A Strategic Blueprint for Explainable Generative Frameworks Using Diffusion Models

Part 1: The Architectural Core: Diffusion as a Generative Paradigm

The efficacy of any synthetic data framework is contingent on the robustness of its underlying generative model. Denoising Diffusion Probabilistic Models (DDPMs) have emerged as the state-of-the-art paradigm, surpassing Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) in generation quality for many domains.[1] A comprehensive framework must begin with a foundational analysis of this architecture, its variants, and its computational trade-offs.

1.1 Deconstructing the Denoising Process: A Foundational Analysis of DDPM

DDPMs are a class of latent variable models inspired by non-equilibrium thermodynamics.[1] The generative process is modeled as the reverse of a fixed diffusion process that gradually destroys data by adding Gaussian noise.

The Forward Process (Diffusion)
This is a fixed, non-learned Markov chain, q, that incrementally adds Gaussian noise to an initial data sample x_0 over T timesteps. The data at timestep t is given by:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I})$$

Where β_t is a predefined variance schedule. As $t \rightarrow T$, this process

transforms any complex data sample x_0 into a sample x_T from a simple, tractable prior distribution (an isotropic Gaussian).6

The Reverse Process (Generation)

Generation is achieved by a learned reverse Markov chain, p_{θ}, that starts with pure noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises it, step-by-step, to produce a new data sample x_0.6 The core of the model is a neural network (often a U-Net architecture for image data) parameterized by θ, which is trained to predict the parameters of the reverse transition $p_{\theta}(x_{t-1}|x_t)$.

Methodological Deep Dive

The authors of DDPM (Ho et al., 2020) demonstrated that the variational bound (VLB) on the data log-likelihood can be simplified.6 Instead of predicting the mean of the denoised sample x_{t-1}, the optimal strategy is to re-parameterize the model ϵ_{θ} to predict the noise ϵ that was added at step t.10 The training objective simplifies to a mean-squared error between the true noise ϵ and the predicted noise $\epsilon_{\theta}(x_t, t)$.6

This pixel-space methodology, while powerful, has a significant limitation: computational cost. Both training and inference require the full U-Net to be evaluated hundreds or thousands of times *per sample*, operating on the full, high-dimensional data (e.g., $256 \times 256 \times 3$).[13] This prohibitive cost is the primary motivation for subsequent architectural innovations.

1.2 Accelerating Generation: Denoising Diffusion Implicit Models (DDIM)

The slow sampling speed of DDPMs stems from their stochastic, Markovian reverse process, which requires the full $T=1000$ (or more) steps to generate a sample.[13] Denoising Diffusion Implicit Models (DDIMs) address this inference bottleneck.[16]

The key innovation of DDIM (Song et al., 2020) is the formulation of a more general, *non-Markovian* diffusion process that leads to the *exact same training objective* as DDPM.[13] This means a DDPM-trained model ϵ_{θ} can be used with a DDIM sampling procedure. This "implicit" or deterministic process allows for sampling by "skipping" steps, generating high-quality samples in as few as 10-50 steps—a 10x to 100x speedup over DDPM sampling.[13] Furthermore, DDIM's deterministic nature enables "consistency": the same latent noise x_T will always produce the same image x_0, a property that DDPMs lack and which allows for semantically meaningful image interpolation.[13]

1.3 The Latent Space Revolution: High-Resolution Synthesis with LDMs

While DDIM accelerates *inference*, it does not solve the problem of high *training* cost. Latent Diffusion Models (LDMs), the architecture underpinning systems like Stable Diffusion, solve this by moving the entire diffusion process out of the high-dimensional pixel space.[15]

The LDM (Rombach et al., 2022) methodology is a two-stage process:

1. **Stage 1: Perceptual Compression.** An autoencoder (typically a VAE) is trained to compress high-dimensional images x into a lower-dimensional *latent representation* z.[22] This latent space is designed to be "perceptually equivalent" to the pixel space, meaning it captures the semantic content while discarding high-frequency, imperceptible details.[21]
2. **Stage 2: Latent Diffusion.** The DDPM/DDIM diffusion process is then trained *entirely within this small, learned latent space*.[14] The U-Net now operates on a much smaller tensor (e.g., $32 \times 32 \times 4$ instead of $256 \times 256 \times 3$), dramatically reducing the computational load for both training and inference.[14]

A high-resolution image is recovered only at the final step, with a single, fast forward pass through the VAE's decoder.[23] This LDM framework also introduces cross-attention layers into the U-Net, allowing the denoising process to be conditioned on general inputs like text, enabling text-to-image synthesis.[14]

The foundational models (DDPM, LDM) were developed for *images*, a homogeneous data modality. Their success relies on properties that other data domains lack: (1) spatial locality, which is effectively exploited by U-Net architectures, and (2) the ability to learn a smooth, compressed semantic latent space, which is exploited by the LDM's VAE.

Tabular and time-series data do not share these properties. Tabular data is heterogeneous (mixed numerical/categorical) with no spatial relationships.[27] Time-series data is sequential and autoregressive.[29] A naive application of an image-based LDM to a table (e.g., training a VAE on its feature vector) would fail, as tabular data lacks the "perceptual" redundancy that the VAE is designed to compress. This fundamental domain-architecture mismatch necessitates the development of bespoke architectures, explored in the next section, and has profound implications for how these models can be explained.

Part 2: Domain-Specific Architectures: Adapting Diffusion Beyond the Pixel

Given that canonical diffusion models are ill-suited for non-image data, a successful framework must employ an architecture specifically designed for its target domain. The following analysis presents the state-of-the-art (SOTA) generative architectures for the two most common and challenging non-image domains: tabular and time-series.

2.1 Path A: Modeling Heterogeneous Data (The Tabular Challenge)

The primary challenge in tabular data generation is its heterogeneous nature. A single table contains a mix of continuous numerical features, discrete numerical features, and non-ordinal categorical features, all governed by complex, non-linear inter-column correlations.[27]

2.1.1 Baseline Methodology: TabDDPM

The TabDDPM framework (Kotelnikov et al., 2023) was a landmark paper demonstrating that diffusion models can be successfully adapted to tabular data, outperforming SOTA GAN (CTGAN) and VAE (TVAE) alternatives.[1]

Its methodology avoids a complex latent space and instead applies a hybrid diffusion process directly to the preprocessed features [1]:

- **Feature Handling:** Numerical features are normalized (e.g., via quantile transformation) and modeled using a standard **Gaussian diffusion** process. Categorical features are one-hot encoded and modeled using a **multinomial diffusion** process, a discrete-data variant of diffusion.[1]
- **Architecture:** As tabular data lacks spatial locality, the denoising network is *not* a U-Net. Instead, TabDDPM employs a simple Multi-Layer Perceptron (MLP) to predict the noise (for numericals) and the original class probabilities (for categoricals).[33]

2.1.2 SOTA Methodology: TabDiff

The TabDiff model (Shi et al., ICLR 2025) is a next-generation tabular diffusion model explicitly designed to improve upon the limitations of TabDDPM. It demonstrates superior generative capacity, achieving up to a 22.5% improvement on pair-wise column correlation estimations, a

key metric for tabular data fidelity.[27]

TabDiff introduces several critical innovations:

1. **Unified Continuous-Time Process:** Unlike TabDDPM's hybrid Gaussian/Multinomial approach, TabDiff models *all* data types (numerical and categorical) within a single, unified continuous-time diffusion framework.[27]
2. **Categorical Handling:** Categorical columns are converted to one-hot vectors, and a special `` class is added. The continuous diffusion process is then applied directly to these one-hot vectors.[31]
3. **Feature-wised Learnable Diffusion:** This is the model's most significant contribution. TabDiff recognizes that different features (e.g., a 'salary' column with high variance and a 'gender' column with low variance) have vastly different distributions. Applying the *same* noise schedule to all features is suboptimal. TabDiff instead implements *learnable, feature-wised noise schedules* that allow the diffusion process to adapt to the unique heterogeneity of each column.[31]
4. **Conditional Generation:** The framework integrates classifier-free guidance, a SOTA conditioning mechanism [38], to perform tasks like *missing value imputation* with high fidelity.[31]

A direct comparison highlights the architectural advantages of TabDiff for a practitioner seeking a SOTA generative foundation.

Table 1: Comparative Analysis of Tabular Diffusion Architectures

Feature	TabDDPM (Kotelnikov et al., 2023)	TabDiff (Shi et al., 2025)
Core Process	Hybrid (Gaussian + Multinomial Diffusion)	Unified Continuous-Time Diffusion
Numerical Handling	Gaussian Diffusion	Gaussian Diffusion
Categorical Handling	Multinomial Diffusion (Discrete)	Continuous Diffusion on One-Hot Vectors
Noise Schedule	Fixed, Shared Schedule	**Feature-wised, Learnable Schedules**
Denoising Network	Multi-Layer Perceptron	MLP / Transformer-based [38]

	(MLP)	
Conditional Task	Class-Conditioned Generation	Classifier-Free Guidance for Imputation

2.2 Path B: Capturing Temporal Dynamics (The Time Series Challenge)

For time-series data, the architectural challenge is different. The model must capture temporal dependencies, auto-correlations, and complex patterns like seasonality and trends.[29]

2.2.1 Autoregressive Methodology: TimeGrad

TimeGrad (Rasul et al., 2021) is a SOTA model for multivariate probabilistic time-series forecasting.[30] Its core idea is to combine a recurrent autoregressive model with a diffusion-based generative process.[30]

- **Methodology:** An RNN (e.g., GRU or LSTM) consumes the time-series history $x_{0, t-1}$ and any relevant covariates c_t to produce a hidden state h_t.[30] This state h_t encapsulates the temporal context. This hidden state is then used to *condition* a diffusion model (DDPM), which generates the sample $x_{0, t}$ for the *next* time step by sampling from $p_\theta(x_{0, t} | h_t)$.[30]
- **Explainability:** TimeGrad is a powerful probabilistic model, but its step-by-step conditional generation process, where a diffusion model is nested inside an RNN, makes it an exceptionally challenging "black box" to interpret from a post-hoc XAI perspective.

2.2.2 Intrinsically Interpretable Methodology: Diffusion-TS

In response to the opacity of models like TimeGrad, recent research has produced architectures with *intrinsic explainability*. Diffusion-TS (ICLR 2024) is a prime example of this "interpretable-by-design" philosophy.[40]

- **Core Idea:** Diffusion-TS is a diffusion framework designed specifically to generate high-quality time series while *simultaneously* providing human-understandable

explanations. It achieves this by focusing on *disentangled temporal representations*.[40]

- **Intrinsic XAI:** The model's architecture features an *interpretable decoder*. This decoder explicitly decomposes the generated time series into two distinct, additive components [44]:
    1. **Seasonal Part:** Modeled using trigonometric representations (i.e., Fourier series).
    2. **Trend Part:** Modeled using polynomial regressors.
- **Training:** The model is trained to reconstruct the original sample (rather than predict the noise) and utilizes a Fourier-based loss term to enforce this decomposition.[40]

The analysis of these two paths reveals a fundamental strategic crossroad. The choice of generative architecture (Part 2) *dramatically* constrains and informs the subsequent choice of an explainability framework (Parts 3 & 4).

If the target domain is tabular, the SOTA model (TabDiff) is a high-performance "black box," which *forces* the adoption of a purely *post-hoc* explainability framework.

If the target domain is time series, the practitioner faces a choice: select a black-box model like TimeGrad (and face a difficult post-hoc interpretation challenge) or select an intrinsically interpretable model like Diffusion-TS, where the XAI is a built-in feature. This co-design of the generative and explanatory components is a central theme in building a modern, trustworthy framework.

Part 3: Unlocking the Black Box: A Framework for Generative Explainability (GenXAI)

The request for "post-hoc explainability" implies an understanding that the generative model is a "black box." As AI models, particularly deep learning systems, grow in complexity, their internal decision-making processes become opaque. This lack of transparency hinders trust and adoption, especially in high-stakes fields like healthcare and finance.[45]

The field of eXplainable Artificial Intelligence (XAI) aims to provide this transparency, generally through two approaches: *intrinsic* methods (designing models that are interpretable by design, like Diffusion-TS) and *post-hoc* methods (developing external techniques to analyze an already-trained model).[45]

3.1 The Paradigm Shift: Why GenAI Demands New XAI (GenXAI)

The rise of Generative AI (GenAI) has created entirely new challenges for explainability.[47] For a traditional classifier, the "decision" is a simple, low-dimensional output (e.g., a class label or a probability). For a generative model, the "decision" is a high-dimensional, complex artifact: a 512x512 image, a 1000-token text document, or a complete synthetic database.

This new paradigm, sometimes called GenXAI [48], must answer new questions:

- Not just "Why was this classified as 'cat'?" but "Why was *this specific* cat generated instead of another?"
- "How did the model *internally* represent the concept of 'cat' during generation?"
- "Which training data is most responsible for this generated output?"

3.2 The Failure of Traditional XAI: Limitations of LIME & SHAP

It is tempting to apply popular, model-agnostic XAI techniques like LIME and SHAP to this problem. This approach is fundamentally flawed.

- **LIME (Local Interpretable Model-Agnostic Explanations):** LIME explains a single prediction by training a simple, interpretable linear model on a *local neighborhood* of the instance.[46] It creates this neighborhood by *perturbing* the input (e.g., turning off words or image patches) and observing the change in the model's output probability.[51]
- **SHAP (SHapley Additive exPlanations):** SHAP is a game-theoretic approach that computes the precise contribution of each feature to a prediction, guaranteeing that the contributions "add up".[46]

These methods are ill-suited for generative models. How does one "perturb" a complex text prompt to explain a generated table? How can SHAP, which requires a single scalar output (like a class probability), be applied to a 512x512 image output? The core assumptions of these frameworks break down.[56]

More critically, research has demonstrated that LIME and SHAP can be *actively fooled*. A malicious model can be constructed to be *intentionally biased* (e.g., making decisions based on gender or race) but designed to *pass* LIME and SHAP explanations as "unbiased".[60] This is achieved by having the model detect the unique statistical signature of the perturbed inputs created by LIME/SHAP and route them to a separate, unbiased model, while routing real inputs to the biased one.[60]

Conclusion: A robust framework cannot rely on LIME or SHAP. It requires a new post-hoc

paradigm designed *specifically for generative models*.

3.3 A New Post-Hoc Paradigm: The PXGen Framework

The PXGen framework (Huang et al., 2025) offers a
novel and powerful solution, providing a post-hoc, model-agnostic, example-based XAI framework built for generative models.62 It is perfectly suited to explain a black-box model like TabDiff.
PXGen's methodology is based on probing the trained model using real data and user-defined criteria.

Methodological Components:

1. **The Model:** A pre-trained, black-box generative model (e.g., the user's trained TabDiff model).[64]
2. **The "Anchor" Set:** The core of PXGen. Instead of random perturbations, PXGen uses a set of *real data samples* (e.g., the original training dataset) as "anchors" or reference points.[62] The model is explored by observing how it understands, reconstructs, or represents these real-world examples.
3. **The "Criteria" (Intrinsic & Extrinsic):** The user defines a set of *interpretable functions* or "criteria" that are used to "score" each anchor.[64]
    - **Intrinsic Criteria:** Measure the model's *internal* understanding of an anchor. For a VAE, this could be the Kullback-Leibler (KL) Divergence of its latent representation.[64] For a diffusion model, this could be the total reconstruction error after a full forward-reverse cycle.
    - **Extrinsic Criteria:** Measure *external*, human-understandable properties of the anchor, *independent* of the model. For a tabular dataset, this could be a binary function for "Age > 65" or "Income < 30,000".

The PXGen Process:

1. **Preparation Phase:** The trained model, the anchor set, and the criteria functions are defined and prepared.[67]
2. **Calculation Phase:** Every anchor in the set is processed by the model and scored by *every* criterion. This generates a feature vector for each anchor, mapping it into a new "explainability space".[62]
3. **Discovery Phase:** The user can now explore the model's behavior by querying this explainability space. This provides "example-based explanations" that reveal the model's internal logic.[67]
    - **Example Query 1:** "Show me all anchors with *High Intrinsic Error* and *Extrinsic 'Digit = 0'*." This might reveal **"Model Delusion"**—a phenomenon where the model, despite

being trained on digit '0', consistently fails to correctly reconstruct or understand a specific *sub-group* of '0' anchors (e.g., those with a loop).[67]
    - **Example Query 2:** "Show me anchors for 'Digit 0' that have *Low Intrinsic Error* and are *representationally similar* to 'Digit 1' anchors." This might reveal **"Aligned Conception"**—a phenomenon where the model's internal representations for two distinct concepts have become confusingly similar.[67]

PXGen provides tractable visualization algorithms (like k-dispersion, to find the most *diverse* examples in a group, or k-center, to find the most *typical* ones) to present these findings.[62] This framework directly provides the deep, post-hoc interpretability requested, moving far beyond the superficial and potentially misleading explanations of LIME/SHAP.

Part 4: Actionable Interpretation Strategies for Diffusion Models

While PXGen provides a complete *global framework* for understanding the model, a practitioner can add "flavour" by implementing other, more specific XAI *techniques* designed explicitly for the diffusion process. These strategies can be used as "criteria" within PXGen or as standalone analysis tools.

4.1 Strategy 1: Interpreting the Denoising Trajectory (The "When")

The generative process in diffusion is not one-shot; it is a *sequence* of T denoising steps.[5] This temporal dimension provides a rich surface for explainability: we can analyze *how* the generated sample *evolves* from noise to data.[68]

A 2024 framework by Park et al. for "Explaining generative diffusion models via visual analysis" does exactly this.[68]

- **Key Finding:** The model generates *semantic* elements (e.g., the overall shape and pose of an object) in the *early* (high-noise, high t) denoising steps. It then refines *details* (e.g., fur texture, eye color) in the *late* (low-noise, low t) steps.[69]
- **Proposed Tools:** The paper introduces tools like **DF-RISE** (a saliency method to visualize the semantic/detail levels at each step) and **DF-CAM** (a Class Activation Mapping variant to interpret the specific visual concepts being generated at *each* step t).[69]

Adaptation for Tabular (A Novel "Flavour"):
This image-based technique can be adapted for a tabular model like TabDiff. The "sample" in TabDiff is a vector of normalized numericals and one-hot categoricals.36 The denoising process reverses noise on this vector.
A novel XAI technique would be to track the *entropy* of the predicted categorical distributions (the one-hot vectors) at each step t during the reverse process.

1. At $t=T$, the model's prediction for a categorical feature 'Gender' would be pure noise (high entropy, e.g., $[0.5, 0.5]$).
2. At $t=0$, the prediction must be a firm 1-in-K choice (zero entropy, e.g., $[1.0, 0.0]$).
3. The *step t* at which this entropy *collapses* reveals the model's internal logic.
4. For example, a model might "decide" the 'Zip Code' (entropy collapses at $t=500$) *long before* it "decides" the 'Income' (entropy collapses at $t=50$). This would provide a direct, quantitative measure of a learned dependency: "The model's choice of Income is dependent on its prior choice of Zip Code."

4.2 Strategy 2: Semantic Analysis of the Latent Space (The "What")

This strategy seeks to understand the "concepts" the model has learned by analyzing its *latent space* for semantically meaningful directions.[72] For LDMs, this is often done by analyzing the intermediate feature maps of the U-Net (e.g., self-attention or cross-attention), which can be used to control the generated image.[74]

A primary challenge is that these discovered "directions" often require manual interpretation.[72] A SOTA framework (arXiv:2410.21314) for *unsupervised* analysis of diffusion latent spaces solves this by *directly leveraging natural language prompts and captions* to automatically map and understand latent directions, making the process scalable and interpretable.[72]

This is a powerful method for uncovering latent biases, such as spurious correlations or the under-representation of marginalized identities.[72]

Caveat for Tabular: This strategy must be applied with care. As established in Part 1, tabular models like TabDiff *do not* use the same VAE-based semantic latent space as LDMs. Their "latent space" is simply the hidden-layer activations of an MLP. The applicability of these image-based latent-space analysis techniques to a non-VAE tabular model is an open and valuable research question.

4.3 Strategy 3: Concept-Guided & Counterfactual Explanations (The "What If")

This is perhaps the most powerful post-hoc technique. A counterfactual (CF) explanation answers the question, "What is the *minimal change* to this input that would *flip* the model's decision?".[79]

Methodology: DiME (Diffusion Models for Counterfactual Explanations)
The DiME framework (Jeanneret et al., 2022) cleverly leverages the diffusion process itself to generate the counterfactual explanation.79

- **Process:**
    1. Start with a query image x (e.g., a face classified by a target classifier as "Not Smiling").
    2. Add a small amount of noise by running the *forward process* for τ steps, yielding x_τ.[83]
    3. Run the *reverse (denoising) process* from x_τ back to x_0.
    4. **Crucially:** *Guide* this reverse diffusion process using the gradients of the target classifier.[80] The guidance gradient *pushes* the generation *away* from the original class ("Not Smiling") and *towards* the target class ("Smiling").[80]
- **Result:** DiME produces a new, realistic image that is minimally changed from the original but is now classified as "Smiling." The *difference* between the two images is the counterfactual explanation, visually showing *exactly what* the classifier associates with a "smile."
- **CoLa-DCE (Concept-guided Latent Diffusion CF Explanations):** This is a more recent extension of DiME that operates in the latent space of LDMs and provides more granular, concept-guided control over the generation.[85]

Adaptation for Tabular (A Novel "Flavour"):
This "guided diffusion" method is a highly promising "flavour" for the tabular framework. A novel "TabDiME" system could be built:

1. Train the SOTA **TabDiff** model to generate high-fidelity synthetic patient data.
2. Train a downstream classifier (e.g., a risk model) on this synthetic data to predict "High-Risk of Readmission."
3. To explain this classifier, apply the DiME methodology: Take a real patient vector x classified as "High-Risk." Add noise to τ steps.
4. Run the *guided reverse diffusion* of TabDiff, using the classifier's gradients to push the generation towards the "Low-Risk" class.
5. The result is a new, synthetic patient vector x_{CF} that is minimally different from x but is now classified as "Low-Risk." Comparing x and x_{CF} provides a powerful, actionable, and human-readable explanation (e.g., "This patient would be Low-Risk if

their 'Blood Pressure' was 10 points lower and 'Medication Adherence' was 'High'").

Table 2: Taxonomy of Post-Hoc XAI Techniques for Diffusion Models

Technique	Source/Paper	Core Question	Methodology
Global Framework	**PXGen** [64]	"Why does the model generalize this way?"	Example-Based (Anchors + Criteria)
Trajectory Analysis	Park et al. [69]	"*When* does the model decide?"	Step-wise Saliency (DF-CAM) / Entropy Tracking
Latent Analysis	arXiv:2410.21314 [72]	"*What* concepts does the model know?"	Mapping Latent Directions w/ NLP
Perturbation (CF)	**DiME / CoLa-DCE** [79]	"*What if* the output were different?"	Guided-Diffusion Counterfactuals

Part 5: A Systems-Level Approach: Integration, Tooling, and Validation

A successful framework requires not only theoretical alignment but a practical blueprint for implementation and validation. This involves selecting the right software libraries and, most importantly, defining a rigorous, multi-stage evaluation protocol.

5.1 Technical Implementation: Leveraging Core Libraries

The framework can be assembled using a stack of modular, open-source libraries.

- The Generative Core: diffusers
  The Hugging Face diffusers library is the SOTA toolkit for diffusion models.24 It is

modular and designed for custom pipelines.
    - **DiffusionPipeline:** The primary API for inference.[87]
    - **Models:** Provides building blocks like the VAE and U-Net.[24]
    - Schedulers: Provides interchangeable noise schedulers (e.g., DDPM, DDIM) that control the speed/quality trade-off.87
      A custom model like TabDiff would be implemented by creating a new DiffusionPipeline that uses a custom MLP/Transformer as its "model" component and its custom learnable scheduler. The library supports custom training loops 89 and integration with accelerate for multi-GPU environments.90
- **Domain-Specific Toolkits:**
    - **Tabular:** The rtdl (Research on Tabular Deep Learning) repository [91] is a critical hub, providing links to the official code for tab-ddpm [32] and other SOTA tabular models. The TabDiff [37] repository provides the SOTA implementation.
    - **Time Series:** Repositories for Diffusion-TS [44], SSSD [92], and TSDiff [93] provide complete implementations. The denoising-diffusion-pytorch library also offers a generic GaussianDiffusion1D class suitable for time-series data.[94]
- **Explainability Toolkits:**
    - **Baselines:** The lime [95] and shap [52] libraries should be implemented *first* to establish a baseline and demonstrate their (expected) limitations.
    - **GenXAI:** The GenAISHAP library [96] offers a unique approach. It adapts SHAP not to explain the *generative output*, but to explain *metrics about the output* (e.g., "why is the *faithfulness* score for this generation low?"). It does this by training a surrogate regression model on features of the input prompt [96], providing a "meta-level" explanation. The PXGen framework would need to be implemented from its paper.[64]

5.2 Validation Protocol I: Quantifying Synthetic Data Quality

Before any explanation can be trusted, the quality of the underlying synthetic data generator *must* be rigorously validated. This is a crucial **two-stage validation process**: data quality first, then explanation quality.

For tabular data, evaluation rests on three pillars [98]:

1. **Fidelity (Resemblance):** Does the synthetic data *look like* the real data?
    - **Distributional Metrics:** Compare the marginal distributions of each column (real vs. synthetic) using metrics like Hellinger Distance [101] or Wasserstein Distance.[102]
    - **Correlation Metrics:** Assess the pair-wise column correlation matrix. This is a key metric where TabDiff is known to excel.[27]
2. **Utility:** Can the synthetic data be *used* for real-world tasks?

    - **"Train on Synthetic, Test on Real" (TSTR):** This is the gold standard.[98] A downstream ML model (e.g., a classifier) is trained *only* on synthetic data and then tested on a held-out set of *real* data. Its performance (e.g., F1-score) is compared to a model trained on real data (TRTR).[98]
    - **Domain Classifier Test:** A binary classifier is trained to distinguish between real and synthetic samples. If the data is high-fidelity, the classifier should fail (i.e., achieve an AUC score $\approx$ 0.5).[102]
3. **Privacy:** Does the synthetic data leak private information?
    - **Exact Match Score:** Count the number of synthetic samples that are exact, verbatim copies of records from the training set.[99] This score should be zero.

For time-series data, similar principles apply using domain-specific metrics like Dynamic Time Warping (DTW).[103] For images, standard metrics include Fréchet Inception Distance (FID) [105] and Inception Score (IS).[107]

5.3 Validation Protocol II: Measuring Explanation Quality (GenXAI)

Once Protocol I is passed (i.e., the data is proven to be high-quality), the framework must proceed to Stage 2: validating the XAI outputs. How do we know an explanation is *correct*?

This is a notoriously difficult problem [109], but the field has converged on several key metrics [110]:

1. **Faithfulness:** This is the most critical metric. Does the explanation *accurately* reflect the model's *true* internal reasoning process, or is it just a plausible-sounding "story"?.[62] The PXGen framework, for example, is designed to ensure faithfulness by basing its "intrinsic criteria" on the model's actual internal state.[62]
2. **Robustness (Stability):** If a tiny, irrelevant perturbation is made to the input, the explanation should not change dramatically.[62]
3. **Plausibility / Human-Agreement (HA):** Does the explanation make sense to a human domain expert? This measures the *usefulness* of the explanation.[62]

The following table synthesizes this two-stage validation process into a comprehensive test plan.

Table 3: Comprehensive Validation Metrics Suite

Part A: Synthetic Data Quality	

(Validation Protocol I)	
Pillar	**Metric**
Fidelity (Resemblance)	Hellinger Distance [101] or Wasserstein Distance [102] Pair-wise Correlation Difference [27]
Utility (Usefulness)	Train on Synthetic, Test on Real (TSTR) [98] Domain Classifier AUC [104]
Privacy	Exact Match Score [99]
Part B: Explanation Quality (Validation Protocol II)	
Pillar	**Metric**
Accuracy	Faithfulness [108]
Reliability	Robustness / Stability [111]
Usability	Human-Reasoning Agreement (HA) [111]

Part 6: Strategic Recommendations for a Novel Framework

The synthesis of this analysis yields two concrete, novel architectural proposals. These represent two different, SOTA philosophies for building an explainable generative framework, providing the "flavour" requested by the practitioner.

6.1 Recommended Architecture 1 (Tabular): The "SOTA-Hybrid" Framework

This architecture is recommended for the **tabular domain** and follows a "Separation of Concerns" philosophy. It pairs the highest-performance generative model (a black box) with the most powerful post-hoc XAI framework.

- **Component 1 (Generation): Implement TabDiff**.[36]
  - **Rationale:** As the ICLR 2025 SOTA, its use of *feature-wised learnable noise schedules* [31] makes it uniquely equipped to handle the deep heterogeneity of real-world tabular data, resulting in superior correlation modeling.[27] The diffusers [88] library should be used as the scaffolding for this custom implementation.
- **Component 2 (Explainability): Implement the PXGen Framework**.[64]
  - **Rationale:** It is a 2025 post-hoc framework designed *specifically* for generative models. Its "Anchor" + "Criteria" methodology [62] is the ideal tool for probing the black-box TabDiff model.
- **Framework Integration (The "Flavour"):**
  1. The trained TabDiff model serves as the "Model" for PXGen.
  2. The entire real training dataset serves as the "Anchor Set".[67]
  3. **Intrinsic Criteria** are defined for TabDiff (e.g., "reconstruction MSE" after a full $x_0 \rightarrow x_T \rightarrow x_0$ cycle; "step t of entropy collapse" from Strategy 4.1).
  4. **Extrinsic Criteria** are defined for the domain (e.g., "Age > 65," "Income < 30k").
  5. The PXGen "Discovery Phase" is used to uncover "Model Delusion" (e.g., "TabDiff generates high-error samples for the 'Age > 65' group") or "Aligned Conception" (e.g., "TabDiff's internal representations for 'Zip Code A' and 'Zip Code B' are indistinguishable").

6.2 Recommended Architecture 2 (Time Series): The "Intrinsic-XAI" Framework

This architecture is recommended for the **time-series domain** and follows an "Interpretable-by-Design" philosophy. It uses a model with built-in XAI and supplements it with a powerful post-hoc technique.

- **Component 1 (Generation + Intrinsic XAI): Implement Diffusion-TS**.[44]

    - **Rationale:** It is a SOTA (2024) model that *already includes* a powerful XAI mechanism.[40]
    - **Primary Explanation:** The model's *interpretable decoder* is used at inference time to output not only the synthetic time series but also its constituent *seasonal* (Fourier) and *trend* (polynomial) components.[44] This is the primary, "by-design" explanation.
- **Component 2 (Post-Hoc XAI): Implement DiME-based Counterfactuals**.[79]
    - **Rationale:** To add a sophisticated post-hoc "flavour" that goes beyond the intrinsic XAI.
- **Framework Integration (The "Flavour"):**
    1. The intrinsic decoder is used to explain a generated forecast (e.g., "The predicted energy spike on Monday is 70% trend-driven and 30% seasonal").
    2. A DiME-like guided diffusion [80] is implemented to ask counterfactual questions of a downstream forecaster (e.g., "What is the *minimal change* in the *past week's trend component* that would *prevent* the Monday spike?"). This allows for probing second-order interactions.

6.3 The Research Frontier: Open Challenges

Both proposed frameworks are at the state-of-the-art, but they also highlight open research challenges that are ideal for a practitioner-researcher to investigate.

1. **XAI Faithfulness:** The rigorous, quantitative validation of *faithfulness* remains the single greatest challenge in XAI.[109] Proving that an explanation *is* the model's true reasoning and not just a plausible-sounding artifact is an unsolved problem.
2. **The Tabular "Latent Space":** The theoretical understanding of the "latent space" in non-VAE-based tabular diffusion models (like TabDiff) is completely unexplored. Applying latent-space analysis techniques (Strategy 4.2) to these models and defining *what* that space represents is a novel, high-impact research direction.
3. **Counterfactual Fidelity:** While powerful, guided-diffusion counterfactuals (like "TabDiME") risk creating out-of-distribution (OOD) samples. Ensuring the "minimal change" is also a *plausible* change within the data manifold is a key challenge.

Works cited

1. TabDDPM: Modelling Tabular Data with Diffusion Models - Proceedings of Machine Learning Research, accessed November 8, 2025, https://proceedings.mlr.press/v202/kotelnikov23a/kotelnikov23a.pdf
2. TabDDPM: Modelling Tabular Data with Diffusion Models - arXiv, accessed November 8, 2025, https://arxiv.org/html/2209.15421v2
3. [2107.03006] Structured Denoising Diffusion Models in Discrete State-Spaces -

arXiv, accessed November 8, 2025, https://arxiv.org/abs/2107.03006
4. Denoising Diffusion Probabilistic Models in Six Simple Steps - arXiv, accessed November 8, 2025, https://arxiv.org/pdf/2402.04384
5. Directly Denoising Diffusion Models - arXiv, accessed November 8, 2025, https://arxiv.org/html/2405.13540v2
6. Denoising Diffusion Probabilistic Models, accessed November 8, 2025, https://arxiv.org/abs/2006.11239
7. Biomedical Image Segmentation Using Denoising Diffusion Probabilistic Models: A Comprehensive Review and Analysis - MDPI, accessed November 8, 2025, https://www.mdpi.com/2076-3417/14/2/632
8. InDepth Guide to Denoising Diffusion Probabilistic Models DDPM - LearnOpenCV, accessed November 8, 2025, https://learnopencv.com/denoising-diffusion-probabilistic-models/
9. Generative AI Research Spotlight: Demystifying Diffusion-Based Models - NVIDIA Developer, accessed November 8, 2025, https://developer.nvidia.com/blog/generative-ai-research-spotlight-demystifying-diffusion-based-models/
10. Step by Step visual introduction to Diffusion Models. - Blog by Kemal Erdem, accessed November 8, 2025, https://erdem.pl/2023/11/step-by-step-visual-introduction-to-diffusion-models/
11. Denoising Diffusion Probabilistic Models - arXiv, accessed November 8, 2025, https://arxiv.org/pdf/2006.11239
12. Step by Step visual introduction to Diffusion Models - Medium, accessed November 8, 2025, https://medium.com/@kemalpiro/step-by-step-visual-introduction-to-diffusion-models-235942d2f15c
13. Denoising Diffusion Implicit Models - arXiv, accessed November 8, 2025, https://arxiv.org/pdf/2010.02502
14. High-Resolution Image Synthesis with Latent Diffusion Models - Computer Vision & Learning Group - Ommer-Lab, accessed November 8, 2025, https://ommer-lab.com/research/latent-diffusion-models/
15. CVPR 2022 paper on High-Resolution Image Synthesis with Latent Diffusion Models, accessed November 8, 2025, https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html
16. [PDF] Denoising Diffusion Implicit Models - Semantic Scholar, accessed November 8, 2025, https://www.semanticscholar.org/paper/Denoising-Diffusion-Implicit-Models-Song-Meng/014576b866078524286802b1d0e18628520aa886
17. [2010.02502] Denoising Diffusion Implicit Models - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2010.02502
18. ermongroup/ddim: Denoising Diffusion Implicit Models - GitHub, accessed November 8, 2025, https://github.com/ermongroup/ddim
19. High-Resolution Image Synthesis with Latent Diffusion Models - GitHub, accessed November 8, 2025, https://github.com/CompVis/latent-diffusion

20. [2112.10752] High-Resolution Image Synthesis with Latent Diffusion Models - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2112.10752
21. High-Resolution Image Synthesis With Latent Diffusion Models - CVF Open Access, accessed November 8, 2025, https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf
22. Latent Diffusion Models: A Review — Part I | by Sertis - Medium, accessed November 8, 2025, https://sertiscorp.medium.com/latent-diffusion-models-a-review-part-i-d0feacc4906
23. Denoising Diffusion Models on Model-Based Latent Space - MDPI, accessed November 8, 2025, https://www.mdpi.com/1999-4893/16/11/501
24. Using diffusers at Hugging Face, accessed November 8, 2025, https://huggingface.co/docs/hub/diffusers
25. Diffusion and Denoising: Explaining Text-to-Image Generative AI - Exxact Corporation, accessed November 8, 2025, https://www.exxactcorp.com/blog/deep-learning/diffusion-and-denoising-explaining-text-to-image-generative-ai
26. arXiv:2112.10752v2 [cs.CV] 13 Apr 2022, accessed November 8, 2025, https://arxiv.org/pdf/2112.10752
27. TabDiff: a Multi-Modal Diffusion Model for Tabular Data Generation - arXiv, accessed November 8, 2025, https://arxiv.org/html/2410.20626v1
28. [2504.16506] A Comprehensive Survey of Synthetic Tabular Data Generation - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2504.16506
29. Auto-Regressive Moving Diffusion Models for Time Series Forecasting, accessed November 8, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/33838/35993
30. Autoregressive Denoising Diffusion Models for Multivariate ... - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2101.12072
31. [Quick Review] TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation - Liner, accessed November 8, 2025, https://liner.com/review/tabdiff-a-mixedtype-diffusion-model-for-tabular-data-generation
32. yandex-research/tab-ddpm: [ICML 2023] The official implementation of the paper "TabDDPM: Modelling Tabular Data with Diffusion Models" - GitHub, accessed November 8, 2025, https://github.com/yandex-research/tab-ddpm
33. TabDDPM: modelling tabular data with diffusion models - Yandex Research, accessed November 8, 2025, https://research.yandex.com/blog/tabddpm-modelling-tabular-data-with-diffusion-models
34. arXiv:2209.15421v1 [cs.LG] 30 Sep 2022, accessed November 8, 2025, https://arxiv.org/abs/2209.15421
35. [2410.20626] TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2410.20626
36. TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation - arXiv, accessed November 8, 2025, https://arxiv.org/html/2410.20626v3

37. [ICLR 2025] TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation - GitHub, accessed November 8, 2025, https://github.com/MinkaiXu/TabDiff
38. Diffusion Models for Tabular Data Imputation and Synthetic Data Generation - arXiv, accessed November 8, 2025, https://arxiv.org/html/2407.02549v1
39. Diffusion Models for Tabular Data Imputation and Synthetic Data Generation - OpenReview, accessed November 8, 2025, https://openreview.net/forum?id=wiYV0KDAE6
40. Diffusion-TS: Interpretable Diffusion for General Time Series Generation - OpenReview, accessed November 8, 2025, https://openreview.net/forum?id=4h1apFjO99
41. Addition of VERY ACCURATE time series diffusion models · Nixtla neuralforecast · Discussion #495 - GitHub, accessed November 8, 2025, https://github.com/Nixtla/neuralforecast/discussions/495
42. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting - arXiv, accessed November 8, 2025, https://arxiv.org/pdf/2101.12072
43. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting - Proceedings of Machine Learning Research, accessed November 8, 2025, http://proceedings.mlr.press/v139/rasul21a/rasul21a.pdf
44. Y-debug-sys/Diffusion-TS: [ICLR 2024] Official ... - GitHub, accessed November 8, 2025, https://github.com/Y-debug-sys/Diffusion-TS
45. Survey of Explainable AI Techniques in Healthcare - PMC - NIH, accessed November 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9862413/
46. LLMs for Explainable AI: A Comprehensive Survey - arXiv, accessed November 8, 2025, https://arxiv.org/html/2504.00125v1
47. Generative AI meets Explainable AI - World Conference on Explainable Artificial Intelligence, accessed November 8, 2025, https://xaiworldconference.com/2025/generative-ai-meets-explainable-ai/
48. [2404.09554] Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2404.09554
49. Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda, accessed November 8, 2025, https://www.semanticscholar.org/paper/Explainable-Generative-AI-(GenXAI)%3A-A-Survey%2C-and-Schneider/3ad3e240cabb3f3471770d25a7414a81175aa0db
50. Explainable Generative AI (GenXAI): a survey, conceptualization, and research agenda, accessed November 8, 2025, https://www.researchgate.net/publication/384058214_Explainable_Generative_AI_GenXAI_a_survey_conceptualization_and_research_agenda
51. LIME: Local Interpretable Model-Agnostic Explanations - C3 AI, accessed November 8, 2025, https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/
52. Two minutes NLP — Explain predictions with LIME | by Fabio Chiusano | Generative AI | Medium, accessed November 8, 2025,

https://medium.com/nlplanet/two-minutes-nlp-explain-predictions-with-lime-aec46c7c25a2
53. 14 LIME – Interpretable Machine Learning, accessed November 8, 2025, https://christophm.github.io/interpretable-ml-book/lime.html
54. Welcome to the SHAP documentation — SHAP latest documentation, accessed November 8, 2025, https://shap.readthedocs.io/
55. shap/shap: A game theoretic approach to explain the output of any machine learning model. - GitHub, accessed November 8, 2025, https://github.com/shap/shap
56. Interpretable generative deep learning: an illustration with single cell gene expression data, accessed November 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9360114/
57. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models - arXiv, accessed November 8, 2025, https://arxiv.org/pdf/2103.04922
58. [1812.05676] A Probe Towards Understanding GAN and VAE Models - arXiv, accessed November 8, 2025, https://arxiv.org/abs/1812.05676
59. [2103.04922] Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2103.04922
60. Bias Goes Undercover Adversarial attacks can fool explainable AI techniques., accessed November 8, 2025, https://www.deeplearning.ai/the-batch/bias-goes-undercover/
61. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME - arXiv, accessed November 8, 2025, https://arxiv.org/html/2305.02012v3
62. PXGen: A Post-hoc Explainable Method for Generative Models - arXiv, accessed November 8, 2025, https://arxiv.org/html/2501.11827v1
63. [2501.11827] PXGen: A Post-hoc Explainable Method for Generative Models - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2501.11827
64. [Revisión de artículo] PXGen: A Post-hoc Explainable Method for, accessed November 8, 2025, https://www.themoonlight.io/es/review/pxgen-a-post-hoc-explainable-method-for-generative-models
65. PXGen: A Post-hoc Explainable Method for Generative Models - ChatPaper, accessed November 8, 2025, https://chatpaper.com/paper/101273
66. accessed November 8, 2025, https://arxiv.org/abs/2501.11827#:~:text=In%20this%20work%2C%20we%20propose,to%20their%20purpose%20and%20requirements.
67. PXGen: A Post-hoc Explainable Method for Generative Models - ResearchGate, accessed November 8, 2025, https://www.researchgate.net/publication/388318025_PXGen_A_Post-hoc_Explainable_Method_for_Generative_Models
68. [2402.10404] Explaining generative diffusion models via visual analysis for interpretable decision-making process - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2402.10404

69. Explaining Generative Diffusion Models via Visual Analysis for Interpretable Decision-Making Process DOI: https://www.sciencedirect.com/science/article/pii/S0957417424000964 - arXiv, accessed November 8, 2025, https://arxiv.org/html/2402.10404v1
70. Explaining generative diffusion models via visual analysis for interpretable decision-making process - Semantic Scholar, accessed November 8, 2025, https://www.semanticscholar.org/paper/Explaining-generative-diffusion-models-via-visual-Park-Ju/5090e3d794430344ea17872c652661c15bfefeba
71. Explaining generative diffusion models via visual analysis for interpretable decision-making process | Request PDF - ResearchGate, accessed November 8, 2025, https://www.researchgate.net/publication/377733836_Explaining_generative_diffusion_models_via_visual_analysis_for_interpretable_decision-making_process
72. Decoding Diffusion: A Scalable Framework for Unsupervised Analysis of Latent Space Biases and Representations Using Natural Language Prompts | OpenReview, accessed November 8, 2025, https://openreview.net/forum?id=EALy79qg9I¬eId=XB1RT92sRP
73. [2307.12868] Understanding the Latent Space of Diffusion Models through the Lens of Riemannian Geometry - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2307.12868
74. Understanding the Latent Space of Diffusion Models through the Lens of Riemannian Geometry - NIPS papers, accessed November 8, 2025, https://papers.neurips.cc/paper_files/paper/2023/file/4bfcebedf7a2967c410b64670f27f904-Paper-Conference.pdf
75. (PDF) Decoding Diffusion: A Scalable Framework for Unsupervised Analysis of Latent Space Biases and Representations Using Natural Language Prompts - ResearchGate, accessed November 8, 2025, https://www.researchgate.net/publication/385354319_Decoding_Diffusion_A_Scalable_Framework_for_Unsupervised_Analysis_of_Latent_Space_Biases_and_Representations_Using_Natural_Language_Prompts
76. [2410.21314] Decoding Diffusion: A Scalable Framework for Unsupervised Analysis of Latent Space Biases and Representations Using Natural Language Prompts - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2410.21314
77. Decoding Diffusion: A Scalable Framework for Unsupervised Analysis of Latent Space Biases and Representations Using Natural Language Prompts - arXiv, accessed November 8, 2025, https://arxiv.org/html/2410.21314v1
78. Stable Bias: Evaluating Societal Representations in Diffusion Models - NIPS papers, accessed November 8, 2025, https://papers.neurips.cc/paper_files/paper/2023/file/b01153e7112b347d8ed54f317840d8af-Paper-Datasets_and_Benchmarks.pdf
79. Diffusion Models for Counterfactual Explanations - CVF Open Access, accessed November 8, 2025, https://openaccess.thecvf.com/content/ACCV2022/papers/Jeanneret_Diffusion_Models_for_Counterfactual_Explanations_ACCV_2022_paper.pdf
80. [2203.15636] Diffusion Models for Counterfactual Explanations - arXiv, accessed

November 8, 2025, https://arxiv.org/abs/2203.15636
81. From Visual Explanations to Counterfactual Explanations with Latent Diffusion - CVF Open Access, accessed November 8, 2025, https://openaccess.thecvf.com/content/WACV2025/papers/Luu_From_Visual_Explanations_to_Counterfactual_Explanations_with_Latent_Diffusion_WACV_2025_paper.pdf
82. [PDF] Diffusion Models for Counterfactual Explanations - Semantic Scholar, accessed November 8, 2025, https://www.semanticscholar.org/paper/Diffusion-Models-for-Counterfactual-Explanations-Jeanneret-Simon/3be678c7d63e66ae5cd7d62a598cbb8f0935fe55
83. (PDF) Diffusion Models for Counterfactual Explanations - ResearchGate, accessed November 8, 2025, https://www.researchgate.net/publication/359574835_Diffusion_Models_for_Counterfactual_Explanations
84. Counterfactual explanations for the Smile. We visualize DiME and DiVE... - ResearchGate, accessed November 8, 2025, https://www.researchgate.net/figure/Counterfactual-explanations-for-the-Smile-We-visualize-DiME-and-DiVE-explanations_fig4_359574835
85. CoLa-DCE – Concept-guided Latent Diffusion Counterfactual ..., accessed November 8, 2025, https://openreview.net/forum?id=IQ0BBfbYR2
86. Discovering Concept Directions from Diffusion-based Counterfactuals via Latent Clustering, accessed November 8, 2025, https://arxiv.org/html/2505.07073v1
87. Diffusers - Hugging Face, accessed November 8, 2025, https://huggingface.co/docs/diffusers/index
88. State-of-the-art diffusion models for image, video, and audio generation in PyTorch. - GitHub, accessed November 8, 2025, https://github.com/huggingface/diffusers
89. Introduction to Diffusers - Hugging Face Diffusion Course, accessed November 8, 2025, https://huggingface.co/learn/diffusion-course/unit1/2
90. Documentation - Hugging Face, accessed November 8, 2025, https://huggingface.co/docs
91. yandex-research/rtdl: Research on Tabular Deep Learning ... - GitHub, accessed November 8, 2025, https://github.com/yandex-research/rtdl
92. AI4HealthUOL/SSSD: Repository for the paper: 'Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models' - GitHub, accessed November 8, 2025, https://github.com/AI4HealthUOL/SSSD
93. GitHub - amazon-science/unconditional-time-series-diffusion: Official PyTorch implementation of TSDiff models presented in the NeurIPS 2023 paper "Predict, Refine, Synthesize, accessed November 8, 2025, https://github.com/amazon-science/unconditional-time-series-diffusion
94. Implementation of Denoising Diffusion Probabilistic Model in Pytorch - GitHub, accessed November 8, 2025, https://github.com/lucidrains/denoising-diffusion-pytorch
95. marcotcr/lime: Lime: Explaining the predictions of any machine learning classifier - GitHub, accessed November 8, 2025, https://github.com/marcotcr/lime

96. microsoft/dstoolkit-genai-shap: SHAP (SHapley Additive ... - GitHub, accessed November 8, 2025, https://github.com/microsoft/dstoolkit-genai-shap
97. steven2358/awesome-generative-ai: A curated list of modern Generative Artificial Intelligence projects and services - GitHub, accessed November 8, 2025, https://github.com/steven2358/awesome-generative-ai
98. Synthetic Tabular Data Evaluation in the Health Domain Covering Resemblance, Utility, and Privacy Dimensions - NIH, accessed November 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10306449/
99. How to evaluate the quality of the synthetic data – measuring from the perspective of fidelity, utility, and privacy | Artificial Intelligence - Amazon AWS, accessed November 8, 2025, https://aws.amazon.com/blogs/machine-learning/how-to-evaluate-the-quality-of-the-synthetic-data-measuring-from-the-perspective-of-fidelity-utility-and-privacy/
100. Structured Evaluation of Synthetic Tabular Data - arXiv, accessed November 8, 2025, https://arxiv.org/html/2403.10424v1
101. Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees - PMC - NIH, accessed November 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12058740/
102. An evaluation framework for synthetic data generation models - arXiv, accessed November 8, 2025, https://arxiv.org/html/2404.08866v1
103. A Comprehensive Evaluation Framework for Synthetic Time Series Data - DiVA portal, accessed November 8, 2025, http://www.diva-portal.org/smash/get/diva2:1954330/FULLTEXT01.pdf
104. Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study, accessed November 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9030990/
105. Fréchet inception distance - Wikipedia, accessed November 8, 2025, https://en.wikipedia.org/wiki/Fr%C3%A9chet_inception_distance
106. How to Implement the Frechet Inception Distance (FID) for Evaluating GANs, accessed November 8, 2025, https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch/
107. Inception score(IS) and Fréchet inception distance(FID) explained | by Niharika Ahuja, accessed November 8, 2025, https://ahujaniharika95.medium.com/inception-score-is-and-fr%C3%A9chet-inception-distance-fid-explained-2bc28a4faea7
108. An Essential Guide for Generative Models Evaluation Metrics - Towards AI, accessed November 8, 2025, https://pub.towardsai.net/an-essential-guide-for-generative-models-evaluation-metrics-255b42007bdd
109. [2207.14160] Do We Need Another Explainable AI Method? Toward Unifying Post-hoc XAI Evaluation Methods into an Interactive and Multi-dimensional Benchmark - arXiv, accessed November 8, 2025, https://arxiv.org/abs/2207.14160

110. Evaluation Metrics Research for Explainable Artificial Intelligence Global Methods Using Synthetic Data - MDPI, accessed November 8, 2025, https://www.mdpi.com/2571-5577/6/1/26
111. A Unified Framework with Novel Metrics for Evaluating the Effectiveness of XAI Techniques in LLMs - arXiv, accessed November 8, 2025, https://arxiv.org/html/2503.05050v2
112. M4: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities, and Models - Article - Faculty & Research, accessed November 8, 2025, https://www.hbs.edu/faculty/Pages/item.aspx?num=65337
113. M4: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities and Models, accessed November 8, 2025, https://proceedings.neurips.cc/paper_files/paper/2023/file/05957c194f4c77ac9d91e1374d2def6b-Paper-Datasets_and_Benchmarks.pdf
114. Finding the Right XAI Method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science in - AMS Journals, accessed November 8, 2025, https://journals.ametsoc.org/view/journals/aies/3/3/AIES-D-23-0074.1.xml
115. F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI - arXiv, accessed November 8, 2025, https://arxiv.org/html/2410.02970v2
116. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering | Transactions of the Association for Computational Linguistics - MIT Press Direct, accessed November 8, 2025, https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00667/121196/Evaluating-Correctness-and-Faithfulness-of