, or token highlights, and expands names on first use, for example, class activation mapping (CAM) and Shapley additive explanations (SHAP). Design encodes the experiment and its control: w for within subjects, b

Table 8

Human effects of explanations. Control is merged into design to save space. Abbrev: w within, b between, fld field; ne no explainer; hb hint baseline; bau business as usual.

Domain	Task	Explainer	Design	N	Effect on trust	Effect on reliance	Effect on accuracy	Effect on time	Notes (measures)
Vision	Triage	Grad-CAM heatmap	w (ne)	48	+ small	neutral	+ small	+ small	Likert, accuracy, time [166]
NLP	Fact check	token highlights	b (hb)	120	neutral	+ small	neutral	+ small	accuracy, time [166]
Public sector	Screening	saliency plus policy notes	fld (bau)	80	+ small	neutral	neutral	+ small	decision quality, time [168]
Clinical	Diagnosis aid	advice with confidence	fld (no advice)	26	+ small	+ small	+ small	+ small	accuracy, time [171]
Mixed	Review with confidence	confidence display	b (no confidence)	160	+ small	lower overuse	+ small	neutral	accuracy, clicks [172]

Table 9

Developer workflow for explanations with hooks, logs, and guardrails.

Phase	Surface	Task	Explainers	Hooks	Logs	Guardrails	Env
Train	Notebook	slice debug	grad, attention, occlusion	grad, attn, layers	seed, input, output, explanations, commit	clip values, seed lock	PyTorch
Eval	Dashboard	compare models	Grad-CAM, shap, rise	grad, mask, forward	dataset id, metric, explanation stats, diffs	input limits, report cache	PyTorch, JAX
Debug	IDE	failure drilldown	saliency, counterfactual mask	layers, mask, grad	case id, steps, diffs, notes	PII strip, export policy	PyTorch
Ship	Service	online audit	light saliency, feature attributions	minimal hooks	sampled requests, summary stats	rate caps, privacy filter	PyTorch
Monitor	Ops	drift and bias watch	concept tests, attribution shift	layers, grad	drift metrics, alert log	rollback, review queue	any

for between subjects, and fld for field deployment, with the baseline in parentheses, such as no explainer or a hint baseline. N is the sample size. Four stacked columns report direction and magnitude for trust, reliance, accuracy, and time using each study's tests, where "+ small" denotes a small positive change and "neutral" denotes no reliable change, and Notes lists the main measures. Evidence shows that explanations can increase trust and acceptance of AI recommendations, while complementary accuracy gains are not reliably achieved by explanations alone [166]. Learning to defer frameworks route hard cases to human experts, with algorithmic guarantees of improved combined team accuracy [167]. Field studies in public services and healthcare report faster work and modest accuracy gains when advice includes provenance and confidence cues [168–171]. Confidence displays help users act selectively [172]. Interface wording can shift trust without changing accuracy, so preregistration and blinded analyses are required [173]. Table 9 turns these results into an engineering plan. The phases mark where work happens: Train, Eval, Debug, Ship, and Monitor. Surface identifies the user interface, for example, a notebook, dashboard, integrated development environment (IDE), service, or operations. Task states the activity. Explainers list the methods to expose. Hooks lists the signals to enable, for example, gradients, attention, forward passes, and masks. Logs name artifacts to capture like seeds, inputs, outputs, explanations, and diffs. Guardrails list enforceable rules such as rate caps, privacy filters, and removal of personally identifiable information (PII). Env provides the runtime, such as PyTorch, JAX, or a production stack in operations. Recommended practices include power analysis, reporting effect sizes with confidence intervals, releasing prompts and seeds, fixed color scales, class-aligned explanations, uncertainty labels, a simple control map, out-of-scope warnings, and complete user interface toggle logs.

5.4. Efficiency and robustness

Table 10 summarizes practical adaptation choices that improve the efficiency of predictors and serving stacks while preserving task performance. The table is intentionally conservative: it reports only effects directly supported by the cited studies and avoids fixed claims about latency, memory reduction, energy savings, or robustness scores unless such values are measured under the same model, dataset, input size, batch size, precision, hardware, and evaluation protocol.

The selected levers cover six common efficiency directions. FlashAttention and FlashAttention-2 reduce attention memory traffic while preserving exact attention computation [174,175]. GPTQ

Table 10

Adaptation menu for efficient predictors and serving stacks.

