CVPR XAI4CV WORKSHOP

Evaluation Cards for XAI Metrics

A documentation template for XAI evaluation metrics.

CVPR JUNE 3-7, 2026 DENVER COLORADO

Rokas Gipiškis[1,2] **Olga Kurasova**[1]

[1]Institute of Data Science and Digital Technologies, Vilnius University · [2]AI Standards Lab

rokas.gipiskis@mif.vu.lt

1 The evaluation gap

*Our meta-review of **11 surveys** on XAI evaluation (2021–2025) finds three recurring structural problems in how evaluation metrics are introduced and reported.*

Terminological fragmentation

Identical metric names denote different operationalizations across papers, and the same underlying property can be renamed by different research groups.

Pawlicki et al. 2024 · Kadir et al. 2023 · Nauta et al. 2023

Skew toward proxy tasks

Functionally-grounded proxies dominate. Human- and application-grounded methods are underrepresented despite better real-world utility.

Lopes et al. 2022 · Mohseni et al. 2021 · Mangold et al. 2025

Inconsistent implementations

Many proposed metrics remain at the level of theoretical definitions, with no accompanying code. Validity can be model- and modality-dependent.

Coroama & Groza 2022 · Banerjee & Barnwal 2022 · Zhou et al. 2021

FIVE RECURRING DOCUMENTATION FAILURES

01	Metrics introduced without declaring target properties.	*I*
02	Results reported without specifying the evaluation context.	*II*
03	No sensitivity or stability analysis at proposal time.	*III*
04	Metric disagreements rarely acknowledged or interpreted.	*IV*
05	Implementation availability inconsistently reported.	*III*

2 The XAI Evaluation Card

Table 1. *A four-section template addressing the documentation failures from §1: **I**→1, **II**→2, **III**→3 and 5, **IV**→4.*

FIELD	DESCRIPTION
***I.* Identity**	
Metric Name	Unique, descriptive name for the metric.
Target Property / Properties	All explainability properties the metric operationalizes (e.g., fidelity, robustness, clarity), with references to definitions used.
Grounding Level	One or more of: functionally-grounded / human-grounded / application-grounded [Doshi-Velez & Kim, 2017].
***II.* Scope and Context**	
Evaluation Context	Model architecture, data modality, and explanation scope (local / global) under which results are reported.
Assumptions	All assumptions required by the metric (e.g., feature independence, locality, linearity, calibrated probabilities, meaningful baselines).
***III.* Implementation and Validation**	
Implementation Available?	Yes / No. If yes, provide URL or repository reference.
Validation Evidence	Sensitivity analysis, stability analysis, correlation with related metrics. Computational cost where relevant.
Gaming Risk	How a method could achieve a high score on this metric without improving the target property.
Known Failure Cases	Conditions under which the metric is known to fail or produce misleading results.
***IV.* Relationships and Limitations**	
Relationship to Other Metrics	Metrics targeting the same property. Known agreements or disagreements in results.
Disagreement Handling	If this metric conflicts with others reported, state which property is prioritized for the target deployment scenario and why.
Limitations	Main limitations as an operationalization of the target property. Contexts where the metric should not be used.

3 Where this fits

*Structured documentation is established for AI models and datasets. The Evaluation Card extends this to the **metrics used to evaluate** those explanations.*

FRAMEWORK	DOCUMENTS
Model Cards [Mitchell 2019]	*a model*
Datasheets [Gebru 2021]	*a dataset*
XAI ToolSheet [Karunagaran 2022]	*an XAI tool*
Explainability Fact Sheets [Sokol & Flach 2020]	*an XAI method*
XAI Evaluation Card (this work)	***an XAI metric***

Adoption

- Integrate with the reproducibility checklists used by major venues.
- Markdown / LaTeX templates and a JSON schema for parsing and indexing completed cards.
- LLM-assisted drafting for factual fields. Judgement-heavy fields (gaming risk, failure cases) left to authors.
- Workshop-level pilot with reviewer rubrics as soft enforcement.

CONTRIBUTION

A four-section template — ***Identity***, ***Scope & Context***, ***Implementation & Validation***, ***Relationships & Limitations*** — mapping onto five documentation failures from 11 XAI evaluation surveys.

READ THE PAPER

KEY REFERENCES Mitchell et al. 2019. Model cards for model reporting. FAccT, 220–229. · Gebru et al. 2021. Datasheets for datasets. CACM 64(12):86–92. · Sokol & Flach 2020. Explainability fact sheets. FAccT, 56–67. · Karunagaran, Lucic & Custis 2022. XAI ToolSheet. IJCAI Workshop on XAI. · Doshi-Velez & Kim 2017. Towards a rigorous science of interpretable ML. arXiv:1702.08608. · Nauta et al. 2023. From anecdotal evidence to quantitative evaluation. ACM Computing Surveys 55(13s):1–42. · *Full bibliography via QR code.*