Methodologies for Diffusion Model Interpretability: A Systematic Review

Tina Lakhani, Jing Wu, *Member, IEEE*, Yu-Kun Lai, *Senior Member, IEEE*, and Ze Ji, *Member, IEEE*

***Abstract*—Diffusion generative models have gained rapid traction since 2020 due to their expressiveness and high-quality outputs. Explaining and interpreting these models is essential for enabling further improvements and fostering trustworthiness. This systematic review identifies and analyzes interpretability methods applied to diffusion models across domains, highlighting key trends, outlining strategies, and identifying emerging research directions. We screened 1,489 papers published between 2020–2025 across IEEE, Scopus, DBLP, arXiv, and Elicit, and included 81 studies that met predefined criteria. Most methods target latent space analysis *(n = 35)*, followed by data attribution *(n = 16)* and denoising dynamics *(n = 14)*. Image generation and text-to-image synthesis dominate application areas *(n = 73)*, with limited coverage in robotics, audio, and neuroscience *(n = 8)*. This review offers a structured taxonomy, quantifies interpretability research trends, and identifies domain-specific and architectural gaps. Supplementary material and processing code are available *here*.**

***Impact Statement*—Diffusion models demonstrate impressive performance across a range of applications, yet their deployment in safety-critical and high-stakes areas remains limited without the trustworthiness that deeper interpretability provides. This systematic review establishes the first cross-domain knowledge base for interpretability in diffusion models by consolidating scattered findings across all application areas into a coherent synthesis. It illuminates foundational trends and converging insights from diverse perspectives, identifies methodological links, and compares techniques and target areas, therefore equipping researchers with the knowledge needed to refine their approaches and transfer successful strategies across domains. The findings of this review creates bridges between disciplines, promoting technical innovation toward informed decision-making and regulatory compliance, and supporting the development of reliable, transparent, and accountable AI systems based on diffusion models.**

***Index Terms*—Diffusion model, interpretability, explainability, trustworthy AI.**

I. Introduction

DIFFUSION models have rapidly emerged as a transformative tool in data generation since their early inception and refinement [1], [2]. Their strength lies in their capacity to learn complex, multi-modal feature spaces and relationships, yielding high-fidelity and novel outputs. Their applications span diverse disciplines, from image synthesis [3] and time-series forecasting [4], to robotic action planning [5], showcasing versatility and power.

However, with the expansion of research avenues into diffusion model performance, so too comes the challenge of understanding and interpreting them. Unlike traditional generative models such as Variational Autoencoders (VAEs) [6] or Generative Adversarial Networks (GANs) [7], diffusion models rely on a sequential denoising process from inputs perturbed entirely with Gaussian noise [1]. Subsequently, their internal decision-making requires a targeted research approach addressing their characteristic denoising dynamics.

To our knowledge, no prior survey provides a comprehensive and systematic review of interpretability methods across the full range of diffusion model applications and which spans analyses of training dynamics, latent structure, data attribution, and architectural conditioning. A related survey is by Lin et al. [8], who offer valuable insights into mechanistic interpretability in multi-modal foundation models, including text-to-image diffusion. Their focus is primarily on neuron-level circuit tracing adapted from language models. In contrast, our work is the first to concentrate on diffusion model interpretability across all domains and relevant methodologies.

A. Research Questions

This review is structured to provide researchers with a clear overview of the current landscape of diffusion model interpretability and is guided by the following research questions:

1) *What interpretability methodologies have been proposed for diffusion generative models across different domains?*
2) *How do these approaches target different aspects of interpretability, such as internal representations, training dynamics, and output-level analyses?*
3) *What methodological patterns, gaps, and future directions can be identified in current research on diffusion model interpretability?*

Tina Lakhani, Jing Wu and Yu-Kun Lai are with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, U.K. (email: lakhaniif@cardiff.ac.uk).

Ze Ji is with the School of Engineering, Cardiff University, Cardiff CF24 3AA, U.K.

B. Paper Structure

The main body of the paper is organised into the following main sections:

Section II: Provides the preliminaries for Diffusion Models, and introduces discussion on interpretability; providing common taxonomies, distinctions and boundaries, including those captured within this work.

Section III: Describes the reproducible process of paper selection for the content of the survey. This section outlines the inclusion/exclusion criteria and methodology for literature selection.

Section IV: Contains the material survey of the literature, providing details resulting from the searches and a narrative overview of the included works.

section V: Provides discussion relating back to the original research questions and identifies future directions that can be explored based on the findings from the full review.

II. Diffusion Model Preliminaries

Diffusion models are generative models that learn to reconstruct data from pure noise. This is achieved by approximating the reverse of a Markovian forward process, which incrementally adds Gaussian noise to data, then iteratively denoises a sample from a prior using the learned reverse process. Originally inspired by non-equilibrium thermodynamics [1], diffusion models have since become foundational in modern generative AI, particularly in high-fidelity image synthesis.

This section outlines the forward and reverse processes of diffusion models, describes the typical training objective, and introduces the widely adopted U-Net architecture used to implement the denoising function and the dominant architecture found in articles for this review.

A. Diffusion Process Overview

Forward Process: Let $\mathbf{x}_0$ be a data sample from the real distribution $q(\mathbf{x}_0)$. The forward process gradually adds Gaussian noise to $\mathbf{x}_0$ over T timesteps, producing a sequence $\mathbf{x}_1, \dots, \mathbf{x}_T$ that converges toward pure noise. Formally, the transition at each step is given by:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\big), \tag{1}$$

where $\beta_t \in (0,1)$ is a fixed variance schedule controlling the noise intensity. Using a closed-form expression, $\mathbf{x}_t$ can be sampled directly from the original data:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\big), \tag{2}$$

where α is the ratio of signal retained at each step, i.e. (1 - β_t) and $\bar{\alpha}_t$ = total signal retention after t noise steps ($\prod_{s=1}^{t} \alpha_s$). This process defines a smooth trajectory from data space into noise space.

Training Objective: To reverse the forward noising process, a neural network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ is trained to predict the noise component $\boldsymbol{\epsilon}$ that was added to $\mathbf{x}_0$ to yield $\mathbf{x}_t$. The widely used loss function is:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right] \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian noise, and t denotes the diffusion timestep.
This trains the model to recover the full noise vector, effectively learning to denoise progressively corrupted samples at arbitrary timesteps. Some variations instead predict the clean image $\mathbf{x}_0$, but both formulations are equivalent under suitable reparameterizations.

Reverse Process and Sampling: At inference, the model generates new samples by starting from Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and applying the learned reverse denoising steps iteratively. A typical sampling update is:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \boldsymbol{\eta},$$
$$\text{where} \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

This equation shows how the model's output governs the denoising trajectory, progressively refining a noisy input into a coherent sample.

B. Network Architecture and Feature Spaces

The denoising network in diffusion models, typically denoted $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, is most often implemented as a time-conditional U-Net [8]. This architecture, originally developed for biomedical image segmentation, is well-suited to diffusion due to its encoder-decoder structure with skip connections, enabling multi-scale feature extraction and preserving high-frequency details.

Fig. 1 illustrates a representative diffusion U-Net used in either a Denoising Diffusion Probabilistic Model (DDPM) or Latent Diffusion Model (LDM), highlighting structural components and feature stages frequently analyzed in interpretability studies. At each timestep t, the network inputs a noisy image $\mathbf{x}_t$ and a learned embedding of the timestep, and outputs either a noise estimate $\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t)$ which is the same shape as $\mathbf{x}_t$, or an estimate of the original clean sample $\hat{\mathbf{x}}_0$. This prediction is then used in the sampling equation (equation 4), to compute $\mathbf{x}_{t-1}$. Conditioning embedding $\mathbf{c}$ can be injected into the U-Net via cross-attention or feature modulation, and may represent class labels (e.g., in class-conditional DDPMs), text prompts (as in LDM), or reference images. Skip connections facilitate gradient flow and retainment of fine details.

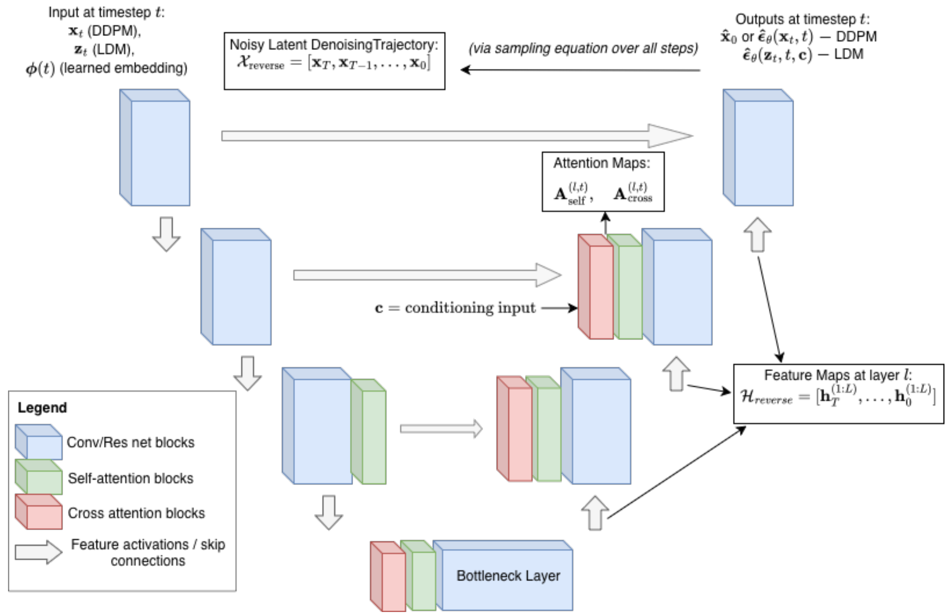

Fig. 1. Diffusion U-Net schematic for DDPM and LDM model networks. Highlighted are common areas for interpretability focus as included in this paper. Note that specific U-Net architectures vary depending on model design and conditioning method.

Modern U-Net variants commonly incorporate attention mechanisms [9], especially near the bottleneck. Self-attention captures long-range spatial dependencies, while cross-attention, crucial in text-to-image models such as Guided Language to Image Diffusion for Generation and Editing (GLIDE) [10] and Latent Diffusion Models [11], enables alignment with conditioning inputs (e.g., text prompts). These attention layers play a central role in semantic progression and have been the focus of many interpretability studies reviewed in this work.

The bottleneck layer, located at the junction of the encoder and decoder, represents the most compressed and abstract state of the image. This region has been found to be a key target in interpretability research and is commonly probed to reveal how visual concepts emerge, transform, and propagate across timesteps.

C. Clarifying Latent Space Terminology

The term *latent space* is used inconsistently across the diffusion model interpretability literature, often referring to different stages or representations in the generative process. For clarity in this review, we distinguish three primary representational spaces, each with a distinct role in diffusion models and the interpretability techniques applied to them:

- $\mathcal{X}$**-space** refers to the sequence of noisy image-like tensors in models such as DDPMs with U-Net backbones. These high-dimensional tensors represent the data as it evolves under the forward noising and the learned reverse denoising process. While the forward process $\mathbf{x}_0 \to \mathbf{x}_1 \to \cdots \to \mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is used during training, interpretability methods typically focus on the reverse generation path $\mathbf{x}_T \to \mathbf{x}_{T-1} \to \cdots \to \mathbf{x}_0$ This reverse trajectory, which we denote $\mathcal{X}_{\text{reverse}} = \{\mathbf{x}_t\}_{t=T}^{0}$, reflects the stepwise reconstruction of semantic structure from noise.
- $\mathcal{Z}$**-space** denotes the lower-dimensional latent variables used in LDMs and can be thought of as the latent diffusion analogy to $\mathbf{x_t}$. These are obtained by encoding pixel-space data into compressed representations $\mathbf{z_0}$ using an external encoder. The diffusion process then occurs within this latent space as $\mathbf{z}_T \to \mathbf{z}_{T-1} \to \cdots \to \mathbf{z}_0$, which is subsequently decoded back to image space. $\mathcal{Z}$-space is thus the domain in which noise is added and removed in LDMs, e.g., [12]–[14].
- $\mathcal{H}$**-space** comprises the internal feature activations computed layer-wise by the U-Net at each timestep [15]. These are distinct from the inputs and outputs of the diffusion trajectory, in that they are the intermediate compressed representations reflecting the model's internal transformations during denoising. Typically, those of the decoding side of the U-Net are used for interpretability analysis where $\mathcal{H}_{reverse} = \{\mathbf{h}_t^{(\ell)}\}_{t=T}^{0}$ where ℓ indexes layers, although bottleneck features are often the most semantically informative. Many interpretability methods, such as probing or attribution techniques, target

this space to understand how the model encodes and manipulates semantic information, e.g., [15]–[17].

This disambiguation is critical for interpretability research, as various methods target distinct components of the model to yield different insights despite often relying on similar terminology. In summary, analyses of temporal semantics typically focus on $\mathcal{X}$- or $\mathcal{Z}$-space, while investigations into internal semantics target $\mathcal{H}$-space and we provide these distinctions in the full survey narrative (Section IV-A), tabular results (Tables IV, VI, V & VIII) & taxonomy tree (Fig. 5).

D. *Diffusion Model Variants*

Diffusion models have evolved rapidly since their original formulation as denoising score-matching processes [2], with advances targeting training stability, sampling speed, architectural flexibility, and cross-domain generalization.

Early improvements focused on training dynamics and baseline architectures. DDPMs [2] employ U-Nets to iteratively denoise Gaussian noise into data and extensions such as DDPM++ [18] and iDDPM [19] introduce enhanced noise schedules, normalization, and classifier-free guidance (CFG). CFG is a sampling technique that combines predictions from conditional and unconditional models to improve fidelity and reduce over-reliance on conditioning signals. The Ablated Diffusion Model framework [18] (ADM) further demonstrates that strong sample quality can be achieved even with simplified training objectives and architectures. For example, it demonstrates that the use of more attention heads with fewer channels per head can improve Fréchet Inception Distance (FID) [20], highlighting the sensitivity of performance to transformer-like design factors.

Accelerated sampling is a progression achieved with the use of Denoising Diffusion Implicit Models (DDIM) [21], which modify the forward process into a non-Markovian schedule, allowing for fewer sampling steps with minimal quality degradation. LDMs [11] further reduce computational cost by operating in compressed latent spaces via autoencoding. This framework underpins models like Stable Diffusion [22]–[24]. Recent Latent Consistency Models (LCMs) [25], [26] build on this by learning one-step mappings for fast inference.

Conditional and multi-modal diffusion has also seen major innovation. GLIDE [10] introduces classifier-free guidance and cross-attention for aligning image generation with text prompts, paving the way for text-to-image models like DALL·E [27] and Stable Diffusion [11]. Domain-specific variants extend diffusion to other modalities, including speech synthesis (Grad-TTS [28]) and voice conversion (DiffWaveNetSVC [29], [30]).

Architecturally, some models replace convolutional backbones entirely. Diffusion Transformers (DiTs) [31] use Vision Transformers [32] to capture global context. At the time of writing, the vast majority of models investigated for interpretability relate to the U-Net, but Transformer architecture is a growing area of research interest, especially since Stable Diffusion moved from U-Net to a Transformer framework in V.3 [24].

Finally, score-based generative models [33] adopt a continuous-time perspective, learning the gradient of the data log-density and sampling via reverse-time SDEs (Statistical Differential Equations) or ODEs (Ordinary Differential Equations). Though formally distinct, they are closely related to DDPMs. Other directions include energy-based formulations [34]–[36] and generalized conditional models [27], which expand the design space beyond standard denoising objectives.

Together, these variants reflect trade-offs in fidelity, efficiency, and domain alignment, shaping the versatility of diffusion-based generation.

E. *Interpretability of Diffusion Models*

Interpretability in machine learning lacks universally rigid definitions, especially across various domains [37], but common themes exist. We adopt the broad taxonomies of [38], [39], where *interpretability* refers to human-understandable insights into a model's internal mechanisms, while *explainability* denotes post-hoc analyses that clarify outputs without requiring transparency. Following [39], we also distinguish between *global* interpretability, understanding model-wide patterns and *local* interpretability, which focuses on specific decisions. These perspectives are often complementary; we include work where aggregating local explanations yields broader behavioral insights.

Diffusion models pose unique interpretability challenges unseen in other generative models stemming from their iterative and stochastic nature; instead of generating outputs in a single pass, denoising occurs over tens to many hundreds of steps. This raises questions about when and where semantic structure emerges, how internal representations evolve, and which components (e.g., noise schedules or architectures), influence specific behaviors. Interpreting diffusion models therefore requires not only traditional machine learning interpretability tools, but also methods to understand this temporal evolution. Dombrowski *et al.* [40] demonstrate that fine-tuning for improved performance can reduce attribution fidelity, highlighting sensitivity to training dynamics and optimization choices. Thus, interpretability can emerge either as an explicit design goal or as a byproduct of architectural modifications. Cold Diffusion [41], for instance, replaces stochastic noise with deterministic, semantically meaningful corruptions (e.g., blurring, masking, downsampling), making each generation step interpretable. By aligning the generative trajectory with human-understandable transformations, such models support tasks like data attribution. For example, Popov & Tuba [42] introduce a noise variant schedule tailored specifically to provide attribution and citation tracing in synthetic content.

In summary, this review collates the diverse range of approaches aimed at interpreting and explaining diffusion models across domains. By focusing on techniques that probe the internal structures, training dynamics, and generative processes of these models, we aim to scope research that advances our understanding of how diffusion models operate beneath the surface. We also include relevant works on explainability that offer holistic insights into model behavior and works that promote intrinsic transparency. In doing so, we recognize interpretability as a multifaceted construct, comprising insights that arise through probing, modifying, or abstracting the model's internal workings

III. Methodology for Paper Survey

This review follows guidelines for systematic literature reviews in software engineering by Kitchenham *et al*. [43], and is also influenced by PRISMA [44]. A structured protocol based on the Joanna Briggs Institute (JBI) Evidence Synthesis guidelines [45] was followed, which can be linked to as supplementary material. Please see Appendix A for more details.

A. Source Search Strategy

The databases chosen to enable the most comprehensive capture of relevant works were: IEEE [46], Scopus [47], ArXiv [48] and dblp [49], covering a major publisher (IEEE) for relevant papers, commonly used indexing databases (Scopus and dblp), and widely used preprint platform. Although papers on ArXiv may not have yet gone through a rigorous peer review process, it is an increasingly common practice for paper preprints to be posted on ArXiv. It is necessary to include these papers to ensure the most up-to-date research is included, although they should be considered with extra care. In addition, the AI research assistant Elicit [50] was used to identify studies potentially missed by database searches and citation chaining. Included works (January 2020–March 2025, English) reflect a predominance of research since seminal U-Net DDPMs in 2020. Search strings, developed *a priori* with the subject librarian, balanced coverage and specificity, and were kept as uniform as possible across databases, adapting to each platform's terminology. Searches occurred 20–28 March 2025. Exact search strings appear in Appendix B.

B. Sources of Evidence

Following database searches (yielding 1489 records), initial de-duplication was performed using a script developed in Python (v.3.10) [51]. This involved a 3-stage process:

1) Removal of any records not relating to papers (e.g., populated with conference details).
2) Removal of records where both DOI and Title were identical, and according to the priority order specified below.
3) Application of fuzzy matching on title and authors with a threshold of 0.80. These were manually examined for any errors (two found and those records kept).

The priority order decided upon for database sources was: **IEEE** > **Scopus** > **arXiv** > **dblp** > **Elicit**

After de-duplication, we began a two-stage screening process. The first stage involved examining title, abstract and keywords of each paper to ascertain if the focus of the work is likely related to our core inclusion criteria of unveiling the internal decision-making mechanisms or behaviors of diffusion models. Works were excluded at this first stage if they were clearly not describing diffusion models (but rather, physics or social diffusion processes, for example), or if they were not pertaining to interpretability. After this point, the second stage screening involved full text evaluation against the inclusion criteria of Table I, with data for the charting table being extracted. Reviews were performed primarily by the first author, TL, with additional discussions including the co-authors. Edge-case citations were discussed as a consensus with a minimum of 2 people, and in addition, the primary author performed random re-checking of paper samples from the results of the second stage screening to ensure consistency with decisions and data extraction.

The full list of papers from the second stage rejection is found in Appendix C. The list of papers included through citation chaining and their source is found in Appendix D.

C. Data Extraction, Charting, and Quality Assessment

1) Data Extraction: The included works were examined for their contributions to the interpretability of diffusion models, with attention paid to where in the pipeline each method is applied and the specific problem it aims to address. The following information was extracted for each paper: authorship and year, focus area of interpretability, core methodological approach, target domain, and the architecture that was used. An example of the data charting format is available in the protocol, accessible via Appendix A.

TABLE I
INCLUSION AND EXCLUSION CRITERIA

Criteria	Include	Exclude
Scope & Focus	Internal mechanisms, training dynamics, holistic outputs.	Diffusion as a black box.
Evaluation & Analysis	Empirical evaluation of instantiated models.	Purely theoretical or highly simplified models.
Domain & Context	Any domain and discipline.	Papers lacking interpretability focus.
Model-Agnostic Analysis	Empirical application on diffusion models.	Tools not evaluated on diffusion models.
Language	English-language papers.	Non-English full-text.
Publish Status	Peer-reviewed or pre-prints.	Rejected works still unpublished at the time of writing.
Date Range	Jan 2020 – May 2025.	Outside the specified date range.

2) Quality Assessment: In alignment with guidelines for systematic reviews in software engineering [52], a separate quality appraisal was conducted on all included studies. This assessment is distinct from the inclusion/exclusion criteria, and was used to evaluate the methodological rigor and reporting clarity of each study. Example charting is found in Appendix A.

Each study was assessed according to the following predefined criteria:

- **Q1. Clarity of Objectives**: Does the study clearly state its objectives or research questions related to diffusion model interpretability?
- **Q2. Evaluation Methods**: Is the experimental design (e.g., dataset, metrics, baselines) well described?
- **Q3. Interpretability Evidence**: Does the study provide empirical evidence (e.g., visualizations, metrics) to support interpretability claims?
- **Q4. Reproducibility**: Are code, datasets, or implementation details made available for further study/evaluation?
- **Q5. Analysis of Limitations**: Does the study reflect on limitations or scope of its interpretability findings?

Each was scored as **1 (yes)**, **0.5 (partial)**, or **0 (no)**. Scores were used to support structured analysis and discussion of methodological strengths and gaps across the field.

3) Data Charting: To provide ease of assimilation, the extracted data are presented in various complementary formats, each offering a distinct level of detail and perspective. These include graphical representations of major trends in domain and architecture, a breakdown of the major themes, a taxonomy tree, and a description of the collected works, organized by methodological approach and highlighting notable and seminal contributions. All can be found in Section IV.

We then evaluate opportunities and gaps in section V, and revisit the research questions proposed in the introduction (Section I). Finally, summary tables present an at-a-glance view of the extracted data, including paper links and key information.

D. Use of Elicit for Paper Search

As previously mentioned, the AI research tool Elicit [50] was used to support the paper search process. The prompt used was: *"Papers relating to methods for interpreting diffusion models, since 2020"*. The first 168 results were included in the screening process. Papers sourced via Elicit were assigned the lowest priority during duplicate processing, which means that any papers identified from Elicit would not have been retrieved through any database searches. This resulted in 9 papers and upon inspection of their metadata, it was found that all 9 did include the primary keyword "diffusion", but their absence from the traditional databases stemmed from the use of the words related to interpretability:

- 6 papers used "interpret*" rather than "interpretab*"
- 1 paper used "explain*" rather than "explainab*"
- 2 papers used synonyms such as "analytical", "evaluating", or "diagnosing", which were not used in the search strings due to their broadness.

As an experiment, we modified the original search terms to reflect a broader truncation to see the difference in output.

TABLE II
KEYWORD SEARCH RESULTS FROM IEEE XPLORE AND ARXIV

Repository	Search Term	Results
IEEE Xplore	`interpretab*`	339
	`interpret*`	2,472
arXiv	`explainab*`	537
	`explain*`	2,028

These comparisons highlight the trade off between precision and recall in keyword-based searches. While using narrower terms (e.g., "interpretab*") improves precision, it may exclude relevant works that use synonyms more loosely. In this context, particularly noting the encapsulating broadness of the search terms, Elicit served as a useful complement to the traditional searches by capturing semantically relevant but terminologically more diverse papers, and reduced the need for an excessively wide and resource-intensive screening process.

E. Data Selection Process Flow

In summary, from **1489** records returned from database searches, 64 met the full inclusion criteria, with 17 additional works found through citation chaining, resulting in **81** works in total included in this review. The full process flow depicting numbers and reasons for paper rejection (or addition) at each stage is shown in the PRISMA flowchart (Fig. 2).

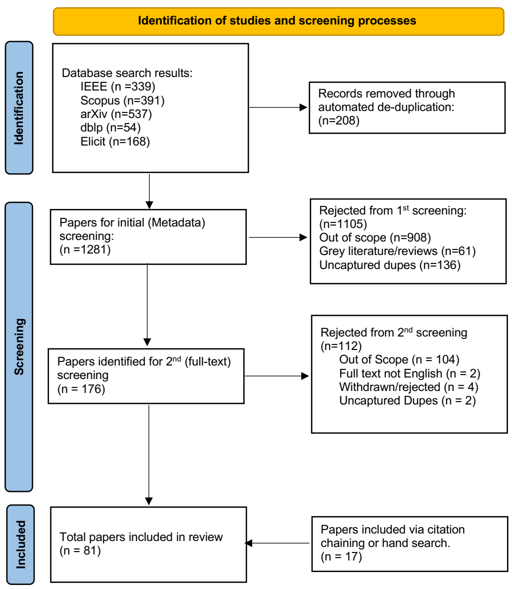

Fig. 2. PRISMA based flowchart showing paper selection process

IV. Results & Findings

In this section, we begin by presenting high-level summaries of the included papers collated from the charting tables, noting that image generation and text-to-image tasks dominated the domain areas (Fig. 3), the U-Net was the most prevalent architecture across domains (Fig. 4), and latent space semantics presented the most common study focus (Table III). We then provide a descriptive narrative, summarizing themes, notable works and developments, and finally organize the included studies thematically, displaying them in tables and our taxonomy tree (Fig. 5), thus offering a multi-perspective view of the current landscape of diffusion model interpretability methodologies.

A. Interpreting Denoising Dynamics

Research on the temporal dynamics of the denoising process explores how semantic features are gradually

TABLE III
Distribution of Works Across Themes

Theme	% of Works
Latent Space Analysis	43%
Data Attribution	20%
Denoising Dynamics	19%
Holistic Output Analysis	9%
Other	9%

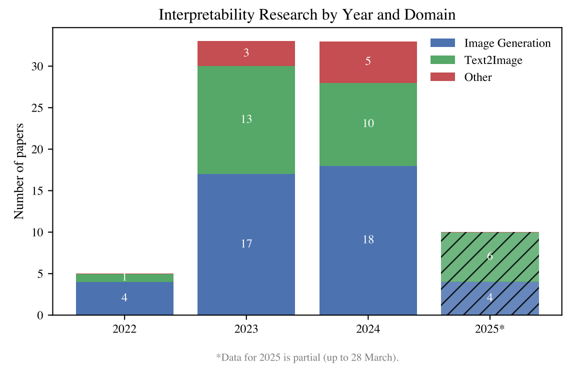

Fig. 3. Counts of papers per year showing the spread across domains.

constructed and refined across sampling steps. The subsequent insights promote intrinsic modifications to improve output. Works for this research area are listed in Table IV.

Foundational contribution to this perspective came from Choi *et al*. [53], who introduced the concept of perceptual phases in denoising by analyzing signal-to-noise ratios across timesteps. They identified distinct *perception bands*, each corresponding to a different level of semantic abstraction, thus framing generation as a staged refinement process. This framing prompted further research into identifying and characterizing the stages of semantic development; classifier probes [54], attention flow analysis [55], [56], mutual information measures [57], uncertainty weighting [58], and dimensionality reduction [59] and even architectural modification [60] have all been employed to examine the emergence and evolution of meaningful representations over time. Other approaches also integrate interpretability into the training objective or model structure, such as uncertainty-aware loss by [58].

Complementary studies examine how learned features evolve across timesteps, identifying *regime shifts* that reflect non-linear transitions in representational flow. These works provide a means to characterize memorization, interpolation, and projection onto learned manifolds [61]–[64], sometimes grounded in broader theories from physics and optimization [65]–[67]. Ambrogioni *et al*.'s [68] formal framework exemplifies this direction, using Bayesian tools to infer latent dynamics and iden-

TABLE IV
INCLUDED PAPERS: INTRINSIC INTERPRETABILITY

Authors and Link	Domain	Core Focus Area	Contribution Summary
Choi *et al.,* 2022 [53]	Image Generation	Denoising dynamics	Seminal work using SNR to reveal coarse-to-fine denoising phases.
Abu-Hussein & Giryes, 2023 [60]	Text2Image	Semantic structuring	Hierarchical latent-space up/downsampling for multi-scale feature analysis.
Ambrogioni, 2023 [68]	Image Generation	Associative memory	Hopfield network analogy as high-dimensional associative memory system.
Bakr *et al.,* 2023 [63]	Image Generation	Denoising dynamics	*ToddlerDiffusion:* Modular decomposition & editing of representations.
Go *et al.,* 2023 [58]	Image Generation	Denoising dynamics	Multi-task learning & Uncertainty weighting to mitigate negative transfer.
Prasad *et al.,* 2023 [61]	Text2Image	Denoising dynamics	Systematic evaluation of timestep and architectural component importance.
Permenter & Yuan, 2023 [67]	Image Generation	Denoising dynamics	Use Manifold theory to frame models as iterative optimization processors.
Raya & Ambrogioni, 2023 [64]	Image Generation	Denoising dynamics	Show hierarchical organization of information & sampling phase transition.
Wang *et al.,* 2023 [72]	Image Generation	Semantic structuring	Show semantically structured latent space using information maximization.
Xu *et al.,* 2023 [73]	Image Generation	Semantic structuring	*Versatile Diffusion*: multi-modal tasks due to modified multi-flow pathways.
Zach *et al.,* 2023 [74]	Text2Image	Semantic structuring	Analytic denoising via GMM expert priors under orthogonality constraints.
Biroli *et al.,* 2024 [65]	Image Generation	Denoising dynamics	Physics-based analysis of distinct training dynamics e.g., memorization.
Jun *et al.,* 2024 [75]	Image Generation	Semantic structuring	Promotes disentangled features using Dynamic Gaussian Anchoring.
Kadkhodaie *et al.,* 2024 [62]	Image Generation	Denoising dynamics	Geometrical analysis of denoising and showing generalized manifold.
McCart *et al.,* 2024 [76]	Neuroscience	Semantic structuring	Disentangled representations of behavior in neural signals from monkeys.
Popov & Tuba, 2024 [42]	Image Generation	Data attribution	Data attribution via Cold Diffusion and transformer embeddings.
Prasad *et al.,* 2024 [59]	Image generation	Denoising dynamics	*EvolvED*: Embed intermediate outputs into 2D space, preserving semantics.

tify structured semantic regimes within the generative process.

Several methods pinpoint the timing of semantic decisions by attributing specific features to individual timesteps. Li *et al*. [69] use Partial Information Decomposition (PID) to isolate the unique, redundant, and synergistic information contributions of different stages. Others use feature-activation visualization [70], where internal features are isolated to reveal mono-semantic properties, and hierarchical belief modeling [71] to map how and when semantic content becomes established and propagated through the denoising chain.

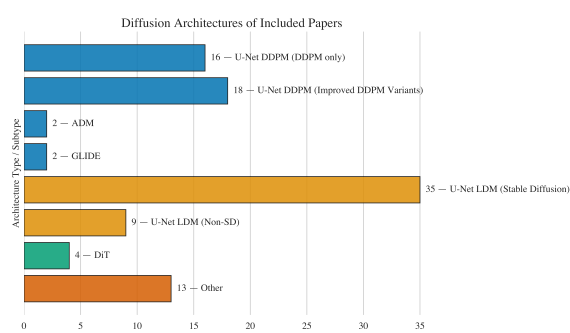

Fig. 4. Counts of papers showing the spread of model architecture types investigated in the contained literature.

Within the text-to-image setting, a growing body of work investigates the temporal unfolding of alignment between text and image during denoising. For example, Qu *et al*. [77] introduce a discriminative probing approach to identify which timesteps are most responsible for semantic alignment, revealing that meaningful text-image associations emerge only in specific mid-to-late stages of denoising. Similarly, Gandikota *et al*. [78] analyze how concept-specific attention evolves, showing that early layers contribute to object layout and spatial structure, while later steps refine texture and appearance, often entangled with specific prompt elements.

Mahajan *et al*. [79] extend this by analyzing when prompts activate attention heads, uncovering patterns of delayed or phase-specific conditioning, especially for complex or compositional prompts. These findings also indicate that semantic alignment is not uniformly imposed through the sampling trajectory, but instead emerges in discrete phases, shaped by both architectural factors and prompt design. Understanding how and when prompts shape the generated output provides insights into failure modes such as prompt forgetting, as well as into broader questions of how internal representations align with human intent.

Video generation has also received attention in this context. Xiao *et al*. [80] demonstrate that pretrained diffusion models implicitly encode motion, and that intermediate denoising states can be interpreted as latent motion trajectories. This suggests that denoising dynamics can serve as a form of unsupervised motion parsing, offering a temporally grounded interpretability signal even in higher-dimensional generative settings.

B. Interpreting Semantic Evolution in $\mathcal{X}$ and $\mathcal{Z}$-space

While the previous section examined the global behaviors of denoising dynamics, this section focuses on studies that analyze the latent *semantic content* represented at each timestep from the intermediate outputs. In diffusion models, generation progresses via a reverse-time trajectory denoted by $\{\mathbf{x}_t\}_{t=T\to 0}$ (or $\{\mathbf{z}_t\}_{t=T\to 0}$ in latent diffusion models), where $\mathbf{x}_T$ is pure Gaussian noise and $\mathbf{x}_0$ is the resultant coherent sample. A sequence of intermediate latent states is constructed by iteratively applying the denoising model. For pixel-space models, each $\mathbf{x}_t$ lies in $\mathcal{X} = \mathbb{R}^{H\times W\times C}$; (for latent diffusion models, each $\mathbf{z}_t$ lies in $\mathcal{Z} = \mathbb{R}^{h\times w\times c}$). At each timestep t, the model receives $\mathbf{x}_t$ (or $\mathbf{z}_t$) as input and predicts either $\hat{\mathbf{x_0}}$, or the noise component $\hat{\epsilon}_\theta(\mathbf{x}_t, t)$, depending on the parametrization. Successive states are generated by drawing the next sample $\mathbf{x}_{t-1}$ (or $\mathbf{z}_{t-1}$) using these predictions within Equation 4. Together, these states form the generation trajectory, which serves as the basis for analyzing how semantic structure unfolds over time in $\mathcal{X}$ and $\mathcal{Z}$ space. Table V lists papers relating to all latent space analyses.

1) $\mathcal{X}$-space: Several studies analyze how semantics become embedded in the intermediate states $\mathbf{x}_t$. Zhu *et al.* [84] demonstrate that high-level attributes such as gender, age, or presence of objects become linearly separable within $\mathcal{X}$ at specific timesteps. By training linear Support Vector Machines (SVMs) on inverted trajectories, they identified semantic directions that could be exploited for boundary-guided manipulation without fine-tuning. Wang *et al.* [12] extended this by using geometric probes such as Singular Value Decomposition (SVD) and clustering to analyze how semantic axes become increasingly aligned with principal components during denoising, suggesting a progressive disentanglement of content from noise.

Further geometric insights involve unsupervised methods [81], [82], [85], to extract interpretable directions in $\mathcal{X}$ based on curvature and geodesic structure. Here, Riemannian analysis reveals that semantic transformations correspond to low-curvature paths, suggesting the model's latent manifold is structured to support efficient semantic traversal. Tumanyan *et al.* [83] offer a complementary perspective by intervening in $\mathcal{X}$ with external guidance, treating it as a rich target space for plug-and-play editing.

2) $\mathcal{Z}$-space: In LDMs, the intermediate latent variables $\mathbf{z}_t$ serve an analogous role to $\mathbf{x}_t$ in pixel-based models, though they reside in a compressed feature space. Several works examine the structure and organization of $\mathcal{Z}$ to uncover how semantic concepts are encoded and disentangled during generation. For example, Burgess *et al.* [91] investigate viewpoint disentanglement, demonstrating that spatial transformations such as camera rotation are captured in structured subspaces of $\mathcal{Z}$ that evolve predictably over denoising steps. Tagaki & Nishimoto, [88] focus on identity representation by mapping brain activity to diffusion features, revealing that identity-specific information is preserved in distinct latent channels during generation.

Zhu *et al.* [13] introduce *Latent Explainer*, a model agnostic contrastive probing framework that aligns user-defined semantic concepts with directions in $\mathcal{Z}$, enabling controlled manipulation and interpretation of latent codes. Concept vector alignment is further explored by Wang *et al.* [89], who show that guidance signals, such as classifier gradients or textual prompts, correspond to semantically meaningful directions, particularly at early and intermediate denoising stages.

Other recent works use generative probing methods [14], [86], [87], [90], [92], in which specific neurons, channels, or token embeddings are activated or perturbed within $\mathcal{Z}$ to generate diagnostic outputs. This technique helps localize semantic attributes to specific spatial regions within the latent representation, and track their temporal emergence across timesteps, yielding a fine-grained interpretability map of the latent space of the model, albeit for certain attributes only.

These studies demonstrate progressive and structured evolution of semantic encoding along the denoising trajectory, with later timesteps containing more linearly accessible and disentangled features. While $\mathcal{X}$ provides a direct view of the model's reconstruction behavior, $\mathcal{Z}$ offers a compressed basis for probing the semantics of generation.

C. $\mathcal{H}$-space: U-Net Feature Representations

In most image and text-to-image interpretability papers reviewed, the U-Net architecture serves as the primary backbone for the denoising process. The hidden feature maps within this network, collectively referred to as $\mathcal{H}$-space, capture transient semantic information across spatial layers and diffusion timesteps. Examining feature trajectories over time and probing specific architectural layers makes $\mathcal{H}$-space a rich area for interpretability research.

A foundational study by Kwon *et al.* [15] showed that aggregated U-Net features contain global, linearly editable directions. By performing vector arithmetic in $\mathcal{H}$-space, they enabled semantic edits such as altering object identity or pose. This study highlighted the U-Net bottleneck as a key region of interest and laid the groundwork for much subsequent interpretability research.

Building on this, several supervised methods explicitly align U-Net features with labeled attributes to extract disentangled concept directions. Baranchuk *et al.* [93] first showed that attribute classifiers trained

TABLE V
INCLUDED PAPERS: POSTHOC - LATENT SPACE INTERPRETABILITY

Authors and Link	**Domain**	**Core Focus Area**	**Contribution Summary**
$\mathcal{X}$-space Interpretability			
Park *et al.,* 2023 [81]	Image Generation	Riemannian Geometry	Use of pull-back metrics to interpret latent dynamics across timesteps.
Park *et al.,* 2023 [82]	Image Generation	Riemannian Geometry	Demonstrate a curved manifold of disentangled semantic directions.
Tumanyan *et al.,* 2023 [83]	Image2Image	Dimension Reduction	Intermediate representations drive semantic and structural consistency.
Zhu *et al.,* 2023 [84]	Image generation	Semantic Discovery	*BoundaryDiffusion*: Distance analysis in latent space for one-step editing.
Park *et al.,* 2024 [85]	Text2Image	Temporal Denoising	Target areas of saliency maps, feature maps and exponential sampling.
$\mathcal{Z}$-space Interpretability			
Brack *et al.,* 2023 [86]	Text2Image	Noise Prediction	*SEGA*: sparse subsets of latent dimensions encode distinct concepts.
Chen, *et al.,* 2023 [87]	Image generation	Scene Geometry	Linear representations of 3D depth in early denoising & causal roles.
Tagaki & Nishimoto, 2023 [88]	Neuroscience	Semantic Reconstruction	Uses brain fMRI and representations in LDM, to reconstruct imagery.
Wang, *et al.,* 2023 [89]	Text2Image	Semantic Combinations	Show linear composition of semantic attributes via concept encoding.
Wu *et al.,* 2023 [90]	Text2Image	Attribute pairing	Semantic disentanglement of content and style at different timesteps.
Burgess, *et al.,* 2024 [91]	Text2Image	Textual Inversion	*ViewNeTI*: (Viewpoint Neural Textual Inversion), encodes 3D from 2D.
Kong *et al.,* 2024 [14]	Text2Image	Hierarchical Analyses	Concept-learning as a task of discrete latent hierarchical model.
Zhu *et al.,* 2024 [13]	Text2Image	Semantic Evolution	*LatentExplainer*: Multi-modal LLMs for perturbation explanations.
Huang, *et al.,* 2025 [92]	Text2Image	Semantic Evolution	*TIDE*: Temporal-aware Interpretable Diffusion transformErs.
Wang *et al.,* 2025 [12]	Text2Image	Semantic Evolution	Discovers editable semantic properties in $\mathcal{Z}$-space using SVD.
$\mathcal{H}$-space Interpretability			
Baranchuk *et al.,* 2021 [93]	Image segmentation	Feature Discovery	Enable high-level semantic segmentation using latent representations.
Gandikota *et al.,* 2023 [78]	Image Generation	LoRA Directions	*ConceptSliders*: LoRA to identify semantic attributes and apply control.
Ismail *et al.,* 2023 [94]	Image Generation	Concept Discovery	Uses concept bottleneck layer to describe and steer output samples.
Kwon *et al.,* 2023 [15]	Image Generation	Semantic Discovery	*Asyrp*: Seminal work identifying semantic representations and properties.
Liu *et al.,* 2023 [95]	Text2Image	Concept Discovery	*Cones*: Semantic modular encoding based on human brain analogy.
Yang *et al.,* 2023 [96]	Image Generation	Score-based Analysis	*DissDiff*: Unsupervised disentanglement via sub-gradient field analyses.
Zhang *et al.,* 2023 [97]	Image Generation	Semantic Discovery	Application of GAN Latent discovery methods to diffusion models.
Chefer *et al.,* 2024 [98]	Text2Image	Semantic Discovery	*Conceptor*: Textual concepts as combinations of interpretable elements.
Dalva & Yanardag, 2024 [99]	Image Generation	Semantic Discovery	*NoiseCLR*: Contrastive method method for semantic representations.
Haas *et al.,* 2024 [16]	Image Generation	Semantic Discovery	Global and local features from images using PCA and spectral analysis.
He *et al.,* 2024 [100]	Image Generation	Semantic Discovery	*LatentFace*: Self-supervised learning for facial features and identity.
Kim *et al.,* 2024 [70]	Image Generation	Semantic Discovery	Isolate monosemantics across differing model states and architectures.
Kouzelis *et al.,* 2024 [101]	Image Generation	Semantic Discovery	Use of Joint and Individual Variation Explained (JIVE) for local edits.
Li *et al.,* 2024 [102]	Text2Image	Semantic Discovery	Identification of desired concepts for responsible image generation.
Varshavsky *et al.,* 2024 [103]	Text2Speech	Semantic Discovery	Supervised & unsupervised semantic discovery applied to Text2Speech.
Yang *et al.,* 2024 [104]	Image reconstruction	Semantic Discovery	Use of semantic representations for content recovery and haze removal.
Zeng *et al.,* 2024 [105]	Text2Image	Semantic Discovery	Discover clusters of vectors to facilitate natural language descriptions.
Gandikota *et al.,* 2025 [106]	Text2Image	Semantic Structure	*SliderSpace*: Directions from a text prompt, using low-rank adaptors.
Park *et al.,* 2025 [17]	Image Generation	Semantic Discovery	Optimizing latent directions for disentangled multi-attribute editing.
Shi *et al.,* 2025 [107]	Image Generation	Semantic Discovery	*DiffLens*: Mechanistic interpretability for identifying bias pathways.

on bottleneck features could enable semantic control, such as manipulating facial expressions. He *et al*. [100] and Shi *et al*. [107] each extend this approach using fine-grained supervision for part-level segmentation and class-conditioned editing. Other works, including Park *et al*. [17], Varshavsky *et al*. [103], and Haas *et al*. [16], employ auxiliary classifiers or concept-aligned probes to identify and steer latent directions, enabling controlled interventions over both spatial regions and semantic attributes.

In parallel, unsupervised methods reveal that $\mathcal{H}$-space naturally organizes around interpretable axes, even without external labels. Yang *et al.* [96] use clustering and linear probes to identify semantic directions across U-Net layers and timesteps. Li *et al.* [102] and Zhang *et al.* [97] apply reconstruction and variational objectives to extract consistent representations along the diffusion trajectory. Dalva & Yanardag [99] employ contrastive learning to discover latent edit paths, while Zeng *et al.* [105] incorporate large language models to annotate clustered features, translating internal activations into natural language descriptions.

Building on these insights, recent work has developed post-hoc tools to identify and manipulate interpretable directions within the latent space $\mathcal{H}$. Chefer et al. [98] propose the *Conceptor* framework, which learns pseudo-token vectors aligned with specific textual concepts. A lightweight multi-layer perceptron is trained to reconstruct images generated from a concept prompt; it then assigns weights to vocabulary tokens, revealing an interpretable decomposition of the concept. This reflects how the model internally represents semantic meaning, exposing interesting associations, e.g., the model encodes a snake as a twisted gecko. This method supports both semantic editing and compositional analysis of generated content.

In parallel, Gandikota et al. [106] introduce *SliderSpace*, a framework for uncovering semantically meaningful directions within the denoising trajectory of diffusion models. Instead of relying on hand-crafted attributes, they extract low-dimensional subspaces by applying principal component analysis to residual activations computed from prompt pairs (e.g., a neutral face vs. smiling face). These Slider Spaces allow smooth interpolation between concepts and generalize across architectures, including both U-Net and transformer-based diffusion models.

Most interpretability methods focus on the latent or activation spaces of diffusion models, particularly within the U-Net architecture. In contrast, Dravid et al. [108] propose the *w2w* framework, which treats the model's *weight space* itself as a generative manifold. By interpolating between trained model checkpoints, *w2w* enables semantic editing, inversion, and controlled sampling—extending interpretability beyond feature activations to the parameters of the model itself.

Gandikota et al. [78] explore cross-model semantic alignment by transferring concept directions from StyleGAN to diffusion models. Using external priors, they demonstrate that diffusion models share structural representations with other generative architectures, suggesting common underlying semantics across model families.

Although the U-Net bottleneck is established as a rich source of global and editable semantic vectors, recent work provides additional nuance by examining how concepts emerge and propagate across both time *and* architectural depth. Studies such as *DisDiff* [96] and *DiffLI*^{2}*D* [104] apply layer-wise probing to trace a hierarchical evolution of features: early layers encode coarse patterns, mid layers capture high-level semantics, and later layers refine visual details. These findings complement earlier analyses of denoising dynamics (Section IV-A) and reinforce that while the bottleneck concentrates semantic information, surrounding layers also play distinct roles. Kim *et al*. [70] and Zeng *et al*. [105] further support this by showing that semantic richness peaks at mid-layers and mid-timesteps, where abstract concepts are most clearly represented.

D. Attribution-Based Interpretability

Attribution-based methods aim to identify how different sources of influence, either inputs at sampling time or training data, shape the outputs of diffusion models. Input attribution can be in the form of classifier guidance or text input. In text-to-image settings, this often involves understanding how individual prompt tokens or phrases drive the emergence of specific visual features. In training data attribution, approaches investigate which training instances are most influential for particular generations and some aim to identify memorization. We structure this section into two subcategories: input data attribution and training data attribution.

1) Input Attribution via Cross-Attention and Saliency: Text-to-image diffusion models rely heavily on cross-attention to align linguistic and visual representations. Studies in this domain seek to map prompt tokens to generated features or spatial regions, often leveraging attention scores, gradient-based attribution, or geometric probing of internal representations.

A foundational method in this area is *DAAM* (Diffusion Attentive Attribution Maps) by Tang *et al*. [109], which visualizes high-resolution spatial heatmaps linking prompt tokens to image regions. Applied to an early release of Stable Diffusion [23], DAAM traces how language concepts influence denoising over time, revealing the evolving impact of cross-attention during generation.

Later works extend and deepen attention-based interpretability. Chen *et al*. [110] introduce *Time-aware Dual Cross-Attention* (TDCA), which visualizes evolving attention patterns across both U-Net layers and denoising steps. By separating coarse and fine textual cues across early and late layers, it reveals how semantic attributes are resolved hierarchically.

Voynov *et al*. [111] propose P^+, which injects structured prompt representations into cross-attention layers at different U-Net stages, showing how different layers encode specific semantic attributes and providing insight into the spatial and hierarchical role of cross-attention in generation. Similarly, Liu *et al*. [95] identify modular *concept neurons* in attention maps, enabling the suppression or enhancement of specific visual concepts via gradient attribution.

Wang *et al*. [89] introduce *Concept Algebra*, a framework for algebraically manipulating cross-attention vectors to perform semantic operations such as addition, subtraction, and abstraction. This reveals that cross-attention layers form a structured and linearly navigable latent space analogous to the $\mathcal{H}$-space concept directions proposed by [15], [16].

Further interpretability is enabled by mechanistic tracing methods such as *DIFF-QUICKFIX* [112] and *LocoGen* [113] , which localize semantic concepts to specific neuron clusters or architectural regions within the U-Net, allowing for targeted edits and conceptual debugging.

Principled perspectives are adopted by Kong *et al.* [57], who apply mutual information to quantify dependencies between prompt tokens and image features, and by Dewan *et al.* [114], who use partial information decomposition to isolate the unique and redundant influences of different input tokens. Both works focus on principled frameworks for measuring semantic contributions.

The preceding works are based on U-Net-based models. In the context of Diffusion Transformers (DiTs), *ConceptAttention* [115] explores attention output space by injecting concept embeddings during inference. This approach yields concept-level saliency maps with sharper, more semantically aligned visualizations than conventional cross-attention. The authors show that DiT's attention heads encode contextually rich, localized semantic features, highlighting architectural implications for interpretability. Huang *et al.* [92] support understanding and control of concept representations during sampling by extracting conceptual edits in an analogous way to identifying semantic edits in U-Net $\mathcal{H}$-space.

Added as a footnote to this section, although not strictly input attribution are Cardenuto *et al.* [116] who provide a method for identifying the model-type responsible for an output. A type of forensic attribution, they identify model-specific artifacts left in synthetic output. Primarily designed for detecting fraudulent biomedical figures (e.g., Western blots), their findings reveal that diffusion models imprint structured textures and reduce high-frequency noise, which may hint at broader architectural commonalities in diffusion pipelines.

2) Training Data Attribution: While input attribution methods explain how prompts shape outputs, training data attribution traces the influence of specific training samples, which could be images, captions or other inputs, on the generative behavior of diffusion models. These approaches aim to uncover how information is internalized during training and reused during inference. Related papers can be found listed in Table VI.

Dai & Gifford [117] propose a counterfactual generation method that approximates influence of individual training samples on a target output. By analyzing the Jacobians of the score function and optimizing counterfactual generations that match the query, they identify training instances that are semantically and visually aligned with the final image.

Several methods extend existing techniques in discriminative analysis to diffusion models. For example, Georgiev *et al.* [55] adapt the TRAK framework [123], which was originally designed for classifiers to estimate training data influence by propagating gradients through snapshots of the model taken during training. The adaptation to diffusion was achieved by computing influence scores at each denoising timestep, providing a temporally resolved view of how training examples affect final predictions. Their work reveals how certain examples are memorized early and exert persistent influence throughout the optimization trajectory. Xie *et al.* [119] also investigate training data influence by adapting the TracIn framework [124] which was originally developed for supervised models to estimate the influence of training points via gradient similarity across training checkpoints. The diffusion-aware approach introduces *Diffusion-ReTrac*, which computes and aggregates gradients at each denoising timestep, enabling temporally resolved estimates of how individual training examples contribute to the final generation. This adaptation captures the evolving influence of data across the diffusion trajectory.

Influence from text prompts in Text2Image models is investigated by Wen *et al.* [118]. They propose a detection-based framework to identify cases where the model has overfit to a prompt by assessing the strength of the association between the input prompt and the generated output, higher magnitudes suggesting a stronger likelihood that the prompt is tied to memorized content.

Mlodozeniec *et al.* [120] extend classical influence functions to diffusion models by proposing a principled and scalable Hessian approximation. They introduce a Generalized Gauss–Newton (GGN) estimator that linearizes the model output rather than the loss, enabling more theoretically grounded influence estimates. Their method retains consistency with the diffusion training objective and is amenable to efficient Monte Carlo approximation.

These methods support a more accountable and transparent view of generative model behavior by revealing internal pathways from data input to generation and impacting application areas such as copyright and bias. Popov & Tuba [42] put copyright at the core of their novel cold-diffusion approach which re-engineers the generative process to allow complete traceability from output to training data. Rather than use stochastic noise, they employ structured operators derived from citation data. The model outputs both the generated visual and accompanying metadata, enabling fully accountable content generation.

TABLE VI
INCLUDED PAPERS: DATA ATTRIBUTION ANALYSES

Authors and Link	**Domain**	**Core Focus Area**	**Contribution Summary**
Model-level Attribution			
Cardenuto *et al.*, 2023 [116]	Image Generation	Forensic attribution	Identified Diffusion-specific artifacts in synthetic output.
Training Data Attribution			
Dai & Gifford, 2023 [117]	Image Generation	Data ablation analysis	Ensemble-based method to trace influential training data.
Georgiev *et al.,* 2023 [55]	Image generation	Counterfactual analysis	Identify influential samples at points along sampling trajectory.
Wen, *et al.,* 2024 [118]	Text2Image	Memorisation	Identify trigger-tokens with adapted data-attribution correlation metric.
Xie, *et al.,* 2024 [119]	Image Generation	Gradient based attribution	Diagnose & correct timestep-induced bias in data influence estimation.
Mlodozeniec *et al.,* 2025 [120]	Image generation	Influence function analysis	Approximates how output would change if data were removed.
Input Attribution			
Tang, *et al.,* 2022 [109]	Text2Image	Cross attention mapping	*DAAM*: Diffusion Attentive Attribution Maps. Seminal work.
Basu, *et al.,* 2023 [112]	Text2Image	Causal mediation	*DIFF-QUICKFIX*: Causal tracing localizing knowledge for editing.
Voynov, *et al.,* 2023 [111]	Text2Image	Textual input space	*XTI*: Extended Textual Inversion - injects tokens into x-attn layers.
Basu, *et al.,* 2024 [113]	Text2Image	Causal tracing	*LOCOGEN*: Localised Generation. Concept identification.
Dewan, *et al.,* 2024 [114]	Text2Image	Information-theoretic	Information decomposition to decompose prompt token influence.
Kong, *et al.,* 2024 [57]	Text2Image	Information-theoretic	Partial information decomposition applied to prompt tokens.
Park & Jang, 2024 [121]	Image2Image	Bi-directional attribution	I^2AM: Image to Image bidirectional Attribution Mapping.
Pennisi, *et al.,* 2024 [122]	Image Generation	Hierarchical modeling	*Diffexplainer*: Synthesized visuals reveal automatic cross-modal bias.
Chen, *et al.,* 2025 [110]	Text2Image	Cross attention analysis	*TDCA*: Time-aware Dual Cross Attention (plug and play).
Helbling, *et al.,* 2025 [115]	Text2Image	Attention saliency maps	*ConceptAttention*: Investigate attention outputs in DiTs.
Park, *et al.,* 2025 [56]	Text2Image	Cross attention relevance	Head Relevance Vectors for cross-attention head importance.

E. Holistic Interpretability and Diffusion Specific Evaluation Metrics

Interpretability approaches in this category focus on global behaviors of diffusion models, assessing generation quality and diversity at the distributional level. These works provide insights that go beyond individual samples or model components. Further, the unique mechanisms underlying diffusion-based generation pose challenges for conventional evaluation. While standard generative metrics like FID [20], Inception Score (IS) [125], and the Disentanglement-Completeness-Informativeness (DCI) score [126] are widely used, they may offer only limited insight into the temporal, semantic, or conditioning-dependent dynamics specific to diffusion processes. These metrics, originally developed for GANs or latent variable models, may overlook failure modes and behavioral signals crucial for the interpretability of diffusion models. Evaluation metrics, both new and modified for diffusion, are listed in Table VII.

Several studies aim to understand how well diffusion models capture the overall data distribution or reproduce meaningful semantic boundaries; the discriminative probing framework developed by Qu *et al.* [77], not only contributes to denoising dynamics to evaluate the internal semantic understanding of text-to-image diffusion models but also helps assess the model's performance on overall image-text matching (ITM) and referring expression comprehension (REC) tasks, treating these as proxies for compositional and referential understanding. Distributional boundaries is also investigated by Lee *et al.* [127], who propose a method to interrogate the distributional reliability of diffusion-based synthesis in robotics. Drawing from the iterative generative mechanisms of diffusion models to "stitch" together plausible yet unseen trajectories [128], they inspect how models extrapolate beyond the training distribution in robotic navigation plans. By detecting infeasible or structurally implausible outputs, they quantify a model's tendency to generate out-of-distribution solutions using their diffusion-bespoke metric: *restoration gap*, to reveal areas of uncertainty in learned generative dynamics. Uncertainty is also exploited by Wu *et al.* [129] for image anomaly detection with their *Masked Diffusion Posterior Sampling* (MDPS) method. Unsupervised and grounded in Bayesian inference, they condition denoising on partially masked inputs and sample from the posterior $p(\mathbf{x}_0 \mid \mathbf{y})$. Their method localizes model uncertainty across an image, thereby mapping regions of low model familiarity and visualizing with heat-maps. This spatial decomposition highlights deviation from the learned data manifold.

A model agnostic focus is taken by Ravuri *et al.* [132] and Kim *et al.* [130], in particular to address the need for evaluation methods and metrics which are particularly suited to generative models. Ravuri *et al.* [132] propose

TABLE VII
DIFFUSION-SPECIFIC EVALUATION METRICS

Metric(s)	Description	Source
Restoration Gap	Trajectory deviation in robotic planning	Lee *et al.* [127]
MDPS	Bayesian anomaly detection via masked conditioning	Wu *et al.* [129]
DF-GRAM, DF-RISE	Adapted Grad-CAM/RISE for diffusion saliency	Park *et al.* [85]
SaD, PaD, HCS	Attribute-aware KL and multi-modal CLIPScore variants	Kim *et al.* [130]
IRS	Real-image recall diversity via synthetic queries	Dombrowski *et al.* [131]
GELs	Model-agnostic eval via moment conditions	Ravuri *et al.* [132]
Noise Magnitude Signal	Overfitting detection via noise strength	Wen *et al.* [118]

a principled framework based on Generalized Empirical Likelihood (GEL) to diagnose and evaluate deep generative models (including diffusion), by recasting evaluation metrics as moment conditions. Their method enables the construction of interpretable tests that can identify key failure modes such as mode dropping, mode imbalance, and improper label conditioning, without requiring access to model likelihoods. By defining flexible, kernel-based moment constraints, there is capability for both aggregate and per-sample diagnostic scores. Kim *et al.* [130] address the limitations of standard evaluation metrics such as FID and CLIPScore [133] by proposing a new suite of attribute-based interpretability metrics for generative models. By using their modified metric, Heterogeneous CLIPScore (HCS), they improve the sensitivity and interpretability of similarity scores between image and text embeddings. Additionally, they define two further divergence-based metrics: *Single-attribute Divergence (SaD)*, which quantifies over or under representation of specific attributes via KL divergence between attribute distributions in generated and training data; and *Paired-attribute Divergence (PaD)*, which measures how well a model preserves joint attribute relationships. Using these metrics, they observe that increasing the number of sampling steps in LDMs can improve FID but worsen SaD, suggesting a trade-off between fine-grained detail and attribute consistency. Dombrowski *et al.* [131] also highlight the limitations of existing metrics in assessing the diversity of generative models, particularly diffusion models. They introduce the *Image Retrieval Score* (IRS), an interpretable and hyperparameter-free metric that quantifies diversity by measuring how many real images can be retrieved using synthetic data as queries. Their evaluation reveals that current diffusion models capture at most 77% of the diversity present in training data. To address this, they propose *Diversity-Aware Diffusion Models* (DiADM), which incorporate a diversity-aware module utilizing pseudo-unconditional features to enhance output diversity without compromising image quality. This approach disentangles diversity from fidelity, offering a more nuanced understanding of generative model performance. Memorization is the focus for Wen *et al.* [118] who demonstrate that the magnitude of noise predictions in text-conditioned diffusion models serves as a measurable proxy for content memorization. Their quantitation helps to assess the strength of the model's response to particular inputs and allows for the detection of overfitting to specific training samples.

Xue *et al.* [134], introduce *SingVisio*, an interactive visual analytics (IVA) tool designed to examine the generation of Singing Voice Conversion (SVC) using the DiffWaveNetSVC diffusion model. SingVisio allows users to explore five interactive and hierarchical visualizations of denoising, facilitating detailed analysis and diagnosis of various stages of generation. This visualization-centred approach is a powerful mechanism for assimilating the vast quantitative outputs required to comprehensively interpret diffusion models.

V. DISCUSSION

Below, we collate our findings in response to the guiding research questions in Section I-A.

A. Interpretable Diffusion Models Across Domains (RQ1)

Interpretability research for diffusion models remains concentrated in computer vision tasks (See Table III and Fig. 3); over 90% of surveyed studies focus on visual domains, including image generation and editing (52%) and text-to-image synthesis (38%). This distribution reflects a broader application landscape: Ma *et al.* [135] estimate that over 85% of diffusion-related papers between 2022 and 2024 address visual problems. As diffusion continues expanding into new fields [4], interpretability research is anticipated to follow.

In image generation, the most prevalent interpretability techniques involve latent space analysis. Researchers probe intermediate trajectories or representations to trace how semantic features emerge and evolve over diffusion steps. These insights enable users to understand and manipulate outputs by identifying disentangled attributes and concept alignment. Most methods target model-specific structures such as U-Nets (Fig. 4), reflecting the tight coupling between architecture and interpretability strategy in this space.

For Text2Image (T2I) tasks, interpretability efforts naturally focus on cross-modal components of diffusion models. Cross-attention analysis is central, with methods that visualize how text prompts influence generated content, or align prompt tokens to image regions [56], [90], [91], [95], [109]–[111]. Some works also adapt post-hoc attention tools from transformer models to diffusion

TABLE VIII
INCLUDED PAPERS: OTHER POSTHOC - HOLISTIC OUTPUTS & DENOISING DYNAMICS

Authors and Link	Domain	Core Focus Area	Contribution Summary
Holistic Analyses			
Lee, *et al.,* 2023 [127]	Robotic Planning	Distributional Feasibility	Introduce a metric to identify likely out-of-distribution outputs.
Ravuri, *et al.,* 2023 [132]	Image Generation	Model Agnostic Analysis	Generalized empirical likelihood to evaluate generative models.
Xue, *et al.,* 2024 [134]	Audio Conversion	Interactive Visual Analytics	Interactive dashboard containing 5 hierarchical views of denoising.
Dombrowski, *et al.,* 2024 [131]	Image Generation	Diversity analysis	*DiADM*: Diversity Aware Diffusion Models, using pseudo-labels.
Kim, *et al.,* 2024 [130]	Image Generation	Model Agnostic Analysis	Diffusion metrics to reveal significant attributes and relationships.
Qu, *et al.,* 2024 [77]	Text2Image	Text-Image Alignment	Image-Text Matching & Referring Expression Comprehension.
Wu, *et al.,* 2024 [129]	Image Generation	Anomaly Detection	*MDPS*: Masked Diffusion Posterior Sampling, evaluates uncertainty.
Weight Space Analysis			
Dravid, *et al.,* 2024 [108]	Image Generation	Weight Space Analysis	*w2w*: weights2weights, weight space as a low-dimensional manifold.
Denoising Dynamics			
Deja, *et al.,* 2022 [66]	Image Generation	Performance Analysis	Tracks generation quality across denoising steps to uncover biases.
Mahajan, *et al.,* 2023 [79]	Text2Image	Text-image Alignment	Interpretable language for target images using prompt inversion.
Li & Chen, 2024 [69]	Image Generation	Hierarchical Modelling	Decision windows of feature hierarchies for multimodal distributions.
Xiao, *et al.,* 2024 [80]	Video Generation	Motion Features Analysis	*MOFT*: MOtion FeaTure, motion information from temporal features.
Sclocchi, *et al.,* 2025 [71]	Image Generation	Hierarchical Modelling	Show phase transition of low to high-level features during sampling.

architectures [115], accounting for the iterative nature of denoising.

Beyond image and T2I domains, interpretability research is limited but growing. Existing studies take a holistic view, analyzing model outputs or learned dynamics in domains such as audio, video, and multimodal synthesis [80], [127], [134]. Although small in number, these works demonstrate the adaptability of interpretability tools and the need for generalizable methods as diffusion models move into new application areas.

Finally, across all domains, intrinsically interpretable diffusion architectures remain comparatively rare (21%), possibly due to ongoing trade-offs between transparency and model complexity [40], [136]. As diffusion models are increasingly deployed in high-stakes contexts, embedding interpretability at the architectural level, rather than relying solely on post-hoc tools, represents a key direction for future research.

B. Target Areas for Diffusion Model Interpretability (RQ2)

Interpretability research on diffusion models spans a wide range of targets across the generative process. Based on our review, we organize these into a taxonomy (Fig. 5) covering both *post-hoc* and *intrinsic* approaches. Among 81 papers, the vast majority (n=65) take a post-hoc view, with the remaining 17 focusing on intrinsic model design.

Within post-hoc methods, the most active area is *Latent Space analysis* (n=35). These studies interrogate hidden representations at various levels, including $\mathcal{H}$-space (U-Net hidden states), $\mathcal{Z}$-space (VAE-style latents), and $\mathcal{X}$-space (input/output embeddings), to understand how semantic concepts emerge or transform during denoising. This line of work is particularly dominant in vision tasks, and often aims to link latent trajectories to interpretable features for editing, steering, or conceptual discovery.

The next most common post-hoc category is *Data Attribution* (n=16), where researchers examine how specific inputs or training examples influence generated outputs. This includes input-level saliency and prompt attribution [109], [111], as well as methods to trace training data contributions [119] or detect memorization [118]. These studies provide critical insights into model behavior and failure modes, especially in tasks involving prompt-to-image translation.

Denoising Dynamics (n=5) is another distinctive post-hoc category, where the focus is on how model outputs evolve across timesteps. These works track the sharpening or disappearance of features during generation [66], [69], helping to identify bottlenecks or semantic emergence over time.

On the intrinsic side, the most active category is *Training & Denoising Interpretability* (n=10), where interpretability is integrated into the model architecture or learning process. These studies propose mechanisms such as phase-based learning [65], constrained representations [53], or transparency-aware objectives [58], [61].

Other intrinsic categories include *Semantic Structuring* (n=5), which focuses on organizing latent spaces to reflect human-interpretable dimensions [73], [75], and a small set of *Theoretical or bespoke frameworks* (n=2), which propose normative definitions or analytical tools. The collective aim of this group of research is to build interpretability into the denoising aspect of the model from the ground up, rather than relying on post-hoc

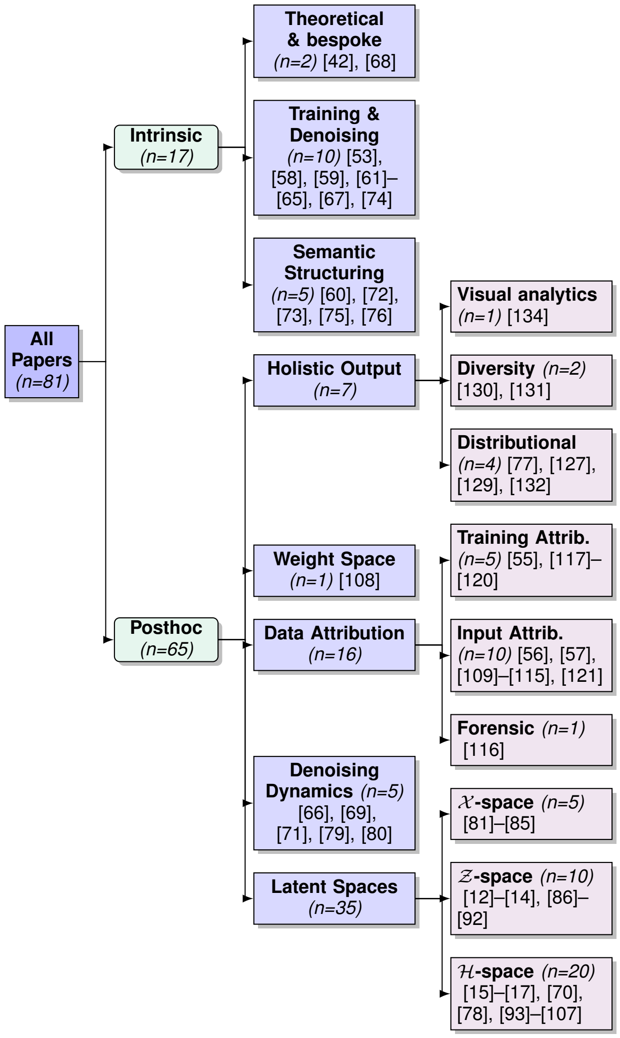

Fig. 5. Taxonomy of interpretability methods for all included works. Each citation appears only once, and is placed in the category with which it most aligns.

probing.

Finally, we observe a small but growing interest in more holistic methods (n=7), such as *Distributional analyses*, *Diversity metrics*, or *Visual analytics* [134]. These approaches infer broader model properties contributing to an explanation of a model's performance.

Overall, the field is currently dominated by post-hoc analyses, especially of latent representations, but there is a visible shift toward intrinsic interpretability and cross-domain generalization. As diffusion models are deployed in more sensitive settings, we anticipate greater demand for built-in interpretability and robust attribution techniques.

C. Patterns, Gaps, and Future Opportunities (RQ3)

As diffusion models rapidly proliferate across domains, extending from image synthesis [3] to robotics [5], [127], planning, and audio [4], [134], there is significant opportunity to keep pace in terms of interpretability. We present a non-exhaustive summary of several potential future opportunities based on areas of missing-ness and limitations identified from examination of the included citations:

- **Emerging need for domain-specific interpretability:** The need for tailored interpretability methods will become increasingly important across domains, for example, identifying whether there are differences in timestep ranges for different applications or tasks would be hugely beneficial to diffusion understanding, not only to accommodate different data modalities and task structures, but also to build trust in high-stakes or embodied deployments.
- **Need for interpretability in non-U-Net architectures:** Most existing interpretability research has focused on diffusion models based on the U-Net architecture. Alternative architectures, especially diffusion transformers, are gaining traction [31], [32]. The release of Stable Diffusion 3.0 in March 2024 [24], which adopts a transformer-based design, further highlights the growing importance of interpreting these newer model classes.
- **Latent space understanding is still nascent:** While examples shown by [16], [99], [104] demonstrate methods to identify global semantic directions within the latent spaces of the U-Nets, robust methods to map these semantics or generalize across models are lacking. For example, limitations identified by [16] demonstrate that disentangled global and local semantic categories are not straightforward to predict or identify and differ between subject content in pretrained models.
- **Out-of-distribution (OOD) behavior is underexplored:** Particularly relevant to applications where diffusion output is temporal (e.g., robot planning [128]), the capacity of diffusion models to stitch together novel samples from suboptimal subsequences is observed but poorly understood or adequately evaluated. Greater examination of the diffusion dynamics that give rise to this phenomenon is an area of potentially keen interest.
- **Lack of standard diffusion based evaluation protocols:** Evaluation still remains heavily dependent on FID and IS metrics [20], [125]. Some metrics are being adapted [85], [118], [130], [131] and some are created from new [127], [129], [132], but identify a lack of breadth and descriptiveness. More standardized interpretability benchmarks, especially for intermediate representations, could offer greater optimization opportunities.
- **End-to-end visual analytics:** Many interpretability techniques are developed for specific facets of the

generation process. Moving towards trustworthy AI will benefit from consolidated frameworks such as [134], allowing comprehensive diagnoses and interpretability at multiple levels of granularity.

D. Limitations

This review has several limitations that we sought to mitigate but could not fully eliminate. First, due to the rapid pace of diffusion model research, we included a number of pre-prints (approx. 21%), allowing us to capture emerging methods at the risk of variability in study rigor.

Second, designing a precise database search strategy was challenging. Interpretability terms are inconsistently applied across fields, and "diffusion" appears in unrelated contexts. As a result, citation chaining played a key role in supplementing our coverage, though it carries a risk of missing relevant studies.

Finally, while we adhered strictly to our inclusion criteria to maintain a focused scope, decisions made were still subjective, based on nuanced interpretations of each study's objectives and contributions. This subjectivity invites debate on many of the edge-case exclusions particularly as a result of second stage screening.

VI. Conclusion

This review provides a broad-scope overview of interpretability methods applied to diffusion models. Scoping has been applied across domains and methodological approaches. By collating trends, categorizing techniques, and highlighting representative works, we identify both the strengths of current research and notable gaps facing researchers concerned with diffusion model interpretability. Our review offers researchers a structured and informed view of the existing landscape and opportunities to expand interpretability research.

Acknowledgments

We would like to extend our sincere thanks to Cathy Parker-Hopkins, for her extensive and sage advice on evidence review protocols, standards and compliance.

References

[1] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, Jul. 2015, pp. 2256–2265.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arxiv:2006.11239*, 2020.

[3] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[4] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[5] Z. Zhu, H. Zhao, H. He, Y. Zhong, S. Zhang, H. Guo, T. Chen, and W. Zhang, "Diffusion models for reinforcement learning: A survey," *arXiv preprint arXiv:2311.01223*, 2023.

[6] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," 2013.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020.

[8] Z. Lin, S. Basu, M. Beigi, V. Manjunatha, R. A. Rossi, Z. Wang, Y. Zhou, S. Balasubramanian, A. Zarei, K. Rezaei *et al.*, "A survey on mechanistic interpretability for multi-modal foundation models," *arXiv preprint arXiv:2502.17516*, 2025.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] L. Wang, B. Gao, Y. Li, Z. Wang, X. Yang, D. A. Clifton, and J. Xiao, "Exploring the latent space of diffusion models directly through singular value decomposition," *arXiv preprint arXiv:2502.02225*, 2025.

[13] M. Zhu, R. Kanjiani, J. Lu, A. Choi, Q. Ye, and L. Zhao, "LatentExplainer: Explaining latent representations in deep generative models with multi-modal foundation models," *arXiv preprint arXiv:2406.14862*, 2024.

[14] L. Kong, G. Chen, B. Huang, E. Xing, Y. Chi, and K. Zhang, "Learning discrete concepts in latent hierarchical models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 36 938–36 975, 2024.

[15] M. Kwon, J. Jeong, and Y. Uh, "Diffusion models already have a semantic latent space," in *11th International Conference on Learning Representations, ICLR 2023*, 2023.

[16] R. Haas, I. Huberman-Spiegelglas, R. Mulayoff, S. Graßhof, S. S. Brandt, and T. Michaeli, "Discovering interpretable directions in the semantic latent space of diffusion models," in *18th IEEE International Conference on Automatic Face and Gesture Recognition*, 2024.

[17] J. Park, M. Shaheryar, S. Lee, and S. K. Jung, "Navigating h-Space for multi-attribute editing in diffusion models," in *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 2025, pp. 1129–1133.

[18] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

[19] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794.

[20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6629–6640.

[21] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations (ICLR)*, 2021.

[22] CompVis, S. AI, and LAION, "Stable diffusion v1.4," 2022.

[23] Stability AI, "Stable diffusion v2," 2022.
[24] ——, "Stable diffusion 3: Research paper," 2024.
[25] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.
[26] S. Luo, Y. Tan, S. Patil, D. Gu, P. von Platen, A. Passos, L. Huang, J. Li, and H. Zhao, "Lcm-lora: A universal stable-diffusion acceleration module," *arXiv preprint arXiv:2311.05556*, 2023.
[27] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021.
[28] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, Jul. 2021, pp. 8599–8608.
[29] X. Zhang, Y. Gu, H. Chen, Z. Fang, L. Zou, L. Xue, and Z. Wu, "Leveraging content-based features from multiple acoustic models for singing voice conversion," *arXiv preprint arXiv:2310.11160*, 2023.
[30] X. Zhang, L. Xue, Y. Gu, Y. Wang, J. Li, H. He, C. Wang, S. Liu, X. Chen, J. Zhang *et al.*, "Amphion: An open-source audio, music, and speech generation toolkit," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 879–884.
[31] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022, pp. 4195–4205.
[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.
[33] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
[34] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.
[35] S. Li, Y. Du, G. Van de Ven, and I. Mordatch, "Energy-based models for continual learning," in *Conference on Lifelong Learning Agents*. PMLR, 2022, pp. 1–22.
[36] R. Gao, Y. Song, B. Poole, Y. N. Wu, and D. P. Kingma, "Learning energy-based models by diffusion recovery likelihood," *arXiv preprint arXiv:2012.08125*, 2020.
[37] Z. C. Lipton, "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.
[38] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. Turin, Italy: IEEE, Oct. 2018, pp. 80–89.
[39] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
[40] M. Dombrowski, H. Reynaud, J. P. Müller, M. Baugh, and B. Kainz, "Trade-offs in fine-tuned diffusion models between accuracy and interpretability," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, M. J. Wooldridge, J. G. Dy, and S. Natarajan, Eds. AAAI Press, 2024, pp. 21 037–21 045.
[41] A. Bansal, E. Borgnia, H.-M. Chu, J. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping, and T. Goldstein, "Cold diffusion: Inverting arbitrary image transforms without noise," *Advances in Neural Information Processing Systems*, vol. 36, pp. 41 259–41 282, 2023.
[42] M. Popov and E. Tuba, "Credible diffusion: Improving diffusion models interpretability with transformer embeddings," in *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, 2024, pp. 1–6.
[43] B. A. Kitchenham, P. Brereton, and D. Budgen, "Evidence-based Software engineering and systematic reviews," in *Evidence-Based Software Engineering and Systematic Reviews*, 1st ed., ser. Chapman & Hall/CRC Innovations in Software Engineering and Software Development. Boca Raton, FL: Chapman and Hall/CRC, an imprint of Taylor and Francis, 2015.
[44] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, M. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ (Clinical research ed.)*, vol. 372, p. n71, 2021.
[45] Joanna Briggs Institute, "Systematic and scoping review protocol template," 2020.
[46] "IEEE xplore digital library."
[47] "Scopus."
[48] "Arxiv.org e-Print archive."
[49] "DBLP computer science bibliography."
[50] Elicit, "Elicit: The AI Research Assistant," Jan. 2023.
[51] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.
[52] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University Joint Report, Tech. Rep. EBSE-2007-01, 2007.
[53] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, and S. Yoon, "Perception prioritized training of diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 472–11 481.
[54] R. Marjieh, K. Q. Weinberger, P. Gucik-Derigny, T. B. Hashimoto, and D. Krueger, "Analyzing diffusion as serial reproduction," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, vol. 202. PMLR, 2023, pp. 24 166–24 186.
[55] K. Georgiev, J. Vendrow, H. Salman, S. M. Park, and A. Madry, "The journey, not the destination: How data guides diffusion models," *arXiv preprint arXiv:2312.06205*, 2023.
[56] M. Park, D. Lee, J. Kim, and B. Han, "Cross-attention head position patterns can align with human visual concepts in text-to-image generative models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.
[57] X. Kong, O. Liu, H. Li, D. Yogatama, and G. Ver Steeg, "Interpretable diffusion via information decomposition," in *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.
[58] H. Go, Y. Lee, S. Lee, S. Oh, H. Moon, and S. Choi, "Addressing negative transfer in diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27 199–27 222, 2023.
[59] V. Prasad, H. van Gorp, C. Humer, R. J. van Sloun, A. Vilanova, and N. Pezzotti, "EvolvED: Evolutionary embeddings to understand the generation process of diffusion models," *arXiv preprint arXiv:2406.17462*, 2024.
[60] S. Abu-Hussein and R. Giryes, "Udpm: Upsampling diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 27 616–27 646, 2024.
[61] V. Prasad, C. Zhu-Tian, A. Vilanova, H. Pfister, N. Pezzotti, and H. Strobelt, "Unraveling the temporal dynamics of the unet in diffusion models," in *Deployable AI@ AAAI*, 2023.
[62] Z. Kadkhodaie, F. Guth, E. P. Simoncelli, and S. Mallat, "Generalization in diffusion models arises from geometry-adaptive

harmonic representations," in *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[63] E. M. BAKR, L. Zhao, V. T. Hu, M. Cord, P. Perez, and M. Elhoseiny, "ToddlerDiffusion: Interactive structured image generation with cascaded schrödinger bridge," in *The Thirteenth International Conference on Learning Representations*, 2024.

[64] G. Raya and L. Ambrogioni, "Spontaneous symmetry breaking in generative diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 66 377–66 389, 2023.

[65] G. Biroli, T. Bonnaire, V. de Bortoli, and M. Mézard, "Dynamical regimes of diffusion models," *Nature Communications*, vol. 15, no. 1, p. 9957, Nov. 2024.

[66] K. Deja, A. Kuzina, T. Trzcinski, and J. Tomczak, "On analyzing generative and denoising capabilities of diffusion-based deep generative models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 218–26 229, 2022.

[67] F. Permenter and C. Yuan, "Interpreting and improving diffusion models from an optimization perspective," *arXiv preprint arXiv:2306.04848*, 2023.

[68] L. Ambrogioni, "In search of dispersed memories: Generative diffusion models are associative memory networks," in *Associative Memory {\&} Hopfield Networks in 2023*, 2023.

[69] M. Li and S. Chen, "Critical windows: Non-asymptotic theory for feature emergence in diffusion models," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 2024, pp. 27 474–27 498.

[70] D. Kim, X. Thomas, and D. Ghadiyaram, "Revelio: Interpreting and leveraging semantic information in diffusion models," *arXiv preprint arXiv:2411.16725*, 2024.

[71] A. Sclocchi, A. Favero, and M. Wyart, "A phase transition in diffusion models reveals the hierarchical nature of data," *Proceedings of the National Academy of Sciences*, vol. 122, no. 1, p. e2408799121, 2025.

[72] Y. Wang, Y. Schiff, A. Gokaslan, W. Pan, F. Wang, C. De Sa, and V. Kuleshov, "InfoDiffusion: Representation learning using information maximizing diffusion models," in *ICML*, 2023.

[73] X. Xu, Z. Wang, G. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7754–7765.

[74] M. Zach, T. Pock, E. Kobler, and A. Chambolle, "Explicit diffusion of gaussian mixture model based image priors," in *International Conference on Scale Space and Variational Methods in Computer Vision*, 2023, pp. 3–15.

[75] Y. Jun, J. Park, K. Choo, T. E. Choi, and S. J. Hwang, "Disentangling disentangled representations: Towards improved latent units via diffusion models," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 3559–3569.

[76] J. D. McCart, A. R. Sedler, C. Versteeg, D. Mifsud, M. Rigotti-Thompson, and C. Pandarinath, "Diffusion-based generation of neural activity from disentangled latent codes," *ArXiv*, pp. arXiv–2407, 2024.

[77] L. Qu, W. Wang, Y. Li, H. Zhang, L. Nie, and T.-S. Chua, "Discriminative probing and tuning for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7434–7444.

[78] R. Gandikota, J. Materzynska, T. Zhou, A. Torralba, and D. Bau, "Concept sliders: LoRA adaptors for precise control in diffusion models," *CoRR*, 2023.

[79] S. Mahajan, T. Rahman, K. M. Yi, and L. Sigal, "Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6808–6817.

[80] Z. Xiao, Y. Zhou, S. Yang, and X. Pan, "Video diffusion models are training-free motion interpreter and controller," in *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024.

[81] Y.-H. Park, M. Kwon, J. Choi, J. Jo, and Y. Uh, "Understanding the latent space of diffusion models through the lens of riemannian geometry," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24 129–24 142, 2023.

[82] Y.-H. Park, M. Kwon, J. Jo, and Y. Uh, "Unsupervised discovery of semantic latent directions in diffusion models," *arXiv preprint arXiv:2302.12469*, 2023.

[83] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1921–1930.

[84] Y. Zhu, Y. Wu, Z. Deng, O. Russakovsky, and Y. Yan, "Boundary guided learning-free semantic control with diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 319–78 346, 2023.

[85] J.-H. Park, Y.-J. Ju, and S.-W. Lee, "Explaining generative diffusion models via visual analysis for interpretable decision-making process," *Expert Systems With Applications*, vol. 248, p. 123231, 2024.

[86] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski, and K. Kersting, "Sega: Instructing text-to-image models using semantic guidance," *Advances in Neural Information Processing Systems*, vol. 36, pp. 25 365–25 389, 2023.

[87] Y. Chen, F. Viégas, and M. Wattenberg, "Beyond surface statistics: Scene representations in a latent diffusion model," *arXiv preprint arXiv:2306.05720*, 2023.

[88] Y. Takagi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 453–14 463.

[89] Z. Wang, L. Gui, J. Negrea, and V. Veitch, "Concept algebra for (score-based) text-controlled generative models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 35 331–35 349, 2023.

[90] Q. Wu, Y. Liu, H. Zhao, A. Kale, T. Bui, T. Yu, Z. Lin, Y. Zhang, and S. Chang, "Uncovering the disentanglement capability in text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1900–1910.

[91] J. Burgess, K.-C. Wang, and S. Yeung-Levy, "Viewpoint textual inversion: Discovering scene representations and 3D view control in 2D diffusion models," in *European Conference on Computer Vision*, 2024, pp. 416–435.

[92] V. S.-J. Huang, L. Zhuo, Y. Xin, Z. Wang, P. Gao, and H. Li, "TIDE: Temporal-aware sparse autoencoders for interpretable diffusion transformers in image generation," *CoRR*, 2025.

[93] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," *arXiv preprint arXiv:2112.03126*, 2021.

[94] A. A. Ismail, J. Adebayo, H. C. Bravo, S. Ra, and K. Cho, "Concept bottleneck generative models," in *The Twelfth International Conference on Learning Representations*, 2023.

[95] Z. Liu, R. Feng, K. Zhu, Y. Zhang, K. Zheng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao, "Cones: Concept neurons in diffusion models for customized generation," in *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, vol. 202. PMLR, 2023, pp. 21 548–21 566.

[96] T. Yang, Y. Wang, Y. Lu, and N. Zheng, "DisDiff: Unsupervised disentanglement of diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 69 130–69 156, 2023.

[97] Z. Zhang, L. Liu, Z. Lin, Y. Zhu, and Z. Zhao, "Unsupervised discovery of interpretable directions in h-space of pre-trained diffusion models," *CoRR*, 2023.

[98] H. Chefer, O. Lang, M. Geva, V. Polosukhin, A. Shocher, M. Irani, I. Mosseri, and L. Wolf, "The hidden language of diffusion models," in *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[99] Y. Dalva and P. Yanardag, "Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 209–24 218.

[100] R. He, Z. Xing, W. Tan, and B. Yan, "A generative framework for self-supervised facial representation learning," *arXiv preprint arXiv:2309.08273*, 2024.

[101] T. Kouzelis, M. Plitsis, M. A. Nicolaou, and Y. Panagakis, "Enabling local editing in diffusion models by joint and individual component analysis," *BMVC2024*, 2024.

[102] H. Li, C. Shen, P. Torr, V. Tresp, and J. Gu, "Self-discovering interpretable diffusion latent directions for responsible text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 006–12 016.

[103] M. Varshavsky-Hassid, R. Hirsch, R. Cohen, T. Golany, D. Freedman, and E. Rivlin, "On the semantic latent space of diffusion-based text-to-speech models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2024, pp. 246–255.

[104] Z. Yang, H. Yu, B. Li, J. Zhang, J. Huang, and F. Zhao, "Unleashing the potential of the semantic latent space in diffusion models for image dehazing," in *European Conference on Computer Vision*. Springer, 2024, pp. 371–389.

[105] E. Z. Zeng, Y. Chen, and A. Wong, "Decoding diffusion: A scalable framework for unsupervised analysis of latent space biases and representations using natural language prompts," *arXiv preprint arXiv:2410.21314*, 2024.

[106] R. Gandikota, Z. Wu, R. Zhang, D. Bau, E. Shechtman, and N. Kolkin, "SliderSpace: Decomposing the visual capabilities of diffusion models," *arXiv preprint arXiv:2502.01639*, 2025.

[107] Y. Shi, C. Li, Y. Wang, Y. Zhao, A. Pang, S. Yang, J. Yu, and K. Ren, "Dissecting and mitigating diffusion bias via mechanistic interpretability," *arXiv preprint arXiv:2503.20483*, 2025.

[108] A. Dravid, Y. Gandelsman, K.-C. Wang, R. Abdal, G. Wetzstein, A. Efros, and K. Aberman, "Interpreting the weight space of customized diffusion models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 137 334–137 371, 2024.

[109] R. Tang, L. Liu, A. Pandey, Z. Jiang, G. Yang, K. Kumar, P. Stenetorp, J. Lin, and F. Türe, "What the DAAM: Interpreting stable diffusion using cross attention," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5644–5659.

[110] X. Chen, C. Bai, Z. Wu, X. Wu, Q. Zou, Y. Xia, and S. Wang, "Coarse-to-fine text injecting for realistic image super-resolution," *Neurocomputing*, p. 129591, 2025.

[111] A. Voynov, Q. Chu, D. Cohen-Or, and K. Aberman, "P+: Extended textual conditioning in text-to-image generation," *arXiv preprint arXiv:2303.09522*, 2023.

[112] S. Basu, N. Zhao, V. I. Morariu, S. Feizi, and V. Manjunatha, "Localizing and editing knowledge in text-to-image generative models," in *The Twelfth International Conference on Learning Representations*, 2023.

[113] S. Basu, K. Rezaei, P. Kattakinda, V. I. Morariu, N. Zhao, R. A. Rossi, V. Manjunatha, and S. Feizi, "On mechanistic knowledge localization in text-to-image generative models," in *Forty-First International Conference on Machine Learning*, 2024.

[114] S. Dewan, R. Zawar, P. Saxena, Y. Chang, A. Luo, and Y. Bisk, "Diffusion PID: Interpreting diffusion via partial information decomposition," *Advances in Neural Information Processing Systems*, vol. 37, pp. 2045–2079, 2024.

[115] A. Helbling, T. H. S. Meral, B. Hoover, P. Yanardag, and D. H. Chau, "ConceptAttention: Diffusion transformers learn highly interpretable features," *CoRR*, vol. abs/2502.04320, 2025.

[116] J. P. Cardenuto, S. Mandelli, D. Moreira, P. Bestagini, E. Delp, and A. Rocha, "Explainable artifacts for synthetic western blot source attribution," in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2024, pp. 1–6.

[117] Z. Dai and D. K. Gifford, "Training data attribution for diffusion models," *arXiv preprint arXiv:2306.02174*, 2023.

[118] Y. Wen, Y. Liu, C. Chen, and L. Lyu, "Detecting, explaining, and mitigating memorization in diffusion models," in *The Twelfth International Conference on Learning Representations*, 2024.

[119] T. Xie, H. Li, A. Bai, and C.-J. Hsieh, "Data attribution for diffusion models: Timestep-induced bias in influence estimation," *Transactions on Machine Learning Research*, 2024.

[120] B. K. Mlodozeniec, R. Eschenhagen, J. Bae, A. Immer, D. Krueger, and R. E. Turner, "Influence functions for scalable data attribution in diffusion models," in *The Thirteenth International Conference on Learning Representations*, 2025.

[121] J. Park and H. Jang, "I2AM: Interpreting image-to-image latent diffusion models via attribution maps," *arXiv preprint arXiv:2407.12331*, 2024.

[122] M. Pennisi, G. Bellitto, S. Palazzo, M. Shah, and C. Spampinato, "Diffexplainer: Towards cross-modal global explanations with diffusion models," *arXiv preprint arXiv:2404.02618*, 2024.

[123] S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, and A. Madry, "TRAK: Attributing model behavior at scale," in *International Conference on Machine Learning*. PMLR, 2023, pp. 27 074–27 113.

[124] G. Pruthi, F. Liu, S. Kale, and M. Sundararajan, "Estimating training data influence by tracing gradient descent," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 920–19 930, 2020.

[125] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[126] C. Eastwood and C. K. Williams, "A framework for the quantitative evaluation of disentangled representations," in *6th International Conference on Learning Representations*, 2018.

[127] K. Lee, S. Kim, and J. Choi, "Refining diffusion planner for reliable behavior synthesis by automatic detection of infeasible plans," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24 223–24 246, 2023.

[128] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," *arXiv e-prints*, pp. arXiv–2205, 2022.

[129] D. Wu, S. Fan, X. Zhou, L. Yu, Y. Deng, J. Zou, and B. Lin, "Unsupervised anomaly detection via masked diffusion posterior sampling," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, pp. 2442–2450.

[130] D. Kim, M. Kwon, and Y. Uh, "Attribute based interpretable evaluation metrics for generative models," in *International Conference on Machine Learning*. PMLR, 2024, pp. 24 271–24 293.

[131] M. Dombrowski, W. Zhang, S. Cechnicka, H. Reynaud, and B. Kainz, "Image generation diversity issues and how to tame them," *arXiv preprint arXiv:2411.16171*, 2024.

[132] S. Ravuri, M. Rey, S. Mohamed, and M. P. Deisenroth, "Understanding deep generative models with generalized empirical likelihoods," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 395–24 405.

[133] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[134] L. Xue, C. Wang, M. Wang, X. Zhang, J. Han, and Z. Wu, "SingVisio: Visual analytics of diffusion model for singing voice conversion," *Computers & Graphics*, vol. 124, p. 104058, 2024.

[135] Z. Ma, Y. Zhang, G. Jia, L. Zhao, Y. Ma, M. Ma, G. Liu, K. Zhang, N. Ding, J. Li *et al.*, "Efficient diffusion models: A comprehensive survey from principles to practices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[136] A. Assis, J. Dantas, and E. Andrade, "The performance-interpretability trade-off: A comparative study of machine

learning models," *Journal of Reliable Intelligent Environments*, vol. 11, no. 1, p. 1, 2025.

Appendices & Supplementary Information

Appendix A
Systematic Review Protocol

The systematic review protocol was created *a priori* to ensure a structured and systematic approach to information gathering.
Protocol Title: Methodological Approaches for Diffusion Generative Model Interpretability: A Systematic Review Protocol
Version: 3.0
Date Finalised: August 2025

This document is available via the gituhub link.

Appendix B
Database Search Strings

TABLE IX
Search queries used in different academic databases.

Database	Date Searched	Search Query
Elicit	25/03/2025	”Papers relating to methods for interpreting diffusion models, since 2020”.
IEEE Xplore	20/03/2025	diffusion OR ”score based” OR ”energy based” NEAR/3 model AND interpretab* OR explainab* OR semantic NEAR/3 ”latent space” OR understand NEAR/3 ”latent space” OR analy* NEAR/3 ”latent space”
Scopus	20/03/2025	(diffusion OR ”score based” OR ”energy based” W/3 model*) AND (interpretab* OR explainab* OR (semantic W/3 ”latent space” OR understand* W/3 ”latent space” OR analy* W/3 ”latent space”)) in (Title, Abstract & Keywords).
ArXiv	28/03/2025	https://arxiv.org/search/advanced?advanced=&terms-0-operator=AND&terms-0-term=diffusion+OR+%22energy-based%22+OR+%22score-based%22&terms-0-field=abstract&terms-1-operator=AND&terms-1-term=interpretab*+OR+explainab*+OR+%22semantic+latent%22&terms-1-field=abstract&classification-physics_archives=all&classification-include_cross_list=include&date-year=&date-filter_by=date_range&date-from_date=2020-01-01&date-to_date=2025-03-28&date-date_type=submitted_date&abstracts=show&size=50&order=-announced_date_first.
DBLP	28/03/2025	diffusion AND (interpretab* — explainab*).

Appendix C
Excluded Papers from Full Text Screening (Unreferenced)

TABLE X: Out of scope papers from second screening.

Title	**Authors**	**Year**	**Reason/Notes**
A Geometric Perspective on Diffusion Models	Defang Chen, Zhenyu Zhou, Jianhan Mei, Chunhua Shen, Chun Chen, C. Wang	2023	Theoretical/Conceptual - not on operational diffusion models
A Hierarchical Diffusion-Convolutional Network with Node-wise Localization for EEG-NIRS-based Brain-Computer Interface	W. Huang, X. Song, D. Kuang	2024	Not diffusion model specific
A latent diffusion approach to visual attribution in medical imaging	Siddiqui A.A., Tirunagari S., Zia T., Windridge D.	2025	Not interpreting diffusion mechanisms
An analytic theory of creativity in convolutional diffusion models	Mason Kamb, Surya Ganguli	2024	Theoretical/Conceptual - not on operational diffusion models
An Interpretable Latent Denoising Diffusion Probabilistic Model for Fault Diagnosis Under Limited Data	T. Zhang, J. Lin, J. Jiao, H. Zhang, H. Li	2024	Edge case exclusion - more downstream than interpretability

Title	Authors	Year	Reason/Notes
Analyzing Diffusion as Serial Reproduction	Raja Marjieh, Ilia Sucholutsky, Thomas A. Langlois, Nori Jacoby, T. Griffiths	2022	Theoretical/Conceptual - not on operational diffusion models
Bayesian Dumbbell Diffusion Model for RGBT Object Tracking With Enriched Priors	S. Fan, C. He, C. Wei, Y. Zheng, X. Chen	2023	Not interpreting diffusion mechanisms
Bayesian MRI reconstruction with joint uncertainty estimation using diffusion models	Guanxiong Luo, Moritz Blumenthal, Martin Heide, M. Uecker	2022	Edge case exclusion - more downstream than diffusion interpretability
Characterizing the Features of Mitotic Figures Using a Conditional Diffusion Probabilistic Model	Bahadir C.D., Liechty B., Pisapia D.J., Sabuncu M.R.	2024	Edge case exclusion - more downstream than diffusion interpretability
Closed-Loop Unsupervised Representation Disentanglement with VAE Distillation and Diffusion Probabilistic Feedback	Xin Jin, Bohan Li, BAAO Xie, Wenyao Zhang, Jinming Liu, Ziqiang Li, Tao Yang, Wenjun Zeng	2024	UnCaptured dupe.
Closed-Loop Unsupervised Representation Disentanglement with VAE Distillation and Diffusion Probabilistic Feedback	Jin X., Li B., Xie B., Zhang W., Liu J., Li Z., Yang T., Zeng W.	2025	No interpretability of diffusion process
Concept Steerers: Leveraging K-Sparse Autoencoders for Controllable Generations	Dahye Kim, Deepti Ghadiyaram	2025	More focus on control than interpretation
Deep generative priors for biomolecular 3D heterogeneous reconstruction from cryo-EM projections	Shi B., Zhang K., Fleet D.J., McLeod R.A., Dwayne Miller R.J., Howe J.Y.	2024	Not interpreting diffusion mechanisms
DEPICT: Diffusion-Enabled Permutation Importance for Image Classification Tasks	Jabbour S., Kondas G., Kazerooni E., Sjoding M., Fouhey D., Wiens J.	2025	Edge case exclusion - more downstream than diffusion interpretability
DifCluE: Generating Counterfactual Explanations with Diffusion Autoencoders and modal clustering	Suparshva Jain, Amit Sangroya, Lovekesh Vig	2025	Edge case exclusion - more downstream than diffusion interpretability
Diff-Props: is Semantics Preserved within a Diffusion Model?	Bonechi S., Andreini P., Corradini B.T., Scarselli F.	2024	Edge case exclusion - more downstream than diffusion interpretability
DiffDGSS: Generalizable Retinal Image Segmentation with Deterministic Representation from Diffusion Models	Xie Y., Qu J., Xie H., Wang T., Lei B.	2024	Edge case exclusion - more downstream than diffusion interpretability
Diffexplainer: Towards Cross-modal Global Explanations with Diffusion Models	Matteo Pennisi, Giovanni Bellitto, Simone Palazzo, Mubarak Shah, Concetto Spampinato	2024	Edge-case exclusion (uses LDM to explain classifier)
Diffuse, Sample, Project: Plug-And-Play Controllable Graph Generation	Sharma K., Kumar S., Trivedi R.S.	2024	Not interpreting diffusion mechanisms
DiffuseGAE: Controllable and High-fidelity Image Manipulation from Disentangled Representation	Leng Y., Huang Q., Wang Z., Liu Y., Zhang H.	2023	Edge case exclusion - more downstream than diffusion interpretability
DiffuseGAE: Controllable and High-fidelity Image Manipulation from Disentangled Representation	Yipeng Leng, Qiangjuan Huang, Zhiyuan Wang, Yangyang Liu, Haoyu Zhang	2023	Edge case exclusion - more downstream than diffusion interpretability
DiffuseReg: Denoising Diffusion Model for Obtaining Deformation Fields in Unsupervised Deformable Image Registration	Zhuo Y., Shen Y.	2024	Optimisation rather than interpretability
DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents	Kushagra Pandey, Avideep Mukherjee, Piyush Rai, Abhishek Kumar	2022	Not interpreting diffusion mechanisms
Diffusion Autoencoders for Few-shot Image Generation in Hyperbolic Space	Lingxiao Li, Kaixuan Fan, Boqing Gong, Xiangyu Yue	2024	More focus on control than interpretation
Diffusion Explainer: Visual Explanation for Text-to-image Stable Diffusion	Seongmin Lee, Benjamin Hoover, Hendrik Strobelt, Zijie J. Wang, Sheng-Hsuan Peng, Austin P. Wright, Kevin Li, Haekyu Park, Haoyang Yang, Duen Horng Chau	2023	Not interpreting diffusion mechanisms
Diffusion Lens: Interpreting Text Encoders in Text-to-Image Pipelines	Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, Yonatan Belinkov	2024	Edge case exclusion - more downstream than diffusion interpretability
Diffusion Map Autoencoder	Julio Candanedo	2025	Not interpreting diffusion mechanisms
Diffusion Models for Counterfactual Explanations	Jeanneret G., Simon L., Jurie F.	2024	Edge case exclusion - more downstream than diffusion interpretability
Diffusion Models for Counterfactual Generation and Anomaly Detection in Brain Images	A. Fontanella, G. Mair, J. Wardlaw, E. Trucco, A. Storkey	2024	Edge case exclusion - more downstream than diffusion interpretability
Diffusion Models for Counterfactual Explanations	Jeanneret G., Simon L., Jurie F.	2023	UnCaptured dupe.
Diffusion Random Feature Model	Esha Saha, Giang Tran	2023	Rejected by ICLR, not yet published elsewhere
Diffusion-Based Visual Counterfactual Explanations - Towards Systematic Quantitative Evaluation	Váth P., M. Frühwald A., Paassen B., Gregorova M.	2025	Edge case exclusion - more downstream than diffusion interpretability
DIFFUSION-TS: INTERPRETABLE DIFFUSION FOR GENERAL TIME SERIES GENERATION	Yuan X., Qiao Y.	2024	Not interpreting diffusion mechanisms
DMCVR: Morphology-Guided Diffusion Model for 3D Cardiac Volume Reconstruction	He X., Tan C., Han L., Liu B., Axel L., Li K., Metaxas D.N.	2023	Not interpreting diffusion mechanisms
Encoding physics to learn reaction-diffusion processes	Chengping Rao, Pu Ren, Qi Wang, Oral Buyukozturk, Hao Sun, Yang Liu	2023	Not diffusion model specific
Energy-Based Model for Accurate Estimation of Shapley Values in Feature Attribution	Cheng Lu, Jiusun Zeng, Yu Xia, Jinhui Cai, Shihua Luo	2025	Not diffusion model specific
Enhancing Conditional Image Generation with Explainable Latent Space Manipulation	Kshitij Pathania	2024	Masters dissertation (unpublished at time of writing)

Title	**Authors**	**Year**	**Reason/Notes**
Enhancing high-resolution reconstruction of flow fields using physics-informed diffusion model with probability flow sampling	Guo Y.; Cao X.; Zhou M.; Leng H.; Song J.	2024	Not interpreting diffusion mechanisms.
Evaluating Diffusion Models for the Automation of Ultrasonic Nondestructive Evaluation Data Analysis	Torenvliet N.; Zelek J.	2024	Not interpreting diffusion mechanisms.
Explainable, Multi-modal Wound Infection Classification from Images Augmented with Generated Captions	Palawat Busaranuvong, Emmanuel Agu, Reza Saadati Fard, Deepak Kumar, Shefalika Gautam, Bengisu Tulu, Diane Strong	2025	Not interpreting diffusion mechanisms
Exploiting Interpretable Capabilities with Concept-Enhanced Diffusion and Prototype Networks	Alba Carballo-Castro, Sonia Laguna, Moritz Vandenhirtz, Julia E. Vogt	2024	Not interpreting diffusion mechanisms
Exploring Behavior-Relevant and Disentangled Neural Dynamics with Generative Diffusion Models	Yule Wang, Chengrui Li, Weihan Li, Anqi Wu	2024	Not interpreting diffusion mechanisms
Exploring how deep learning decodes anomalous diffusion via Grad-CAM	Jaeyong Bae, Yongjoo Baek, Hawoong Jeong	2024	Not diffusion model specific
Factorized Diffusion Autoencoder for Unsupervised Disentangled Representation Learning	Wu A.; Zheng W.-S.	2024	Not interpreting diffusion mechanisms
From Points to Functions: Infinite-dimensional Representations in Diffusion Models	Sarthak Mittal, Guillaume Lajoie, S. Bauer, Arash Mehrjou	2022	Edge case exclusion - more downstream than diffusion interpretability
Fuzzy-Conditioned Diffusion and Diffusion Projection Attention Applied to Facial Image Correction	M. E. Helou	2023	Not interpreting, improving performance only
Generating and evaluating synthetic data in digital pathology through diffusion models	Pozzi M.; Noei S.; Robbi E.; Cima L.; Moroni M.; Munari E.; Torresani E.; Jurman G.	2024	Not interpreting diffusion mechanisms
Generating Counterfactual Trajectories with Latent Diffusion Models for Concept Discovery	Varshney P.; Lucieri A.; Balada C.; Dengel A.; Ahmed S.	2025	Edge case exclusion - more downstream than diffusion interpretability
Generative Modelling With Inverse Heat Dissipation	Severi Rissanen, Markus Heinonen, A. Solin	2022	Not interpreting diffusion mechanisms
Generative Time Series Forecasting with Diffusion, Denoise, and Disentanglement	Li Y.; Lu X.; Wang Y.; Dou D.	2022	Not interpreting diffusion mechanisms
Good Seed Makes a Good Crop: Discovering Secret Seeds in Text-to-Image Diffusion Models	Katherine Xu, Lingzhi Zhang, Jianbo Shi	2024	Not interpreting diffusion mechanisms
Hierarchical Diffusion Autoencoders and Disentangled Image Manipulation	Lu Z.; Wu C.; Chen X.; Wang Y.; Bai L.; Qiao Y.; Liu X.	2024	Not interpreting diffusion mechanisms
Hierarchically branched diffusion models leverage dataset structure for class-conditional generation	Tseng A.M.; Shen M.; Biancalani T.; Scalia G.	2024	Not interpreting diffusion mechanisms
High-Precision Face Generation and Manipulation Guided by Text, Sketch, and Mask	Q. Guo; X. Gu	2025	Not interpreting diffusion mechanisms
How Much Is Enough? A Study on Diffusion Times in Score-Based Generative Models	Giulio Franzese, Simone Rossi, Lixuan Yang, A. Finamore, Dario Rossi, M. Filippone, Pietro Michiardi	2022	Not interpreting diffusion mechanisms
Hyperbolic Geometric Latent Diffusion Model for Graph Generation	Fu X.; Gao Y.; Wei Y.; Sun Q.; Peng H.; Li J.; Li X.	2024	Not interpreting diffusion mechanisms
ICE: Intrinsic Concept Extraction from a Single Image via Diffusion Models	Fernando Julio Cendra, Kai Han	2025	Edge case exclusion - more downstream than diffusion interpretability
Illumination and Shadows in Head Rotation: Experiments with Denoising Diffusion Models	Asperti A.; Colasuonno G.; Guerra A.	2024	Not interpreting diffusion mechanisms
IMAGE TRANSLATION AS DIFFUSION VISUAL PROGRAMMERS	Han C.; Liang J.C.; Wang Q.; Rabbani M.; Dianat S.; Rao R.; Wu Y.N.; Liu D.	2024	Edge case exclusion - more downstream than diffusion interpretability
Improving Fairness using Vision-Language Driven Image Augmentation	M. D'Incà; C. Tzelepis; I. Patras; N. Sebe	2024	Not interpreting diffusion mechanisms
Interpretable Alzheimer's Disease Classification Via a Contrastive Diffusion Autoencoder	Ayodeji Ijishakin, Ahmed Abdulaal, Adamos Hadjivasiliou, Sophie Martin, James Cole	2023	Edge case exclusion - more downstream than diffusion interpretability
Interpretable Matching of Optical-SAR Image via Dynamically Conditioned Diffusion Models	Gou S.; Wang X.; Wang X.; Chen Y.	2024	Edge case exclusion - more downstream than diffusion interpretability
Interpretable Measures of Conceptual Similarity by Complexity-Constrained Descriptive Auto-Encoding	A. Achille; G. V. Steeg; T. Y. Liu; M. Trager; C. Klingenberg; S. Soatto	2024	Not interpreting diffusion mechanisms
Interpretable-through-prototypes deepfake detection for diffusion models	A. Aghasanli; D. Kangin; P. Angelov	2023	Edge case exclusion - more downstream than diffusion interpretability
Interpreting Deep Neural Networks through Prototype Factorization	Subhajit Das, Panpan Xu, Zeng Dai, A. Endert, Liu Ren	2020	Not diffusion model specific
Interventional and Counterfactual Inference with Diffusion Models	Patrick Chao, Patrick Blöbaum, S. Kasiviswanathan	2023	Not interpreting diffusion mechanisms
ISPDiff: Interpretable Scale-Propelled Diffusion Model for Hyperspectral Image Super-Resolution	W. Dong; S. Liu; S. Xiao; J. Qu; Y. Li	2024	Not interpreting diffusion mechanisms
Iterative Search Attribution for Deep Neural Networks	Zhu Z.; Chen H.; Wang X.; Zhang J.; Jin Z.; Xue J.; Shen J.	2024	Not diffusion model specific
JADE: Joint-aware Latent Diffusion for 3D Human Generative Modeling	Haorui Ji, Rong Wang, Taojun Lin, Hongdong Li	2024	Edge case exclusion - more downstream than diffusion interpretability
KGDiff: towards explainable target-aware molecule generation with knowledge guidance	Qian H.; Huang W.; Tu S.; Xu L.	2024	Edge case exclusion - more downstream than diffusion interpretability

Title	**Authors**	**Year**	**Reason/Notes**
Knowledge tracing via multiple-state diffusion representation	Zhang K.; Ji T.; Zhang H.	2024	Not interpreting diffusion mechanisms
Label-free and interpretable hyperspectral imaging for intraoperative clinical applications	Zhang Y.; Yu S.; Wang C.; Zhu X.; Zheng Y.; Bao J.	2021	Not interpreting diffusion mechanisms
Language-Oriented Semantic Latent Representation for Image Transmission	Giordano Cicchetti, Eleonora Grassucci, Jihong Park, Jinho Choi, Sergio Barbarossa, Danilo Comminiello	2024	Not interpreting diffusion mechanisms
Latent Diffusion Energy-Based Model for Interpretable Text Modeling	Yu P.; Xie S.; Ma X.; Jia B.; Pang B.; Gao R.; Zhu Y.; Zhu S.-C.; Wu Y.N.	2022	Not interpreting diffusion mechanisms
Let us Build Bridges: Understanding and Extending Diffusion Generative Models	Xingchao Liu, Lemeng Wu, Mao Ye, Qiang Liu	2022	Edge case exclusion - more downstream than diffusion interpretability
Linear Spaces of Meanings: Compositional Structures in Vision-Language Models	M. Trager; P. Perera; L. Zancato; A. Achille; P. Bhatia; S. Soatto	2023	Not interpreting diffusion mechanisms
Localizing Object-level Shape Variations with Text-to-Image Diffusion Models	O. Patashnik; D. Garibi; I. Azuri; H. Averbuch-Elor; D. Cohen-Or	2023	Not interpreting diffusion mechanisms
Mapping the Mind of an Instruction-based Image Editing using SMILE	Zeinab Dehghani, Koorosh Aslansefat, Adil Khan, Adín Ramírez Rivera, Franky George, Muhammad Khalid	2024	Edge case exclusion - more downstream than diffusion interpretability
Masked Completion via Structured Diffusion with White-Box Transformers	Druv Pai, Ziyang Wu, Sam Buchanan, Yaodong Yu, Yi Ma	2024	Not interpreting diffusion mechanisms
MedDiffusion: Boosting Health Risk Prediction via Diffusion-based Data Augmentation	Zhong Y.; Cui S.; Wang J.; Wang X.; Yin Z.; Wang Y.; Xiao H.; Huai M.; Wang T.; Ma F.	2024	Edge case exclusion - more downstream than diffusion interpretability
MIDGArD: Modular interpretable diffusion over graphs for articulated designs	Leboutet, Quentin; Wiedemann, Nina; Cai, Zhipeng; Paulitsch, Michael; Yuan, Kai	2024	Edge case exclusion - more downstream than diffusion interpretability
Modeling Causal Mechanisms with Diffusion Models for Interventional and Counterfactual Queries	Patrick Chao, Patrick Blobaum, Sapan Patel, S. Kasiviswanathan	2023	Uncaptured dupe
Neural Message Passing Induced by Energy-Constrained Diffusion	Qitian Wu, David Wipf, Junchi Yan	2024	Not interpreting diffusion mechanisms
Noise Crystallization and Liquid Noise: Zero-shot Video Generation using Image Diffusion Models	Muhammad Haaris Khan, Hadrien Reynaud, Bernhard Kainz	2024	Not interpreting diffusion mechanisms
On the notion of Hallucinations from the lens of Bias and Validity in Synthetic CXR Images	Gauri Bhardwaj, Yuvaraj Govindarajulu, Sundaraparipurnan Narayanan, Pavan Kulkarni, Manojkumar Parmar	2023	Not interpreting diffusion mechanisms
Plug-and-Play Interpretable Responsible Text-to-Image Generation via Dual-Space Multi-facet Concept Control	Basim Azam, Naveed Akhtar	2025	Not interpreting diffusion mechanisms
Probabilistic and semantic descriptions of image manifolds and their applications	Tu P.; Yang Z.; Hartley R.; Xu Z.; Zhang J.; Fu Y.; Campbell D.; Singh J.; Wang T.	2023	Edge case exclusion - more downstream than diffusion interpretability
Product of Gaussian Mixture Diffusion Model for non-linear MRI Inversion	Laurenz Nagler, Martin Zach, Thomas Pock	2025	Not interpreting diffusion mechanisms
Quantifiable Quantization Sensitivity of Diffusion Models	Keith G. Mills, Mohammad Salameh, Ruichen Chen, Negar Hassanpour, Wei Lu, Di Niu	2024	Edge case exclusion - more downstream than diffusion interpretability
Reconstruction of patient-specific confounders in AI-based radiologic image interpretation using generative pretraining	Han T.; Žigutyte L.; Huck L.; Huppertz M.S.; Siepmann R.; Gandelsman Y.; Blüthgen C.; Khader F.; Kuhl C.; Nebelung S.; Kather J.N.; Truhn D.	2024	Edge case exclusion - more downstream than diffusion interpretability
Residual Denoising Diffusion Models	J. Liu; Q. Wang; H. Fan; Y. Wang; Y. Tang; L. Qu	2024	Not interpreting diffusion mechanisms
RightSizing: Disentangling Generative Models of Human Body Shapes with Metric Constraints	Wu Y.; Shu C.; Pai D.K.	2024	Not interpreting diffusion mechanisms
SAeUron: Interpretable Concept Unlearning in Diffusion Models with Sparse Autoencoders	Bartosz Cywiński, Kamil Deja	2025	Edge case exclusion - more downstream than diffusion interpretability
Sequential Data Generation with Groupwise Diffusion Process	Sangyun Lee, Gayoung Lee, Hyunsu Kim, Junho Kim, Youngjung Uh	2023	Rejected by ICLR, not yet published elsewhere
SkillDiffuser: Interpretable Hierarchical Planning via Skill Abstractions in Diffusion-Based Task Execution	Z. Liang; Y. Mu; H. Ma; M. Tomizuka; M. Ding; P. Luo	2024	Not interpreting diffusion mechanisms
StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models	Z. Wang; L. Zhao; W. Xing	2023	Not interpreting diffusion mechanisms
SymmCD: Symmetry-Preserving Crystal Generation with Diffusion Models	Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, Siamak Ravanbakhsh	2025	Not interpreting diffusion mechanisms
Temporal Knowledge Graph Reasoning Based on Diffusion Probability Distribution	Zhou G.-Y.; Li P.-F.; Xie P.-H.; Luo C.-Y.	2024	Could not access full text version in English

Title	Authors	Year	Reason/Notes
TIDE: Training Locally Interpretable Domain Generalization Models Enables Test-time Correction	Aishwarya Agarwal, Srikrishna Karanam, Vineet Gandhi	2024	Not interpreting diffusion mechanisms
TrackDiffuser: Nearly Model-Free Bayesian Filtering with Diffusion Model	Yangguang He, Wenhao Li, Minzhe Li, Juan Zhang, Xiangfeng Wang, Bo Jin	2025	Not interpreting diffusion mechanisms
Trade-Offs in Fine-Tuned Diffusion Models between Accuracy and Interpretability	Dombrowski M.; Reynaud H.; Müller J.P.; Baugh M.; Kainz B.	2024	Not interpreting diffusion mechanisms
Transient Stability Assessment Based on Imbalanced Sample Enhancement of Denoising Diffusion Probabilistic Model	Li Y.; Liu J.; Liu J.; Wang G.; Mo T.; Lin K.	2024	Could not access full text version in English
Unpacking SDXL Turbo: Interpreting Text-to-Image Models with Sparse Autoencoders	Viacheslav Surkov, Chris Wendler, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre	2024	Rejected and withdrawn post peer-review
Unveiling Concept Attribution in Diffusion Models	Quang H. Nguyen, Hoang Phan, Khoa D. Doan	2025	Rejected and withdrawn post peer-review
Unveiling Deepfakes with Latent Diffusion Counterfactual Explanations	C. Yang; B. Peng; J. Dong; X. Zhang	2025	Edge case exclusion - more downstream than diffusion interpretability
Variational Diffusion Method for Remote Sensing Image Fusion	C. Zhang; J. Han; J. Zhu; Z. Wang	2024	Not interpreting diffusion mechanisms
Wasserstein proximal operators describe score-based generative models and resolve memorization	Benjamin J. Zhang, Siting Liu, Wuchen Li, Markos A. Katsoulakis, Stanley J. Osher	2024	Theoretical insight, but not primarily aimed at interpretability
X-IQE: eXplainable Image Quality Evaluation for Text-to-Image Generation with Visual Large Language Models	Yixiong Chen, Li Liu, Chris Ding	2023	Not interpreting diffusion mechanisms
XMOL: Explainable Multi-property Optimization of Molecules	Aye Phyu Phyu Aung, Jay Chaudhary, Ji Wei Yoon, Senthilnath Jayavelu	2024	Not interpreting diffusion mechanisms

Appendix D
Citation chained Papers

TABLE XI
Papers included from citation chaining or free searches.

Included paper and ref.	Source
Baranchuk *et al.,* 2021 [93]	Citation chained from Burgess *et al.,* 2024 [91]
Choi *et al.,* 2022 [53]	Google Scholar search result
Deja *et al.,* 2022 [66]	Google Scholar search result
Basu *et al.,* 2023 [112]	Citation chained from Helbling *et al.,* 2025 [115]
Brack *et al.,* 2023 [86]	Citation chained from Kim *et al.,* 2024 [70]
Kadkhodaie *et al.,* 2023 [62]	Citation chained from Helbling *et al.,* 2025 [115]
Park *et al.,* 2023 [81]	Google Scholar search result
Prasad *et al.,* 2023 [61]	Citation chained from Prasad *et al.,* 2024 [59]
Raya & Ambrogioni, 2023 [64]	Citation chained from Li & Chen, 2024 [69]
Tumanyan *et al.,* 2023 [83]	Citation chained from Kim *et al.,* 2024 [70]
Voynov *et al.,* 2023 [111]	Citation chained from Burgess *et al.,* 2024 [91]
Wang *et al.,* 2023 [89]	Citation chained from Zeng *et al.,* 2024 [105]
Wu *et al.,* 2023 [90]	Google Scholar search result
Xu *et al.,* 2023 [73]	Google Scholar search result
Zhu *et al.,* 2023 [84]	Citation chained from Park *et al.,* 2025 [17]
Basu *et al.,* 2024 [113]	Google Scholar search result
Sclocchi *et al.,* 2024 [71]	Citation chained from Li & Chen, 2024 [69]