From Embeddings to Exploration: Engineering Interactive Latent Space Visualizations for AI Model Sensemaking

SEBE VANBRABANT, Hasselt University - Flanders Make, Belgium
JARNE THYS, Hasselt University - Flanders Make, Belgium
GILLES EERLINGS, Hasselt University - Flanders Make, Belgium
KRIS LUYTEN, Hasselt University - Flanders Make, Belgium
GUSTAVO ROVELO RUIZ, Hasselt University - Flanders Make, Belgium
DAVY VANACKEN, Hasselt University - Flanders Make, Belgium

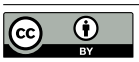
Machine learning systems are often inspected through 2D projections of high-dimensional representations using techniques such as t-SNE or UMAP. While these visualizations provide useful overviews of clustering and similarity, they are inherently static: they display only the existing data points and do not allow users to interactively explore a model's decision space. We present an interactive exploration system — LAPEX— that uses a Variational Autoencoder (VAE) as a generative proxy over a model's training distribution, turning the latent space into a navigable workspace for model sensemaking. Unlike static embeddings, the proxy provides an explicit decoding path from latent coordinates to inputs, enabling interaction patterns such as continuous sampling, interpolation between anchors, and region probing. We operationalize these capabilities through a set of interactive probes that augment a familiar scatter-plot overview with generative overlays for comparing transitions between classes and examining sparsely populated regions. A within-subject formative study ($N = 16$) comparing an interactive VAE-based method to a static t-SNE baseline shows that generative interaction substantially improves counterfactual reasoning and influences how users assess model behavior in sparse or uncertain regions, while static embeddings sometimes provide clearer boundary perception. From these findings, we derive concrete design guidelines and architectural considerations for engineering interactive AI model exploration systems using generative latent representations.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; *Graphical user interfaces*; • **Computing methodologies** → *Artificial intelligence*; *Machine learning approaches*; Supervised learning.

Additional Key Words and Phrases: Explainable AI, AI Model Sensemaking, Latent Space Exploration, Multidimensional Visual Exploration, Variational Autoencoders, Dimensionality Reduction

ACM Reference Format:
Sebe Vanbrabant, Jarne Thys, Gilles Eerlings, Kris Luyten, Gustavo Rovelo Ruiz, and Davy Vanacken. 2026. From Embeddings to Exploration: Engineering Interactive Latent Space Visualizations for AI Model Sensemaking. *Proc. ACM Hum.-Comput. Interact.* 10, 4, Article EICS012 (June 2026), 31 pages. https://doi.org/10.1145/3816764

Authors' Contact Information: Sebe Vanbrabant, Digital Future Lab, Hasselt University - Flanders Make, Diepenbeek, Limburg, Belgium, sebe.vanbrabant@uhasselt.be; Jarne Thys, Digital Future Lab, Hasselt University - Flanders Make, Diepenbeek, Limburg, Belgium, jarne.thys@uhasselt.be; Gilles Eerlings, Digital Future Lab, Hasselt University - Flanders Make, Diepenbeek, Limburg, Belgium, gilles.eerlings@uhasselt.be; Kris Luyten, Digital Future Lab, Hasselt University - Flanders Make, Diepenbeek, Limburg, Belgium, kris.luyten@uhasselt.be; Gustavo Rovelo Ruiz, Digital Future Lab, Hasselt University - Flanders Make, Diepenbeek, Limburg, Belgium, gustavo.roveloruiz@uhasselt.be; Davy Vanacken, Digital Future Lab, Hasselt University - Flanders Make, Diepenbeek, Limburg, Belgium, davy.vanacken@uhasselt.be.



ACM 2573-0142/2026/6-ARTEICS012
https://doi.org/10.1145/3816764

1 Introduction

A central challenge in modern AI is the black-box problem: unlike early approaches, such as decision trees that are human-readable, today's deep neural networks are far less transparent [38, 62]. Explainable AI (XAI) addresses this by coupling predictions with explanations that clarify the reasoning behind those predictions [15]. However, approaches often remain static, cover only a narrow set of questions, and primarily support developers [1, 9, 60, 68]. A common interface for inspecting model behavior at scale is a 2D embedding view obtained via dimensionality reduction (e.g., t-SNE [58] or UMAP [44]) that shows a scatter plot of existing data points. While such views can be effective for overview and cluster-oriented inspection, they are inherently static and constrained to the sampled training dataset: users can look at what is present, but have limited means to probe plausible in-between inputs that support counterfactual-style questions (e.g., *How to* change an instance so that the model's prediction flips). As a result, embedding techniques provide weak affordances for iterative sensemaking, hypothesis formation, and counterfactual reasoning: processes that inherently require interaction [21, 60, 70]. This motivates the engineering of interactive, human-in-the-loop exploration interfaces that let users actively generate and test hypotheses about model behavior through interactive visualization rather than static inspection.

Recently, generative methods have been introduced in the field of XAI. Notably, Large Language Model (LLM)-based approaches can produce personalized explanations and handle diverse intelligibility queries, which are common user questions in interactive systems [16, 33–36, 42, 60]. Beyond language models, generative models such as (variational) autoencoders (VAEs) are frequently used to guide counterfactual generation [2, 12, 14, 24, 40, 48, 59]. Furthermore, a line of work has investigated how VAEs can yield latent representations that are more structured or interpretable, and how these representations can be visualized [4, 19, 45, 56]. However, their role remains underexplored in the context of *interactive* systems for exploring the decision behavior of a separate black-box classifier. Current embedding tools force users to inspect what the model has seen (training data) in relation to its predictions. We engineer a system that allows users to inspect *what the model does* as part of the decision process. From a sensemaking perspective, such proxy spaces should support an iterative workflow: users start from familiar anchors (e.g., typical data samples), explore nearby cases, and use lightweight probes (e.g., small latent interpolations) to test hypotheses about model behavior, rather than relying on one-off explanations [28, 51].

We address the aforementioned limitations of embedding-based inspection by engineering a VAE as the backbone of an interactive ***La****tent-space* ***P****roxy* ***Ex****ploration* interface for black-box model sensemaking: **LAPEX**. Whereas prior VAE-based XAI work commonly uses the latent space to generate *static* counterfactuals, or uses interaction primarily to explain the VAE itself, we instead use the VAE as an *interactive, generative proxy* to probe how a *separate* black-box classifier responds to plausible inputs. This yields a 2D scatter plot of encoded data points and model predictions, similarly to techniques that leverage dimensionality reduction; however, a VAE models a *continuous* latent space with generation of in-between points, whereas embeddings from t-SNE, for instance, have no inverse mapping and are, thus, limited to the dataset. To fairly assess the impact of such continuous interaction, we compare LAPEX, the **Interactive** VAE-based approach, to a **Static** t-SNE variant in a formative user study. We operationalize exploration in the **Interactive** interface through a set of *interactive probes*: toggleable masks in the latent space that use generated data to interactively visualize classifier predictions. In summary, we present the following contributions:

- **C1**: We reframe latent spaces of VAEs as interactive human-in-the-loop interfaces that support exploration, sensemaking, and counterfactual reasoning by probing a separate classifier's predictions, and detail an implementation of LAPEX, an interactive system for such latent proxy exploration.

- **C2**: A set of *interactive probes* that leverage the VAE's latent space to generate novel views on the model's decision space that are not attainable with non-generative dimensionality reduction techniques, enabling continuous decision-space sampling, boundary tracing, and counterfactual interpolation.
- **C3**: A formative user study conducted with $N = 16$ computer scientists that contrasts a **Static** t-SNE-based approach with an **Interactive** VAE-based proxy condition, characterizing how continuous probing (sampling/interpolation) changes users' exploration strategies and when point-based embeddings remain beneficial for global orientation.
- **C4**: A set of actionable engineering guidelines, grounded in observed user behavior, task outcomes, order effects, and failure cases, for designing interactive systems that expose black-box model behavior through generative latent proxies.

Section 2 positions VAEs within XAI and dimensionality reduction literature, on which we base LAPEX's interactive probes detailed in Section 3 (**C2**). We report a formative user study comparing **Static** and **Interactive** approaches in Section 4 (**C3**); the resulting guidelines (**C4**) and the LAPEX system architecture (**C1**) are described in Section 5. Section 6 summarizes the conclusions.

2 Related Work

We briefly discuss developments in VAEs, the state of XAI, and the role and benefits of VAEs and their latent spaces in exploratory and explanatory systems.

2.1 Variational Autoencoders

VAEs are generative encoder-decoder models that learn a latent representation of training data and, crucially for interactive exploration, provide a *decoder* that maps latent coordinates back to valid inputs [27]. This explicit "inverse mapping" enables interfaces to sample and interpolate between data points in latent space and immediately visualize the corresponding reconstructed inputs. VAEs are typically trained by balancing a reconstruction objective with a regularization term that encourages a structured latent distribution, commonly via a Kullback-Leibler divergence to a prior such as a standard normal distribution [30]. The β-VAE variant scales the regularization strength by a factor β to bias the latent space toward a smoother, more factorized structure with individual dimensions capturing independent features at the expense of reconstruction quality [19].

2.2 Explainable AI and Counterfactual Explanations

XAI methods are commonly distinguished between *intrinsically interpretable* models (e.g., rules or decision trees) and *post-hoc* techniques that help people reason about an already-trained black-box model [3, 13]. Our work focuses on post-hoc support for model sensemaking in interactive settings. Well-known post-hoc techniques such as LIME, SHAP, and Anchors primarily target *Why?*/*Why not?* questions by attributing a prediction to features or local rules [39, 52, 53]. Moreover, counterfactual explanations address *How to?* by identifying changes that would flip a model's prediction, often formulated as finding an alternative input x' close to an original input x that yields a different prediction [11, 61, 63]. One common approach to obtain x' is to optimize an objective over an input sample x so that it is minimally changed while still altering the prediction [63], with extensions encouraging sparsity [8], or plausibility via *prototypes* [59]. Prototypes are representative inputs that reliably lead to a given outcome, addressing *When?* [36], and can be acquired by reconstructing the mean encoding of the instances belonging to a class [57, 59]. On the other hand, *criticisms* highlight borderline or atypical cases with the same prediction. Both prototypes and criticisms can be obtained using the MMD-critic method [26]. Rather than optimizing the input directly, counterfactuals can also be generated by optimizing the sample's *latent representation* in a VAE:

REVISE finds the smallest possible change in latent space whose reconstruction crosses the decision boundary [24], C-CHVAE samples faithful counterfactuals uniformly in a sphere around the input sample in latent space [48], while Latent-CF leverages a regular autoencoder without regularization for faster and sparser counterfactuals [2].

2.3 Dimensionality Reduction for Interactive Systems

Despite disentanglement offering interpretable individual dimensions, latent spaces as a whole become increasingly more complex as their dimensionality increases, hindering visualization and interpretation. Dimensionality reduction techniques can be used to reduce high-dimensional structures into lower-dimensional representations while preserving essential geometric or statistical properties. PCA [49], t-SNE [58], and UMAP [44] are commonly used to visualize latent spaces, for instance, in accessible 2D formats [45, 56]. PCA is a linear projection that is often used as a first step because it is fast and provides a well-defined linear mapping; importantly, it also offers a straightforward approximate inverse mapping back to the original space. In contrast, t-SNE and UMAP are nonlinear and primarily optimized for preserving local neighborhood structure, which often yields visually clear clusters but typically does not provide a meaningful inverse mapping.

