Benchmarking XAI Explanations with Human-Aligned Evaluations

Rémi Kazmierczak[1], Steve Azzolin[2], Eloïse Berthier[1], Anna Hedström[3], Delhomme [4], David Filliat[1], Nicolas Bousquet[5], Goran Frehse[1], Massimiliano Mancini[2], Baptiste Caramiaux[6], Andrea Passerini[2], Gianni Franchi[1]

[1]Unité d'Informatique et d'Ingénierie des Systèmes, ENSTA, Institut Polytechnique de Paris, Palaiseau, France
[2]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
[3]Understandable Machine Intelligence Lab, TU Berlin, Berlin, Germany
[4]Laboratory of Applied Ergonomics and Psychology, Université Gustave Eiffel, Versailles, France
[5]SINCLAIR Laboratory, Palaiseau, France
[6]Institute of Intelligent Systems and Robotics, Sorbonne Université, Paris, France
{remi.kazmierczak, eloise.berthier, goran.frehse, gianni.franchi}@ensta-paris.fr
{steve.azzolin, massimiliano.mancini, andrea.passerini}@unitn.it
hedstroem.anna@gmail.com
patricia.delhomme@univ-eiffel.fr
nicolas.bousquet@edf.fr
baptiste.caramiaux@sorbonne-universite.fr

Abstract

We introduce PASTA (Perceptual Assessment System for explanaTion of Artificial Intelligence), a novel human-centric framework for evaluating eXplainable AI (XAI) techniques in computer vision. Our first contribution is the creation of the **PASTA-dataset**, the first large-scale benchmark that spans a diverse set of models and both saliency-based and concept-based explanation methods. This dataset enables robust, comparative analysis of XAI techniques based on human judgment. Our second contribution is an automated, data-driven benchmark that predicts human preferences using the **PASTA-dataset**. This scoring called **PASTA-score** offers scalable, reliable, and consistent evaluation aligned with human perception. Additionally, our benchmark allows for comparisons between explanations across different modalities, an aspect previously unaddressed. We then propose to apply our scoring method to probe the interpretability of existing models and to build more human-interpretable XAI methods.

Code — https://github.com/RemiKaz/PASTA
Datasets — https://github.com/RemiKaz/PASTA

Introduction

As Deep Neural Networks (DNNs) are increasingly deployed in high-stakes domains such as law and medicine (Surden 2021; Litjens et al. 2017), understanding their decision-making process has become essential (Bender et al. 2021). Their opacity often earns them the label "black boxes" (Castelvecchi 2016), raising trust and accountability concerns in critical applications (Vereschak et al. 2024). This has given rise to the field of eXplainable AI (XAI) (Gunning et al. 2019).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

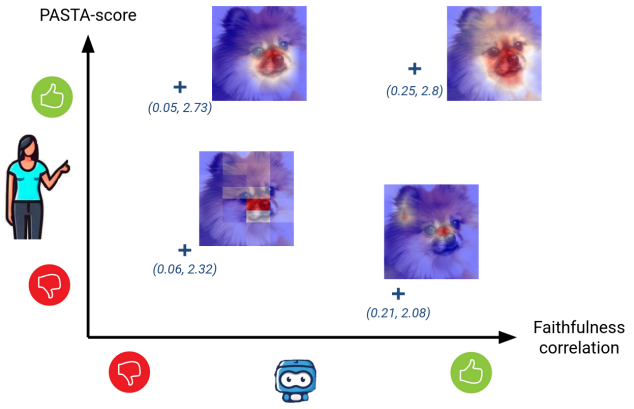


Figure 1: **PASTA automates the evaluation of human perception** of provided explanations by computing a PASTA-score. By integrating PASTA-score (y-axis) with existing faithfulness metrics (x-axis), we aim to foster the development of explanations that are not only aligned with the model's behavior but also comprehensible to human evaluators. Samples reported in the figure correspond to the label *dog* for a ResNet50 classifier trained on PascalPART. XAI methods: top left: GradCAM; Top right: FullGrad; Bottom left: SHAP; Bottom right: AblationCAM

A wide variety of XAI techniques have been proposed (Speith 2022; Saeed and Omlin 2023), notably saliency-based methods (Muhammad and Yeasin 2020; Böhle et al. 2024), which highlight relevant input features, and concept-based methods (Yan et al. 2023; Díaz-Rodríguez et al. 2022), which associate predictions with high-level semantic concepts. However, comparing such heterogeneous approaches remains an open problem.

Evaluating XAI methods is particularly challenging for

two main reasons. First, the diversity of explanation types complicates the definition of a common evaluation framework. Second, the notion of a "good explanation" is inherently subjective. This creates a dichotomy between *non-perceptual* evaluations—focused on model-centric metrics using toolkits like `OpenXAI`, `Quantus`, and `Xplique` (Agarwal et al. 2022; Hedström et al. 2023; Fel et al. 2022)—and *perceptual* evaluations, which assess human understanding. While the latter is often explored through anecdotal examples (Selvaraju et al. 2017; Wang et al. 2020), user studies (Dawoud et al. 2023; Colin et al. 2022), or region-of-interest alignment (Liu et al. 2024a; Arras, Osman, and Samek 2022), there is still a lack of standardization in assessing explanations from the human perspective (Nauta et al. 2023). Yet, this dimension is crucial—explanations faithful to the model's reasoning may still be unintelligible to human users, limiting their actionability. In this sense, our proposed PASTA-score allows for evaluating explanations based on a combination of faithfulness and human preferences, as shown in Figure 1.

To address these challenges, we propose **PASTA**—the *Perceptual Assessment System for explanaTion of Artificial intelligence*—which aims to automate the human-aligned evaluation of XAI methods. PASTA has two core components. First, a benchmark, **PASTA-dataset**, composed of four diverse image-based classification datasets with aligned concept annotations, enabling the comparison of 20 XAI techniques across multiple architectures. Second, the **PASTA-score**, a data-driven metric designed to predict human preferences, providing a scalable way to evaluate explanations from a perceptual standpoint. Unlike prior benchmarks focused solely on saliency or user studies (Colin et al. 2022; Dawoud et al. 2023), PASTA unifies both saliency-based and concept-based methods under a single evaluation framework.

Our contributions are: **(1) Comprehensive XAI Benchmark**: We introduce the PASTA-dataset, enabling the evaluation of both visual and concept-based explanations. **(2) Large-scale Method Evaluation**: We assess 20 XAI methods across multiple datasets and models, including both post-hoc and ante-hoc methods. Our first result suggests that human annotators tend to prefer saliency and perturbation-based techniques, like LIME and SHAP. **(3) Human-aligned Explanation Scoring**: We propose **PASTA-score**, an automated yet perceptually grounded assessment of explanations, trained on the PASTA-dataset to replicate human preferences. **(4) Practical Applications**: We present three use cases of PASTA-score, showing how it can guide the design of more interpretable models, and serve as a proxy for visual human assessment in practical deployments. The pipeline of the global workflow is presented in Figure 2.

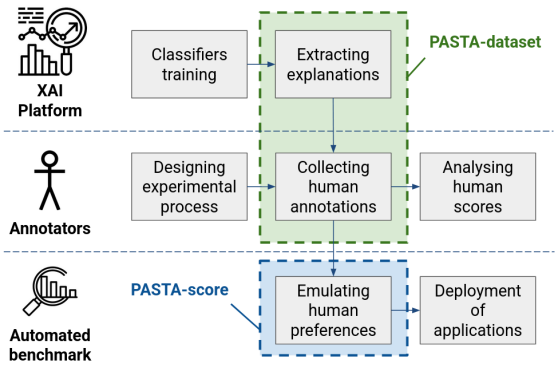


Figure 2: Pipeline of the PASTA framework. We first collect a dataset of human preferences called PASTA-dataset. This dataset is used to learn to emulate human preferences for the new samples on the test set using the PASTA-score. The trained PASTA-score can then be deployed to downstream applications as a consistent-over-time, quick and cost-effective replacement for human feedback.

Related Work

Explainable AI. To address the challenge of explaining DNNs, several specialized tools have been proposed, often categorized into *post-hoc* and *ante-hoc* methods (Arrieta et al. 2020; Rudin et al. 2022). *Post-hoc* methods encompass any tool external to the model, allowing us to gain insights from any pre-trained DNN. Popular examples are GradCAM (Selvaraju et al. 2017), LIME (Ribeiro, Singh, and Guestrin 2016), and SHAP (Lundberg and Lee 2017). While most post-hoc explainers agree in providing input regions most responsible for a certain prediction, they differ in many non-trivial details, and selecting and evaluating the most appropriate explainer for each task can be challenging (Leavitt and Morcos 2020; Roy et al. 2022). *Ante-hoc* methods, instead, aim at modifying the underlying model architecture to provide explanations by design. This can be done in the framework of Concept Bottleneck Models (CBMs) (Koh et al. 2020) by prompting the model to first predict a set of human-understandable high-level concepts, and then making the final prediction using a shallow and interpretable classifier that supports human inspection, or by decomposing the *reasoning* of the model into smaller and more actionable steps (Ge et al. 2023).

Evaluating explainability. While several methods have been proposed to quantitatively measure explanation quality, such as faithfulness (Petsiuk, Das, and Saenko 2018; Dasgupta, Frost, and Moshkovitz 2022; Azzolin et al. 2025), sparsity (Chalasani et al. 2020; Bénard et al. 2021), robustness (Alvarez-Melis and Jaakkola 2018; Montavon, Samek, and Müller 2018), sensitivity (Adebayo et al. 2018; Hedström et al. 2024) and alignment to an assumed ground truth (Colin et al. 2022; Mohseni, Block, and Ragan 2021; Dawoud et al. 2023), they inherently overlook the perceptual aspect with respect to the human, who is the expected consumer of such explanations. Evaluating explanations via user studies, e.g., where annotators are asked to rate and evaluate explanations (Chen et al. 2018; Shu et al. 2019; Yang et al. 2022; Kares et al. 2025), are, however, very costly, prone to unreproducibility issues (Nauta et al. 2023), and often unfeasible for tasks that require trained users, like in the medical domain (Miró-Nicolau, Moyà-Alcover, and Jaume-i Capó 2022; Muddamsetty, Jahromi, and Moeslund 2021). In this work, we take the first step towards standard-

izing the evaluation of human perception preferences of explanations (Nauta et al. 2023). We propose to overcome the issues of hard-to-reproduce large-scale user studies by automating the evaluation of XAI techniques through a multi-value scoring method that mimics human preferences while taking into account the users' diverse expectations, which naturally emerge in user-based studies.

Automated scoring. Automated scoring involves developing models that assign scores to inputs based on a reference dataset, often derived from human ratings. A particularly active area of research in this domain is automated essay scoring. Traditionally, this has been addressed through handcrafted feature extraction (Yannakoudakis, Briscoe, and Medlock 2011), but modern methods tend to be closer to model as a judge (Lee et al. 2024; Taghipour and Ng 2016; Chiang et al. 2024). More recently, there has been a growing interest in using embeddings from large language models (LLMs) as features for scoring. The first successful attempt in this direction was made by Yang et al. (2020). Building on this trend, other approaches have incorporated LLM embeddings with models like LSTMs (Wang et al. 2022), integrated text generation into the training loop (Xiao et al. 2024), or introduced multi-scale aspects to enhance performance (Li et al. 2023).

