Interpretable Generative Stress Testing

A Framework to Audit Decision Boundaries through Synthetic Data Generation

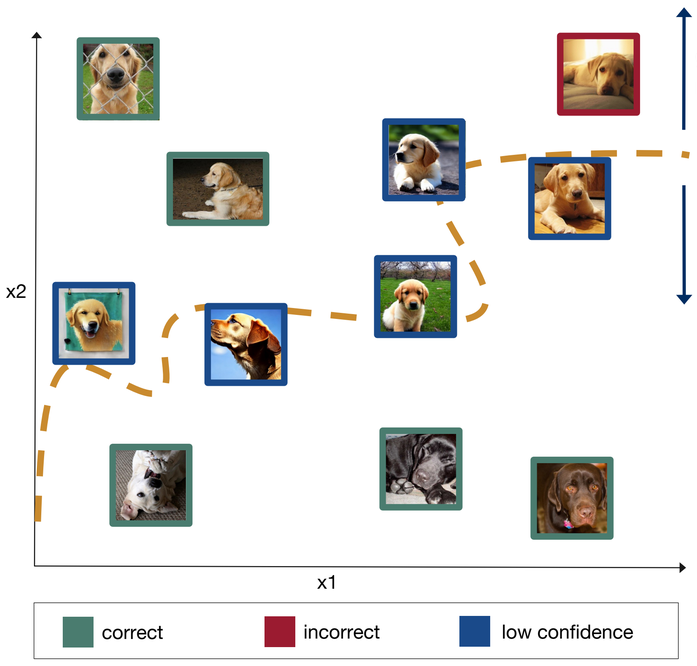
Me!

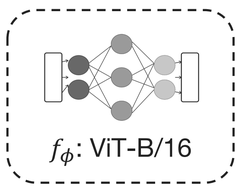
Inês Gomes[1,2,4], Luís F. Teixeira[1,3] , Jan N. van Rijn[4] and Thomas Bäck[4]

[1]Faculty of Engineering, University of Porto, Portugal [2] Artificial Intelligence and Computer Science Laboratory, Portugal

[3]INESC TEC, Portugal [4]Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands

MOTIVATION





98% accuracy

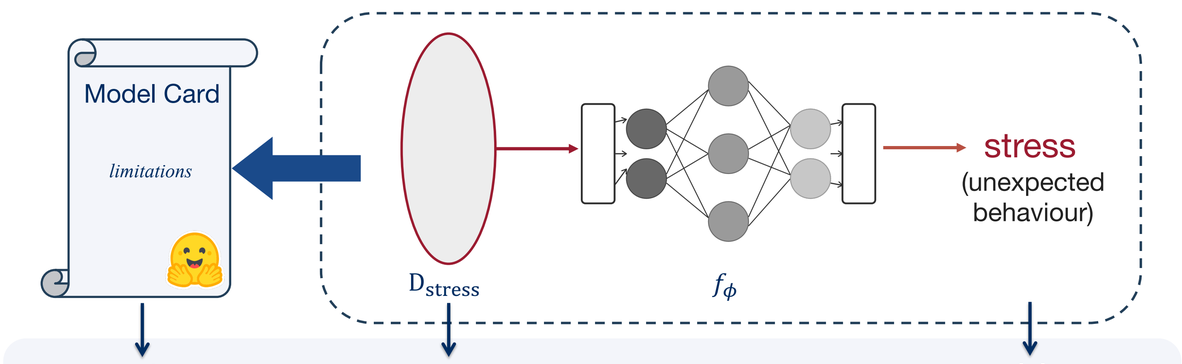
high test set performance ≠ **model trustworthiness**

low-confidence regions can hinder model limitations

GOAL

Develop a framework to expose model limitations associated with low-confidence regions in image classifiers

STRESS TESTING FRAMEWORK



1.3 finding patterns + explain low-confidence

1.2 synthetic data generation

1.1 low-confidence (*arbitrary decision boundaries*)