Lever (with citation)	Technique	Setting	Supported efficiency effect	Supported fidelity or accuracy effect	Validation needed before use
Attention kernel [174, 175]	FlashAttention / FlashAttention-2	exact attention kernel	Reduces memory traffic in attention and improves throughput, especially for long sequences; FlashAttention-2 further improves parallelism and work partitioning over FlashAttention	Exact attention computation; no approximation-induced accuracy loss is expected apart from numerical precision differences	Validate numerical agreement, batch size, sequence length, precision, and hardware-specific throughput
Quantization [176]	GPTQ	one-shot 3/4-bit weight PTQ	Compresses GPT-family models to 3 or 4 bits per weight without retraining; the paper reports quantizing 175B-parameter models in about four GPU hours and running OPT-175B on a single A100 after compression	Reports negligible degradation relative to FP16 baselines, mainly measured by perplexity and downstream zero-shot accuracy; the paper also reports end-to-end inference speedups of about 3.25× on A100 and 4.5× on A6000 in its tested setup	Requires representative calibration data and task-level validation; activation quantization is not included, and multiplication speedup is limited by mixed-precision hardware support
Sparsity [177]	SparseGPT	one-shot 50–60% sparsity	Prunes GPT-family models to at least 50% sparsity without retraining; the paper reports execution on OPT-175B and BLOOM-176B in under 4.5 hours and up to 60% unstructured sparsity in large models	Reports minimal accuracy loss, measured mainly through perplexity and zero-shot accuracy; larger models are shown to be more compressible	Actual inference speedup depends on sparse-kernel support and hardware compatibility; semi-structured 2:4 or 4:8 sparsity is more hardware-friendly than unstructured sparsity
Token reduction [178]	Token Merging (ToMe)	r tokens merged per block	Merges similar tokens inside transformer blocks without requiring retraining; the paper reports about 2× throughput for ViT-L/ViT-H image models and 2.2× throughput for ViT-L video in tested settings	Reports small accuracy drops of about 0.2–0.3% in the reported image and video settings; when used during training, it also reports 2× throughput for ViT-B audio with about 0.4% mAP drop	Tune r for each model, input resolution, modality, and dataset; exact latency and memory gains should be measured on the target hardware
Token pruning [179]	DynamicViT	hierarchical dynamic token sparsification	Uses lightweight prediction modules to estimate token importance and dynamically prune redundant tokens at different layers; the paper reports pruning 66% of input tokens, reducing FLOPs by 31–37%, and improving throughput by over 40% in tested vision transformers	Reports an accuracy drop within 0.5% on ImageNet for various vision-transformer models under the reported pruning setting	Requires training or fine-tuning of the token prediction modules; actual speedup should be validated for the target model, batch size, input resolution, and hardware
Early exit [180]	FastBERT	uncertainty-based adaptive exit	Adds student classifiers to intermediate Transformer layers and exits early for easier samples; the paper reports tunable inference speedups from about 1× to 12× over BERT, depending on the selected threshold	Reports competitive accuracy with small performance loss through self-distillation across twelve English and Chinese NLP datasets	Requires task-specific fine-tuning and self-distillation of branch classifiers; the exit threshold must be calibrated for the required speed and accuracy trade-off

compresses GPT family models through one-shot 3/4-bit post-training weight quantization and reports efficient inference with minimal degradation in the tested settings [176]. SparseGPT performs one-shot pruning of large GPT-style models to high sparsity without retraining, with accuracy evaluated primarily through perplexity and zero-shot performance [177]. ToMe improves the throughput of vision, video, and audio transformers by merging similar tokens within transformer blocks [178]. DynamicViT reduces computation by dynamically pruning less informative tokens using learned token-importance prediction modules [179]. FastBERT reduces average inference cost by allowing easy samples to exit from intermediate Transformer layers through uncertainty-based adaptive inference [180].

For each method, the table separates the supported efficiency effect from the supported fidelity or accuracy effect. This distinction matters because an efficiency method may reduce computation in the original paper yet still require local validation before deployment. Quantization requires representative calibration data, sparsity requires hardware or library support for sparse inference, token merging and token pruning require tuning for the model, resolution, and dataset, and early exit methods require task-specific threshold calibration. Robustness is therefore treated as a validation requirement rather than a universal fixed score. In this work, an adaptation is accepted only after rechecking task performance, numerical agreement where applicable, and stability

under the target deployment conditions. This makes Table 10 a practical guide for selecting efficient configurations without overstating results beyond the evidence in the cited papers. The levers in Table 10 act on the predictor and the serving stack, which set the compute budget within which explanations are produced; they do not by themselves reduce the cost of an explanation method. Explanation cost is a separate concern. Gradient and CAM methods are inexpensive, requiring one or a few backward passes, whereas perturbation and removal-based methods, such as RISE, occlusion, and Shapley estimators, are dominated by the number of forward passes and scale poorly to large inputs. Documented strategies for reducing explanation cost include amortized estimators that predict attribution in a single forward pass [61] and structure-exploiting computation for tree ensembles [76]. The stability of the resulting explanations across input variation, random seeds, and distribution shifts is treated as a validation requirement rather than a fixed score, and the patterns observed across the surveyed studies are summarized in Section 5.5.

5.5. Governance and reproducibility

Governance expectations for explainable AI are increasingly shaped by risk management and documentation regimes such as the NIST AI Risk Management Framework, ISO/IEC 42,001, and the EU AI Act,

which together call for documented risk assessment, human oversight, and post-deployment monitoring of high-risk systems. These regimes define obligations but leave open the concrete artifacts that make an explanation pipeline auditable. Peer-reviewed proposals supply several of these artifacts: Datasheets for Datasets, Data Cards, and Data Statements structure the documentation of a dataset's source, license, preprocessing, and known limitations [181–183], and model cards summarize a model's intended use, evaluation conditions, and limitations [184]. For explanations specifically, explainability fact sheets propose criteria for assessing explainable approaches [185], and evaluation toolkits such as Quantus provide standardized faithfulness and robustness metrics that can be logged with an explainer's configuration [23]. Empirical surveys of AI research find that key experimental, data, and method details are frequently under-reported, which limits reproducibility [186]; widely recommended practices therefore include preregistering and releasing the protocol, pinning code and dependencies, fixing and logging random seeds, versioning datasets, and publishing an executable environment with the code [187]. For human subject studies, preregistered analysis plans and de-identified data sharing are advisable where permitted, and auditing frameworks situate such practices within broader accountability processes [188,189]. Across the studies we reviewed, governance and reproducibility reporting are uneven, and the following patterns are qualitative observations rather than measured effects. Methods that report strong faithfulness, such as perturbation-based approaches and exact Shapley estimators, tend to be the most computationally demanding and rarely report energy use or edge device latency, whereas efficiency-oriented methods, such as gradient saliency and attention heatmaps, are more often reported as sensitive to seed variation and distribution shift. Human-subject evidence is comparatively limited, with small sample sizes and inconsistent reporting of effect sizes, and documentation or licensing artifacts appear more frequently in vision and NLP work than in graph or reinforcement learning. These observations point to a need for a shared evaluation infrastructure that records faithfulness, robustness, computational cost, human outcomes, and governance artifacts within a single reproducible protocol.