You et al. [68, 69] use t-SNE to project ResNet-34 [17] embeddings from CIFAR-10 [29] into 2D, augmenting points with LLM-generated descriptions. Similarly, ActiVis and Summit rely on t-SNE and UMAP, respectively, to reduce high-dimensional neuron activations to a 2D space where each point represents a predicted class and spatial proximity reflects similarity [22, 25]. One key distinction between VAE-based approaches and conventional dimensionality reduction techniques is that VAEs learn a latent space with an explicit decoding function, allowing not only projection but also the generation of novel samples between existing points.

2.4 Interactive Explainable AI for Latent Spaces and Embeddings

Several interactive XAI and visualization systems aim to help users understand autoencoders and their latent spaces. For instance, VAE Explainer adds interactions to a VAE with interactive model inputs, the learned latent space, and outputs [4]. It uses a 2D latent space visualized directly as a 2D scatter plot, with each dot representing a sample from the MNIST dataset [31], color-coded by label (digits 0-9). Sercu et al. [56] use a similar approach that enables developers to select and compare different types of autoencoders; by using dimensionality reduction techniques such as PCA and t-SNE, the latent space is projected onto a 2D scatter plot with each dot colored by the model's predicted class. Similarly, Metta et al. [45] introduce the *Identikit Game*, using a visual interface based on β-VAEs to navigate the latent space. Users are given a starting image x and a target image x', and the goal is to transform x into x' by adjusting the latent dimensions. Instead of projecting the entire latent space into 2D, they select the two most relevant dimensions and display their reconstructions on a discretized grid, where each dimension can be implicitly associated with a feature. This iterative process helps users understand the latent structure and the features associated with each dimension, intuitively teaching them complex, abstract patterns.

Prior work primarily treats the latent space as an object of understanding in its own right, or as a means to generate example instances. In contrast, LAPEX leverages VAEs and their latent spaces as a post-hoc technique, offering users an interactive, navigable overview of black-box model behavior. Whereas approaches to counterfactual generation produce an optimized x' from x via latent space, we use the latent space itself for *interactive exploration*. Concretely, we use the decoder to attach meaning to intermediate locations (via reconstructed inputs and corresponding classifier predictions), enabling exploration actions such as continuous sampling, interpolation-based comparison, and region probing. To our knowledge, no prior system uses a generative latent

model as an interactive proxy for a separate black-box classifier whose primary purpose is to sensemake about the classifier rather than to explain the generative model itself.

3 Interactive Probes for Navigating a Generative Proxy Space

LAPEX uses a β-VAE trained on a black-box classifier's training data to construct a *generative proxy space* as shown in Fig. 1. Like other approaches to counterfactual generation, this proxy is data-driven; it is not a faithful representation of the classifier's decision space, and some decoded samples may be out of distribution. We therefore use the proxy for interactive exploration and hypothesis generation *relative to the training data*, not as a literal depiction of decision boundaries. This guarantees that interpolations between existing data points always lie within a relevant part of the data distribution, whereas extrapolations beyond known points may yield out-of-distribution artifacts via the learned decoder. This follows the broader caution that post-hoc explanations can give imperfect representations of black-box behavior in parts of the feature space [55]. Conceptually, we treat this space as a sensemaking workspace in which users iteratively form and revise *frames* about model behavior rather than consuming a single static explanation. Following Klein et al. [28], sensemaking can be understood as a reciprocal process in which data is interpreted through existing frames, while those frames are continuously adapted in response to new data, shaping what is considered relevant, how it is understood, and how it informs the evolving interpretation.

We show an overview of LAPEX in Fig. 2. In both the **Static** and **Interactive** versions, users can inspect individual data points via interactive visualization (e.g., hover tooltips), while a linked view shows the corresponding image (original in **Static**; reconstructed in **Interactive**) and the target classifier's predicted label (and its color) and confidence. Standard navigation interactions (pan/zoom) support moving between overview and detail. This workflow supports an information-foraging strategy where users follow proximal cues (e.g., color/confidence) to decide where to look next [51]. We operationalize these capabilities through a set of *interactive probes* that augment the scatter plot with additional, generative views of regions between data points (implemented via latent-space masks). While approaches differ on multiple dimensions, each dimension is necessitated or influenced by another, and differences between both approaches are minimized to the extent possible: PCA is needed to obtain a 2D view like in t-SNE, generative probes are only possible in generative (VAE-based) approaches, and the layout of both interfaces is identical, with only toggles for available probes (see tab with "Latent Space Insights" in Fig. 2) and the positioning of the scatter plot varying. From a sensemaking perspective, these probes act as engineering scaffolds: they provide low-cost probes that let users test whether observations match their current frame and update it when they do not [28], while also increasing the yield of exploration by reducing the cost of sampling sparsely populated regions [51]. Fig. 3 shows the three generative interactive probes side-by-side. We discuss each interactive probe in turn in the remainder of this section.

3.1 Scatter Plot

The base view of both the **Static** and **Interactive** conditions is a 2D scatter plot of embedded dataset instances. Point positions are determined by t-SNE projection in the **Static** version, or directly when the latent space is 2D (otherwise by a PCA projection of the VAE latents) in the **Interactive** approach. Each data point corresponds to an input image x from the dataset, and is colored in up to three distinct slices by (1) ground-truth label, (2) the target (black-box) classifier's prediction on x, and (3) its prediction on the VAE reconstruction of x, when available. This enables quick comparisons between ground truth labels and (mis)classifications of images. To support example-based reasoning, data points are further divided into three categories: (1) prototypes (typical data points), (2) criticisms (atypical data points), and (3) points on a convex hull surrounding each class to indicate the boundaries of each class. These categories provide concrete anchors for selection

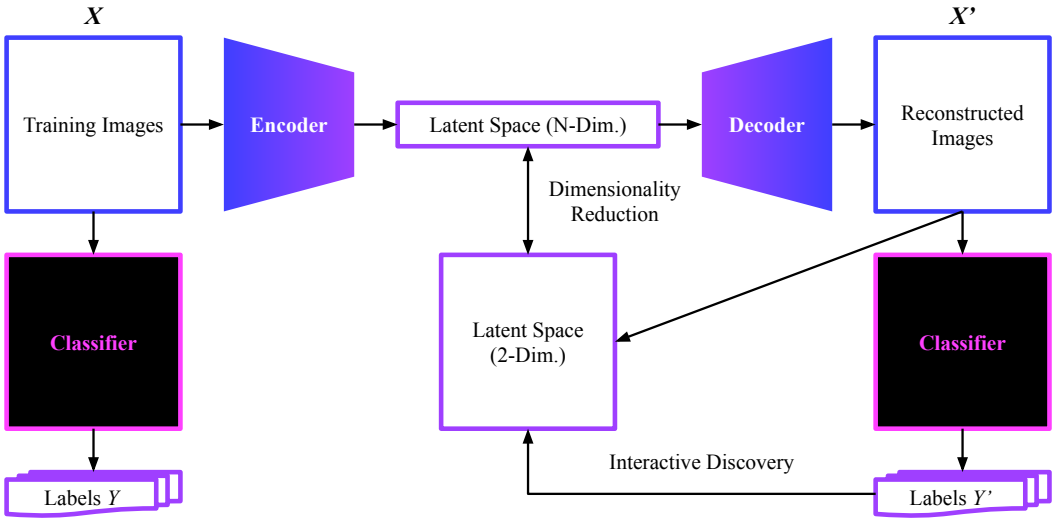


Fig. 1. Schematic overview of the VAE-based proxy. Assume a classifier was trained on training data X to predict corresponding labels Y. We train a VAE on this same set X, learning to reconstruct it as X' using an N-dimensional latent space. Users can see a 2D representation of this learned latent space, on which we impose a set of interactive probes. Interacting with a point in the 2D space finds its corresponding representation in N-dimensional latent space, decodes it, and shows the reconstructed image $x' \in X'$ along with the original classifier's prediction $y' \in Y'$ for this sample, with which the user can update their frame. Note the loop in the image's right half, showing the sensemaking process as a human-in-the-loop workflow.

and comparison, but the scatter plot itself remains point-based: it visualizes the distribution of observed data and associated predictions, without generating new samples.

3.2 Convex Hulls

We overlay convex hulls for each predicted class to provide an at-a-glance summary of where that class's data points lie in 2D and how strongly different classes overlap. Importantly, they are not a decision boundary of the classifier in input space, and they are sensitive to projection distortions. As an interactive probe they help users quickly identify: (1) regions with substantial overlap (often with less confident predictions), (2) regions dominated by a single class (often associated with more consistent predictions), and (3) areas outside all hulls (low data density in the projection), which can be useful for discussing what it means to explore beyond observed examples. This probe is entirely based on the position of points and does not rely on the VAE's generative aspects. We included this visualization to assess how users would reason about overlapping, unambiguous, and empty areas, respectively, in **Static** vs. **Interactive** approaches.

3.3 Latent Grids and Interactive Counterfactual Interpolation

To enable continuous probing in the **Interactive** version, the latent grid densely samples the 2D projection and decodes each sampled location into an input that can be re-evaluated by the target classifier. Concretely, we first fit PCA on the VAE's learned latent space. Next, we uniformly sample a 2D space in the range $[-N, N]$, where N defines the visible extent of the latent space in both "height" and "width". For each grid location, we apply the inverse PCA transform to obtain a vector in the VAE's latent space. Finally, we decode these latent vectors into their corresponding

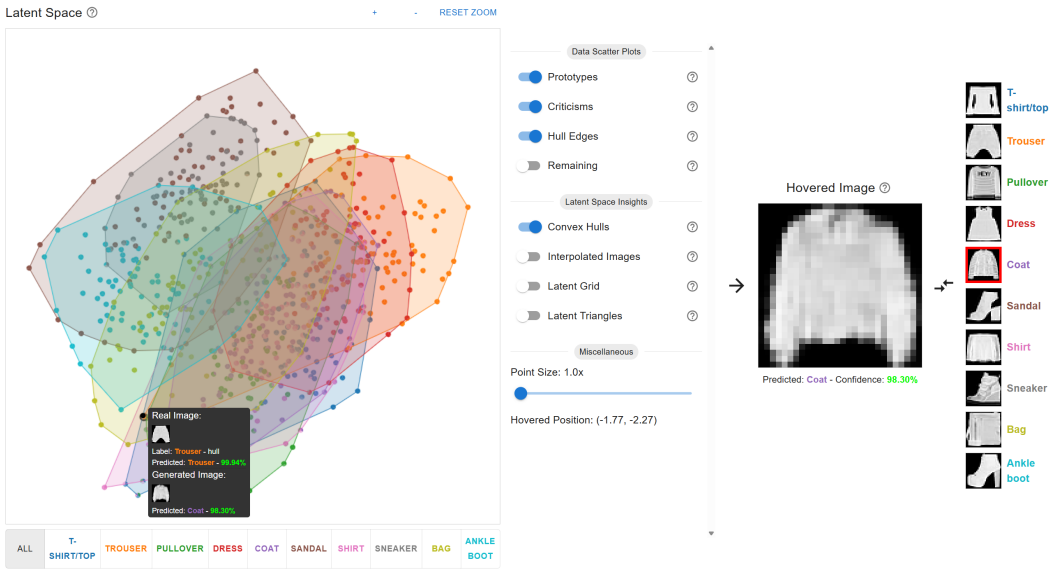


Fig. 2. **Interactive** LAPEX user interface overview for a FashionMNIST classifier. Users can navigate the latent space in the 2D plane on the left, where we show prototypes, criticisms, and hull edges contoured by their convex hulls. Additional interactive probes can be selected under *Latent Space Insights*, discussed in Section 3 and illustrated in Fig. 3. We show a sample that has been misclassified following reconstruction in the latent space. The image's nearest neighbors are shown in the vertical stack on the far right.