1.1. ARBITRARY DECISION BOUNDARIES

submitted under review

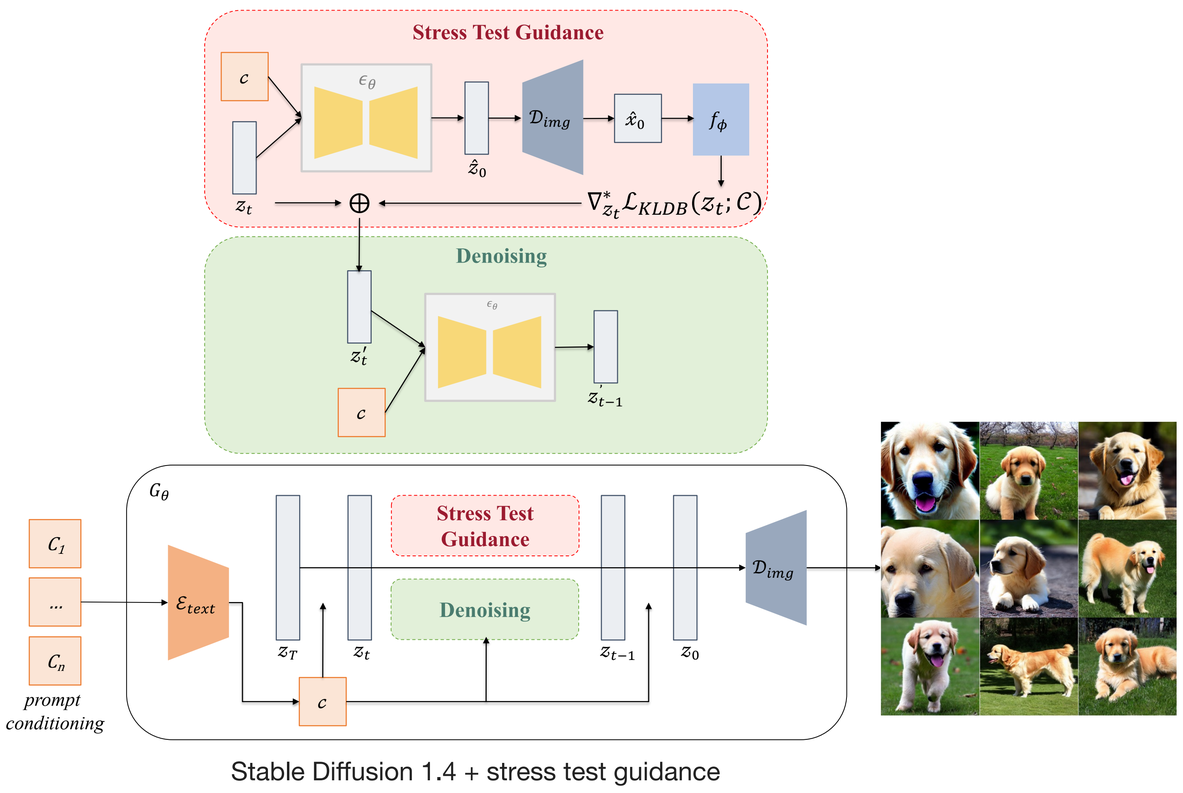
	golden retriever	labrador retriever	cocker spaniel	australian shepherd	border collie	
decision boundary (B)	**0.50**	**0.50**	**0**	**0**	**0**	
	0.48	0.42	0.02	0.08	0	→ KLDB = 0.11
	0.10	0.88	0.01	0.01	0	→ KLDB = 0.52

probability scores

$$\mathrm{KLDB}(B||f_\phi(x)) = \sum_{i=1}^{n} B_i \log\left(\frac{B_i}{\mathrm{softmax}(f_\phi(x)_i)}\right), \qquad \mathrm{KLDB} \in [0,+\infty),\ \downarrow \text{better}$$