6. Conclusion

This survey establishes a comprehensive framework for explainable artificial intelligence (XAI) from 2020 to 2026, linking model architectures and explanation methods across seven domains to standardized datasets, evaluation protocols, human outcomes, computational characteristics, and governance requirements. The synthesis shows that faithfulness and stability depend critically on matched baselines, fixed random seeds, and controlled sampling budgets. Human outcomes are nuanced: explanations can increase trust and acceptance of AI recommendations, yet complementary accuracy gains are not reliably produced by explanations alone and emerge only under specific conditions and calibrated interface design. The survey offers actionable guidance through lifecycle workflows, an adaptation menu that documents efficiency and fidelity trade-offs for each lever, and governance artifacts that map policy to technical implementation.

Several gaps remain open. The field needs unified benchmark suites that pair architectures with explanation methods under standardized protocols and contamination detection. Faithfulness evaluation should converge on documented baselines, sampling budgets, and shared toolchains that report deletion and insertion AUC alongside robustness metrics. Human studies should scale to realistic workloads with pre-registered analyses and confidence intervals. Computational research should extend to long sequences and streaming settings and document energy constraints for edge deployment. Robustness protocols should cover seed variation, noise, augmentation, retraining, and distribution shift. Governance integration should be tightened to ensure that every release ships with SPDX licenses, risk registers, and software bills of materials, as well as commit-pinned repositories and validation-ready containers. By turning explanations from ad hoc demonstrations into

standardized, auditable components, this survey helps practitioners select methods with evidence-based confidence, researchers run rigorous evaluations, and auditors verify compliance through systematic validation. The path forward rests on shared benchmarks, transparent reporting, and governance-aware design.

CRedit authorship contribution statement

Habib Khan: Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Visualization, Writing – original draft. **Faleh Alqahtani:** Conceptualization, Methodology, Investigation, Formal analysis, Validation, Writing – review & editing. **Weiwei Jiang:** Conceptualization, Methodology, Investigation, Validation, Visualization, Writing – review & editing. **Muhammad Hamza Ali:** Investigation, Data curation, Visualization, Validation, Writing – review & editing. **Sareer Ul Amin:** Investigation, Visualization, Validation, Resources, Writing – review & editing. **JaKeoung Koo:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All research artifacts are openly available.

References

- [1] T. Miller, Explanation in artificial intelligence: insights from the social sciences, *Artif. Intell.* 267 (2019) 1–38.
- [2] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion* 58 (2020) 82–115.
- [3] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): toward medical XAI, *IEEE Trans. Neural Netw. Learn. Syst.* 32 (2020) 4793–4813.
- [4] V. Belle, I. Papantonis, Principles and practice of explainable machine learning, *Front. Big Data* 4 (2021) 688969.
- [5] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, *Data Min. Knowl. Discov.* 38 (2024) 3043–3101.
- [6] D. Minh, H.X. Wang, Y.F. Li, T.N. Nguyen, Explainable artificial intelligence: a comprehensive review, *Artif. Intell. Rev.* 55 (2022) 3503–3568.
- [7] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: *International Conference on Machine Learning*, PMLR, 2017, pp. 3319–3328.
- [8] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [9] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [10] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 782–791.
- [11] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [12] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the GDPR, *Harv. J.L. & Tech.* 31 (2017) 841.
- [13] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, C. Shah, Counterfactual explanations and algorithmic recourses for machine learning: a review, *ACM Comput. Surv.* 56 (2024) 1–42.
- [14] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV), in: *International Conference on Machine Learning*, PMLR, 2018, pp. 2668–2677.
- [15] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J.K. Su, This looks like that: deep learning for interpretable image recognition, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [16] M. Nauta, R. Van Bree, C. Seifert, Neural prototype trees for interpretable fine-grained image recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14933–14943.
- [17] M. Nauta, J. Schlöterer, M. Van Keulen, C. Seifert, PIP-net: patch-based intuitive prototypes for interpretable image classification, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2744–2753.

