EMERGENCE AND EVOLUTION OF INTERPRETABLE CONCEPTS IN DIFFUSION MODELS

Berk Tinaz *
Dept. of Electrical and Computer Engineering
University of Southern California
Los Angeles, CA
tinaz@usc.edu

Zalan Fabian *
Dept. of Electrical and Computer Engineering
University of Southern California
Los Angeles, CA
fabian.zalan@gmail.com

Mahdi Soltanolkotabi
Dept. of Electrical and Computer Engineering
University of Southern California
Los Angeles, CA
soltanol@usc.edu

ABSTRACT

Diffusion models have become the go-to method for text-to-image generation, producing high-quality images from noise through a process called reverse diffusion. Understanding the dynamics of the reverse diffusion process is crucial in steering the generation and achieving high sample quality. However, the inner workings of diffusion models is still largely a mystery due to their black-box nature and complex, multi-step generation process. Mechanistic Interpretability (MI) techniques, such as Sparse Autoencoders (SAEs), aim at uncovering the operating principles of models through granular analysis of their internal representations. These MI techniques have been successful in understanding and steering the behavior of large language models at scale. However, the great potential of SAEs has not yet been applied toward gaining insight into the intricate generative process of diffusion models. In this work, we leverage the SAE framework to probe the inner workings of a popular text-to-image diffusion model, and uncover a variety of human-interpretable concepts in its activations. Interestingly, we find that *even before the first reverse diffusion step* is completed, the final composition of the scene can be predicted surprisingly well by looking at the spatial distribution of activated concepts. Moreover, going beyond correlational analysis, we show that the discovered concepts have a causal effect on the model output and can be leveraged to steer the generative process. We design intervention techniques aimed at manipulating image composition and style, and demonstrate that (1) in early stages of diffusion image composition can be effectively controlled, (2) in the middle stages of diffusion image composition is finalized, however stylistic interventions are effective, and (3) in the final stages of diffusion only minor textural details are subject to change.

1 Introduction

Diffusion models (DMs) [15, 41] have revolutionized the field of generative modeling. These models iteratively refine images through a denoising process, progressively transforming Gaussian noise into coherent visual outputs. DMs have established state-of-the-art in image [8, 28, 36, 34, 16], audio [22], and video generation [17]. The introduction of text-conditioning in diffusion models [34, 35], i.e. guiding the generation process via text prompts, enables careful customization of generated samples while simultaneously maintaining exceptional sample quality.

While DMs excel at producing images of exceptional quality, the internal mechanisms by which they ground textual concepts in visual features that govern generation remain opaque. The time-evolution of internal representations through

*Equal contribution.

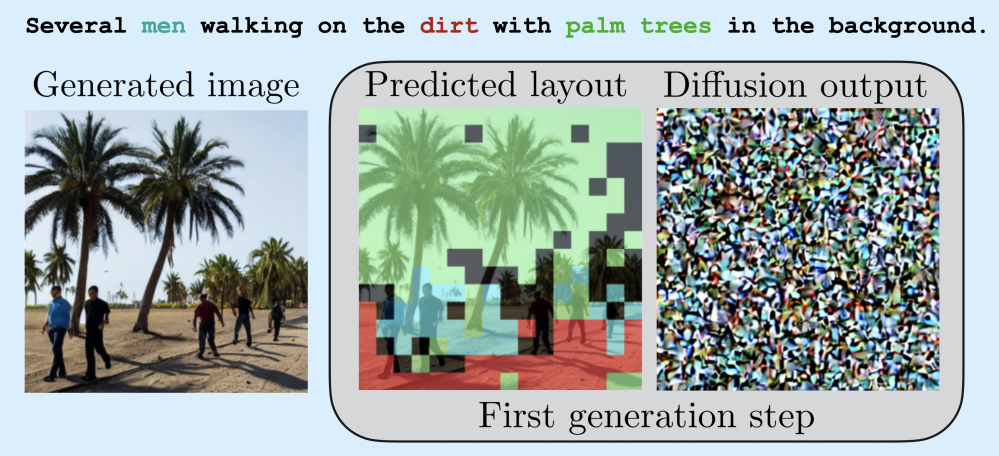


Figure 1: Coarse image composition emerges during the very first generation step in the diffusion process. We generate an image with the prompt `"Several men walking on the dirt with palm trees in the background"`. Our interpretability framework can predict segmentation masks for each object mentioned in the input prompt, solely relying on model activations cached during the first diffusion step. At this early stage, the posterior mean predicted by the diffusion model does not contain any visual clues about the final generated image.