(a) Scatter plot showing prototypes, criticisms, hull edges, and interpolated samples at the barycentric coordinates resulting from Delaunay triangulation.

(b) Latent grid colored according to predictions on the learned latent space. Boundaries appear smooth, but the 2D projection can be out of bounds.

(c) Latent triangles colored according to the prediction and confidence on the reconstructed weighted average of their nearest neighbors.

Fig. 3. Side-by-side comparison of global, PCA-based interactive probes for the generative proxy latent space of the **Interactive** LAPEX approach for a classifier trained on the FashionMNIST dataset.

reconstructions. The grid is colored according to the classifier's prediction of the reconstructed image, as shown in Fig. 3b, with color saturation indicating the confidence of the prediction. Latent grids not only mitigate gaps in unknown regions, but also overlapping regions: whereas multiple data points can map to (roughly) the same place in a learned latent space, a point in the latent space itself can only decode into one specific reconstruction, mitigating the noisy and overlapping data, as seen in the center of Fig. 2. Grids fill the latent space in its entirety, but they lack discrete anchors to select, which complicates fine-grained comparisons. To facilitate controlled traversal, we add an *interactive counterfactual interpolation* mechanism: users select two grid locations and move a slider that linearly interpolates between their corresponding latent vectors, decoding and re-evaluating intermediate points on demand. This can be used to discover decision boundaries between areas interactively, illustrating how to visually morph one class into another.

3.4 Latent Triangles and Interpolated Images

Latent grids can become visually noisy as the latent dimensionality increases, and inverse-PCA sampling may produce locations that are weakly supported by the data (e.g., due to projection and reconstruction error). To provide a more selective, anchor-driven alternative, we introduce *latent triangles*. We perform Delaunay triangulation [7] over a set of selected anchor points in the 2D projection (e.g., prototypes and criticisms), and color each triangle according to the prediction on its decoded centroid, as illustrated in Fig. 3c. This summarizes how predictions vary across the continuous proxy space while keeping the visualization tied to concrete data points. Building on this triangulation, the *interpolated images* probe generates additional comparison anchors. For an edge between two anchors, we decode latent interpolations between the corresponding points and display intermediate reconstructions as shown in Fig. 3a. These images act as "bridges" for visual comparison: because they are generated between known samples, they convey which features appear to change between two existing, adjacent anchors. Extending this idea, barycentric interpolation within a triangle provides three-way interpolation among adjacent anchors, yielding intermediate reconstructions supported by nearby data points rather than uniform sampling alone.

4 Formative User Study to Assess the Impact of Interactive Latent Proxy Exploration

This section reports a formative user study that examines how an **Interactive** generative proxy interface affects the ways people explore and reason about a black-box classifier compared to a **Static** embedding view. We focus on interaction capabilities and their appropriation, rather than on optimizing a particular explanation metric. The **Static** condition applies t-SNE to the activations of the output layer of the classifier, whereas the **Interactive** condition uses a VAE-based generative proxy: it provides the same 2D overview but additionally supports the generative interactive probes described in Section 3. Participants completed the same set of tasks in both conditions (see Section 4.3) to elicit exploration strategies, sensemaking behaviors, and use of specific interactive probes. We used a within-subjects design, balanced with a Latin square to mitigate order effects between conditions and datasets. The full study procedure is outlined in Fig. 5. Our study was approved by the Social and Societal Ethics Committee (SMEC) of Hasselt University.

4.1 Participants

We recruited 16 participants, a sample size in line with prior HCI research [5], via email invitations to researchers at our institute. We specifically sought participants with a computer science background and practical experience in AI/ML, data analysis, or programming, as both conditions require understanding fundamental concepts such as model predictions and embeddings. At the same time, we included participants with varying levels of experience to capture a range of exploration strategies from users with different degrees of familiarity. Participants completed a

pre-study questionnaire assessing their technical background and AI-related familiarity/skills (see Appendix A). Participants reported relatively high Tech Comfort ($M = 0.79, SD = 0.07$) and Need for Cognition (NCS-6; $M = 0.72, SD = 0.16$), with moderate AI Familiarity ($M = 0.68, SD = 0.18$), AI Proficiency ($M = 0.56, SD = 0.18$), and AI Experience ($M = 0.56, SD = 0.19$); aggregated backgrounds are shown in Fig. 4. Participants' highest degree attained was collected: 3 participants held a Bachelor's degree, 11 held a Master's degree, and 2 held a Doctoral degree.

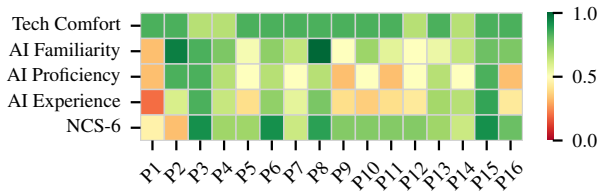


Fig. 4. Individual participants' normalized self-rated level of technological comfort, AI familiarity, AI proficiency, AI experience, and NCS-6 ratings, averaged within each scale and linearly rescaled to the range $[0, 1]$.

4.2 Apparatus

We evaluated two interface conditions across two image classification datasets (MNIST [31] and FashionMNIST [67]) that support quick, non-domain-specific visual validation of reconstructions and predictions; both consist of grayscale 28×28 images with 10 classes, while FashionMNIST is visually more complex. For the **Static** version, we computed 2D t-SNE embeddings; for the **Interactive** version, we trained a β-VAE on an MNIST classifier using a 2-dimensional latent space (to avoid additional dimensionality reduction) with $\beta = 16$ (further encouraging decoupled latent dimensions). For the FashionMNIST classifier, we trained a 16-dimensional latent space with $\beta = 8$.

4.3 Tasks

Although no single standardized user-study protocol exists for evaluating XAI interfaces [38], closely related work commonly uses task-based comparisons against a baseline (e.g., You et al. [68]), or case studies with experts (e.g., Kahng et al. [25]). We implicitly blended these by recruiting participants with moderate to high technical backgrounds (including experts). Participants completed the same task set in each condition: one familiarization task (Task 0; T0) and four main tasks aligned with recurring intelligibility needs and common XAI evaluation emphases [21, 46, 54]. In Task 0, participants were asked to explore the interface and make sense of the latent space by assessing model confidence, reading the tooltips, and using the available probes. Task 1 (T1) asked participants to identify when, in general, the model classifies an image as one class vs. another (*When?*-question). Here, we scored participants' task performance on (1) their ability to draw a decision boundary in the latent space, and (2) their explanation of why the decision boundary was located there. Answers were rated as correct (1/1), partially correct (0.5/1), or incorrect (0/1); both answers were then weighted equally. Task 2 (T2) evaluated participants' ability to discover counterfactual information between classes (*How to?*-question), and their answers were scored similarly to those in Task 1. Task 3 (T3) assessed participants' ability to perceive similarity between different classes. Here, we asked participants to rank three pairs of classes from most similar to least similar, receiving 0.25 points per correct pairwise comparison, and an additional 0.25 if the overall ranking was correct. This yielded a score between 0 and 1. Task 4 (T4) explored how users calibrate their trust based on generative insights. The regions consisted of one out-of-bounds area named *Pilea*, one area

with high overlap in data points named *Soricida*, and one area with one single, unambiguous class named *Arenaria*[1]. The full task descriptions are included in Appendix B. To keep sessions feasible, we limited the study to these tasks, and each task was allowed to take up to 5 minutes (4 minutes for Task 0); sessions lasted 98.1 minutes on average ($SD = 22.4$). Since this was a formative study to assess reasoning processes, participants were not instructed to complete tasks as quickly as possible. Nonetheless, we did collect and investigate timings of tasks, visualizations, and datasets, but found no significant differences in timing between the two approaches.

4.4 Procedure

Participants first completed an informed consent form and provided some demographic information (see Appendix A). Participants also completed the NCS-6 questionnaire to measure their *need for cognition*: the extent to which they are inclined towards effortful cognitive activities [37]. Participants completed the familiarization task and four main tasks (see Appendix B) for each condition with either the **Static** or the **Interactive** version, and either the MNIST or FashionMNIST dataset (thus also switching datasets between conditions to further mitigate potential learning effects). Participants were encouraged to think aloud, enabling the facilitator to take notes on qualitative insights [20]. After each task, participants rated answer confidence and indicated which interactive probes they used and whether these supported their answer (Appendix D).

After each condition, participants completed standardized questionnaires: (1) the *System Causability Scale (SCS)*, which aims to determine whether and to what extent an explainable user interface, explanation, or explanation process itself is suitable for its intended purpose [23]; (2) the *Explanation Satisfaction Scale (ESS)*, which collects participants' judgments after having worked with the explainer [20]; and (3) the *XAI Trust Scale*, which asks participants whether they are confident in the explainer, and whether it is predictable, reliable, efficient, and believable [20]. Participants also rated each interactive probe on a Likert scale to assess whether it was 'understandable', 'useful', 'efficient', 'enjoyable', and 'trustworthy', adapted from prior work on evaluating intelligible interfaces [6]. The study concluded with a semi-structured interview, in which participants' experiences, answers, and understanding were further discussed, guided by the open questions found in Appendix E.4.

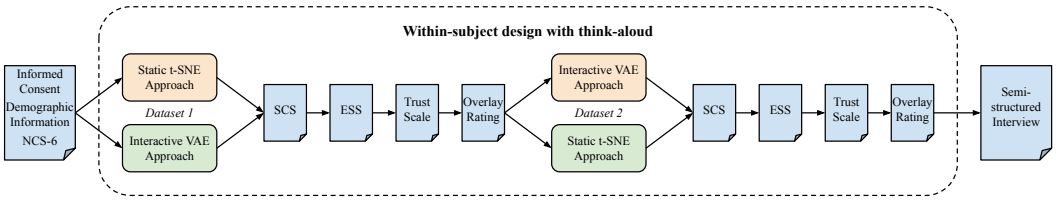


Fig. 5. Within-subject study procedure. Participants completed tasks in both conditions (**Static** t-SNE and **Interactive** VAE; order counterbalanced) using MNIST and FashionMNIST (dataset order counterbalanced). After each condition, participants completed standardized questionnaires and rated interactive probes. The study concluded with a semi-structured interview; we used a think-aloud protocol during task execution.