Creating the PASTA-dataset

To assess the quality of XAI explanations for image classification decisions from a human-centric perspective, we constructed a comprehensive dataset comprising images, predictions, explanations, and evaluations of these explanations, as depicted in Figure 3. To account for the heterogeneity of different XAI methods, model backbones, and training scenarios, we constructed the PASTA-benchmark to include 1000 images sampled across 4 available datasets, 7 classification backbones, 20 XAI methods, 6 questions, and 5 annotations per explanation, question) pair. The challenge in annotating such a dataset resides in its multiplicative nature, where each question requires an annotation across multiple backbones, datasets, XAI methods, images, and human annotators. Consequently, the PASTA-dataset contains an overall number of 633,000 samples, each corresponding to a unique Likert-like rating, which is the largest benchmark available of this kind.

To construct such a dataset, the initial phase involved developing a unified platform designed to integrate various models, XAI methods, and datasets in a streamlined manner. This platform encompasses a diverse array of models and XAI methods. Given the potential utility of this platform as a baseline for future research, we intend to release it publicly upon publication of this paper. Further details regarding the overall procedure, including details about the datasets employed, model training, and explanation extraction, are available in Section A.1 of the appendix. The subsequent phase in dataset creation is dedicated to annotating the explanations, which involved 24 participants in an online process. A pivotal insight from existing user study literature (Xuan et al. 2025; Liao et al. 2022) is that human perception of explanations is not unidimensional; rather, it encompasses a range

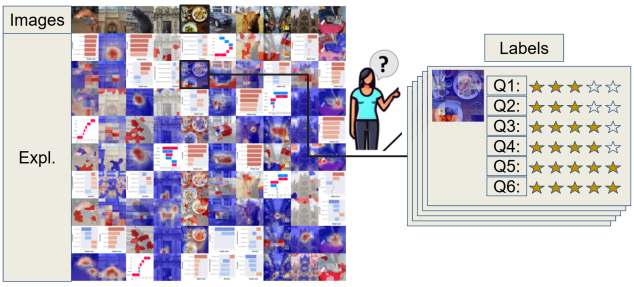


Figure 3: Overview of the human annotation process in the PASTA-dataset. We compute a total number of 46 explanations for each image, out of which 21 are sampled and rated by humans according to six questions. Further details available in Appendix A.2.

of potentially unaligned desiderata. For instance, a saliency-based explanation that highlights a dog to predict a cat can be entirely clear, thereby satisfying complexity desiderata, while simultaneously not fulfilling plausibility desiderata. Consequently, we pose multiple questions designed to encompass a spectrum of human assessment as broad as possible. Specifically, the following questions were asked:

- Q1: Is the provided explanation consistent with how I would explain the predicted class?
- Q2: Overall the explanation provided for the model prediction can be trusted?
- Q3: Is the explanation easy to understand?
- Q4: Can the explanation be understood by a large number of people, independently of their demographics (age, gender, country, etc.) and culture?
- Q5: With this perturbed image, to what extent has the explanation changed ? (Examples with good predictions and light perturbations)
- Q6: With this perturbed image, to what extent has the explanation changed? (Examples with bad predictions and strong perturbations)

The selection of these questions reflects a broader discourse on user studies and desiderata. To ensure that annotators comprehensively understand the task and expectations, we implemented an evaluation protocol developed with the assistance of a psychologist. This protocol includes annotator training and continuous monitoring throughout the process. Discussions regarding the desiderata, a detailed evaluation protocol, and information about the annotators are available in Section A.2 of the Appendix.

Analysis of Human Preference

Having collected a large number of human preferences for different XAI models and backbones in the PASTA-dataset, we now proceed to analyze human scores in relation to each method. The full analysis is available in Appendix B.3.

Human preferences for output format: As illustrated in Figure 4, results indicate that humans tend to prefer image-based explainers in relation to questions Q1-Q5, meaning

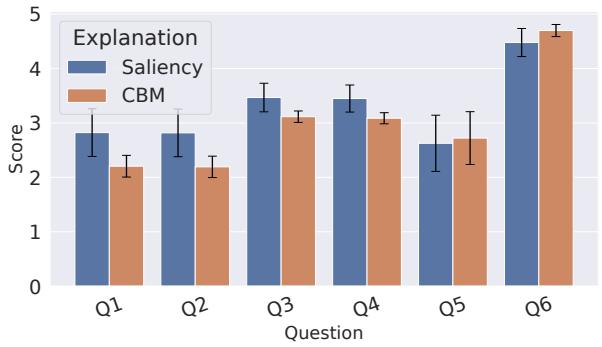


Figure 4: Scores for each question, for saliency-based and CBM-based explanation. Overall, saliency-based explanations are preferred over CBM-based ones.

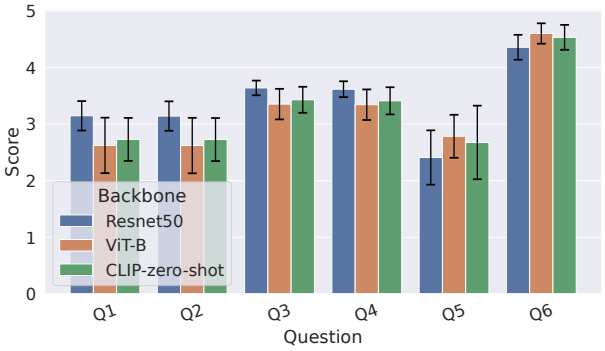


Figure 5: Scores for each question, for different backbones of saliency methods. As backbones for saliency methods, ViT-B and CLIP obtain overall similar results, while ResNet50 has generally better scores.

that saliency maps are perceived as more interpretable than concept-based explanations. Although several factors may contribute to this behavior—such as the lower cognitive load of image-based explanations compared to concept-based ones—a comprehensive investigation of the underlying psychological causes is left for future work. The sole exception to this observation pertains to Q6 (note that for Q5, the ratings are inverted, with a low score indicating favorable behavior). This phenomenon can be explained by the fact that presenting explanations as a heatmap overlaid on the image reduces the perceptual impact of perturbations.