- [18] H. Chefer, S. Gur, L. Wolf, Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 397–406.
- [19] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du, Explainability for large language models: a survey, *ACM Trans. Intell. Syst. Technol.* 15 (2024) 1–38.
- [20] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, *arXiv preprint arXiv:1702.08608*, 2017.
- [21] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, *Inf. Fusion* 76 (2021) 89–106.
- [22] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötter, M. Van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI, *ACM Comput. Surv.* 55 (2023) 1–42.
- [23] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M.M.-C. Höhne, Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations and beyond, *J. Mach. Learn. Res.* 24 (2023) 1–11.
- [24] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (XAI): a survey, *arXiv preprint arXiv:2006.11371*, 2020.
- [25] M.A. Kadir, A. Mosavi, D. Sonntag, Evaluation metrics for XAI: a review, taxonomy, and practical applications, in: *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, IEEE, 2023, pp. 000111–000124.
- [26] J. Adebayo, J. Gilmer, M. Muehly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, *Adv. Neural Inf. Process. Syst.* 31 (2018).
- [27] R.C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3429–3437.
- [28] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, SmoothGrad: removing noise by adding noise, *arXiv preprint arXiv:1706.03825*, 2017.
- [29] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nat. Mach. Intell.* 1 (2019) 206–215.
- [30] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, Openxai: towards a transparent evaluation of model explanations, *Adv. Neural Inf. Process. Syst.* 35 (2022) 15784–15799.
- [31] L. Yuan, Y. Gu, Y. Han, T. Du, M. Grzegorzec, C. Li, Exploring transparency in pathological image analysis: a comprehensive review of explainable artificial intelligence (XAI) techniques, *Comput. Sci. Rev.* 60 (2026) 100863.
- [32] P. Dube, M. Kumar, Integrating explainable AI with federated learning for next-generation IoT: a comprehensive review and prospective insights, *Comput. Sci. Rev.* 56 (2025) 100697.
- [33] X. Liu, Y.Y. Mariano, The fire-vit model for tunnel fire detection with vision transformer improvement, *J. Comput. Cogn. Eng.* 4 (2025) 89–96.
- [34] M.L. Rethlefsen, S. Kirtley, S. Waffenschmidt, A.P. Ayala, D. Moher, M.J. Page, J.B. Koffel, PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews, *Syst. Rev.* 10 (2021) 39.
- [35] I. Covert, S. Lundberg, S.-I. Lee, Explaining by removing: a unified framework for model explanation, *J. Mach. Learn. Res.* 22 (2021) 1–90.
- [36] C. Frye, D. de Mijolla, T. Begley, L. Cowton, M. Stanley, I. Feige, Shapley explainability on the data manifold, *arXiv preprint arXiv:2006.01272*, 2020.
- [37] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-CAM: score-weighted visual explanations for convolutional neural networks, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 24–25.
- [38] J. DeYoung, S. Jain, N.F. Rajani, E. Lehman, C. Xiong, R. Socher, B.C. Wallace, ERASER: a benchmark to evaluate rationalized NLP models, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4443–4458.
- [39] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local rule-based explanations of black box decision systems, *arXiv preprint arXiv:1805.10820*, 2018.
- [40] W. Xu, Interpretable machine learning-based fraud detection model and knowledge discovery in financial transactions, in: *International Conference on Computational Finance and Business Analytics*, Springer, 2025, pp. 334–346.
- [41] A.-H. Karimi, B. Schölkopf, I. Valera, Algorithmic recourse: from counterfactual explanations to interventions, in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 353–362.
- [42] M. Pawelczyk, C. Agarwal, S. Joshi, S. Upadhyay, H. Lakkaraju, Exploring counterfactual explanations through the lens of adversarial examples: a theoretical and empirical analysis, in: *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 4574–4594.
- [43] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, P. Ravikumar, On completeness-aware concept-based explanations in deep neural networks, *Adv. Neural Inf. Process. Syst.* 33 (2020) 20554–20565.
- [44] C.-K. Yeh, B. Kim, P. Ravikumar, Human-centered concept explanations for neural networks, in: *Neuro-Symbolic Artificial Intelligence: The State of the Art*, IOS Press, 2021, pp. 337–352.
- [45] O. Li, H. Liu, C. Chen, C. Rudin, Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.
- [46] N. Cammarata, G. Goh, S. Carter, C. Voss, L. Schubert, C. Olah, Curve circuits, *Distill* 6 (2021) e00024–006.
- [47] A. Gangwal, A. Lavecchia, Explainable AI methods for drug discovery: a survey of interpretability, metrics and mechanistic insight, *Comput. Sci. Rev.* 61 (2026) 100943.
- [48] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable AI (XAI): core ideas, techniques, and solutions, *ACM Comput. Surv.* 55 (2023) 1–33.
- [49] C. Molnar, *Interpretable Machine Learning*, Lulu.com, 2020.
- [50] P. Mahya, J. Fürnkranz, An empirical comparison of interpretable models to post-hoc explanations, *AI* 4 (2023).