4.5 Results

In this section, we analyze the results for each questionnaire, task, and interactive probe. A post-hoc sensitivity analysis estimated the minimum detectable effect size for the paired comparisons. Given $N = 16$, $\alpha = .05$, and 80% power, the analyses were sensitive to standardized paired differences

[1]Pilea, Soricida, and Arenaria are Latin for pancake plant, shrew, and sandstone, respectively. Names were chosen randomly to not imply a ranking or elicit an emotional reaction that might bias the ranking.

of $d_z \geq 0.75$. Given $N = 8$, $\alpha = .05$, and 80% power, the analyses were sensitive to standardized paired differences of $d_z \geq 1.16$. Thus, smaller within-subject effects may not have been reliably detectable with this sample size. For the between-group order comparisons, with $N = 8$ per group, $\alpha = .05$, and 80% power, the analyses were sensitive only to large standardized between-group differences of $d \geq 1.51$. Consequently, smaller effects may not have been reliably detectable with our current sample size. The median of the SCS score increased from 0.75 in the **Static** version to 0.81 for the **Interactive** version; Fig. 6 shows individual ratings. For the question "I could change the level of detail on demand", the median scores of **Interactive** and **Static** versions were both 4, but the means were 4.19 and 3.75, respectively. A Wilcoxon Signed-rank test shows a significant effect ($W = 21$, $Z = 2.44$, $p < 0.05$, $r = 0.61$, $d_z = 0.70$, mean difference = 0.44, 95% CI [0.10, 0.77]), which we attribute to the various generative insights of the **Interactive** version, not present in the **Static** version, although d_z is below the sensitivity threshold, indicating that this analysis was underpowered for an effect of this magnitude. P6 noted during Task 2 that "*doing this in the baseline approach was doable, but required more interpretation*". Similarly, P9 stated for Task 4 that "*without this tool, I was not sure how the images in that region would look like*". Moreover, P3 stated that "*the slider and the generated images were useful when the complexity was higher*", implying that generative features provided internal visual references. These remarks suggest that the **Interactive** version's generative elements can help participants quickly develop an intuitive understanding of how to navigate and reason about the latent space. For the ESS, the median score decreased slightly,

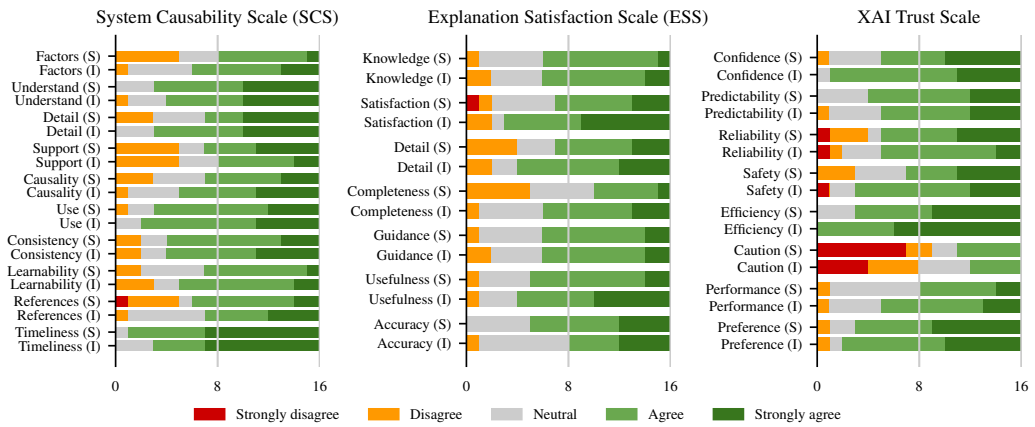


Fig. 6. Level of agreement on Likert scale ratings of the **Static** (S) and **Interactive** (I) versions on the SCS, ESS, and XAI Trust Scale. Participants rated both approaches highly, agreeing positively with most statements in both conditions, with some significant differences in favor of the **Interactive** approach. Note that "Caution" in the XAI Trust Scale is reverse-scored.

from 0.76 in the **Static** approach to 0.74 for the **Interactive** one. Despite a noticeable increase in both average perceived user satisfaction and completeness for the **Interactive** approach, we did not find any significant differences in paired tests for any of the questions, orders, or conditions. The median ratings of the XAI Trust Scale increased slightly from 0.81 in the **Static** version to 0.83 for the **Interactive** one. For the question "The explainer is efficient in that it works very quickly", the medians of the **Static** and **Interactive** approaches were 4 and 5, respectively. A Wilcoxon Signed-rank test shows a significant effect ($W = 21$, $Z = 2.45$, $p < 0.05$, $r = 0.61$, mean difference

= 0.375, 95% CI [0.11, 0.64]), indicating that participants perceived the **Interactive** version as significantly more efficient.

Using a two-way ANOVA, we found a non-significant interaction between order of presentation and the condition on the mean XAI Trust Scale scores ($F(1, 14) = 3.56$, $p = 0.08$, $\eta^2_{\text{partial}} = 0.20$). Given that this result approached conventional significance thresholds, we further explored the potential impact of presentation order. We compared participants who first used the **Interactive** version (Group A) with those who first used the **Static** version (Group B). Using a Mann-Whitney U test, we found a significant effect on the average difference in XAI Trust scores between the two groups. The mean ranks of Group A and Group B were 10.88 and 6.13, respectively ($U = 51$, $Z = 2.01$, $p < 0.05$, $r = 0.50$, $d = 0.94$), however, the observed Cohen's d was below the sensitivity threshold, indicating that this analysis was underpowered for an effect of this magnitude. The medians of the difference in participants' mean 5-point Likert ratings were 0.25 and -0.13 for Group A and Group B, and the mean difference between groups was 0.38 on the 5-point scale, with a Welch 95% CI of [-0.05, 0.80]. Concretely, when participants used the **Static** version *after* the **Interactive** one, they rated it significantly worse. Participants frequently attributed this to missing morphing functionality — especially the latent grid and its interpolation — which they had used to explore transitions across decision boundaries and reliability in regions with limited data coverage (P6, P16, P11). This was already apparent in Task 0 (P4: "*now I would have liked interpolation*"), and recurred in Task 2 (P6: "*Not having the grid interpolation tool is really noticeable here*"; P16: "*here you don't have the cool morphing, which I think is a shame*").

Furthermore, we found a significant effect on the difference in scores for "I am confident in the explainer. I feel that it works well" between the two groups, using a Mann-Whitney U-test. The mean ranks for participants who first used the **Interactive** version (Group A) and those who first used the **Static** version (Group B) were 11.25 and 5.75, respectively ($U = 54$, $Z = 2.56$, $p < 0.05$, $r = 0.64$, $d = 1.58$). The medians of the difference in participants' 5-point Likert ratings were 0.5 and 0 for Group A and Group B, and the mean difference between groups was 1.25 on the 5-point scale, with a Welch 95% CI of [0.38, 2.13]. This indicates that participants were less confident when using the **Static** version after the **Interactive** one. However, we found no significant difference in reported confidence in the tasks themselves between the two conditions, regardless of order, as shown in Fig. 7. Nonetheless, task confidence was generally high, with a mean overall task confidence of 3.9 for the **Static** version and 3.8 for the **Interactive** one. During the semi-structured interview,

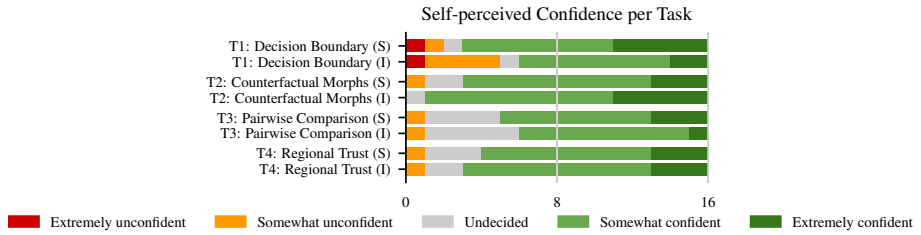


Fig. 7. Level of agreement on Likert scale ratings comparing participants' perceived confidence per task between the **Static** version (S) and the **Interactive** version (I). We found no significant differences between the two approaches in this regard.

participants P1, P4, P5, P6, P7, P8, P9, P13, and P16 all reported experiencing the biggest difference in usefulness between the two conditions in tasks that required feature-based interpretation and comparison. They consistently preferred the **Interactive** version over the **Static** one in such use

cases, as it facilitated more efficient reasoning about comparisons between samples. For Task 2, P9 noted: "*For this task, the latent grid interpolation method helped me to identify the changes that needed to happen. For these kinds of tasks, I would prefer this method*". While P12 and P14 would have liked more time than the foreseen 4 minutes to get acquainted with the condition, most participants were quick to pick up either version, regardless of AI experience.

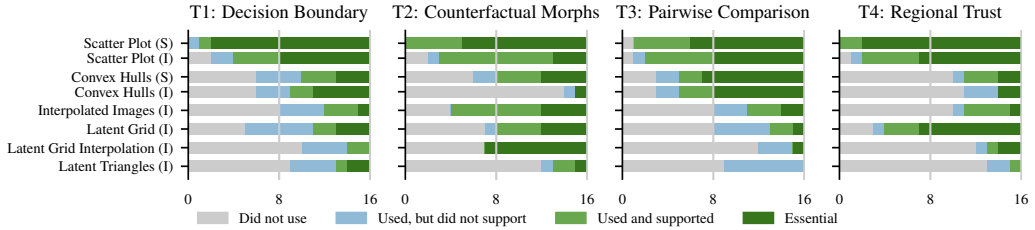
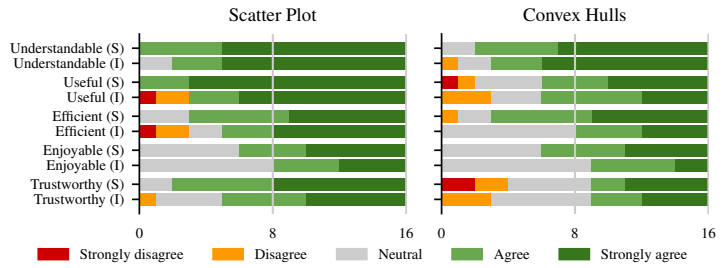


Fig. 8. Interactive probe usage throughout the four tasks in the user study in the **Static** version (S) and the **Interactive** version (I). Participants almost always used the scatter plot to support their answer, regardless of conditions. The generative interactive probes were primarily used for tasks that involved interpolating across unknown regions, such as morphing counterfactuals between regions or assessing trust in uncertain regions.

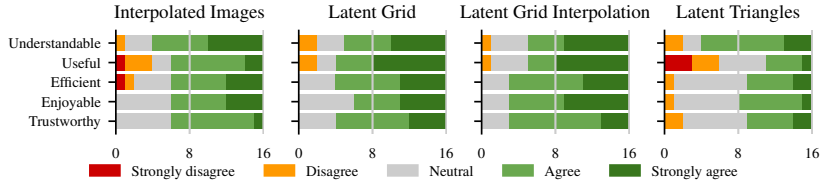
Fig. 8 shows tool usage across tasks. In **Task 1**, participants primarily relied on the scatter plot and convex hulls; generative probes were often used but deemed less essential. Several participants (P4, P6, P7, P10, P13, P14) expressed interest in more granular interactive probes, like a grid for two classes rather than all classes. For instance, P4 stated that they "*expected to see an explicit line between the two classes*" but, instead, the classes were separated by a third class. We found no significant difference between task scores; however, participants reported significantly higher confidence in the **Static** version when using it after the **Interactive** one ($N = 8$). The medians of the **Static** version and the **Interactive** one were 5 and 4, respectively, and a Wilcoxon Signed-rank test shows a significant effect on the reported confidence ($W = 0$, $Z = -2.35$, $p < 0.05$, $r = 0.83$, $d_z = 1.21$, mean difference $= -1.25$, 95% CI $[-2.12, -0.39]$). The primary reason appears to be that t-SNE produces clearer cluster boundaries, enabling more confident responses, whereas PCA projections of the latent space exhibit more class overlap. In **Task 2**, participants who used the **Interactive** version performed significantly better than those who used the **Static** version, with the task accuracy increasing from 63% in the **Static** version to 90% for the **Interactive** one. The medians of the **Static** and **Interactive** versions were 0.5 and 1, respectively. A Wilcoxon Signed-rank test shows a significant effect on the achieved score ($W = 56$, $Z = 2.18$, $p < 0.05$, $r = 0.54$, $d_z = 0.63$, mean difference $= 0.28$, 95% CI $[0.044, 0.52]$), although the Cohen's d_z was below the sensitivity threshold, indicating that this analysis was underpowered for an effect of this magnitude. In the semi-structured interviews, the **Interactive** version's ability to morph images was brought up by 10/16 participants as the most useful aspect distinguishing it from the **Static** approach. For **Task 3**, we did not find a significant difference between the **Static** and **Interactive** conditions, nor between conditions on either MNIST or FashionMNIST. However, scores were significantly higher on FashionMNIST compared to MNIST, regardless of the condition. The medians of MNIST and FashionMNIST were 0.375 and 0.75, respectively. A Wilcoxon Signed-rank test shows a significant effect on the achieved score ($W = 6.5$, $Z = -2.90$, $p < 0.05$, $r = 0.73$, $d_z = 0.98$, mean difference $= -0.39$, 95% CI $[-0.60, -0.18]$). The higher scores on FashionMNIST likely reflect the greater visual distinctiveness of its classes compared to MNIST, making similarity judgments easier regardless of the visualization. In **Task 4**, Pilea, the area that was completely out of bounds and

contained information only when using generative interactive probes, influenced the rankings of 9 participants: five participants changed their ranking in the **Interactive** version compared to the **Static** one from trusting Soricida over Pilea to trusting Pilea more, and four participants switched from trusting Arenaria over Pilea to trusting Pilea more. According to Fig. 8, this was mainly due to the grid visualization, which sampled beyond existing data points. For instance, P13 initially ranked Soricida above Pilea in the scatter plot, but explicitly switched their order after exploring the latent grid. These results suggest that classifications of fully generated, yet unambiguously classified, images were trusted more than ambiguous classifications of real images.