Human preferences for model architecture: Figure 5 illustrates the average scores across XAI methods that use the same backbone, highlighting a general preference for ResNet50. CLIP and ViT achieve similar scores, likely due to the architectural similarities between the two models. ResNet50, which played a pivotal role in the development of many XAI methods, consistently scores higher. This could suggest a potential bias toward ResNet50 in the design and effectiveness of current XAI methods. The results for the last two questions may be due to ViT being more sensitive to the perturbations used than Resnet50. Among the methods we studied, those based on feature factorization—like Eigen-CAM (Muhammad and Yeasin 2020) and Deep Feature Factorization (Collins, Achanta, and Susstrunk 2018)—tend to give more consistent and preferred explanations. This may be because they remove complex components that can create confusing artifacts, making the explanations easier to understand.

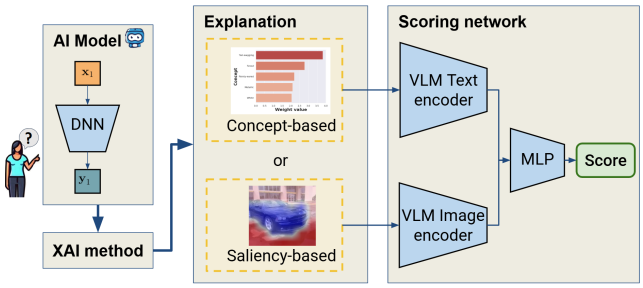


Figure 6: Functioning of the PASTA-score. First, we extract the embeddings for each explanation using a frozen image encoder of a VLM. Then, we employ a scoring network trained on the labels provided by the PASTA-dataset to generate a final score.

Developing the PASTA-score

To provide a tool for measuring human assessment of XAI techniques, we introduce the PASTA-score, which simulates a human evaluation. The global pipeline is illustrated in Figure 6. More precisely, the PASTA-score is composed of an embedding network, that processes both CBM outputs or saliency maps, and a scoring network, that computes scores from the embeddings. Using the data collected in the PASTA-dataset, the PASTA-score aims at predicting the human scores for questions Q1 to Q6 for new explanations, playing the role of an automated benchmark.

Computation of the embeddings

Drawing inspiration from recent literature in automated essay scoring (Yang et al. 2020; Wang et al. 2022), which encounters similar challenges due to working with a dataset of thousands of samples (21,110 samples per question, precisely) and requiring a DNN to automatically learn the score, we opt for a foundation model that we will fine-tune using a multi-linear perceptron. However, unlike automated essay scoring, we deal with both image and textual inputs, making the use of a Vision Language Model (VLM) mandatory. We tested multiple candidates, such as CLIP (Yan et al. 2023), BLIP (Li et al. 2022a) and LLaVa (Liu et al. 2024b). This choice allows for a unified integration of both concept-based explanations, which can be transformed into text, and saliency map-based explanations, which can be projected into the same embedding space. Let $\boldsymbol{x}_i \in \mathbb{R}^{H\times W\times 3}$ be the i-th test image with height H and width W, and let $\boldsymbol{e}_i^{saliency} \in \mathbb{R}^{H\times W}$ be a saliency-based explanation for this image. We denote the image encoder of a Vision-Language Model (VLM) as $\text{VLM}_{\text{image}}$. To embed a saliency explanation, we apply the encoder to the image overlaid with its

heatmap:

$$\phi_{\text{image}}(\boldsymbol{e}_i^{saliency}) = \text{VLM}_{\text{image}}(Heatmap(\boldsymbol{x}_i, \boldsymbol{e}_i^{saliency})), \quad (1)$$

where $Heatmap$ generates the visual overlay of the explanation on the image.

For concept-based methods (CBMs), let $\boldsymbol{e}_i^{CBM} \in \mathbb{R}^K$ be the explanation vector, where K is the number of concepts. This vector is turned into a sentence using a text template, and then encoded with the VLM's text encoder VLM_{text}:

$$\phi_{\text{text}}(\boldsymbol{e}_i^{CBM}) = \text{VLM}_{\text{text}}(Sentence(\boldsymbol{e}_i^{CBM})). \quad (2)$$

Since the PASTA-score is compatible with different VLMs and does not rely on a specific one, we evaluate it using several VLMs: CLIP (Liu et al. 2024b), SIGLIP (Zhai et al. 2023), EVA (Sun et al. 2023), and BLIP (Li et al. 2022a). This results in multiple variants of our metric: PASTA-score$^{\text{CLIP}}$, PASTA-score$^{\text{SIGLIP}}$, PASTA-score$^{\text{EVA}}$, and PASTA-score$^{\text{BLIP}}$.

To support our design choices, Appendix D.2 presents extensive ablations on various factors: the impact of textual templates, the number of concepts K, the way saliency maps are visualized, and whether label information is included. We also explore alternative versions of Equations 1 and 2, and how these choices affect the final score.

Scoring network

Once the embeddings are computed, the label information is concatenated to the embedding, and a scoring network composed of a multi-layer perceptron is used to predict scores. Inspired by Automated Essay Scoring (Yang et al. 2020; Wang et al. 2022), we use a loss L that combines a similarity loss L_s, a mean squared error (MSE) loss L_{mse}, and a ranking loss L_r. From a set of ground truth scores obtained from majority voting $\{m_k\}_{k\in[0,N_s]}$ and the predictions given by the scoring network $\{\hat{m}_k\}_{k\in[0,N_s]}$, the resulting loss is defined as:

$$L(m_k, \hat{m}_k) = \alpha L_s(m_k, \hat{m}_k) + \beta L_{mse}(m_k, \hat{m}_k) + \gamma L_r(m_k, \hat{m}_k), \quad (3)$$

where α, β, and γ are hyperparameters controlling the relative importance of each component. Formulas about the different losses are given in Appendix D.1. Since the PASTA dataset provides $N_s = 5$ ground-truth votes per inference, we explored different aggregation strategies. To mitigate the phenomenon of high non-consensus, the mode was selected as the final choice. Concerning the output range, out-of-range labels ($> 5, < 1$) are rare but clipped when they occur.