- [51] H. Suresh, J.V. Gutttag, A framework for understanding unintended consequences of machine learning, *arXiv preprint arXiv:1901.10002*, 2019, 2, 73.
- [52] Z.C. Lipton, The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery, *Queue* 16 (2018) 31–57.
- [53] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, *Nat. Mach. Intell.* 2 (2020) 56–67.
- [54] J. Hatwell, M.M. Gaber, R.M.A. Azad, GBT-hips: explaining the classifications of gradient boosted tree ensembles, *Appl. Sci.* 11 (2021) 2511.
- [55] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bannetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion* 58 (2020) 82–115, <https://doi.org/10.1016/j.inffus.2019.12.012>, <https://www.sciencedirect.com/science/article/pii/S1566253519308103>.
- [56] Z. Cheng, Y. Wu, Y. Li, L. Cai, B. Ihnaini, A comprehensive review of explainable artificial intelligence (XAI) in computer vision, *Sensors* 25 (2025) 4166.
- [57] R. Gipiškis, C.-W. Tsai, O. Kurasova, Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: a survey, *Ict Express* 10 (2024) 1331–1354.
- [58] F. Ahmed, N.S. Naz, S. Khan, A.U. Rehman, W.M. Ismael, M.A. Khan, Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges, *BMC Med. Imaging* 26 (2026) 37.
- [59] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, *Data Min. Knowl. Discov.* 38 (2024) 2770–2824.
- [60] A. Ghorbani, J. Wexler, J.Y. Zou, B. Kim, Towards automatic concept-based explanations, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [61] N. Jethani, M. Sudarshan, I. Covert, S.-I. Lee, R. Ranganath, Fastshap: real-time Shapley value estimation, in: *ICLR 2022*, 2022.
- [62] Z. Qi, S. Khorram, F. Li, Visualizing deep networks by optimizing with integrated gradients, in: *AAAI*, vol. 34, 2020, pp. 11890–11898.
- [63] G. Erion, J.D. Janizek, P. Sturmels, S.M. Lundberg, S.-I. Lee, Improving performance of deep learning models with axiomatic attribution priors and expected gradients, *Nat. Mach. Intell.* 3 (2021) 620–631.
- [64] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K.T. Schütt, S. Dähne, D. Erhan, B. Kim, The (un) reliability of saliency methods, in: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, 2019, pp. 267–280.
- [65] C. Burger, C. Walter, T. Le, L. Chen, Towards robust and accurate stability estimation of local surrogate models in text-based explainable AI, *arXiv preprint arXiv:2501.02042*, 2025.
- [66] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling lime and SHAP: adversarial attacks on post hoc explanation methods, in: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.
- [67] M.T. Keane, B. Smyth, Good counterfactuals and where to find them: a case-based technique for generating counterfactuals for explainable AI (XAI), in: *International Conference on Case-Based Reasoning*, Springer, 2020, pp. 163–178.
- [68] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, E. Baralis, Concept-based explainable artificial intelligence: a survey, *ACM Comput. Surv.* (2023).
- [69] M. Xue, Q. Huang, H. Zhang, L. Cheng, J. Song, M. Wu, M. Song, Protopformer: concentrating on prototypical parts in vision transformers for interpretable image recognition, *arXiv preprint arXiv:2208.10431*, 2022.
- [70] M. Fontes, J.D.S. De Almeida, A. Cunha, Application of example-based explainable artificial intelligence (XAI) for analysis and interpretation of medical imaging: a systematic review, *IEEE Access* 12 (2024) 26419–26427.
- [71] P. Kumar, Explainable AI/ML testing: ensuring transparency, accountability, and compliance, *J. Artif. Intell. Mach. Learn. & Data Sci.* 1 (2023) 476–482.
- [72] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, A. Preece, Sanity checks for saliency metrics, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 6021–6029.
- [73] P. Hase, M. Bansal, Evaluating explainable AI: which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*, 2020.
- [74] S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim, A benchmark for interpretability methods in deep neural networks, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [75] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D.I. Inouye, P.K. Ravikumar, On the (in) fidelity and sensitivity of explanations, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [76] M. Muschalik, F. Fumagalli, B. Hammer, E. Hüllermeier, Beyond treeshap: efficient computation of any-order Shapley interactions for tree ensembles, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 14388–14396.
- [77] B. Ustun, A. Spangher, Y. Liu, Actionable recourse in linear classification, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 10–19.
- [78] M. Yuksekogul, M. Wang, J. Zou, Post-hoc concept bottleneck models, *arXiv preprint arXiv:2205.15480*, 2022.
- [79] L. Sixt, M. Granz, T. Landgraf, When explanations lie: why many modified BP attributions fail, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 9046–9057.
- [80] H. Lakkaraju, O. Bastani, "How do I fool you?" Manipulating user trust via misleading black box explanations, in: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 79–85.

- [81] T. Speith, A review of taxonomies of explainable artificial intelligence (XAI) methods, in: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 2239–2250.
- [82] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence, *Inf. Fusion* 99 (2023) 101805.
- [83] K. Kalasampath, K. Spoorthi, S. Sajeev, S.S. Kuppa, K. Ajay, A. Maruthamuthu, A literature review on applications of explainable artificial intelligence (XAI), *IEEE Access* 13 (2025) 41111–41140.
- [84] A. Chaddad, Y. Hu, Y. Wu, B. Wen, R. Kateb, Generalizable and explainable deep learning for medical image computing: an overview, *Curr. Opin. Biomed. Eng.* 33 (2025) 100567.
- [85] S. Sun, W. An, F. Tian, F. Nan, Q. Liu, J. Liu, N. Shah, P. Chen, A review of multimodal explainable artificial intelligence: past, present and future, *arXiv preprint arXiv:2412.14056*, 2024.
- [86] S. Mohseni, N. Zarei, E.D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, *ACM Trans. Interact. Intell. Syst. (TiIS)* 11 (2021) 1–45.
- [87] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *BMJ* 372 (2021).
- [88] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P.N. Bennett, K. Inkpen, et al., Guidelines for human-AI interaction, in: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.
- [89] P.P. Angelov, E.A. Soares, R. Jiang, N.I. Arnold, P.M. Atkinson, Explainable artificial intelligence: an analytical review, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 11 (2021) e1424.
- [90] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11976–11986.
- [91] A. Dosovitskiy, An image is worth 16x16 words: transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929*, 2020.
- [92] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 10347–10357.
- [93] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: *European Conference on Computer Vision*, Springer, 2020, pp. 213–229.
- [94] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLO: exceeding YOLO series in 2021, *arXiv preprint arXiv:2107.08430*, 2021.
- [95] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, *Adv. Neural Inf. Process. Syst.* 34 (2021) 12077–12090.
- [96] B. Cheng, I. Misra, A.G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.
- [97] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: decoding-enhanced BERT with disentangled attention, *arXiv preprint arXiv:2006.03654*, 2020.
- [98] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* 21 (2020) 1–67.
- [99] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., LLaMA 2: open foundation and fine-tuned chat models, *arXiv preprint arXiv:2307.09288*, 2023.
- [100] A.Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D.S. Chaplot, D.D.L. Casas, E.B. Hanna, F. Bressand, et al., Mixtral of experts, *arXiv preprint arXiv:2401.04088*, 2024.
- [101] B. Lim, S. Arik, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, *Int. J. Forecast.* 37 (2021) 1748–1764.
- [102] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, W. Zhang, Informer: beyond efficient transformer for long sequence time-series forecasting, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 11106–11115.
- [103] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, M. Long, Timesnet: temporal 2D-variation modeling for general time series analysis, *arXiv preprint arXiv:2210.02186*, 2022.
- [104] Y. Nie, A time series is worth 64Words: long-term forecasting with transformers, *arXiv preprint arXiv:2211.14730*, 2022.
- [105] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, T.-Y. Liu, Do transformers really perform badly for graph representation? *Adv. Neural Inf. Process. Syst.* 34 (2021) 28877–28888.
- [106] D. Kreuzer, D. Beaini, W. Hamilton, V. Létourneau, P. Tossou, Rethinking graph transformers with spectral attention, *Adv. Neural Inf. Process. Syst.* 34 (2021) 21618–21629.
- [107] L. Rampášek, M. Galkin, V.P. Dwivedi, A.T. Luu, G. Wolf, D. Beaini, Recipe for a general, powerful, scalable graph transformer, *Adv. Neural Inf. Process. Syst.* 35 (2022) 14501–14515.
- [108] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, et al., Mastering atari, go, chess and shogi by planning with a learned model, *Nature* 588 (2020) 604–609.
- [109] D. Hafner, J. Pasukonis, J. Ba, T. Lillicrap, Mastering diverse domains through world models, *arXiv preprint arXiv:2301.04104*, 2023.
- [110] R. Wang, R. Shivanna, D. Cheng, S. Jain, D. Lin, L. Hong, E. Chi, DCN v2: improved deep & cross network and practical lessons for web-scale learning to rank systems, in: *Proceedings of the Web Conference 2021*, 2021, pp. 1785–1797.
- [111] K. Zhou, H. Wang, W.X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, J.-R. Wen, S3-rec: self-supervised learning for sequential recommendation with mutual information maximization, in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1893–1902.
- [112] R. Qiu, Z. Huang, H. Yin, Z. Wang, Contrastive learning for representation degeneration problem in sequential recommendation, in: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 813–823.