1.2. SYNTHETIC DATA GENERATION

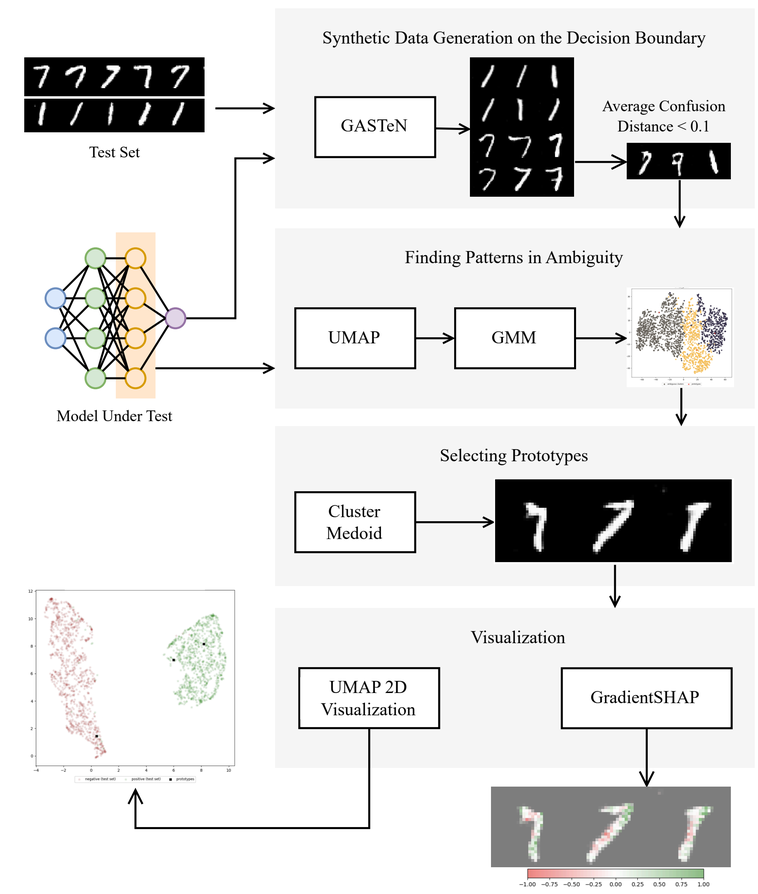
submitted under review



Stable Diffusion 1.4 + stress test guidance

1.3. FINDING PATTERNS

published (CVPRW)



1.3. EXPLAINING LOW-CONFIDENCE

WIP (WACV)

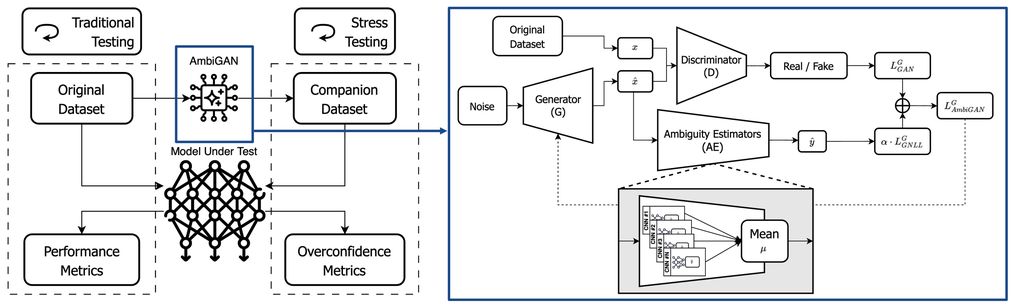
subgroup discovery to identify groups of images driving low-confidence	**OR**	concept-based XAI to identify visual features driving low-confidence
✅ Global textual explanations easily integrated in model cards		✅ Pinpoints specific visual concepts (local and global explanations)
⚠️ Images may belong to multiple subgroups		⚠️ Low confidence may come from missing features
⚠️ Descriptions depend on VLM quality		⚠️ Mixed-class samples hard to detect

Open question: how to characterize the decision boundary? (OOD/hardness/ambiguity/noise/…)

HUBRIS BENCHMARKING

submitted under review

Hubris: model overconfidence when faced with ambiguous data



published datasets

ACKNOWLEDGMENTS

Funded by the European Union under the **Horizon Europe Framework** Programme Grant Agreement Nº: 101095387. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Health and Digital Executive Agency can be held responsible for them.

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI