Computer Vision and Image Understanding 270 (2026) 104815

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Exploring real–synthetic boundaries in diffusion latent space☆

Simone Bonechi [a,*], Paolo Andreini [a], Barbara Toniella Corradini [b]

[a] Department of Information Engineering and Mathematics, University of Siena, Siena 53100, Italy
[b] AI for Good, Italian Institute of Technology (IIT), Genoa 16152, Italy

ARTICLE INFO

Keywords:
Diffusion Model
Stable Diffusion
AI-generated content representation
Latent representation properties

ABSTRACT

This study investigates whether pre-trained Diffusion Models (DMs) inherently encode in a different way real and synthetic objects in an image within their latent representations. While DMs have achieved state-of-the-art performance in generative modeling and demonstrated rich semantic representations despite being trained solely with denoising objectives, their widespread accessibility raises critical concerns about detecting AI-generated content. Motivated by the intuition that a model must implicitly capture the distribution of real data to generate it, we hypothesized that real and synthetic objects are mapped to separable regions of the latent space. To test this hypothesis, we conduct a systematic analysis using a pre-trained DM to encode both real objects from the COCO and the Pascal VOC datasets and synthetic samples from DALL-E 3, Midjourney, and Stable Diffusion. We extract internal representations from different layers and evaluate their discriminative capacity through distance-based metrics and linear probing. Our findings reveal that decoder features, in particular, induce statistically meaningful representational differences between real and synthetic objects. The results show that these representations support high linear-probing accuracy under the evaluated experimental conditions, suggesting that pre-trained diffusion representations contain useful signals for distinguishing real and synthetic content.

1. Introduction

In recent years, Diffusion Models (DMs) (Ho et al., 2020) have emerged as a powerful class of generative models, achieving state-of-the-art performance in synthesizing highly diverse, detailed, and realistic content. Among them, Stable Diffusion Models (SDMs) (Rombach et al., 2022) stand out for their ability to generate images that are nearly indistinguishable from real ones and for its versatility in accepting conditioning (e.g., textual prompts and segmentation masks). By conducting the diffusion process in a compressed latent space, SDMs achieve high-quality image synthesis with substantially reduced computational requirements. This balance of efficiency and generative fidelity has contributed to the widespread deployment of SDMs in research, industry, and creative domains.

Although DMs are primarily designed and trained for image generation, a growing body of work has shown that they possess intrinsic properties that make them effective for a variety of downstream vision tasks. Indeed, despite being optimized primarily for noise prediction, their extensive training on large-scale image–text datasets (Schuhmann et al., 2022, 2021) enables them to learn semantically rich and informative representations. Mirroring phenomena observed in large language models (Wei et al., 2022), these internal representations exhibit emergent properties that facilitate tasks such as semantic segmentation (Baranchuk et al., 2022; Corradini et al., 2025). Motivated by these observations, in this work, we investigate whether diffusion models exhibit an emergent capacity to encode real and synthetic objects in separable regions of their latent space.

Demonstrating the validity of this idea would provide a deeper understanding of generative model behavior, with significant theoretical implications for AI-generated content detection, data curation, and adversarial defense. This exploratory study examines this hypothesis through an extensive experimental investigation, in particular, we investigate whether the object representations extracted from real datasets such as COCO (Lin et al., 2014) and Pascal-VOC (Everingham et al., 2010) are distinguishable from the representations of objects generated by DALL-E 3 (Betker et al., 2023), Midjourney (Midjourney, 2022), and different versions of Stable Diffusion (v1.4 and v1.5) (Rombach et al., 2022). We explore various layers of the diffusion model's U-Net architecture, examining the separability between the internal representations of real and generated objects. Following Bonechi et al. (2025, 2024), we first segment the image to extract each object, and then we employ a pre-trained SDM v1.5 to obtain a prototype vector of each of them in real and AI-generated images. Finally, we compare the

☆ This article is part of a Special issue entitled: 'YCVIU_MF-DISM' published in Computer Vision and Image Understanding.

* Corresponding author.
E-mail address: simone.bonechi@unisi.it (S. Bonechi).

https://doi.org/10.1016/j.cviu.2026.104815
Received 9 December 2025; Received in revised form 12 May 2026; Accepted 16 May 2026
Available online 28 May 2026

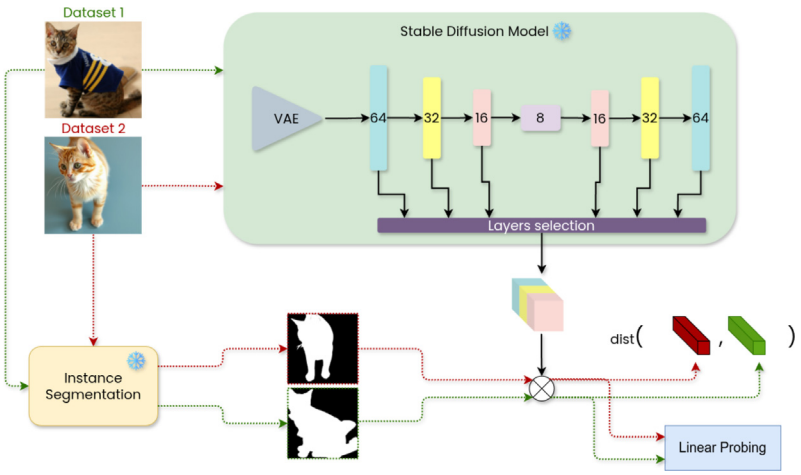


Fig. 1. Overview of the framework employed for analyzing the internal representation of the SDM. For each input image, we extract internal features from multiple SDM layers and filter them using object segmentation masks. The filtered features are then averaged to create object prototypes, which we compare by measuring distances in the feature space and training a linear classifier to assess separability. This approach enables direct comparison of how real and synthetic object instances are represented within the model.

objects' prototypes to evaluate whether there is a difference between the encoding of the two types of data (see Fig. 1). Our analysis shows that among all the examined layers, the decoder features at spatial resolutions 16×16 and 32×32 exhibit the strongest discriminative behavior between real and synthetic objects in our evaluation. Aggregating prototypes from these two layers allows a linear classifier to achieve over 99% accuracy, indicating substantial real–synthetic separability under our experimental conditions.

Moreover, we conduct additional experiments to further assess our approach under alternative setups that commonly arise in real-world scenarios. First, we evaluate the features derived from Stable Diffusion 1.5 in comparison to those extracted via other prominent backbones, specifically transformer-based models (CLIP Radford et al., 2021 and DINO Caron et al., 2021) and Convolutional Neural Networks (CNNs) (EfficientNet-B0 Tan and Le, 2019 and ResNet-50 He et al., 2016) backbones. Second, since new generators continuously emerge, we evaluate generalization to unseen generators via a cross-generator protocol. Then, we relax the assumption that segmentation masks are available by replacing object-level representations with image-level embeddings. We demonstrate that the separability between real and synthetic samples is maintained even at this image level. We also prove that the object-level granularity of our analysis ensures that the resulting insights remain applicable across tasks—especially those involving localized edits (e.g., modifying a person or an object in a particular location) that are typical of real-world scenarios. Indeed, beyond image generation, modern systems increasingly support image editing (Y. Huang et al., 2025; Brooks et al., 2023), reference-based personalization (Zhou et al., 2024), and video synthesis (Ho et al., 2022; J. Huang et al., 2025), with more recent efforts pushing toward unified frameworks that jointly integrate multimodal understanding, generation, and editing (Lu et al., 2024; Qin et al., 2025). Finally, we provide additional evidence that the collected dataset is not substantially affected by bias related to the segmentation masks and that, despite its limited size, the sample remains statistically significant.

To our knowledge, this study presents the first experimental evidence that specific layers within an SDM encode real and synthetic objects into linearly separable latent representations. This finding advances our understanding of how SDMs internally structure semantic information and lays the groundwork for future content-verification and detection methods that exploit the intrinsic properties of diffusion-based feature spaces.