Classifier results

Note that in the PASTA-dataset each sample corresponds to a triplet (input image, explanation, human ratings). The same image thus appears multiple times for different XAI methods, and the same XAI method appears multiple times for different images. To guarantee that no leakage occurs between train-test splits, we design them to ensure that the same image, or the same XAI method, does not appear in different splits. Images and XAI methods included in the training splits are randomly chosen based on the run's random seed. For Q1 to Q6, we calculate the Mean Square Error (MSE), Quadratic Weighted Kappa (QWK), and Spearman Correlation Coefficient (SCC) between the predicted and ground truth labels on the test set. The results are presented in Table 1, where we also ablate different choices of embedding methods. We also report the inter-annotator agreement values, which correspond to the expected deviation of the metrics between a randomly selected annotator's score and the mode. Our network best replicates answers to Q1 and Q2, with similar performance across PASTA-score$^{\text{CLIP}}$ and PASTA-score$^{\text{SIGLIP}}$. This is likely due to greater rating diversity and stronger agreement between annotators, which supports more stable training. In contrast, Q3 to Q5 shows lower agreement, and Q5–Q6 has less diverse ratings. While the MSE stays similar, it becomes harder to learn the ranking patterns, likely due to the more subjective nature of these questions and the added uncertainty from image perturbations in Q5 and Q6.

Applications

In this section, we explore three different applications using the PASTA-score as a replacement for human feedback, which would be difficult or too costly to run at scale without automation We use PASTA-score to guide XAI methods toward better interpretability (in the first and third applications) and to analyze how model size affects interpretability (in the second application). All experiments use the PASTA-score model trained on Q1 for consistency.

Mixture of XAI methods

Our first application is to dynamically select the explainer giving the explanation that best matches human judgments for each specific image, using a mixture of XAI methods. In our experiments, we fix the classifier to be a ResNet50, and we select the explanation with the highest PASTA-score for each image. The distribution of selected XAI methods is shown in Table 7. The results indicate a substantial diversity in the methods employed, with FullGrad emerging as the most frequently used, selected nearly half of the time. This trend is reflective of user ratings within the PASTA-dataset, where FullGrad is identified as providing the most effective explanations according to annotators. In terms of faithfulness, the computation of the average faithfulness correlation across explanations selected by our PASTA-score yields a relatively stable value, with a slight improvement compared to the value obtained by averaging over every explainer (0.0627 for our selection versus 0.0579 for the average over every explainer). This confirms that it is possible to use PASTA to enhance the interpretability of explanations without compromising their faithfulness.

Backbone size influence on the understanding of explanation

We use the PASTA-score to investigate whether model size influences the human perception of explanations, and how this relates to XAI methods. To this end, we compute the average PASTA-score within an identical experimental frame-

Metric	Model	Q1	Q2	Q3	Q4	Q5	Q6
MSE ↓	PASTA-score^{CLIP}	0.990 ± 0.104	**0.993 ± 0.096**	2.111 ± 2.529	**0.811 ± 0.095**	1.476 ± 0.183	0.752 ± 0.127
	PASTA-score^{SIGLIP}	**0.989 ± 0.113**	1.009 ± 0.125	**0.842 ± 0.094**	0.840 ± 0.106	**1.396 ± 0.177**	**0.739 ± 0.140**
	PASTA-score^{BLIP}	3.297 ± 1.840	3.287 ± 1.835	5.938 ± 2.542	4.642 ± 3.135	3.005 ± 1.385	10.710 ± 4.943
	PASTA-score^{EVA}	1.666 ± 1.215	1.747 ± 1.174	3.355 ± 3.099	2.097 ± 2.608	1.767 ± 0.568	3.324 ± 5.091
	Human	0.415 ± 0.037	0.429 ± 0.049	0.562 ± 0.104	0.478 ± 0.080	0.509 ± 0.102	0.299 ± 0.051
QWK ↑	PASTA-score^{CLIP}	0.450 ± 0.066	0.452 ± 0.063	0.199 ± 0.040	0.216 ± 0.052	0.165 ± 0.060	0.159 ± 0.031
	PASTA-score^{SIGLIP}	**0.471 ± 0.055**	**0.459 ± 0.056**	**0.237 ± 0.052**	0.219 ± 0.035	**0.177 ± 0.061**	0.165 ± 0.018
	PASTA-score^{BLIP}	0.328 ± 0.023	0.340 ± 0.020	0.181 ± 0.003	0.173 ± 0.017	0.081 ± 0.081	0.159 ± 0.011
	PASTA-score^{EVA}	0.462 ± 0.050	0.457 ± 0.054	0.160 ± 0.099	**0.230 ± 0.018**	0.163 ± 0.029	**0.185 ± 0.049**
	Human	0.849 ± 0.013	0.845 ± 0.017	0.731 ± 0.050	0.748 ± 0.041	0.848 ± 0.029	0.796 ± 0.048
SCC ↑	PASTA-score^{CLIP}	0.484 ± 0.064	0.484 ± 0.062	0.213 ± 0.040	0.230 ± 0.050	0.197 ± 0.073	0.193 ± 0.030
	PASTA-score^{SIGLIP}	**0.501 ± 0.052**	**0.490 ± 0.057**	**0.247 ± 0.048**	0.223 ± 0.029	0.213 ± 0.071	0.196 ± 0.020
	PASTA-score^{BLIP}	0.397 ± 0.036	0.411 ± 0.033	0.207 ± 0.065	0.194 ± 0.014	0.088 ± 0.104	0.218 ± 0.012
	PASTA-score^{EVA}	0.484 ± 0.046	0.474 ± 0.048	0.150 ± 0.155	**0.247 ± 0.015**	**0.216 ± 0.035**	**0.220 ± 0.057**
	Human	0.844 ± 0.017	0.839 ± 0.019	0.722 ± 0.045	0.742 ± 0.038	0.850 ± 0.023	0.789 ± 0.039

Table 1: Mean Square Error (MSE), Quadratic Weighted Kappa (QWK), and Spearman Correlation Coefficient (SCC) for each question. Each value is the average of 5 runs with standard deviation. *Human* refers to inter-annotator agreement.

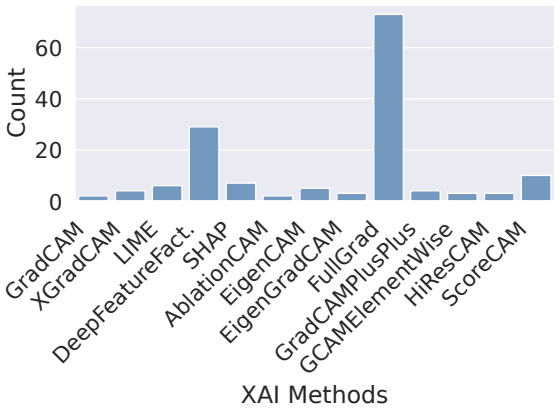


Figure 7: Distribution of methods selected by our mixture of XAI techniques. The x-axis denotes the number of images in the automated benchmark for which the respective XAI method attained the highest performance.

work, varying only the size of the backbone model. Specifically, we employ CLIP as the classifier and select backbones from among its ViT-B-16, ViT-L-14, ViT-H-14, and ViT-g-14 variants. The results of this analysis are presented in Figure 8. Our results show that for activation map-based XAI methods, performance metrics drop as the model size increases. This decline is particularly pronounced when transitioning from the ViT-B to the ViT-L architecture. Several hypotheses may account for this phenomenon. The most plausible explanation is the emergence of artifacts associated with high-norm tokens in the activation maps of larger models, which are used to store information (Darcet et al. 2023). Note that, if this phenomenon is present in our case, such high-norm token artifacts are not universal across all ViTs.. Interestingly, this decrease in score is not perceptible in image perturbation-based XAI techniques, which reinforces the hypothesis that activation artifacts contribute to the reduced interpretability of explanations.

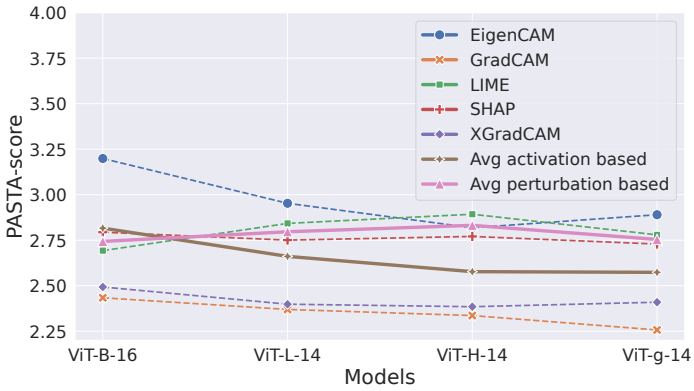
Steering XAI methods towards better alignment

We propose to use the PASTA-score to enhance the interpretability of an off-the-shelf XAI method, namely RISE (Petsiuk, Das, and Saenko 2018). Our approach is as follows: while RISE generates random masks and selects the one that has the best class scores S_{proba}, we introduce a regularization component based on the PASTA-score. Consequently, instead of rating masks using S_{proba}, we employ the following formula:

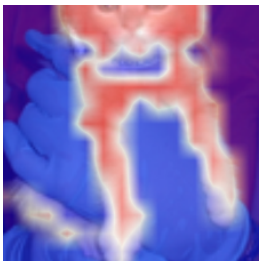
$$w_{RISE+PASTA} = \lambda S_{PASTA} + (1-\lambda) S_{proba}\,, \qquad (4)$$

where $\lambda \in [0,1]$ is a hyperparameter. When $\lambda = 0$, the generated explanation aligns with the original RISE method. Conversely, if $\lambda = 1$, the explanation produced corresponds to a scenario that maximizes the PASTA-score. Note that setting $\lambda = 1$ would result in an explainer optimizing only for human preferences while neglecting the true behavior of the model, which may not yield useful explanations.

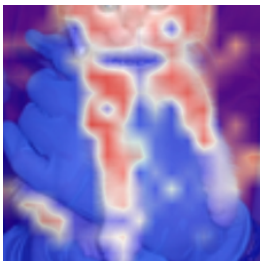
Upon analyzing the samples generated through the optimization process, we initially observe a slight improvement in the localization of highlighted objects. For instance, the explanation depicted in Figure 9e exhibits fewer indecisive zones and demonstrates enhanced precision compared to the explanation shown in Figure 9d. Regarding the case where $\lambda = 1$, we note that the explanation begins to hallucinate zones of interest while omitting others, like in Figure 9c. Additionally, one can observe that the PASTA-score favors large heatmaps. However, the optimized explanations do not systematically overlap with the entire zone of the prediction, suggesting that alignment with the segmentation map of the object to assess the quality of saliency-based explanations, as conducted in previous studies (Karmani et al. 2024; Li et al. 2022b), may prove to be inadequate.



(a) Comparison of various XAI methods across different ViT models.

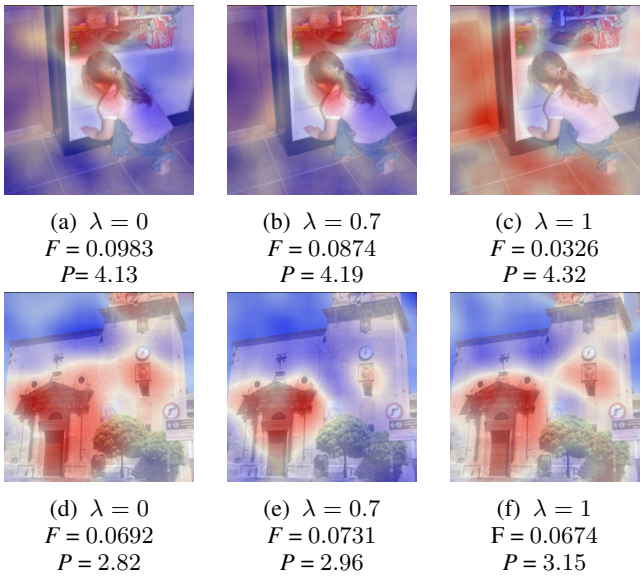


(b) ViT-B-16; PASTA-score of the explanation: 3.12



(c) ViT-g-14; PASTA-score of the explanation: 2.75

Figure 8: Impact of classifier backbone size on the perceived interpretability of image explanations. A notable decrease in the PASTA-score is observed as the model size increases (left). Examination of image samples suggests that artifacts present in the background are likely responsible for this decline.



(a) $\lambda = 0$ $F = 0.0983$ $P = 4.13$

(b) $\lambda = 0.7$ $F = 0.0874$ $P = 4.19$

(c) $\lambda = 1$ $F = 0.0326$ $P = 4.32$

(d) $\lambda = 0$ $F = 0.0692$ $P = 2.82$

(e) $\lambda = 0.7$ $F = 0.0731$ $P = 2.96$

(f) $\lambda = 1$ F = 0.0674 $P = 3.15$

Figure 9: Optimized explanations derived through RISE adjusted with the PASTA-score. Label of the top images: home_or_hotel. Label of the bottom images: Renaissance. F denotes the faithfulness correlation score, P denotes the PASTA-score.

Conclusions

We introduce PASTA, a novel perceptual scoring method designed to benchmark XAI techniques in a human-centric manner. We collect a large-scale benchmark dataset (PASTA-dataset), and use it for an assessment of XAI explanations by human annotators. Based on this dataset, we develop an automated scoring method (PASTA-score) that spans previously unexplored modalities, effectively mimicking human preferences and allowing for circumventing resource-intensive user studies for applications that may benefit so. Deploying PASTA allows new quantitative observations: Our findings reveal a distinct preference for saliency-based explanations, identify a negative impact of backbone size, and demonstrate the potential to generate more human pleasant explanations without compromising faithfulness. These results not only align with human intuition but also corroborate visual examples, affirming the scalability and reliability of PASTA-score.

Limitations: First, the PASTA-score is trained on specific datasets and explanation modalities, which may limit its generalizability to other unseen domains, especially those with domain-specific semantics. Second, the human preferences captured by the PASTA-dataset may inherit the intrinsic biases of human annotators. Third, although PASTA reduces the need for costly user studies, it remains an approximation of subjective human judgment and may overlook nuanced or task-specific interpretability needs, which may justify the need for more resource-intensive ad-hoc human interactions in downstream use cases.

Broader impact: Dynamic scoring approaches could be explored to capture the evolving nature of XAI techniques and their use in real-world applications. PASTA intends to take a step towards creating a transparent and trustworthy AI ecosystem. By aligning AI explanations with human preferences, we aim to foster the development of more interpretable AI systems that can be understood and trusted by users. This work also introduces a perceptual metric, paving the way for future research to implement the PASTA-score as a perceptual loss aimed at enhancing the trustworthiness of networks, drawing for example, inspiration from the emerging use of LPIPS (Zhang et al. 2018) in tasks such as image generation (Jo, Yang, and Kim 2020).

Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 2024 - AD011014675R1). This work is also supported by the EU project ELLIOT (101214398).

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I. J.; Hardt, M.; and Kim, B. 2018. Sanity Checks for Saliency Maps. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 9525–9536.

Agarwal, C.; Krishna, S.; Saxena, E.; Pawelczyk, M.; Johnson, N.; Puri, I.; Zitnik, M.; and Lakkaraju, H. 2022. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35: 15784–15799.

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Arras, L.; Osman, A.; and Samek, W. 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81: 14–40.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

Azzolin, S.; Longa, A.; Teso, S.; and Passerini, A. 2025. Reconsidering Faithfulness in Regular, Self-Explainable and Domain Invariant GNNs. In *The Thirteenth International Conference on Learning Representations*.

Bénard, C.; Biau, G.; Da Veiga, S.; and Scornet, E. 2021. Interpretable random forests via rule extraction. In *International Conference on Artificial Intelligence and Statistics*, 937–945.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Böhle, M.; Singh, N.; Fritz, M.; and Schiele, B. 2024. B-cos alignment for inherently interpretable CNNs and vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6): 4504–4518.

Castelvecchi, D. 2016. Can we open the black box of AI? *Nature News*, 538(7623): 20.

Chalasani, P.; Chen, J.; Chowdhury, A. R.; Wu, X.; and Jha, S. 2020. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, 1383–1391.

Chen, C.; Zhang, M.; Liu, Y.; and Ma, S. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*, 1583–1592.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhang, H.; Zhu, B.; Jordan, M.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132.

Colin, J.; Fel, T.; Cadène, R.; and Serre, T. 2022. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems*, 35: 2832–2845.

Collins, E.; Achanta, R.; and Susstrunk, S. 2018. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 336–352.

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2023. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.

Dasgupta, S.; Frost, N.; and Moshkovitz, M. 2022. Framework for evaluating faithfulness of local explanations. In *International Conference on Machine Learning*, 4794–4815.

Dawoud, K.; Samek, W.; Eisert, P.; Lapuschkin, S.; and Bosse, S. 2023. Human-Centered Evaluation of XAI Methods. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, 912–921. IEEE.

Díaz-Rodríguez, N.; Lamas, A.; Sanchez, J.; Franchi, G.; Donadello, I.; Tabik, S.; Filliat, D.; Cruz, P.; Montes, R.; and Herrera, F. 2022. EXplainable Neural-Symbolic Learning (X-NeSyL) methodology to fuse deep learning representations with expert knowledge graphs: The MonuMAI cultural heritage use case. *Information Fusion*, 79: 58–83.

Fel, T.; Hervier, L.; Vigouroux, D.; Poche, A.; Plakoo, J.; Cadene, R.; Chalvidal, M.; Colin, J.; Boissin, T.; Bethune, L.; Picard, A.; Nicodeme, C.; Gardes, L.; Flandin, G.; and Serre, T. 2022. Xplique: A Deep Learning Explainability Toolbox. *Workshop on Explainable Artificial Intelligence for Computer Vision (CVPR)*.

Ge, J.; Luo, H.; Qian, S.; Gan, Y.; Fu, J.; and Zhang, S. 2023. Chain of thought prompt tuning in vision language models. *arXiv preprint arXiv:2304.07919*.

Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; and Yang, G.-Z. 2019. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37): eaay7120.

Hedström, A.; Weber, L.; Krakowczyk, D.; Bareeva, D.; Motzkus, F.; Samek, W.; Lapuschkin, S.; and Höhne, M. M. M. 2023. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research*, 24(34): 1–11.

Hedström, A.; Weber, L.; Lapuschkin, S.; and Höhne, M. 2024. Sanity checks revisited: An exploration to repair the model parameter randomisation test. *arXiv preprint arXiv:2401.06465*.

Jo, Y.; Yang, S.; and Kim, S. J. 2020. Investigating loss functions for extreme super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 424–425.

Kares, F.; Speith, T.; Zhang, H.; and Langer, M. 2025. What Makes for a Good Saliency Map? Comparing Strategies for Evaluating Saliency Maps in Explainable AI (XAI). *arXiv preprint arXiv:2504.17023*.

Karmani, S.; Sivakaran, T.; Prasad, G.; Ali, M.; Yang, W.; and Tang, S. 2024. KPCA-CAM: Visual Explainability of Deep Computer Vision Models using Kernel PCA. In *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, 1–5. IEEE.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International Conference on Machine Learning*, 5338–5348.

Leavitt, M. L.; and Morcos, A. 2020. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*.

Lee, S.; Kim, S.; Park, S. H.; Kim, G.; and Seo, M. 2024. Prometheusvision: Vision-language model as a judge for fine-grained evaluation. *arXiv preprint arXiv:2401.06591*.

Li, F.; Xi, X.; Cui, Z.; Li, D.; and Zeng, W. 2023. Automatic essay scoring method based on multi-scale features. *Applied Sciences*, 13(11): 6775.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, Y.; Wang, H.; Duan, Y.; Xu, H.; and Li, X. 2022b. Exploring visual interpretability for contrastive language-image pre-training. *arXiv preprint arXiv:2209.07046*.

Liao, Q. V.; Zhang, Y.; Luss, R.; Doshi-Velez, F.; and Dhurandhar, A. 2022. Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, 147–159.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J. A.; Van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88.

Liu, G.; Zhang, J.; Chan, A. B.; and Hsiao, J. H. 2024a. Human attention guided explainable artificial intelligence for computer vision models. *Neural Networks*, 177: 106392.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Miró-Nicolau, M.; Moyà-Alcover, G.; and Jaume-i Capó, A. 2022. Evaluating explainable artificial intelligence for x-ray image analysis. *Applied Sciences*, 12(9): 4459.

Mohseni, S.; Block, J. E.; and Ragan, E. 2021. Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In *26th International Conference on Intelligent User Interfaces*, 22–31.

Montavon, G.; Samek, W.; and Müller, K.-R. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73: 1–15.

Muddamsetty, S. M.; Jahromi, M. N.; and Moeslund, T. B. 2021. Expert level evaluations for explainable AI (XAI) methods in the medical domain. In *International Conference on Pattern Recognition*, 35–46. Springer.

Muhammad, M. B.; and Yeasin, M. 2020. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.

Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; and Seifert, C. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s): 1–42.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144.

Roy, S.; Laberge, G.; Roy, B.; Khomh, F.; Nikanjam, A.; and Mondal, S. 2022. Why don't XAI techniques agree? Characterizing the disagreements between post-hoc explanations of defect predictions. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 444–448. IEEE.

Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; and Zhong, C. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16: 1–85.

Saeed, W.; and Omlin, C. 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263: 110273.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395–405.

Speith, T. 2022. A review of taxonomies of explainable artificial intelligence (XAI) methods. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2239–2250.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Surden, H. 2021. Machine learning and law: An overview. *Research Handbook on Big Data Law*, 171–184.

Taghipour, K.; and Ng, H. T. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891.

Vereschak, O.; Alizadeh, F.; Bailly, G.; and Caramiaux, B. 2024. Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–14.

Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; and Hu, X. 2020. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 24–25.

Wang, Y.; Wang, C.; Li, R.; and Lin, H. 2022. On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. *arXiv preprint arXiv:2205.03835*.

Xiao, C.; Ma, W.; Xu, S. X.; Zhang, K.; Wang, Y.; and Fu, Q. 2024. From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. *arXiv preprint arXiv:2401.06431*.

Xuan, Y.; Small, E.; Sokol, K.; Hettiachchi, D.; and Sanderson, M. 2025. Comprehension is a double-edged sword: Over-interpreting unspecified information in intelligible machine learning explanations. *International Journal of Human-Computer Studies*, 193: 103376.

Yan, A.; Wang, Y.; Zhong, Y.; Dong, C.; He, Z.; Lu, Y.; Wang, W. Y.; Shang, J.; and McAuley, J. 2023. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3090–3100.

Yang, R.; Cao, J.; Wen, Z.; Wu, Y.; and He, X. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1560–1569.

Yang, Y.; Zheng, Y.; Deng, D.; Zhang, J.; Huang, Y.; Yang, Y.; Hsiao, J. H.; and Cao, C. C. 2022. Hsi: Human saliency imitator for benchmarking saliency-based model explanations. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, 231–242.

Yannakoudakis, H.; Briscoe, T.; and Medlock, B. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 180–189.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.