- [113] M.S. Jamil, S.P. Banik, G.A. Rahaman, S. Saha, Advanced Grad-CAM++: improved visual explanations of CNN decisions in diabetic retinopathy, in: *Computer Vision and Image Analysis for Industry 4.0*, Chapman and Hall/CRC, 2023, pp. 64–75.
- [114] A. Chattopadhyay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks, in: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2018, pp. 839–847.
- [115] V. Petsiuk, A. Das, K. Saenko, Rise: randomized input sampling for explanation of black-box models, *arXiv preprint arXiv:1806.07421*, 2018.
- [116] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: *European Conference on Computer Vision*, Springer, 2014, pp. 818–833.
- [117] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: *International Conference on Machine Learning*, PMLR, 2017, pp. 3145–3153.
- [118] Z. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, GNNExplainer: generating explanations for graph neural networks, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [119] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, X. Zhang, Parameterized explainer for graph neural network, *Adv. Neural Inf. Process. Syst.* 33 (2020) 19620–19631.
- [120] S. Greydanus, A. Koul, J. Dodge, A. Fern, Visualizing and understanding atari agents, in: *International Conference on Machine Learning*, PMLR, 2018, pp. 1792–1801.
- [121] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.
- [122] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., DINOv2: learning robust visual features without supervision, *arXiv preprint arXiv:2304.07193*, 2023.
- [123] A. Hatamizadeh, J. Kautz, Mambavision: a hybrid Mamba-transformer vision backbone, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25261–25270.
- [124] Y. Tian, Q. Ye, D. Doermann, Yolov12: attention-centric real-time object detectors, *Adv. Neural Inf. Process. Syst.* 38 (2026) 78433–78457.
- [125] K. Clark, M.-T. Luong, Q.V. Le, C.D. Manning, Electra: pre-training text encoders as discriminators rather than generators, *arXiv preprint arXiv:2003.10555*, 2020.
- [126] I. Beltagy, M.E. Peters, A. Cohan, Longformer: the long-document transformer, *arXiv preprint arXiv:2004.05150*, 2020.
- [127] A. Gu, T. Dao, Mamba: linear-time sequence modeling with selective state spaces, *arXiv preprint arXiv:2312.00752*, 2023.
- [128] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M.S. Kale, J. Love, et al., Gemma: open models based on gemini research and technology, *arXiv preprint arXiv:2403.08295*, 2024.
- [129] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, et al., DeepSeek-r1: incentivizing reasoning capability in LLMs via reinforcement learning, *arXiv preprint arXiv:2501.12948*, 2025.
- [130] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, *arXiv preprint arXiv:2505.09388*, 2025.
- [131] S. Arik, T. Pfister, Tabnet: attentive interpretable tabular learning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 6679–6687.
- [132] Y. Gorishniy, I. Rubachev, V. Khurlov, A. Babenko, Revisiting deep learning models for tabular data, *Adv. Neural Inf. Process. Syst.* 34 (2021) 18932–18943.
- [133] N. Hollmann, S. Müller, K. Eggenberger, F. Hutter, TabPFN: a transformer that solves small tabular classification problems in a second, *arXiv preprint arXiv:2207.01848*, 2022.
- [134] Y. Gorishniy, I. Rubachev, N. Kartashev, D. Shlenskii, A. Kotelnikov, A. Babenko, Tabr: tabular deep learning meets nearest neighbors, in: *International Conference on Learning Representations*, vol. 2024, 2024, pp. 18209–18249.
- [135] Y. Gorishniy, A. Kotelnikov, A. Babenko, Tabm: advancing tabular deep learning with parameter-efficient ensembling, in: *International Conference on Learning Representations*, 2025, 2025, pp. 77899–77935.
- [136] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S.B. Hoo, R.T. Schirmeister, F. Hutter, Accurate predictions on small data with a tabular foundation model, *Nature* 637 (2025) 319–326.
- [137] A. Das, W. Kong, R. Sen, Y. Zhou, A decoder-only foundation model for time-series forecasting, *arXiv preprint arXiv:2310.10688*, 2023.
- [138] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, D. Sahoo, Unified training of universal time series forecasting transformers, in: *Forty-First International Conference on Machine Learning*, 2024.
- [139] A.F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S.S. Rangapuram, S.P. Arango, S. Kapoor, et al., Chronos: learning the language of time series, *arXiv preprint arXiv:2403.07815*, 2024.
- [140] X. Shi, S. Wang, Y. Nie, D. Li, Z. Ye, Q. Wen, M. Jin, Time-moe: billion-scale time series foundation models with mixture of experts, in: *International Conference on Learning Representations*, vol. 2025, 2025, pp. 34635–34667.
- [141] H. Shirzad, A. Vellingker, B. Venkatachalam, D.J. Sutherland, A.K. Sinop, Exphormer: sparse transformers for graphs, in: *International Conference on Machine Learning*, PMLR, 2023, pp. 31613–31632.

- [142] Q. Wu, W. Zhao, C. Yang, H. Zhang, F. Nie, H. Jiang, Y. Bian, J. Yan, Sgformer: simplifying and empowering transformers for large-graph representations, *Adv. Neural Inf. Process. Syst.* 36 (2023) 64753–64773.
- [143] A. Behrouz, F. Hashemi, Graph Mamba: towards learning on graphs with state space models, in: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 119–130.
- [144] C. Wang, O. Tsepa, J. Ma, B. Wang, Graph-Mamba: towards long-range graph sequence modeling with selective state spaces, *arXiv preprint arXiv:2402.00789*, 2024.
- [145] S. Wang, S. Liu, W. Ye, J. You, Y. Gao, Efficientzero v2: mastering discrete and continuous control with limited data, *arXiv preprint arXiv:2403.00564*, 2024.
- [146] J. Bruce, M.D. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, et al., Genie: generative interactive environments, in: *Forty-First International Conference on Machine Learning*, 2024.
- [147] N. Hansen, H. Su, X. Wang, TD-MPC2: scalable, robust world models for continuous control, in: *International Conference on Learning Representations*, vol. 2024, 2024, pp. 47376–47405.
- [148] Google DeepMind, Genie 3: A New Frontier for World Models, Google DeepMind Blog, 2025. Research preview announcement.
- [149] W. Wei, X. Ren, J. Tang, Q. Wang, L. Su, S. Cheng, J. Wang, D. Yin, C. Huang, Llmrec: large language models with graph augmentation for recommendation, in: *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 806–815.
- [150] S. Rajput, N. Mehta, A. Singh, R. Hulikal Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Tran, J. Samost, et al., Recommender systems with generative retrieval, *Adv. Neural Inf. Process. Syst.* 36 (2023) 10299–10315.
- [151] C. Liu, J. Lin, J. Wang, H. Liu, J. Caverlee, Mamba4Rec: towards efficient sequential recommendation with selective state space models, *arXiv preprint arXiv:2403.03900*, 2024.
- [152] Z. Liu, Q. Liu, Y. Wang, W. Wang, P. Jia, M. Wang, Z. Liu, Y. Chang, X. Zhao, SIGMA: selective gated Mamba for sequential recommendation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 12264–12272.
- [153] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, B. Li, Axiom-based Grad-CAM: towards accurate visualization and explanation of CNNs, in: *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.
- [154] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, Y. Wei, LayerCAM: exploring hierarchical class activation maps for localization, *IEEE Trans. Image Process.* 30 (2021) 5875–5888.
- [155] H.G. Ramaswamy, et al., Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 983–991.
- [156] M. Bany Muhammad, M. Yeasin, Eigen-CAM: class activation map using principal components, *arXiv e-prints arXiv:2008.00299*, 2020.
- [157] L. Beyer, O.J. Hénaff, A. Kolesnikov, X. Zhai, A.V.D. Oord, Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.
- [158] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, et al., The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale, *Int. J. Comput. Vis.* 128 (2020) 1956–1981.
- [159] P. Sun, H. Kretschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., Scalability in perception for autonomous driving: waymo open dataset, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446–2454.
- [160] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, SuperGlue: a stickier benchmark for general-purpose language understanding systems, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [161] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, J. Leskovec, Open graph benchmark: datasets for machine learning on graphs, *Adv. Neural Inf. Process. Syst.* 33 (2020) 22118–22133.
- [162] J. Fu, A. Kumar, O. Nachum, G. Tucker, S. Levine, D4rl: datasets for deep data-driven reinforcement learning, *arXiv preprint arXiv:2004.07219*, 2020.
- [163] K. Cobbe, C. Hesse, J. Hilton, J. Schulman, Leveraging procedural generation to benchmark reinforcement learning, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 2048–2056.
- [164] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, et al., Mind: a large-scale dataset for news recommendation, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3597–3606.
- [165] C. Gao, S. Li, W. Lei, J. Chen, B. Li, P. Jiang, X. He, J. Mao, T.-S. Chua, KuaiRec: a fully-observed dataset and insights for evaluating recommender systems, in: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 540–550.
- [166] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M.T. Ribeiro, D. Weld, Does the whole exceed its parts? The effect of AI explanations on complementary team performance, in: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–16.
- [167] B. Wilder, E. Horvitz, E. Kamar, Learning to complement humans, *arXiv preprint arXiv:2005.00582*, 2020.
- [168] M. De-Arteaga, R. Fogliato, A. Chouldechova, A case for humans-in-the-loop: decisions in the presence of erroneous algorithmic scores, in: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.
- [169] M. Janssen, M. Hartog, R. Matheus, A. Yi Ding, G. Kuk, Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government, *Soc. Sci. Comput. Rev.* 40 (2022) 478–493.
- [170] S. Ganesan, N. Somasiri, et al., Navigating the integration of machine learning in healthcare: challenges, strategies, and ethical considerations, *J. Comput. Cogn. Eng.* 4 (2024) 8–23.
- [171] C. Panigutti, A. Beretta, F. Giannotti, D. Pedreschi, Understanding the impact of explanations on advice-taking: a user study for AI-based clinical decision support systems, in: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–9.
- [172] M. Naiseh, A. Simkute, B. Zieni, N. Jiang, R. Ali, C-XAI: a conceptual framework for designing XAI tools that support trust calibration, *J. Responsible Technol.* 17 (2024) 100076.
- [173] F. Poursabzi-Sangdeh, D.G. Goldstein, J.M. Hofman, J.W. Wortman Vaughan, H. Wallach, Manipulating and measuring model interpretability, in: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–52.
- [174] T. Dao, D. Fu, S. Ermon, A. Rudra, C. Ré, Flashattention: fast and memory-efficient exact attention with IO-awareness, *Adv. Neural Inf. Process. Syst.* 35 (2022) 16344–16359.
- [175] T. Dao, Flashattention-2: faster attention with better parallelism and work partitioning, *arXiv preprint arXiv:2307.08691*, 2023.
- [176] E. Frantar, S. Ashkboos, T. Hoeffler, D. Alistarh, GPTQ: accurate post-training quantization for generative pre-trained transformers, *arXiv preprint arXiv:2210.17323*, 2022.
- [177] E. Frantar, D. Alistarh, SparseGPT: massive language models can be accurately pruned in one-shot, in: *International Conference on Machine Learning*, PMLR, 2023, pp. 10323–10337.
- [178] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, J. Hoffman, Token merging: your ViT but faster, *arXiv preprint arXiv:2210.09461*, 2022.
- [179] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, C.-J. Hsieh, Dynamicvit: efficient vision transformers with dynamic token sparsification, *Adv. Neural Inf. Process. Syst.* 34 (2021) 13937–13949.
- [180] W. Liu, P. Zhou, Z. Wang, Z. Zhao, H. Deng, Q. Ju, Fastbert: a self-distilling BERT with adaptive inference time, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6035–6044.
- [181] T. Gebru, J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H.D. Iii, K. Crawford, Datasheets for datasets, *Commun. ACM* 64 (2021) 86–92.
- [182] M. Pushkarna, A. Zaldivar, O. Kjartansson, Data cards: purposeful and transparent dataset documentation for responsible AI, in: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1776–1826.
- [183] E.M. Bender, B. Friedman, Data statements for natural language processing: toward mitigating system bias and enabling better science, *Trans. Assoc. Comput. Linguist.* 6 (2018) 587–604.
- [184] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I.D. Raji, T. Gebru, Model cards for model reporting, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229.
- [185] K. Sokol, P. Flach, Explainability fact sheets: a framework for systematic assessment of explainable approaches, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 56–67.
- [186] O.E. Gundersen, S. Kjensmo, State of the art: reproducibility in artificial intelligence, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.
- [187] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, H. Larochelle, Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program), *J. Mach. Learn. Res.* 22 (2021) 1–20.
- [188] I.D. Raji, A. Smart, R.N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 33–44.
- [189] J. Mökander, Auditing of AI: legal, ethical and technical approaches, *Digit. Soc.* 2 (2023) 49.