The paper is organized as follows: Section 2 reviews relevant literature on diffusion models, while Section 3 details our methodology, including dataset selection and preprocessing. Section 4 describes the experimental setup, while Section 5 presents the obtained results. Finally, Sections 6, and 7 discuss our findings, draw conclusions, and highlight future research directions.

2. Related work

This section reviews the literature on the evolution of DMs (Section 2.1) and the use of their internal representation in tasks beyond image generation (Section 2.2).

2.1. Evolution of diffusion models

First introduced by Sohl-Dickstein et al. in Sohl-Dickstein et al. (2015), DMs have represented a breakthrough in generative AI, as they have surpassed earlier generative approaches in both image quality and diversity, producing outputs that can be nearly indistinguishable from real images. Their success stems from a unique training approach that learns to reverse a gradual noising process, allowing for more stable training compared to adversarial methods (Dhariwal and Nichol, 2021). In particular, DMs rely on a forward process that progressively adds noise to images and a reverse process guided by a U-shaped neural network (U-Net), which systematically removes the noise to produce high-fidelity images from random noise.

Early implementations, such as the Denoising Diffusion Probabilistic Models (DDPMs) by Ho et al. (2020), implemented both the forward and reverse processes as Markov chains. While effective, this approach involved calculating the distribution at each step, which made the computational cost nearly impractical for broader applications. To overcome this limitation, Denoising Diffusion Implicit Models (DDIMs) (Song et al., 2021) introduced a non-Markovian framework, significantly reducing computational demands and allowing the generation of higher-resolution images. Nevertheless, both DDPMs and DDIMs operate directly in the high-dimensional pixel space, resulting in much higher computational costs compared to alternative generative models like Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs). A breakthrough came with Latent Diffusion Models (LDMs) (Rombach et al., 2022), which used a VAE to encode images into a lower-dimensional latent space. In this architecture, the diffusion process occurs in the latent space using a U-Net (also known as Diffusion U-Net), after which the VAE decoder converts the processed vector back to pixel space. This approach not only enhanced computational efficiency and reduced resource requirements but also

enabled new capabilities, such as generating images conditioned on additional inputs (Ho and Salimans, 2021), e.g. textual prompts or reference images. In this work, we aim to explore a pre-trained diffusion model known as SDM (Rombach et al., 2022), a type of text-to-image LDM that has emerged as the state-of-the-art in image generation due to its capability to produce photorealistic images based on any given textual prompt.

2.2. Exploration and use of the DMs internal representation

To achieve state-of-the-art performance, DMs typically rely on extremely large datasets containing hundreds of millions (Schuhmann et al., 2021) or even billions of examples (Schuhmann et al., 2022). The unaffordable scale of data and computational resources required to train a DM often makes re-training these models from scratch infeasible. This limitation has encouraged the development of new research fields focused on exploring *pre-trained diffusion models* to uncover properties within their internal representations that can be leveraged to enhance their generative capabilities (for instance, through textual prompt engineering Hao et al., 2023) or to tackle downstream tasks, including image classification and segmentation, all without the need for re-training. In Baranchuk et al. (2022), it is demonstrated that applying K-Means clustering to DDPM features produces a spatially coherent representation useful for downstream tasks, particularly semantic segmentation. They analyze different U-Net decoder layers across diffusion timesteps to identify the most effective for segmentation, then use their embedding vectors to train a Multi-Layer Perceptron for pixel-wise labeling. FreeSeg-Diff (Corradini et al., 2025) introduces a fully training-free approach to open-vocabulary segmentation by leveraging internal representations from SDM, outperforming several supervised and zero-shot methods. In Mukhopadhyay et al. (2023), the authors demonstrate that DMs can surpass GANs in image classification by leveraging their internal feature representations. Instead of using them solely for generation, the study extracts and pools discriminative features from the U-Net backbone of diffusion models to train a classifier. In ASYRP paper (Kwon et al., 2023), the authors explore the internal representations of pretrained diffusion models and introduce a semantic latent space termed "h-space". This space exhibits homogeneity, linearity, robustness, and consistency across timesteps, making it suitable for straightforward image manipulation. In Pnvr et al. (2023), the authors introduce LD-ZNet, a novel method for text-based segmentation of both real and AI-generated images by leveraging the latent space of LDMs. The key innovation lies in using the latent representations (z-space) learned by LDMs as input features for segmentation tasks, instead of traditional inputs like RGB images or CLIP embeddings. LD-ZNet significantly improves segmentation performance compared to traditional methods. In L. Zhang et al. (2023), a neural network is trained to modify the features of a pre-trained SDM to follow specific multimodal conditioning and improve generated image quality. The proposed model, ControlNet, adds spatial conditioning (e.g., edges, depth maps, segmentation maps, poses) to pretrained text-to-image diffusion models. By freezing the original model's parameters and adding trainable layers, ControlNet enables precise control over image generation while maintaining stability during fine-tuning. More related to this work, in Bonechi et al. (2024, 2025) the authors extract the prototypes of objects belonging to different classes in various datasets to explore the capability of different layers of an SDM to preserve semantics.

3. Methods

The four datasets employed in this study are presented in Section 3.1, whereas the procedure used to extract the prototypes from each image is described in Section 3.2. Finally, Section 3.3 details the data pre-processing steps for generating instance segmentation maps of AI-generated images and selecting the subset of real and synthetic images.

3.1. Datasets

To compare the internal representations of real and AI-generated objects, our investigation leverages four datasets, comprising two sets of authentic images and two sets of synthetically generated images. Finally, to evaluate the proposed approach on edited images, we also include one dataset containing manipulated images. Below, we summarize the key characteristics of each dataset.

Real image datasets.

- Pascal-VOC (Everingham et al., 2010) is a widely used benchmark for image semantic segmentation. Its images feature a limited set of object categories, typically with the primary object centered in the scene. The dataset includes 20 object categories plus a background class and a "do not care" class for uncertain regions. Both training and validation sets provide pixel-level annotations.
- COCO (Lin et al., 2014) is a comprehensive dataset containing over 100,000 images designed for object detection, segmentation, and captioning tasks. Unlike Pascal-VOC, COCO images often capture complex scenes with multiple objects. The dataset provides instance-level annotations for 80 object categories and background labels across its training and validation sets.

Synthetic image datasets.

- GenImage (Zhu et al., 2024) is a large-scale dataset specifically created for detecting AI-generated images. It comprises approximately 1.35 million synthetic images generated by eight different AI models: BigGAN (Brock et al., 2019), GLIDE (Nichol et al., 2022), VQDM (Gu et al., 2022), Stable Diffusion (v1.4 and v1.5) (Rombach et al., 2022), ADM (Dhariwal and Nichol, 2021), Midjourney (2022), and Wukong (2022). The dataset includes the 1000 ImageNet (Deng et al., 2009) classes, with each class containing roughly 1350 fake images (1300 for training and 50 for testing). Images were generated using simple text prompts following the template "photo of [class]" and feature resolutions ranging from 128×128 to 1024×1024, depending on the generator model.
- ImagiNet (Boychev and Cholakov, 2025) contains 200,000 high-resolution images balanced between real and synthetic content. The synthetic images are generated using various methods, including GANs (Goodfellow et al., 2020), DM (Ho et al., 2020), and commercial tools like Midjourney (2022) and DALL-E 3 (Betker et al., 2023), while real images are collected from public datasets. The dataset supports two evaluation tasks: distinguishing between real and synthetic images and identifying the specific generative model used for generation.

Image editing dataset.

- MagicBrush (K. Zhang et al., 2023) is a large-scale, manually annotated dataset designed specifically for instruction-guided real image editing. Created to overcome the limitations of noisy, automatically synthesized training data, MagicBrush features over 10,000 high-quality triplets consisting of a source image, a natural language instruction, and a target image. The dataset is highly versatile, covering diverse, real-world editing scenarios such as single-turn, multi-turn, mask-provided, and mask-free editing.

3.2. Prototype extraction

Fig. 1 illustrates our step-by-step procedure (inspired by Bonechi et al. (2025)) for extracting feature representations and comparing the embeddings of real and AI-generated images. This process is summarized below:

(a) **Image Encoding:** Each input image is encoded in the latent space of the SDM at inference step $t = 0$, i.e., the final denoising step in the diffusion U-Net.

Table 1
Datasets preparation. Number of images in each dataset used in our analysis.

Dataset	Number of images
Pascal-VOC	654
COCO	1000
DALL-E 3	709
Stable Diffusion v1.4	566
Stable Diffusion v1.5	547
Midjourney	561

(b) **Feature Extraction:** Features are extracted from the selected residual blocks of the U-Net.
(c) **Feature Masking:** Instance segmentation masks are applied directly in the feature space to isolate the representations of specific objects.
(d) **Object Prototyping:** An object-level descriptor is obtained by averaging the masked region across spatial dimensions, resulting in a prototype vector that encapsulates the characteristic features of each object instance.

Additionally, to demonstrate that our analysis can be applied even without relying on segmentation masks, we introduce an alternative extraction pipeline that computes prototypes for the entire image. In this setting, we do not apply the feature masking (step (c)), and we directly compute the average of the entire feature map in order to obtain an Image Prototype instead of the Object Prototype (step (d)).

3.3. Data pre-processing

Following the procedure described in Section 3.2, we examine the features extracted from the four datasets described in Section 3.1. Since GenImage and ImagiNet lack instance segmentation labels, we employ a pre-trained segmentation network to generate the required instance segmentation maps. We utilize Mask2Former (Cheng et al., 2022), a universal image segmentation architecture capable of handling multiple segmentation tasks (panoptic, instance, and semantic) within a single framework. The model comprises a feature extractor, a pixel decoder, and a transformer decoder, with its key innovation being masked attention that restricts attention to localized features around predicted segments. We specifically use the large-sized version of Mask2Former pre-trained on COCO instance segmentation with a swin backbone (Liu et al., 2021), available on Hugging Face.[1] To ensure the goodness of the segmentation masks, the predicted label maps are visually inspected to remove those with inaccurate segmentation results. While GenImage and ImagiNet contain images from various generators, we decided to focus our analysis on recent models, excluding outdated architectures. More specifically, we select images generated with four specific models: Midjourney, Stable Diffusion v1.4, Stable Diffusion v1.5 (from GenImage), and DALL-E 3 (from ImagiNet). Since the real and synthetic datasets contain a wide variety of classes, we limit our analysis to seven common objects shared across all datasets: cat, dog, person, car, bird, boat, and train. The background and the object classes excluded from this selection are considered together. To ensure meaningful analysis, we select only images where objects from these classes occupy more than 10% of the image area, ensuring the objects are visible and well-represented. This selection criteria yield a total of 4037 images; the number of images in each dataset is reported in Table 1:

While we acknowledge that the reliance on an external segmentation network and the restriction to a specific subset of classes impose certain limitations, these methodological choices are functional to the exploratory nature of our study, allowing us to isolate the signal of "realness" at object and class level from potential noise introduced by semantics. These constraints could be easily eliminated in real-world applications that will exploit our findings employing the image prototype extracted as described in Section 3.2. Furthermore, to further assess the potential bias introduced by our image selection procedure and the adequacy of the sample size for the proposed analysis, we collect an extended dataset of 32,000 images (4000 synthetic images per generator, alongside 16,000 real images from the COCO and Pascal datasets). For this supplementary set, we do not manually select the segmentation masks for quality assurance, nor do we apply filters for object size or class.

4. Experimental setup

Our primary goal is to investigate the discriminative power of features extracted from different layers of the diffusion U-Net between real and AI-generated objects. We hence develop an experimental framework based on prototype representations computed from both real and synthetic images, following the extraction procedure detailed in Section 3.2.

As our feature extractor, we employ a pre-trained open-source implementation[2] of the Stable Diffusion Model (v1.5), which processes input images standardized to a resolution of 512×512 pixels. We analyze features from both encoder and decoder layers at spatial resolutions of 64×64, 32×32, and 16×16, extracting object-level prototypes from each layer. Once the object-level prototypes are computed, we rely on two complementary approaches to conduct our analyses: *feature distance*, which quantifies how far apart features from different datasets lie in the representation space, and *linear probing*, which tests whether a simple linear classifier can separate those features. While high feature distance indicates strong features' dissimilarity, it does not guarantee that the separation is easy to exploit for classification. Conversely, linear probing directly assesses the separability by checking how well features align with a linear decision boundary.

First, we assess the separability of the features by computing the pairwise distance distribution between prototypes extracted from real and generated objects encoded into different U-Net layers (Section 4.1): the closer the distributions, the more similar the features involved are. To this end, we consider encoder and decoder layers at multiple spatial resolutions (64×64, 32×32, and 16×16), and compute distances across three types of object feature pairs: Real vs. Fake, Fake vs. Fake, and Real vs. Real. Second, we employ a linear classifier to quantify how well these features discriminate between real and AI-generated objects in terms of classification accuracy (Section 4.2). Based on these analyses, we identify the layers with the strongest discriminative potential and carry out a more detailed layer-wise investigation in Section 4.3. Finally, to assess whether our approach remains effective in more realistic settings, we conduct additional tests using features extracted from the selected layers:

- To determine whether the discriminative characteristics of these features are robust across unseen datasets, we employed a cross-generator evaluation approach (Section 4.4);
- We utilized the extended dataset to provide additional evidence that the obtained results are robust to variations in sample size, manual curation of the segmentation masks, and the filters applied to object class and size (Section 4.5);
- To assess whether relying on object prototypes constitutes a practical limitation, we extracted image prototypes and trained a linear classifier to evaluate their linear separability (Section 4.6). Indeed, we provide empirical evidence that object-level analysis is highly beneficial for tasks such as image manipulation detection, where localized evaluation is essential for identifying altered subparts (Section 4.7).
- We compare the features extracted from Stable Diffusion 1.5 with those obtained using alternative transformer-based and CNN-based backbones (Section 4.8);

[1] https://huggingface.co/facebook/mask2former-swin-large-coco-instance.

[2] https://github.com/hkproj/pytorch-stable-diffusion.

4.1. Feature distance analysis

To compare object prototypes across datasets, we employ Distance Correlation (Székely and Rizzo, 2007), a statistical measure that captures both linear and non-linear associations between feature representations, making it suitable for assessing structural similarity in high-dimensional spaces. For two n-dimensional vectors $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ it is defined as:

$$D_{\text{corr}}(\mathbf{X}, \mathbf{Y}) = \sqrt{\frac{dCov^2(\mathbf{X}, \mathbf{Y})}{\sqrt{dVar^2(\mathbf{X}) dVar^2(\mathbf{Y})}}},$$

where the distance covariance $dCov^2(\mathbf{X}, \mathbf{Y})$ and the distance variance $dVar^2(\mathbf{X})$ between $\mathbf{X}$ and $\mathbf{Y}$ are

$$dVar^2(\mathbf{X}) = dCov^2(\mathbf{X}, \mathbf{X})$$

$$dCov^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a})(b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b})$$

If x_i, x_j are two samples of $\mathbf{X}$ and y_k, y_l are two samples of $\mathbf{Y}$, it is possible to define:

$$a_{ij} = D_{euc}(x_i, x_j), \quad b_{kl} = D_{euc}(y_k, y_l), \quad \bar{a}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} a_{ij},$$

$$\bar{a}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} a_{ij}, \quad \bar{b}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} b_{ij}, \quad \bar{b}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} b_{ij}$$

To assess how effectively different layers capture discriminative features between real and synthetic objects, we compute three types of pairwise distances across our datasets:

- *Real vs. Fake* — Distances between object prototypes extracted from real images (Pascal-VOC or COCO) and synthetic images generated by AI models (DALL-E 3, Midjourney, Stable Diffusion v1.4 and v1.5).
- *Fake vs. Fake* — Distances between object prototypes within synthetic images generated by AI models.
- *Real vs. Real* — Distances between object prototypes within real images.

The discriminative power of a layer is evaluated by analyzing the separation among these distributions: an effective layer should exhibit greater *Real vs. Fake* distances compared to both *Fake vs. Fake* and *Real vs. Real* distances. Since the obtained pairwise distance dataset contains approximately 26 million pairwise observations per layer, the standard null-hypothesis significance testing invariably yields uninformative p-values approaching numerical underflow. Consequently, we adopt an effect-size-centric statistical analysis. Global differences across the three distributions are evaluated using the non-parametric Kruskal–Wallis test (Kruskal and Wallis, 1952), relying on Epsilon-squared (ϵ^2) to quantify the global effect size, which indicates the proportion of total variance explained by the pair configuration. For post-hoc pairwise comparisons, we utilized Cliff's Delta (δ) (Cliff, 1993) as a robust, non-parametric measure to assess the practical magnitude of the differences between groups, allowing us to evaluate the actual discriminative utility of each layer independent of sample size constraints.

4.2. Feature linear probing

To further assess the information captured by each layer of the SDM, we employ a linear probing technique. This common approach evaluates the quality of features extracted by deep neural networks by training a simple linear classifier on them (Alain and Bengio, 2017; Chen et al., 2020; Grill et al., 2020; Caron et al., 2021). Our linear classifier consists of a single linear unit followed by a sigmoidal activation function. This classifier is trained for 1000 epochs using binary cross-entropy loss and Stochastic Gradient Descent (SGD) with a learning rate $\lambda = 0.01$ and momentum $\mu = 0.9$. The model is trained to distinguish prototypes of real and generated objects, i.e. derived from real (COCO and Pascal-VOC) datasets or fake (DALL-E 3, Midjourney, Stable Diffusion v1.4, and v1.5) image sources.

For this evaluation, before the prototype extraction, we partition the images into distinct training and test splits (85% training and 15% test) and fit the linear probe for a predetermined, fixed number of epochs, following standard practice. Given the potential for imbalanced datasets, we use both accuracy and balanced accuracy as the performance metric for the linear classifier across all evaluated scenarios.

4.3. Fine-grained study of feature similarities across datasets

Leveraging initial findings that show promising discriminative capabilities, we conduct a fine-grained analysis to identify the most effective SDM layer(s) for distinguishing real from synthetic objects. This phase moves from broad cross-dataset comparisons to focused, dataset-specific analyses, examining features extracted independently from each source. This allows for not only comparing real and synthetic object features but also identifying similarities among objects generated by the same model and comparing them against objects' features from different image generators. This evaluation helps determine the potential of using the features to both identify AI-generated content and attribute it to a specific generator. We utilize both distance metrics, as in Section 4.1 and a linear probing approach as in Section 4.2 to train a *one vs. all classifier* designed to recognize objects generated by DALL-E 3, Midjourney, and Stable Diffusion (grouping versions v1.4 and v1.5). Our intuition is that features from objects created by the same generator would exhibit greater mutual similarity compared to features from real objects or those from other AI tools. Consequently, the linear classifier should be able to accurately identify the source generator.

Furthermore, to provide a visual depiction of these relationships, we employ UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction (McInnes et al., 2018) to plot the feature distributions from real and AI-generated data. With the UMAP analysis, we also investigate how image backgrounds are encoded by plotting their features separately from those of the primary objects. This detailed analysis will help to reveal if different generative models encode objects in distinct ways. This unique encoding pattern, if present, could be potentially used as a "fingerprint" for tasks such as image attribution or AI-generated content detection.

4.4. Cross-generator evaluation

In real-world scenarios, collecting large-scale datasets from every emerging generator is often infeasible. Therefore, it is crucial to evaluate whether the discriminative object features extracted from images of one generator generalize to others in a cross-generator setting. To assess this, we construct a dataset of object features from a single source generator and object prototypes from real images to train a binary classifier (real vs. fake). We then test the classifier's ability to correctly identify fake objects produced by unseen generators. We evaluate this generalization capability using two distinct methods: (i) the linear classifier described in Section 4.2, and (ii) a k-Nearest Neighbor (k-NN) approach.

4.5. Dataset evaluation

To conduct our pairwise distance analysis, we utilize a smaller dataset, as the number of pairwise combinations grows prohibitively large even with a limited number of images. To ensure high-quality feature extraction, we manually curate the segmentation masks generated by Mask2Former and filter the objects based on class and size. To assess the potential impact of these selection criteria and the extent to which our findings generalize beyond the reduced setting, we perform the following experiments:

- We compare the linear probing results using features extracted from the best layer(s) on the extended dataset. In this setup, we omit the manual mask curation but retain the class and size filters. This experiment allows us to assess whether the reduced dataset remains broadly consistent with the larger setting and whether the results are substantially affected by manual selection of the segmentation outputs.
- We remove the class and dimension filters on both datasets (the reduced version from the primary experiments and the extended version) to evaluate the direct impact of these filters on linear probing accuracy.

4.6. Image prototype evaluation

While using a segmentation network to obtain object prototypes allows for fine-grained analysis, it may present a significant bottleneck in real-world applications due to computational overhead and low segmentation quality. To address this potential limitation, we also extract prototypes at the image level (as described in Section 3.3). We then evaluate the linear separability of these image features, following the same methodology used for object-level features in Section 3.2.

4.7. Evaluation on image editing

We utilize approximately 8800 images from the MagicBrush dataset, randomly splitting them into an 85% training set and a 15% testing set, along with their corresponding masks used for manipulation localization. Using these masks, we extract distinct prototypes from the tampered images to isolate the manipulated objects from the authentic regions. Finally, we apply linear probing to assess the separability of the extracted features, aiming to determine the relevance of our object-level analysis in a localized manipulation setting.

4.8. Comparison with other backbones

To evaluate the quality and discriminative power of the features extracted from Stable Diffusion 1.5, we benchmark our approach against a diverse set of widely adopted feature-extraction backbones. Specifically, we compute image prototype representations using our Stable Diffusion method and compare them against corresponding prototypes derived from two transformer-based models (CLIP and DINO) and two CNN architectures (EfficientNet-B0 and ResNet-50). We compare the features extracted from the two variants of our dataset (the reduced version used for all the experiments presented in the paper and the extended version comprising 32,000 images) using linear probing.

5. Results

Following the experimental setup detailed in Section 4, we analyze various layers of the diffusion U-Net to identify those that most effectively separate feature representations of real and AI-generated objects. Section 5.1 presents our analysis of the pairwise distances between object-level prototypes. Next, Section 5.2 details the performance of a linear classifier trained to distinguish between features of real and synthetic objects extracted from different U-Net layers. Additionally, Section 5.3 reports on a focused investigation of the specific layer(s) that exhibit the strongest discriminative capabilities. Furthermore, we present a comprehensive evaluation of the extracted features, showing their robustness and generalizability through cross-generator testing (Section 5.4) and experiments on an extended, uncurated dataset (Section 5.5). Additionally, we validate the versatility of our approach by exploring both image- and object-level prototypes (Section 5.6)—highlighting their utility in localized tasks like manipulation detection (Section 4.7)—and conclude by comparing Stable Diffusion against alternative CNN and transformer architectures as feature extractors for synthetic image detection (4.8).

5.1. Investigating real–synthetic prototype distances

We first characterize real and synthetic objects by analyzing the distribution of pairwise distances between their corresponding representations. Fig. 2 presents box plots of these feature distances (characterized by their median values and interquartile ranges), illustrating how effectively each U-Net layer captures the distinctions between authentic and AI-generated content. Table 2 presents the descriptive statistics—including the mean, median, and standard deviation characterizing the pairwise distance distributions across the evaluated encoder and decoder layers of the diffusion U-Net.

The massive scale of the dataset renders standard p-values uniformly near zero across all Kruskal–Wallis tests; thus, our comparative evaluation relies strictly on effect sizes to assess practical significance. Accordingly, Table 3 details the discriminative capability of each layer, reporting the global effect size (Epsilon-squared, ϵ^2) to quantify the total variance explained by the domain pairing, alongside the absolute values of Cliff's Delta ($|\delta|$) for the post-hoc pairwise comparisons.

Based on these results, the most discriminative features are extracted from the first layer of the U-Net decoder at 16×16 resolution (Decoder, 1st Layer, 16×16). This layer exhibits the highest global effect magnitude ($\epsilon^2 = 0.1153$, indicating that over 11.5% of the total distance variance is explained by the domain configuration), accompanied by a Cliff's δ value of 0.42 indicating a good separation between mixed and purely synthetic pairs.

5.2. Assessing real–synthetic separability via linear probing

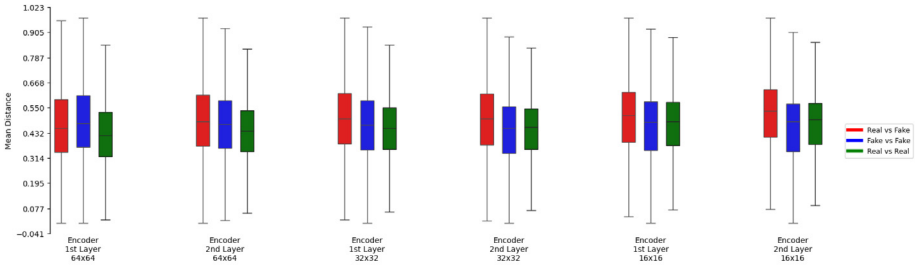
To assess the efficacy of features extracted from different network layers in differentiating real from AI-generated objects, we employed a linear classification methodology, as described in Section 4.2. A comparison between the linear classification results presented in Table 4 and the feature distance analysis (Section 5.1) reveals a subtle but meaningful divergence. Features extracted from the encoder at 16×16 resolution and from the decoder at both 16×16 and 32×32 resolutions exhibit good linear separability. However, this does not perfectly align with the patterns observed through distance-based evaluation.

Specifically, the first decoder layer at 16×16 resolution shows the best separability according to the distance-based evaluation, suggesting that real and generated samples are mapped into distinct latent space regions, nevertheless the highest linear-classification accuracy is achieved by the first decoder layer at 32×32 resolution. This implies that, while the raw distances might be greater in the 16×16 layer, the arrangement of features in the 32×32 layer is more suitable to be divided by a simple linear boundary. These two different analytical perspectives can expose complementary properties of the feature space, highlighting the importance of a multifaceted evaluation strategy.

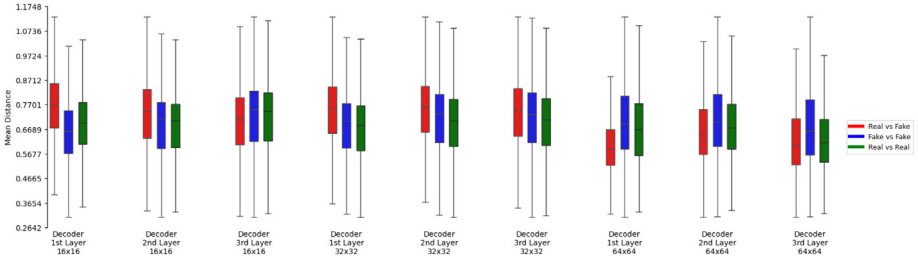
5.3. Feature similarities across datasets

Since distance-based analysis and linear probing capture different aspects of the feature space, with the former quantifying cluster proximity and the latter assessing linear separability, a comprehensive evaluation should incorporate both. We therefore construct a combined feature vector by concatenating outputs from the first decoder layers at 16×16 and 32×32 resolutions. Exploiting this feature vector we perform a detailed evaluation across multiple datasets, using both distance-based metrics and linear probing to assess its discriminative power.

The distance matrix in Fig. 3 shows the mean and standard deviation of inter-dataset feature distances. The diagonal entries represent distances between features of objects from the same dataset; lower values, therefore, imply that the corresponding representations within that dataset exhibit greater similarity. Notably, the self-comparison distances within synthetic objects are substantially lower than their distances to real ones, as evidenced by the values in the first two



(a) Distribution of distances between object features extracted by the diffusion U-Net Encoder.



(b) Distribution of distances between object features extracted by the diffusion U-Net Decoder.

Fig. 2. (Colour online) Assessing feature differentiation of real and AI-generated images across U-Net layers. Box plots show the distribution of pairwise mean distances between feature embeddings extracted from different U-Net layers, computed for three types of image pairs: *Real vs. Fake*, *Fake vs. Fake*, and *Real vs. Real*. Results are provided for encoder layers (a) and decoder layers (b), across resolutions 64 × 64, 32 × 32, and 16 × 16.

Table 2
Measuring the spontaneous separation of real and synthetic images via distance correlation. We report Mean, Median, and Standard Deviation (Std) for pairwise distance correlations within each resolution block of the Encoder and Decoder. Comparisons include homogeneous pairs (F–F: fake vs. fake, R–R: real vs. real) and heterogeneous pairs (R–F: real vs. fake). Higher R–F values indicate points in the network where the model spontaneously separates real from synthetic representations.

Block	Resolution	Group	1st layer			2nd layer			3rd layer		
			Mean	Median	Std	Mean	Median	Std	Mean	Median	Std
Encoder	64 × 64	F–F	0.497	0.479	0.181	0.677	0.674	0.198	–	–	–
		R–R	0.432	0.420	0.154	0.635	0.635	0.169	–	–	–
		R–F	0.478	0.455	0.186	0.702	0.689	0.211	–	–	–
	32 × 32	F–F	0.688	0.691	0.189	0.743	0.758	0.170	–	–	–
		R–R	0.669	0.674	0.162	0.748	0.762	0.147	–	–	–
		R–F	0.727	0.724	0.194	0.799	0.804	0.178	–	–	–
	16 × 16	F–F	0.777	0.802	0.140	0.849	0.882	0.125	–	–	–
		R–R	0.790	0.805	0.120	0.868	0.889	0.098	–	–	–
		R–F	0.817	0.828	0.138	0.899	0.915	0.112	–	–	–
Decoder	16 × 16	F–F	0.659	0.661	0.138	0.853	0.894	0.136	0.832	0.870	0.124
		R–R	0.693	0.696	0.137	0.862	0.887	0.105	0.838	0.866	0.100
		R–F	0.764	0.770	0.140	0.898	0.917	0.115	0.830	0.845	0.103
	32 × 32	F–F	0.798	0.826	0.129	0.789	0.819	0.138	0.764	0.788	0.140
		R–R	0.794	0.811	0.120	0.777	0.793	0.125	0.752	0.768	0.125
		R–F	0.860	0.872	0.125	0.830	0.843	0.127	0.790	0.801	0.131
	64 × 64	F–F	0.318	0.310	0.118	0.445	0.431	0.162	0.356	0.331	0.155
		R–R	0.295	0.290	0.113	0.417	0.405	0.145	0.303	0.288	0.121
		R–F	0.240	0.228	0.088	0.399	0.376	0.154	0.308	0.277	0.146

rows (and columns) of the matrix. For instance, the average distance between DALL-E 3 and Pascal-VOC samples is significantly higher than the distance between DALL-E 3 and Midjourney, illustrating that generated objects tend to cluster together and are more distant from real data. Additionally, given their architectural similarity, features from Stable Diffusion v1.4 and v1.5 exhibit significant overlap, making them difficult to distinguish. In contrast, DALL-E 3 features are highly distinct from both real objects and other synthetic sources, a separation likely attributable to its recognizable and consistent visual style (see Fig. 4).

To complement the distance analysis, we assess the linear separability of features concatenated from the first decoder layers at both 16 × 16 and 32 × 32 resolutions. By training a linear classifier, we achieve an accuracy of 99.26% (balanced accuracy: 99.13%), which represents an improvement of approximately one percentage point over using features solely from the 32 × 32 decoder layer. Additionally, we evaluate these features for source identification. For this task, we trained a series of *one vs. all* linear classifiers, where each is tasked with identifying objects generated by a specific model (e.g., DALL-E 3) from all other samples, including real images. Table 5 reports

Table 3
Quantifying the separation of real and synthetic representations via effect size analysis. We report Kruskal–Wallis global effect size (ϵ^2) and pairwise Cliff's Delta ($|\delta|$) values for representational distance correlations across network layers. Larger effect sizes indicate stronger separation between distributions. The row corresponding to the peak separation (1st layer, 16 × 16 Decoder) is bolded. Abbreviations: F–F (fake vs. fake), R–R (real vs. real), R–F (real vs. fake).

Block	Resolution	Layer	ϵ^2 (Global)	$\|\delta\|$ (R–F vs. F–F)	$\|\delta\|$ (R–F vs. R–R)	$\|\delta\|$ (F–F vs. R–R)
Encoder	64 × 64	1st	0.0120	0.07	0.13	0.21
		2nd	0.00830	0.06	0.17	0.12
	32 × 32	1st	0.01052	0.10	0.17	0.06
		2nd	0.0210	0.17	0.17	0.01
	16 × 16	1st	0.0150	0.15	0.12	0.03
		2nd	0.0370	0.24	0.20	0.05
Decoder	16 × 16	**1st**	**0.1153**	**0.42**	**0.30**	**0.15**
		2nd	0.0286	0.20	0.22	0.02
		3rd	0.0047	0.08	0.07	0.03
	32 × 32	1st	0.0532	0.27	0.31	0.05
		2nd	0.0222	0.16	0.24	0.09
		3rd	0.0088	0.09	0.17	0.09
	64 × 64	1st	0.0992	0.40	0.30	0.10
		2nd	0.0194	0.18	0.09	0.09
		3rd	0.0292	0.20	0.03	0.19

Table 4
Linear probing classification accuracy across U-Net layers for real versus synthetic objects. Accuracy and balanced accuracy of a linear classifier trained on feature embeddings extracted from different depths of the diffusion U-Net to distinguish real from AI-generated objects.

Block	Resolution	Layer	Accuracy	Balanced accuracy
Encoder	64 × 64	1st	87.21%	85.17%
		2nd	92.31%	91.01%
	32 × 32	1st	94.81%	93.95%
		2nd	96.57%	95.79%
	16 × 16	1st	97.87%	97.40%
		2nd	97.50%	96.93%
Decoder	16 × 16	1st	97.41%	97.04%
		2nd	97.68%	97.55%
		3rd	98.05%	97.96%
	32 × 32	1st	**98.33%**	**98.17%**
		2nd	97.12%	96.70%
		3rd	97.50%	96.93%
	64 × 64	1st	96.39%	95.90%
		2nd	94.25%	93.53%
		3rd	94.35%	93.48%

Table 5
Performance of *one vs. all* linear classifiers in identifying the source object generator.

Features	Accuracy	Balanced accuracy
DALL-E 3	98.98%	98.85%
Midjourney	98.05%	96.22%
SDM 1.4	84.06%	68.86%
SDM 1.5	84.99%	69.57%
SDM 1.4 + SDM 1.5	98.70%	98.45%

the classification accuracy for DALL-E 3, Midjourney, and both Stable Diffusion models individually. Given their feature similarity, we also report results for a classifier trained to identify the two Stable Diffusion versions as a single, unified class.

The performance of the linear classifier demonstrates that the concatenated features from the two analyzed decoder layers are sufficiently discriminative to attribute objects to their source generators. Notably, while the feature representations of the two Stable Diffusion versions are too similar to permit linear separation, they are easily distinguished from other generators when treated as a single class. This finding supports the hypothesis that features from specific layers of a pre-trained Diffusion Model can serve as a fingerprint to characterize the AI tool used for generation.

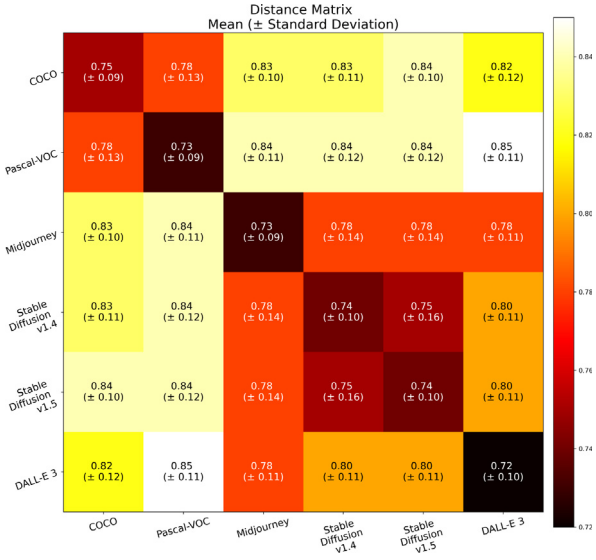


Fig. 3. Average distance matrix between features from different datasets.

To provide a further interpretation, we exploit UMAP, an algorithm that preserves the topological structure of high-dimensional data, to perform a dimensionality reduction and project the 2560-dimensional concatenated features into a 3D space. The resulting 3D plots compare the feature distributions of each synthetic dataset against real objects, with separate analyses for foreground objects (Fig. 6) and background (Fig. 5).

Real and synthetic objects form distinguishable feature clusters, suggesting representational differences. The gap is more evident for foreground objects, where synthetic features show deviations from real ones across all tested generators.

5.4. Cross-generator results

Following the experimental setup in Section 4.4, we trained a linear classifier and a k-NN classifier using, in turn, images from Dalle-3, Midjourney, and Stable Diffusion (combining versions 1.4 and 1.5) as the training set. Tables 6 and 7 present the cross-generator evaluation results for the linear classifier and the k-NN, respectively.

When using a linear classifier, the model learns the specific distribution of the training features and struggles to generalize to unseen

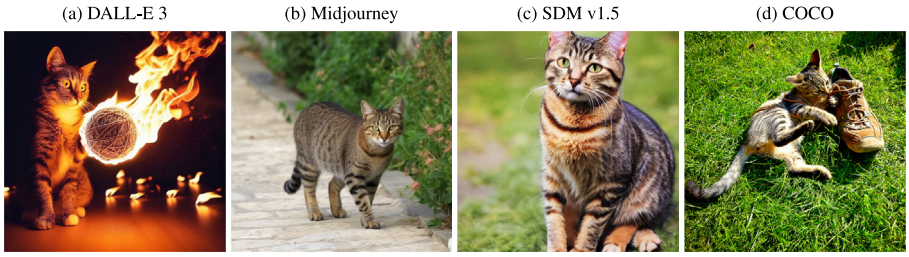


Fig. 4. Examples of cat images generated by different models, compared to a real image from COCO. As can be observed, DALL·E 3 produces images with a distinctive style.

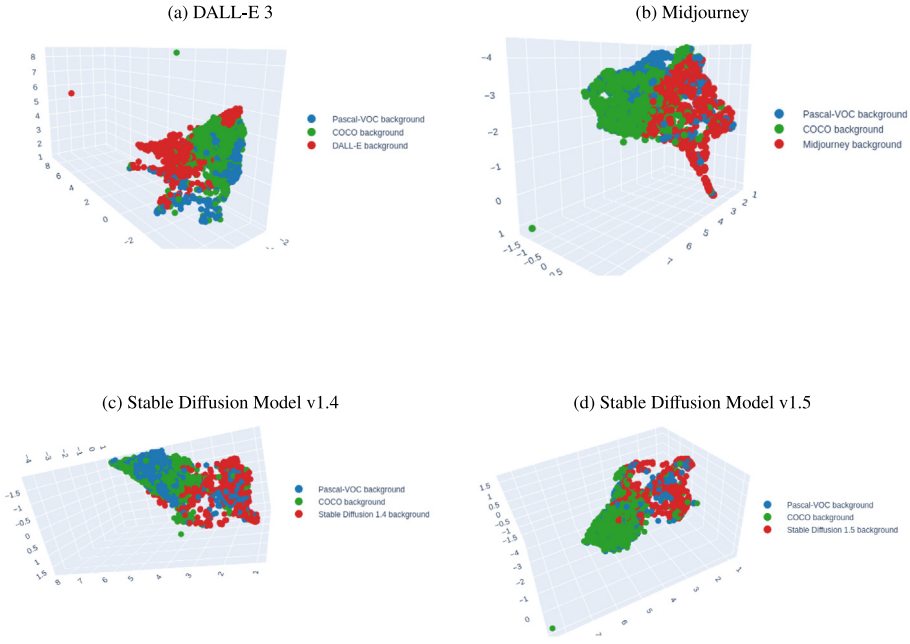


Fig. 5. Assessing background features separation by UMAP projection. 3D feature distributions of real and generated backgrounds obtained using UMAP.

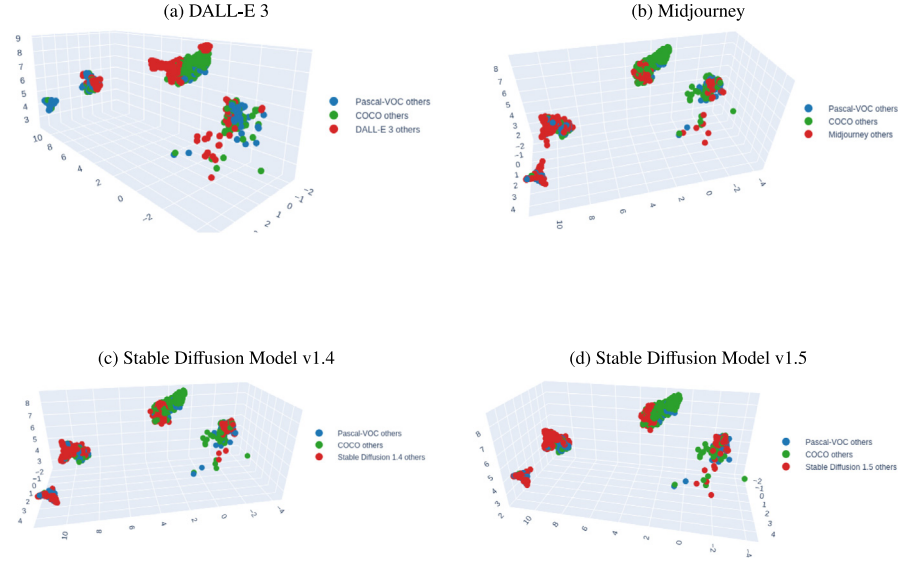


Fig. 6. Assessing foreground objects features separation by UMAP projection. 3D feature distributions of real and generated foreground objects obtained using UMAP.

Table 6
Cross-generator evaluation using a linear classifier. We use object-level features from one generator to train a linear classifier and evaluate the model on features extracted from objects generated by the other models.

Trained on	Accuracy when tested on		
	Midjourney	SDV1.4 + SDV1.5	DALL-E 3
Midjourney	97.96%	54.26%	66.36%
SDV1.4 + SDV1.5	53.01%	98.86%	52.73%
DALL-E 3	65.80%	54.12%	99.07%

Table 7
Cross-generator evaluation using k-NN. We use object-level features from one generator to train a k-NN classifier and evaluate the model on features extracted from objects generated by the other models.

Trained on	Accuracy when tested on		
	Midjourney	SDV1.4 + SDV1.5	DALL-E 3
Midjourney	90.89%	88.49%	80.63%
SDV1.4 + SDV1.5	87.73%	91.90%	78.30%
DALL-E 3	84.20%	77.13%	90.65%

Table 8
Impact of dataset dimension, filters, and manual selection on linear probing accuracy.

Filter on size	Filter on classes	Manual mask selection	Dataset type	Accuracy
Yes	Yes	Yes	Reduced	99.26%
Yes	Yes	No	Extended	98.68%
No	No	Yes	Reduced	97.59%
No	No	No	Extended	97.00%
Yes	No	Yes	Reduced	98.74%
Yes	No	No	Extended	98.27%

generators. However, when employing a simple distance-based k-NN classifier, we observe that the model generalizes significantly better to unseen object features (different generators). The drop in accuracy on unseen distributions is moderate and could potentially be mitigated with further optimization (feature normalization or domain adversarial training).

5.5. Results on the extended and uncurated dataset

Following the setup detailed in Section 4.5, we compare the linear probing results obtained by concatenating features extracted from the first decoder layers at 16×16 and 32×32 resolutions on an extended version of our dataset. Table 8 details the performance across both datasets under varying conditions, allowing us to evaluate the direct impact of our selection and filtering choices.

The findings suggest that our curated, reduced dataset accurately reflects the general case. Comparing the fully curated setup (the reduced dataset with manual mask selection and all filters, yielding 99.26%) to the generalized setup (the extended dataset without mask selection and all filters, yielding 98.68%) reveals a performance drop of merely 0.58%. This minimal difference confirms that the manual curation process does not introduce significant bias, ensuring that conclusions drawn from the smaller dataset scale safely to the extended, uncurated dataset. Of the applied constraints, the object dimension filter has the most significant impact. Removing both the class and size filters drops accuracy from 98.68% to 97.00% on the extended dataset, and from 99.26% to 97.59% on the reduced dataset. This decline is driven almost entirely by the dimension filter; applying this filter in isolation achieves an accuracy of 98.27% on the extended dataset and 98.74% on the reduced dataset. These results confirm that while the curation steps enable us to focus our statistical analysis on well-defined objects, their absence does not introduce biases that would undermine the core findings of our study.

Table 9
Comparison of feature extraction backbones.

	Acc. on reduced dataset	Acc. on extended dataset
Ours (Stable Diffusion)	**98.85%**	**98.92%**
CLIP	97.55%	97.44%
DINO	94.74%	94.88%
EfficientNet B0	92.80%	92.09%
ResNet 50	93.45%	93.79%

5.6. Image level results

In real-world scenarios, reliance on segmentation for feature extraction can be a significant limitation. To address this, we followed the experimental setup in Section 4.6 and evaluated the linear separability of features concatenated from the first decoder layers at 16×16 and 32×32 resolutions. We achieved an accuracy of 98.85% (balanced accuracy: 98.84%), which is comparable to the 99.26% accuracy obtained using object-level features. This demonstrates that real and synthetic images remain nearly linearly separable in the SDM latent space, even using global image-level features. However, it is worth noting that while image-level features are sufficient for detection, the object-level analysis remains valuable for tasks requiring finer granularity, such as localized image editing or manipulation detection.

5.7. Results on image manipulation

When evaluated on the MagicBrush test set, our approach achieved an accuracy of 97.65% using linear probing. This result indicates a clear degree of linear separability between the feature prototypes of manipulated objects and those of authentic regions within the same image. Overall, these findings support the relevance of our object-level analysis in localized manipulation settings, suggesting that the extracted representations capture information that may be useful for distinguishing manipulated from authentic regions.

5.8. Backbones comparison

Following the experimental setup detailed in Section 4.8, we evaluate and compare the performance of various feature extraction backbones. As shown in Table 9, our proposed approach utilizing Stable Diffusion features consistently outperforms all evaluated baselines across both dataset scales. Specifically, our method achieves the highest linear probing accuracies of 98.85% on the Reduced Dataset and 98.92% on the Extended Dataset, demonstrating the highly discriminative nature of the extracted prototypes. Among the baseline models, transformer-based architectures exhibit a clear advantage over traditional CNNs. CLIP emerges as the most competitive alternative, exceeding 97% accuracy on both datasets, followed by DINO at approximately 94.8%. In contrast, the CNN-based feature extractors yield the lowest overall performance, clustering between 92% and 94%. Furthermore, the performance hierarchy across all models remains remarkably stable regardless of dataset size.

6. Discussion

The results show that features extracted from a pre-trained SDM contain strong discriminative signals between real and synthetic data under the evaluated conditions, with the most informative representations coming from the early decoder layers of the U-Net. Both the distance-based analysis and the linear-probing experiments indicate that decoder features at 16×16 and 32×32 resolutions encode statistically meaningful differences between real and generated samples.

A possible interpretation is that early decoder layers are sensitive to differences in low- and mid-level image statistics that arise from the generation process. During training, the diffusion model learns to reverse a noising process and progressively reconstruct spatial structure

from latent representations. The first decoder blocks, which bridge coarse semantic information and emerging spatial detail, may therefore encode activation patterns that are sensitive to the statistical regularities of synthetic images. Under this interpretation, objects produced by the same generator may share partially coherent representational signatures, while real images or samples from different generators may deviate from these patterns. This hypothesis is consistent with the low intra-generator distances and the source-attribution results observed in our experiments.

The cross-generator evaluation indicates that the learned separation is not fully invariant across synthetic distributions and may depend on generator-specific artifacts or dataset-specific correlations. The better performance of k-NN suggests that some local structure in the feature space may transfer across generators, but the observed drop in performance confirms that generalization is only partial. Consequently, our findings support the presence of useful discriminative cues in SDM representations, which may depend on the dataset used to train the classifier, while not establishing the existence of a general real–synthetic boundary shared across all generative models. Broader generalization to new generators therefore remains an important direction for future work.

The experiments on the extended dataset further show that the main trends persist when manual mask curation and object class/size constraints are relaxed. Similarly, the image-level analysis shows that the discriminative signal is not restricted to object-level prototypes, making the approach applicable even when segmentation masks are unavailable. At the same time, the object-level experiments remain valuable for localized manipulation analysis, as shown by the results on the MagicBrush dataset.

Finally, the comparison with CNN- and transformer-based backbones indicates that SDM features are highly competitive for real–synthetic discrimination in the evaluated setting. However, the evaluation remains limited to a finite set of datasets, generators, object categories, and feature-extraction choices. Broader real-world generalization, including robustness to unseen generators, image post-processing, compression, adversarial perturbations, and newer architectures such as SDXL or Flux, remains a matter of future work.

7. Conclusion and future work

In this study, we investigated whether the internal representations of a pre-trained Stable Diffusion Model contain information that can distinguish real from synthetic objects and images. Motivated by the hypothesis that generative models encode aspects of the real-data distribution during training, we conducted a systematic analysis of SDM U-Net features using distance-based statistics and linear probing.

Our experiments, involving various datasets, show that specific decoder layers, particularly at spatial resolutions of 16×16 and 32×32, provide highly real-vs-synthetic discriminative representations. Linear probing achieves high accuracy in our evaluation setting, and the extended-dataset experiments suggest that these results are not primarily driven by manual mask curation or by the object class and size filters used in the main analysis. Moreover, image-level features remain highly discriminative, while object-level prototypes are useful for localized manipulation scenarios.

Nonetheless, despite our experiments support the existence of strong real–synthetic discriminative cues in SDM representations under the evaluated conditions, the performance drop in the cross-generator evaluation suggests caution in considering a universal latent boundary between real and synthetic content.

Overall, our results have potential implications for AI-generated content detection and media forensics, suggesting that generative-model representations may provide useful signals for content verification.

Future work should explore whether these emergent properties persist in newer architectures (e.g., SDXL, Flux) and investigate the robustness of these latent signatures against adversarial attacks or post-processing techniques. Ultimately, this work establishes a foundational understanding of how diffusion models structure semantic information, paving the way for more efficient, intrinsic methods of content verification.

CRediT authorship contribution statement

Simone Bonechi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Paolo Andreini:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Barbara Toniella Corradini:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT, Gemini and Claude in order to improve the quality of English text. After using these tools/services, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Only publicly available data is used.

References

Alain, G., Bengio, Y., 2017. Understanding intermediate layers using linear classifier probes.

Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A., 2022. Label-efficient semantic segmentation with diffusion models. In: International Conference on Learning Representations.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., Ramesh, A., 2023. Improving image generation with better captions. Comput. Sci. 2 (3), 8.

Bonechi, S., Andreini, P., Corradini, B.T., Scarselli, F., 2024. Diff-props: is semantics preserved within a diffusion model? Procedia Comput. Sci. 246, 5244–5253.

Bonechi, S., Andreini, P., Corradini, B.T., Scarselli, F., 2025. An analysis of pre-trained stable diffusion models through a semantic lens. Neurocomputing 614, 128846.

Boychev, D., Cholakov, R., 2025. ImagiNet: A multi-content benchmark for synthetic image detection. In: Workshop on Datasets and Evaluators of AI Safety.

Brock, A., Donahue, J., Simonyan, K., 2019. Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations.

Brooks, T., Holynski, A., Efros, A.A., 2023. Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. ICML '20, JMLR.org.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299.

Cliff, N., 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. Psychol. Bull. 114 (3), 494.

Corradini, B.T., Shukor, M., Couairon, P., Couairon, G., Scarselli, F., Cord, M., 2025. FreeSeg-Diff: Training-free open-vocabulary segmentation with diffusion models. In: 2025 International Joint Conference on Neural Networks. IJCNN, pp. 1–8.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Dhariwal, P., Nichol, A., 2021. Diffusion models beat gans on image synthesis. Adv. Neural Inf. Process. Syst. 34, 8780–8794.

Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88 (2), 303–338.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63 (11), 139–144.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent a new approach to self-supervised learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B., 2022. Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706.

Hao, Y., Chi, Z., Dong, L., Wei, F., 2023. Optimizing prompts for text-to-image generation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (Eds.), Advances in Neural Information Processing Systems. vol. 36, Curran Associates, Inc., pp. 66923–66939.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. 33, 6840–6851.

Ho, J., Salimans, T., 2021. Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J., 2022. Video diffusion models. Adv. Neural Inf. Process. Syst. 35, 8633–8646.

Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S., 2025. Diffusion model-based image editing: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 47 (6), 4409–4437.

Huang, J., Zhang, G., JIE, Z., Jiao, S., Qian, Y., Chen, L., Wei, Y., Ma, L., 2025. M4V: Multimodal mamba for efficient text-to-video generation.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Amer. Statist. Assoc. 47 (260), 583–621.

Kwon, M., Jeong, J., Uh, Y., 2023. Diffusion models already have a semantic latent space. In: The Eleventh International Conference on Learning Representations.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A., 2024. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26439–26455.

McInnes, L., Healy, J., Saul, N., Großberger, L., 2018. UMAP: Uniform manifold approximation and projection. J. Open Source Softw. 3 (29), 861.

Midjourney, 2022. Midjourney. https://www.midjourney.com/home/.

Mukhopadhyay, S., Gwilliam, M., Agarwal, V., Padmanabhan, N., Swaminathan, A., Hegde, S., Zhou, T., Shrivastava, A., 2023. Diffusion models beat gans on image classification. arXiv preprint arXiv:2307.08702.

Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M., 2022. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning. PMLR, pp. 16784–16804.

Pnvr, K., Singh, B., Ghosh, P., Siddiquie, B., Jacobs, D., 2023. Ld-znet: A latent diffusion approach for text-based image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4157–4168.

Qin, J., Huang, J., Qiao, L., Ma, L., 2025. STAR: Stacked AutoRegressive scheme for unified multimodal learning. arXiv preprint arXiv:2512.13752.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al., 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. Adv. Neural Inf. Process. Syst. 35, 25278–25294.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A., 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. PMLR, pp. 2256–2265.

Song, J., Meng, C., Ermon, S., 2021. Denoising diffusion implicit models. In: International Conference on Learning Representations.

Székely, G.J., Rizzo, M.L., 2007. Distance correlation: a measure for dependence between multivariate random variables. Ann. Stat. 35 (6), 2769–2792.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W., 2022. Emergent abilities of large language models. Trans. Mach. Learn. Res. Survey Certification.

Wukong, 2022. Wukong. https://xihe.mindspore.cn/modelzoo/wukong.

Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y., 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. Adv. Neural Inf. Process. Syst. 36, 31428–31449.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847.

Zhou, D., Bai, J., Wang, J., Chen, H., Guangyong, C., Hu, X., Heng, P.-A., 2024. MagicTailor: Component-controllable personalization in text-to-image diffusion models. pp. 10225–10233.

Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y., 2024. Genimage: A million-scale benchmark for detecting ai-generated image. Adv. Neural Inf. Process. Syst. 36.