Fig. 9a shows how the scatter plot and convex hulls, present in both conditions, were rated; Fig. 8 shows their usage across the different tasks. The scatter plot was the most frequently used visualization across all tasks in both conditions, receiving high ratings for understandability, usefulness, and trustworthiness, and largely positive ratings for efficiency and enjoyment. Convex hulls were also positively received, but seen as less useful and less trustworthy than the scatter plot, mainly because they were affected by outliers. For instance, P10 stated in Task 3 that "*my first idea was to use the convex hulls, but they are very strongly influenced by outliers, so now I will look at the points*". We show the level of agreement on features present only in the **Interactive** approach



(a) Level of agreement with Likert scale questions about interactive probes that were available in both the **Static** (S) and **Interactive** (I) approaches.



(b) Level of agreement with Likert scale questions about interactive probes unique to the **Interactive** interface.

Fig. 9. Level of agreement with Likert scale questions about interactive probes used in the **Static** (S) and **Interactive** (I) versions. Overall ratings for the scatter plot were high, and generally positive for the latent grid and its interpolation; usefulness ratings for latent images and triangles were more polarizing.

in Fig. 9b and their usage across the different tasks in Fig. 8. The latent grid and its interpolation were the best received after the scatter plot. Participants valued the grid's simplicity in showing the locations of different classes. The grid's interpolation further enabled detailed comparisons between selected samples, as evidenced by its usefulness reported in Task 2. Interpolated images were also well received and widely used for morphing; for instance, P5 and P13 used them to explore the boundaries between two classes in more detail, with the points serving as implicit counterfactuals. The latent grid also heavily impacted perceived trust in out-of-bounds predictions. P13 commented

that "*Task 4 was the most difficult since I was very uncertain about what would happen in the empty space, the latent grid helped reduce the uncertainty.*" and stated that the grid caused them to change their answer compared to the first condition. The latent triangles were the most polarizing, and participants frequently pitted them against the latent grid. For instance, P9 commented that "*The grid is useful because it gives an overview of all classes. I used this more than the triangles because it was simpler, I think, and I didn't see the difference between the two.*", while P2 commented that the latent triangles were "*too much*". Nonetheless, P6 stated that the triangles "*indicated clear decision boundaries*", and P15 reported them as being especially useful to create "*order in the chaos*".

5 Architecture and Guidelines for Interactive Latent Space Exploration Systems

One general finding is that participants considered the **Static** t-SNE approach to be *very helpful*, with the generative aspects of the **Interactive** VAE-based approach making their experience *even better*. We believe this highlights the overall need and appreciation for these 2D embedding-based interfaces for exploring model decisions: they are fairly accessible to users with varying backgrounds and still support experienced AI researchers in interpreting model behavior. In this section, we discuss the model training process and backend implementation of our interactive system, taking the insights acquired from the exploratory user study into consideration. Hereafter, we distill the insights from the user study into actionable guidelines for developers grounded in concrete failure modes. Finally, we offer considerations for LAPEX's engineering and deployment.

5.1 Human-in-the-loop Model Training Flow and Interactive System Implementation

We summarize our LAPEX system architecture in Fig. 10; this process is further discussed in the remainder of this section. In short: the developer first trains a set of VAE-based proxies and selects one based on evaluation metrics. This proxy is then given to the `API`: an HTTP server that generates the interactive probes discussed in Section 3 via dedicated *Router*s and streams them to the frontend. The backend was implemented in Python 3.13.2, using PyTorch [47] for ML-related tasks and Flask for communication with the frontend, which was in turn implemented in React.

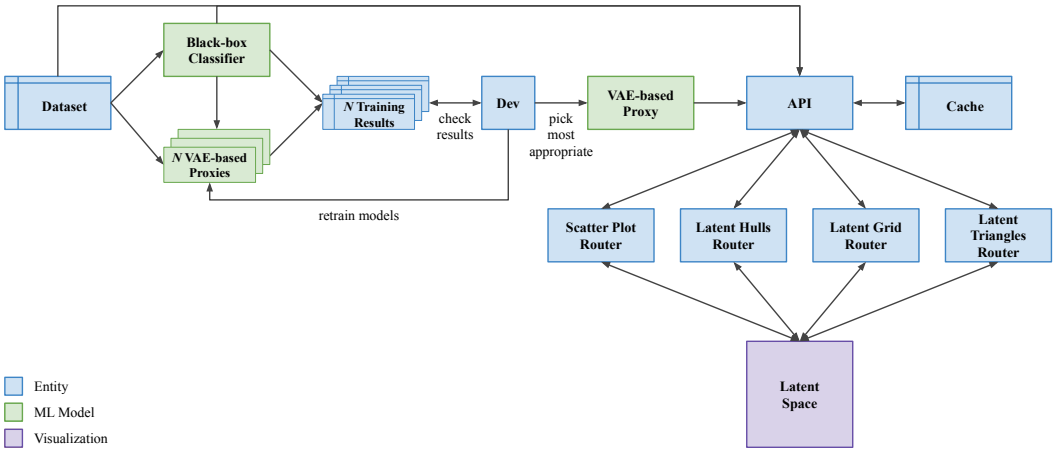


Fig. 10. Overview of the model training and interactive system implementation. Based on the classifier and its dataset, the developer trains *N* VAE-based proxies. These are evaluated on reconstruction quality, fidelity to the classifier, and dimensionality reduction. One proxy is then given to the `API`, which exposes several routers per interactive probe to the frontend visualization. Caching is essential for timely interaction and enables the precomputation and storage of results from immutable, computationally intensive computations.

For obtaining the VAE-based proxy, we train a set of VAEs on the classifier's training data via a grid search over hyperparameter ranges (latent dimension and β), and evaluate them on (1) reconstruction loss and Fréchet Inception Distance (FID) [18] for reconstruction quality; (2) task accuracy, fidelity accuracy, and ShapGAP [41] to determine how well the input image and reconstruction maintain matching classifier predictions, and (3) random triplet accuracy [64] between classifier logits and VAE latents to assess how well the dimensionality reduction accurately maintains relative distances between three data points. To pick the most appropriate VAE-based proxy, we evaluate these models on the aforementioned metrics by analyzing their Pareto front: we pick the model that best matches the classifier (e.g., maximal fidelity accuracy and triplet accuracy) without sacrificing reconstructive quality (e.g., minimal reconstruction loss and FID). Developers should carefully balance these metrics before selecting an appropriate VAE-based proxy or retraining the VAE instead with a larger range of hyperparameters informed by this analysis.

Our LAPEX system backend is implemented as a modular `API` that connects the dataset, a VAE-based proxy, and the target classifier. The proxy encodes input images into a latent space and decodes samples for probing, while the classifier provides predictions for both original and reconstructed inputs. In-between samples needed for generative interactive probes (e.g., for building the latent grid) can be constructed following the techniques outlined in Section 3. The `API` is implemented as an abstract base class that encapsulates shared concerns (caching, metadata, and request handling); concrete subclasses implement either the **Static** t-SNE baseline via scikit-learn or the **Interactive** VAE-based method. Prototypes and criticisms are precomputed using a Python implementation of the MMD-critic method of Kim et al. [26], and `Routers` handle tasks specific to a given interactive probe discussed in Section 3. A `Router` registers endpoints on the `API`'s HTTP server; extending the system, therefore, merely requires adding a new backend `Router` and a corresponding visualization in the frontend. To keep interaction responsive, we cache and reuse immutable, expensive results (e.g., prototypes/criticisms and latent-grid samples), enabling precomputation and substantially shorter request latencies. This reduces the time to start exploring a model from minutes to seconds, limited primarily by the network's speed in fetching results rather than by the processing required to recalculate the visualizations. The frontend is implemented in React using an HTML5 `canvas`. Rendering of the latent space is handled by a controlled draw pipeline with an `InteractiveCanvas` abstraction for panning, zooming, and hit-testing of selectable anchors.

5.2 Developer Guidelines

G1: Enable direct interaction while preserving point-based anchors. Our study results show that latent space exploration can support interactive model sensemaking. Across conditions, scatter plots were generally the most understandable, useful, and trustworthy probes. While generative insights offer additional value, maintaining scatter plots of actual, non-generated data is essential, as they serve as reliable anchors for navigation; otherwise, users can over-weight generated evidence and lose track of what is actually supported by the dataset, as observed in Task 4. Preserving points aligns well with our other visualizations, which, aside from the latent grid, rely directly on them to shape the probe. Furthermore, as evident from Task 2, participants using the **Interactive** VAE-based approach performed significantly better at visual morphing than those using the **Static** approach, and most participants also mentioned this as the most useful aspect of the **Interactive** approach.

G2: Favor simple global overviews with the option of dense local detail. Latent grids were consistently rated as more useful and approachable than more complex constructs such as latent triangles, which some participants described as excessive or confusing. Users preferred overviews that clearly showed where classes were located before engaging with finer-grained structures; when the visualization became too intricate too early, some

participants reported not understanding what additional value it provided and fell back to the scatter plot instead. This supports a design strategy that prioritizes simple, global representations first and introduces denser local detail only when users intentionally zoom in. This conceptually aligns with semantic zooming, a technique in which the level of detail is increased, and different types of information are shown as data is magnified [50].

G3: Support local class-level decision and boundary comparison. In Task 1, we found that participants performed better in the **Static** version than in the **Interactive** approach. A recurring failure mode is that users expect to see decision boundaries even when clear, explicit boundaries do not exist: while PCA provides faithful but ambiguous clusters/boundaries, t-SNE offers visually clear clustering that can lead to overconfidence. Participants also indicated that generative probes would be more actionable if they supported class subselections rather than the entire decision space. We therefore revisit the probes from Section 3 to support *local* exploration in Section 5.3: whereas our interactive probes initially convey a *global* view of the classifier's decision-making, they can be tailored to individual classes instead, better addressing decision boundary ambiguity.

G4: Design for counterfactual comparison, not just inspection. While the **Static** t-SNE embeddings' clear clustering helped participants orient themselves globally, the **Interactive** version better supported counterfactual discovery by enabling *direct* feature-based comparison via smooth reconstructions. Participants perceived the **Interactive** approach as more efficient for reasoning about differences between classes, with sliders, interpolation, and side-by-side morphing enabling direct comparison. Here, task accuracy increased significantly on counterfactual tasks, although the observed effect size was below the post-hoc sensitivity threshold. Participants repeatedly identified latent interpolation and visual morphing as the most valuable features for understanding how changes to input images would affect black-box predictions. We therefore recommend treating interpolation-based comparison as a first-class interaction: make endpoint selection easy, keep endpoints visible as anchors, and present intermediate steps as evidence for counterfactual hypotheses.

G5: Support trust calibration via out-of-distribution visibility. Task 4 shows a clear failure mode for generative probes: participants trusted fully generated, confidently classified samples over noisy regions containing real but overlapping data. Participants did not discriminate between real and generated images; instead, they relied on the classifier's predicted label and confidence, shifting trust judgments when the latent grid made high-confidence behavior visible in areas with no real data coverage. This demonstrates that generative sampling can help *expose* where a model appears (un)stable, but that the same visual smoothness and confidence cues can also *mislead* by making unsupported regions look authoritative. We therefore recommend making out-of-distribution status explicit whenever decoded samples are shown (e.g., visually distinguish generated from real, and overlay local support such as latent density or distance to nearest real neighbor), so that users can calibrate trust with respect to both predicted confidence and proximity to the training distribution.

G6: Treat the full generative system as the default experience. The order effect in XAI Trust Scale ratings suggests a concrete onboarding failure mode: once users have experienced generative probes, removing them is immediately perceived as a loss of explanatory power. Participants who started with the **Interactive** version immediately missed the generative probes when switching to the **Static** version. This is corroborated by the use of the latent grid to rank regional trust. We therefore recommend treating the full generative interface as the default experience and onboarding through controlled emphasis rather than feature withholding: present all interactive probes up front but guide attention (e.g., via semantic

zooming as in G2, sensible defaults, and clear probe labels), so users learn early what each probe provides and how it should be interpreted.

G7: Evaluate latent-space interfaces on reasoning processes, not only task accuracy. Traditional quantitative metrics such as task accuracy, confidence ratings, and aggregate satisfaction scores did not fully capture the benefits and risks of generative exploration. While some tasks showed no significant performance differences, qualitative data revealed substantial shifts in how users reasoned, compared alternatives, and calibrated trust in response to generated evidence. Think-aloud protocols, probe usage patterns, and order effects provided critical insight into these reasoning trajectories. This shows that evaluation should prioritize how users explore, compare, and form hypotheses over time, rather than relying solely on correctness or satisfaction metrics. Finally, as latent space exploration is a complex interaction mechanism, future work should clearly evaluate specific aspects of the system, such as focusing specifically on counterfactuals (*How to?*) or prototypes (*When?*).

5.3 Extending LAPEX for Local Class Comparisons

Thus far, whenever the latent space dimensionality exceeded two, a PCA projection was applied to enable consistent two-dimensional visualization. Based on user study feedback and **G3**, we here extend our implementation to also support local, class-specific visualizations: whereas PCA keeps the two most statistically relevant dimensions, we instead select the two dimensions that exhibit the largest variability for the selected classes, as measured by the Structural Similarity (SSIM), following Metta et al. [45]. Concretely, for each latent dimension, we take a set of anchor latents from the selected classes, perturb that dimension by ±3, decode both reconstructions, and compute the SSIM between them. We average this SSIM score over anchors and pick the two dimensions with the lowest average SSIM, since a lower SSIM indicates a larger visual change. In our implementation, we use 32 anchors sampled randomly from the latents of the selected classes. We illustrate this global vs. local strategy in Fig. 11 and show examples per interactive probe in Fig. 12.

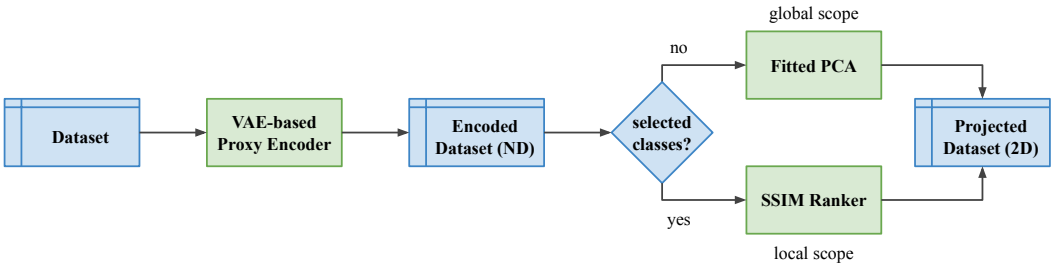
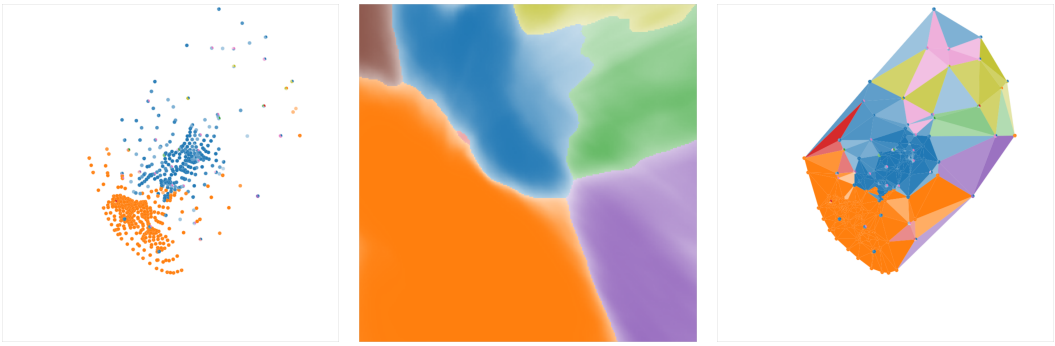


Fig. 11. Overview of the dimensionality reduction strategy. The dataset (or a relevant partition of the selected classes) is encoded into its latent representation using the VAE. When no classes were selected, we assume a global scope and use PCA to find two relevant dimensions to which we project the data. If a subselection of classes was provided by the user, we pick the two most expressive dimensions learned by the VAE instead.

5.4 Limitations and Critical Reflection

Although our results demonstrate the practical value of interactive latent proxy exploration, some considerations must be taken into account in its engineering and deployment:

- While the local interactive probes are grounded in qualitative insights from the user study, we did not conduct a follow-up evaluation to empirically validate them. Local refinements should



(a) Prototypes, criticisms, and hull edges with interpolated images between them. Although overlap still exists, these are primarily ambiguous images, such as very narrow T-shirts resembling Trousers.

(b) Latent grid following the two most relevant dimensions for the selected classes. Although other classes appear outside of selected class boundaries, selected classes dominate the colored space.

(c) Prototypes, criticisms, and hull edges with latent triangles between them. While global projections can be demanding to interpret, local projections show two clusters with finer boundaries and outliers.

Fig. 12. Side-by-side comparison of local, SSIM-based interactive probes for a generative proxy latent space for two selected classes (in orange and blue) for a classifier trained on the FashionMNIST dataset.

be interpreted as design implications derived from observed usage patterns, participant feedback, and task performance. Notably, the latent triangles, being the least preferred in the user study, are most positively affected by switching to SSIM-based ranking, but this change was not validated with users. Nevertheless, local interactive probes are functionally identical to the global visualization but use a different dimensionality-reduction method, highlighting the extensibility of our interactive probes and LAPEX as a whole.

- Latent space exploration can be cognitively demanding. Although participants in our study could work with both the **Static** and **Interactive** approaches, they sometimes expressed hesitation and uncertainty during more complex tasks. We found no clear relationship between participants' technical background or AI expertise and their perceived difficulty or performance. Nevertheless, we do not recommend employing LAPEX for nontechnical end users as is. In the same vein, causal reasoning is currently implicit and part of the user interaction. Future work could explicitly integrate causal relationships between images and predictions, for instance, drawing inspiration from the CLEAR method [65, 66].
- LAPEX supports both global and local latent-space exploration via projections onto two latent dimensions. However, it does not provide human-in-the-loop control over which latent dimensions are shown. Future research could integrate mixed-initiative approaches that enable users to reason about and navigate specific latent factors directly. For instance, Elmqvist et al. [10] show how to transform one projection into another, enabling users to select the dimensions of interest from a matrix of all 2D scatter plots. This can be extended to the dimensions of a VAE-based latent space instead.
- Users explore a learned proxy rather than the classifier's input distribution directly, as discussed in Section 3. While common in counterfactual approaches, smooth reconstructions and latent interpolations are not guaranteed to remain on the learned data distribution or to preserve the causal factors the classifier uses. Thus, although the displayed predictions always come from the classifier, probed inputs may reflect artifacts of the generation rather than valid in-distribution variations. Decoded samples should be treated as hypotheses and accompanied

not only by visual out-of-distribution cues, but also by explicit signals of reconstruction reliability and proximity to real neighboring examples. Safeguards could include quantitative uncertainty estimates, density-aware visualizations, or constraints on latent traversal to regions supported by the data. Future work can address these limitations by incorporating Splatterplots [43], which mitigate overdraw and convey data density, or by exploring models that explicitly integrate latent spaces into decision-making [32].

- The proposed visualizations are primarily intended for model developers and researchers inspecting model behavior. They are less suitable for contexts where data privacy, model confidentiality, or other forms of intellectual property protection are critical. This is a known research challenge in the field of counterfactuals, since explanations can unintentionally expose decision boundaries and undesired shortcuts [61], and may inadvertently reveal sensitive information, opening the model to adversarial attacks. Consequently, LAPEX should be used with caution outside of controlled development or analysis environments.

With these considerations in mind, we envision the foundations provided by LAPEX's engineering to serve as an interactive probing mechanism complementary to existing XAI approaches. Similar tools and systems (see Section 2) often rely on 2D embeddings to visualize model predictions, as does LAPEX, but their interactivity typically remains limited. LAPEX addresses this gap by introducing *interactive* probes grounded in comparable 2D latent spaces. Moreover, the generative properties of the VAE are beneficial across multiple XAI tasks and corresponding intelligibility types [33, 36], while enabling intuitive interaction within these 2D representations. For example, counterfactuals (*How to?*) are often derived from VAE latent spaces, but are rarely communicated through intuitive 2D visualizations. Similarly, prototypes (*When?*) can be extracted from latent representations, yet are rarely displayed in a 2D space. Beyond XAI approaches that rely on latent spaces, LAPEX can also serve as a mechanism to rapidly probe not only model predictions but also corresponding explanations, such as those instantiated by LIME and SHAP. Overall, LAPEX not only introduces interactivity into existing 2D-based XAI tools, but also provides a unified framework for addressing multiple explanatory questions within an interactive latent space.

6 Conclusions

Two-dimensional embedding views (e.g., t-SNE) are a common interface for inspecting black-box classifiers, but they are largely static: they show observed samples but offer limited affordances for probing plausible in-between inputs and iteratively testing hypotheses. We addressed this engineering gap by building LAPEX: an interactive exploration system that uses a VAE as a generative latent proxy, turning the latent space into a navigable interface and enabling the exploration of not only existing instances but also of generated in-between samples (**C1**). We operationalize these capabilities through a set of *interactive probes* that augment a familiar scatterplot overview with generative overlays for comparing transitions between classes and examining sparsely populated regions (**C2**). A formative user study examined how such interactive latent space exploration influences ways people explore and reason about a black-box classifier, comparing **Static** t-SNE-based and **Interactive** VAE-based approaches (**C3**). In particular, generative probes can help users understand model decision boundaries via interactive counterfactual discovery and reason in detail about model predictions in sparsely populated or out-of-bounds regions. The study also revealed complementary strengths: the **Static** version's clearer cluster boundaries support more confident decision boundary identification, while the **Interactive** version's generative capabilities enable richer exploration and more accurate counterfactual reasoning. This suggests that the choice between approaches depends on the specific analytical task at hand. Based on insights from the user study, we formulated actionable guidelines for engineering interactive VAE-based

model-sensemaking systems (**C4**). We detailed our system implementation, showing how it can be built on top of VAEs while maintaining interactivity. Furthermore, we revisited our approach for generating interactive probes to enable local class-level comparisons via SSIM ranking of relevant VAE dimensions. We conclude that proxies are effective exploration tools, well-liked among users with varying levels of AI expertise. The modularity of LAPEX enables a multitude of interactive probes, which can be further extended and improved. Our findings also raise the question of what techniques can further enhance the generative modeling of complex decision-making systems, and what interactive probes can be designed to accommodate them in a human-centered manner.

Acknowledgments

This work was funded by the Special Research Fund (BOF) of Hasselt University, BOF23OWB31 and BOF24OWB28, and by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program, R-13509. Furthermore, we would like to thank all participants who took part in our user study for their time, effort, and valuable feedback.

GenAI Usage Disclosure

GenAI was used for assistance with programming, brainstorming, and improving the readability and language of the manuscript. All concepts and ideas originate from the author(s); genAI was used to refine thoughts and facilitate putting them into practice. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3173574.3174156

[2] Rachana Balasubramanian, Samuel Sharpe, Brian Barr, Jason Wittenbach, and C. Bayan Bruss. 2021. Latent-CF: A Simple Baseline for Reverse Counterfactual Explanations. In *Workshop on Fair AI in Finance at NeurIPS*. arXiv:2012.09301 [cs.LG] https://arxiv.org/abs/2012.09301

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. doi:10.1016/j.inffus.2019.12.012

[4] Donald Bertucci and Alex Endert. 2024. VAE Explainer: Supplement Learning Variational Autoencoders with Interactive Visualization. arXiv:2409.09011 [cs.HC] https://arxiv.org/abs/2409.09011

[5] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. doi:10.1145/2858036.2858498

[6] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: An Intelligible Translation Environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3174098

[7] Boris Delaunay. 1934. Sur la sphère vide. A la mémoire de Georges Voronoï. *Bulletin of the Russian Academy of Sciences. Mathematical Series* 6 (1934), 793–800.

[8] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/c5ff2543b53f4cc0ad3819a36752467b-Paper.pdf

[9] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 82, 19 pages. doi:10.1145/3411764.3445188

[10] Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete. 2008. Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1141–1148. doi:10.1109/TVCG.2008.153

[11] Riccardo Guidotti. 2024. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* 38, 5 (01 Sep 2024), 2770–2824. doi:10.1007/s10618-022-00831-6

[12] Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. 2019. Black Box Explanation by Learning Image Exemplars in the Latent Feature Space. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I* (Würzburg, Germany). Springer-Verlag, Berlin, Heidelberg, 189–205. doi:10.1007/978-3-030-46150-8_12

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018), 42 pages. doi:10.1145/3236009

[14] Riccardo Guidotti, Anna Monreale, Francesco Spinnato, Dino Pedreschi, and Fosca Giannotti. 2020. Explaining Any Time Series Classifier. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*. 167–176. doi:10.1109/CogMI50398.2020.00029

[15] David Gunning and David W. Aha. 2019. DARPA's Explainable Artificial Intelligence Program. *AI Magazine* 40, 2 (2019), 44–58. doi:10.1609/aimag.v40i2.2850 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1609/aimag.v40i2.2850

[16] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 907–924. doi:10.1145/3708359.3712133

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. doi:10.1109/CVPR.2016.90

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Sy2fzU9gl

[20] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* Volume 5 - 2023 (2023). doi:10.3389/fcomp.2023.1096257

[21] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300809

[22] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. 2020. Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 1096–1106. doi:10.1109/TVCG.2019.2934659

[23] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS). *KI - Künstliche Intelligenz* 34, 2 (01 Jun 2020), 193–198. doi:10.1007/s13218-020-00636-z

[24] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. arXiv:1907.09615 [cs.LG] https://arxiv.org/abs/1907.09615

[25] Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, and Duen Horng Chau. 2018. ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 88–97. doi:10.1109/TVCG.2017.2744718

[26] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

[27] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*. arXiv:1312.6114 [stat.ML] https://arxiv.org/abs/1312.6114

[28] Gary Klein, Judith K. Phillips, Elizabeth L. Rall, and David A. Peluso. 2007. A Data-Frame Theory of Sensemaking. In *Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*, Robert R. Hoffman (Ed.). Lawrence Erlbaum Associates Publishers, 113–155.

[29] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images.
[30] Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79 – 86. doi:10.1214/aoms/1177729694
[31] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. doi:10.1109/5.726791
[32] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. 2018. Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). doi:10.1609/aaai.v32i1.11771
[33] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590
[34] Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing* (Orlando, Florida, USA) *(UbiComp '09)*. Association for Computing Machinery, New York, NY, USA, 195–204. doi:10.1145/1620545.1620576
[35] Brian Y. Lim and Anind K. Dey. 2010. Toolkit to support intelligibility in context-aware applications. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (Copenhagen, Denmark) *(UbiComp '10)*. Association for Computing Machinery, New York, NY, USA, 13–22. doi:10.1145/1864349.1864353
[36] Brian Y Lim, Qian Yang, Ashraf M Abdul, and Danding Wang. 2019. Why these Explanations? Selecting Intelligibility Types for Explanation Goals. In *Explainable Smart Systems Workshop at IUI*. https://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-20.pdf
[37] Gabriel Lins de Holanda Coelho, Paul H P Hanel, and Lukas J Wolf. 2018. The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version. *Assessment* 27, 8 (Aug. 2018), 1870–1885.
[38] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* 106 (2024), 102301. doi:10.1016/j.inffus.2024.102301
[39] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
[40] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2020. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. In *Workshop on Do the right thing: Machine learning and Causal Inference for improved decision making at NeurIPS*. arXiv:1912.03277 [cs.LG] https://arxiv.org/abs/1912.03277
[41] Ettore Mariotti, Adarsa Sivaprasad, and Jose Maria Alonso Moral. 2023. Beyond Prediction Similarity: ShapGAP for Evaluating Faithful Surrogate Models in XAI. In *Explainable Artificial Intelligence*, Luca Longo (Ed.). Springer Nature Switzerland, Cham, 160–173.
[42] David Martens, James Hinns, Camille Dams, Mark Vergouwen, and Theodoros Evgeniou. 2025. Tell me a story! Narrative-driven XAI with Large Language Models. *Decision Support Systems* 191 (2025), 114402. doi:10.1016/j.dss.2025.114402
[43] Adrian Mayorga and Michael Gleicher. 2013. Splatterplots: Overcoming Overdraw in Scatter Plots. *IEEE Transactions on Visualization and Computer Graphics* 19, 9 (2013), 1526–1538. doi:10.1109/TVCG.2013.65
[44] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML] https://arxiv.org/abs/1802.03426
[45] Carlo Metta, Eleonora Cappuccio, and Salvatore Rinzivillo. 2025. Interactive Visual Exploration of Latent Spaces for Explainable AI: Bridging Concepts and Features. In *Adaptive XAI: Advancing Intelligent Interfaces for Tailored AI Explanations (2nd Edition) Workshop at IUI*. https://ceur-ws.org/Vol-3957/AXAI-paper10.pdf
[46] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 11, 3–4, Article 24 (Sept. 2021), 45 pages. doi:10.1145/3387166
[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
[48] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing

Machinery, New York, NY, USA, 3126–3132. doi:10.1145/3366423.3380087

[49] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 11 (1901), 559–572. doi:10.1080/14786440109462720

[50] Ken Perlin and David Fox. 1993. Pad: an alternative approach to the computer interface. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques* (Anaheim, CA) *(SIGGRAPH '93)*. Association for Computing Machinery, New York, NY, USA, 57–64. doi:10.1145/166117.166125

[51] Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106, 4 (1999), 643–675. doi:10.1037/0033-295X.106.4.643

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). doi:10.1609/aaai.v32i1.11491

[54] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2024. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (2024), 2104–2122. doi:10.1109/TPAMI.2023.3331846

[55] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (01 May 2019), 206–215. doi:10.1038/s42256-019-0048-x

[56] Tom Sercu, Sebastian Gehrmann, Hendrik Strobelt, Payel Das, Inkit Padhi, Cícero Nogueira dos Santos, Kahini Wadhawan, and Vijil Chenthamarakshan. 2019. Interactive Visual Exploration of Latent Space (IVELS) for peptide auto-encoder model selection. In *Deep Generative Models for Highly Structured Data Workshop at ICLR*.

[57] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf

[58] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[59] Arnaud Van Looveren and Janis Klaise. 2021. Interpretable Counterfactual Explanations Guided by Prototypes. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano (Eds.). Springer International Publishing, Cham, 650–665.

[60] Sebe Vanbrabant, Gilles Eerlings, Gustavo Alberto Rovelo Ruiz, and Davy Vanacken. 2025. ECHO: Enhancing Conversational Explainable AI through Tool-Augmented Language Models. *Proc. ACM Hum.-Comput. Interact.* 9, 4, Article EICS014 (June 2025), 33 pages. doi:10.1145/3734191

[61] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. 2024. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *ACM Comput. Surv.* 56, 12, Article 312 (Oct. 2024), 42 pages. doi:10.1145/3677119

[62] Warren J. von Eschenbach. 2021. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology* 34, 4 (01 Dec 2021), 1607–1622. doi:10.1007/s13347-021-00477-0

[63] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018), 841–887. Spring 2018.

[64] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. *Journal of Machine Learning Research* 22, 201 (2021), 1–73. http://jmlr.org/papers/v22/20-1061.html

[65] Adam White and Artur d'Avila Garcez. 2020. Measurable Counterfactual Local Explanations for Any Classifier. In *24th European Conference on Artificial Intelligence (ECAI)*. http://ecai2020.eu/papers/514_paper.pdf

[66] Adam White, Kwun Ho Ngan, James Phelan, Kevin Ryan, Saman Sadeghi Afgeh, Constantino Carlos Reyes-Aldasoro, and Artur d'Avila Garcez. 2023. Contrastive counterfactual visual explanations with overdetermination. *Machine Learning* 112, 9 (01 Sep 2023), 3497–3525. doi:10.1007/s10994-023-06333-w

[67] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:1708.07747 [cs.LG] https://arxiv.org/abs/1708.07747

[68] Yuzhe You, Helen Weixu Chen, and Jian Zhao. 2025. Enhancing AI Explainability for Non-technical Users with LLM-Driven Narrative Gamification. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 221, 7 pages. doi:10.1145/3706599.3719795

[69] Yuzhe You and Jian Zhao. 2024. Gamifying XAI: Enhancing AI Explainability for Non-technical Users through LLM-Powered Narrative Gamifications. arXiv:2410.04035 [cs.HC] https://arxiv.org/abs/2410.04035

[70] Tong Zhang, Mengao Zhang, Wei Yan Low, X. Jessie Yang, and Boyang Albert Li. 2025. Conversational Explanations: Discussing Explainable AI with Non-AI Experts. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 409–424. doi:10.1145/3708359.3712143

A User Study: Pre-Task Questionnaires

A.1 AI Familiarity

Participants indicated their AI familiarity via the questions in Table 1.

1 = Extremely unfamiliar 2 3 4 = Neutral 5 6 7 = Extremely familiar

Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Artificial Intelligence and Machine Learning	❍	❍	❍	❍	❍	❍	❍
2. Generative AI	❍	❍	❍	❍	❍	❍	❍
3. Neural Networks	❍	❍	❍	❍	❍	❍	❍
4. Latent Spaces	❍	❍	❍	❍	❍	❍	❍
5. White boxes and Black boxes	❍	❍	❍	❍	❍	❍	❍
6. True/false positives/negatives	❍	❍	❍	❍	❍	❍	❍
7. Dimensionality Reduction	❍	❍	❍	❍	❍	❍	❍

Table 1. Questions to estimate participants' AI familiarity.

A.2 AI Experience

Participants indicated their AI experience via the questions in Table 2.

1 = Extremely unfamiliar 2 3 4 = Neutral 5 6 7 = Extremely familiar

Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Computer Programming	❍	❍	❍	❍	❍	❍	❍
2. Data Analysis	❍	❍	❍	❍	❍	❍	❍
3. Training AI Models	❍	❍	❍	❍	❍	❍	❍
4. Evaluating AI Models	❍	❍	❍	❍	❍	❍	❍
5. Interpreting/Explaining AI Model Behavior	❍	❍	❍	❍	❍	❍	❍

Table 2. Questions to estimate participants' AI experience.

A.3 The Need for Cognition Scale (NCS-6)

Participants responded to each item using the scale in Table 3 by marking one circle (❍) per row. Note that questions 3 and 4 are reverse-scored.

1 = Extremely uncharacteristic 2 = Somewhat uncharacteristic 3 = Uncertain 4 = Somewhat characteristic 5 = Extremely characteristic

Item	1	2	3	4	5
1. I would prefer complex to simple problems.	❍	❍	❍	❍	❍
2. I like to have the responsibility of handling a situation that requires a lot of thinking.	❍	❍	❍	❍	❍
3. Thinking is not my idea of fun.	❍	❍	❍	❍	❍
4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.	❍	❍	❍	❍	❍
5. I really enjoy a task that involves coming up with new solutions to problems.	❍	❍	❍	❍	❍
6. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.	❍	❍	❍	❍	❍

Table 3. The Need for Cognition Scale (NCS-6) [37].

B User Study: Tasks

Task 0: Initial Exploration

Task: Explore the interface and its tools[2]; then, use each visualization to get familiar with it. Explore the latent space to find the regions and examples where the model is highly confident in its predictions, and areas where the classifications seem uncertain or mixed. You can always consult the tooltips next to each tool in subsequent tasks.

Max. Duration: 4 minutes

Task 1: Model Behavior Understanding and Decision Boundary Identification

Task: Identify when, in general, the model classifies an image as either [a one or a seven]/[a Coat or a Pullover]. Outline where the decision boundary lies and identify examples that support this reasoning.

Max. Duration: 5 minutes

Task 2: Counterfactual Discovery and Morphing

Task: Find an image that the model predicts as ['7']/['Coat']; discover and explain how to morph the image through minimal changes so that the model would classify it as a ['9']/['Bag'] instead. Write down what features have to change.

Max. Duration: 5 minutes

[2]We simply called interactive probes 'tools' in the user study.

Task 3: Pairwise Class Similarity Assessment

Task: Rank these three class pairs from most similar to least similar based on what you observe: [(2, 9), (5, 6), (4, 9)]/[(Dress, Shirt), (Coat, Pullover), (Dress, Trouser)]. Explain your reasoning.
Max. Duration: 5 minutes

Task 4: Regional Trust Calibration

Task: The facilitator will show three regions in the latent space. Where do you trust the model's predictions, and where do you not? Justify your choices using the visualization.
Max. Duration: 5 minutes

C User Study: Familiarization Task 0 Tooltips

This section lists the tooltips shown when hovering over the question marks seen in Fig. 2.

C.1 Latent Space

```
The visualization below is called the <b>latent space</b>. It is a learned
    representation of the data visualized as a 2-dimensional space.
<br />
<br />
In the latent space, each point corresponds to an image, and the position of the point
    reflects the features of that image as understood by the model.
<br />
<br />
Points that are <b>close together in the latent space</b> represent images that are <b>
    similar in terms of these learned features</b>, while points that are far apart
    represent dissimilar images.
<br />
<br />
Consequently, classes with <b>similar features tend to cluster together</b>, while
    classes with distinct features are further apart.
```

C.2 Prototypes

```
Toggle to show or hide prototypes (i.e., typical images) in the latent space.
<br />
<br />
Images shown on hover {!isGenerated ? <b>are real</b> : <b>have been generated</b>}!
```

C.3 Criticisms

```
Toggle to show or hide criticisms (i.e., atypical images) in the latent space.
<br />
<br />
Images shown on hover {!isGenerated ? <b>are real</b> : <b>have been generated</b>}!
```

C.4 Hull Edges

```
Toggle to show or hide the points part of the convex hull surrounding each class; they
    can help you understand each class's boundaries.
```

```
<br />
<br />
Images shown on hover {!isGenerated ? <b>are real</b> : <b>have been generated</b>}!
```

C.5 Remaining

```
Toggle to show or hide the remaining data points not in any of the previous 3
    categories.
<br />
<br />
Images shown on hover {!isGenerated ? <b>are real</b> : <b>have been generated</b>}!
```

C.6 Convex Hulls

```
Convex hulls show the outlines of predicted classes in the latent space; they can help
    you understand the model's <b>decision boundaries</b>.
```

C.7 Interpolated Images

```
The interpolated images show the average of 2 adjacent images; they can help you
    understand <b>how to</b> morph one image into another.
<br />
<br />
Images shown on hover <b>have been generated</b>!
```

C.8 Latent Grid

```
The latent grid shows the model's predictions across the latent space; it can help you
    understand <b>when</b>, in general, the model predicts which classes are where in
    the latent space.
<br />
<br />
Images shown on hover <b>have been generated</b>!
```

C.9 Latent Triangles

```
Latent triangles show the average of the 3 surrounding images; they can help you
    understand how the model thinks about similar, blended images.
<br />
<br />
Images shown on hover <b>have been generated</b>!
```

C.10 Hovered Image

```
This is the image that corresponds to the point you are currently hovering over in the
    latent space.
It can be a real image or a generated image.
<br />
<br />
To the right, you can see the nearest neighbors for each class in the dataset.
```

```
These are the real images that are most similar to the hovered image in latent space.
<br />
The image with the red border is the most similar image of the same class as the
    hovered image.
```

D User Study: Per-Task Questionnaires

D.1 Confidence

Participants were asked to rate their confidence in their answer after each task for both conditions (except for the exploration task 0) via the question in Table 4. For the **Interactive** condition, we also asked participants which visualization most supported their answer.

1 = Extremely unconfident 2 = Somewhat unconfident 3 = Undecided 4 = Somewhat confident 5 = Extremely confident

Item	1	2	3	4	5
How confident are you in your answer?	❍	❍	❍	❍	❍

Table 4. After each task, participants were asked how confident they were in their answer.

D.2 Interactive Probe Usage

Participants indicated how the interactive probes supported their answers via the questions in Table 5.

0 = I did not use this visualization 1 = I used this visualization, but it did not support my answer 2 = I used this visualization to support my answer 3 = This visualization provided essential information to support my answer

Item	0	1	2	3
Scatter plot	❍	❍	❍	❍
Convex hulls	❍	❍	❍	❍
Interpolated images	❍	❍	❍	❍
Latent grid	❍	❍	❍	❍
Latent grid's interpolation	❍	❍	❍	❍
Latent triangles	❍	❍	❍	❍

Table 5. After each task in the **Interactive** condition, participants were also asked to indicate which interactive probe most supported them in getting their answer. For the **Static** version, participants only had the scatter plot and convex hulls.

E User Study: Post-Task Questionnaires

E.1 The System Causability Scale (SCS)

Participants responded to each item using the scale in Table 6 by marking one circle (❍) per row:

1 = Strongly disagree 2 = Disagree 3 = Neutral 4 = Agree 5 = Strongly agree

Item	1	2	3	4	5
1. I found that the data included all relevant known causal factors with sufficient precision and granularity.	❍	❍	❍	❍	❍
2. I understood the explanations within the context of my work.	❍	❍	❍	❍	❍
3. I could change the level of detail on demand.	❍	❍	❍	❍	❍
4. I did not need support to understand the explanations.	❍	❍	❍	❍	❍
5. I found the explanations helped me to understand causality.	❍	❍	❍	❍	❍
6. I was able to use the explanations with my knowledge base.	❍	❍	❍	❍	❍
7. I did not find inconsistencies between explanations.	❍	❍	❍	❍	❍
8. I think that most people would learn to understand the explanations very quickly.	❍	❍	❍	❍	❍
9. I did not need more references in the explanations.	❍	❍	❍	❍	❍
10. I received the explanations in a timely and efficient manner.	❍	❍	❍	❍	❍

Table 6. The System Causability Scale (SCS) [23].

E.2 The Explanation Satisfaction Scale (ESS)

Participants responded to each item using the scale in Table 7 by marking one circle (❍) per row. We substituted *software*, *algorithm*, and *tool* in the original questionnaire of Hoffman et al. [20] with *underlying AI model* for consistency and clarity in the communication to the participants.

1 = Strongly disagree 2 = Disagree 3 = Neutral 4 = Agree 5 = Strongly agree

E.3 The XAI Trust Scale

Participants responded to each item using the scale in Table 8 by marking one circle (❍) per row. We substituted *tool* and *system* in the original questionnaire of Hoffman et al. [20] with *explainer* for consistency and clarity in the communication to the participants.

1 = Strongly disagree 2 = Disagree 3 = Neutral 4 = Agree 5 = Strongly agree

E.4 Semi-structured Interview: Open-ended Questions

In addition to the quantized questions discussed above, we gave participants the following open-ended questions to initiate an open conversation with the facilitator:

- What task(s) did you find the most difficult? Which one was the easiest? Why?

Item	**1**	**2**	**3**	**4**	**5**
1. From the explanation, I know how the underlying AI model works.	❍	❍	❍	❍	❍
2. This explanation of how the underlying AI model works is **satisfying**.	❍	❍	❍	❍	❍
3. This explanation of how the underlying AI model works has **sufficient detail**.	❍	❍	❍	❍	❍
4. This explanation of how the underlying AI model works seems **complete**.	❍	❍	❍	❍	❍
5. This explanation of how the underlying AI model works **tells me how to use it**.	❍	❍	❍	❍	❍
6. This explanation of how the underlying AI model works is **useful to my goals**.	❍	❍	❍	❍	❍
7. This explanation of the underlying AI model shows me how **accurate** the underlying AI model is.	❍	❍	❍	❍	❍

Table 7. The Explanation Satisfaction Scale (ESS) [20].

Item	**1**	**2**	**3**	**4**	**5**
1. I am confident in the explainer. I feel that it works well.	❍	❍	❍	❍	❍
2. The outputs of the explainer are very predictable.	❍	❍	❍	❍	❍
3. The explainer is very reliable. I can count on it to be correct all the time.	❍	❍	❍	❍	❍
4. I feel safe that when I rely on the explainer, I will get the right answers.	❍	❍	❍	❍	❍
5. The explainer is efficient in that it works very quickly.	❍	❍	❍	❍	❍
6. I am wary of the explainer.	❍	❍	❍	❍	❍
7. The explainer can perform the task better than a novice human user.	❍	❍	❍	❍	❍
8. I like using the explainer for decision-making.	❍	❍	❍	❍	❍

Table 8. The XAI Trust Scale [20].

- In what task did you experience the biggest difference in usefulness between the two approaches? In what scenarios would you choose one explainer over the other? Why?
- Which aspects of the explainers were the most helpful? Are there any aspects of one condition that you missed in the other?
- Did the quality of the generated images impact the usefulness of the explainer interface?
- Do you have any further comments?