the generative process, from pure noise to high-quality images, renders the understanding of DMs even more challenging compared to other deep learning models. A particular blind spot is the early, 'chaotic' stage [46] of diffusion, where noise dominates the generative process. Recently, a flurry of research has emerged towards demystifying the inner workings of DMs. In particular, a line of work attempts to interpret the internal representations by constructing saliency maps from cross-attention layers [44]. Another direction is to find interpretable editing directions directly in the model's feature space that allows for guiding the generation process [23, 13, 31, 5, 30, 10, 9, 4]. However, most existing techniques are aimed at addressing particular editing tasks and are not wide enough in scope to provide a more holistic interpretation on the internal representations of diffusion models.

Mechanistic interpretability (MI) [29] is focused on addressing the above challenges via uncovering operating principles from inputs to outputs that reveal how neural networks process information internally. A line of work within MI uses linear or logistic regression on model activations, also known as probing [12, 27], to uncover specific knowledge stored in model internals. Extensions [18, 1] explore nonlinear variants for improved detection and model steering. Recently, sparse autoencoders have emerged within MI as powerful tools to discover highly interpretable features (or *concepts*) within large models at scale [6]. These learned features enable direct interventions to steer model behavior in a controlled manner. Despite their success in understanding language models, the application of SAEs to diffusion models remains largely unexplored. Recent work [43] leverages SAEs and discovers highly interpretable concepts in the activations of a distilled DM [37]. While the results are promising, the paper focuses on a single-step diffusion model, and thus the time-evolution of visual features, a key characteristic and major source of intrigue around the inner workings of DMs, is not captured in this work.

In this paper, we aim to bridge this gap and address the following key questions:

- What level of image representation is present in the early, 'chaotic' stage of the generative process?
- How do visual representations evolve through various stages of the generative process?
- Can we harness the uncovered concepts to steer the generative process in an interpretable way?
- How does the effectiveness of such interventions depend on diffusion time?

We perform extensive experiments on the features of a popular, large-scale text-to-image DM, Stable Diffusion v1.4 [34], and extract thousands of concepts via SAEs. We propose a novel, scalable, vision-only pipeline to assign interpretations to SAE concepts. Then, we leverage the discovered concepts to explore the evolution of visual representations throughout the diffusion process. Strikingly, we find that the coarse composition of the image emerges *even before the first reverse diffusion update step*, at which stage the model output carries no identifiable visual information (see Figure 1). Moreover, we demonstrate that intervening on the discovered concepts has interpretable, causal effect on the generated output image. We design intervention techniques that edit representations in the latent space of SAEs aimed at manipulating image composition and style. We perform an in-depth study on the effectiveness of such interventions

as reverse diffusion progresses. We find that image composition can be effectively controlled in early stages of diffusion, however such interventions are ineffective in later stages. Moreover, we can manipulate image style at middle time steps without altering image composition. Our work deepens our understanding on the evolution of visual representations in text-to-image DMs and opens the door to powerful, time-adaptive editing techniques.

2 Background

Diffusion models – In the diffusion framework, a forward noising process progressively transforms the clean data distribution $\boldsymbol{x}_0 \sim q_0(\boldsymbol{x})$ into a simple distribution q_T (typically isotropic Gaussian distribution) through intermediary distributions q_t. In general, q_t is chosen such that $\boldsymbol{x}_t$ is obtained by mixing $\boldsymbol{x}_0$ with an appropriately scaled i.i.d. Gaussian noise, $q_t(\boldsymbol{x}_t|\boldsymbol{x}_0) \sim \mathcal{N}\left(\boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}\right)$, where the variance σ_t^2 is chosen according to a variance schedule. Diffusion models [40, 15, 41, 42] learn to reverse the forward process to generate new samples from q_0 by simply sampling from the tractable distribution q_T. Throughout this paper, we assume that the diffusion process is parameterized by a continuous variable $t \in [0, 1]$, where $t = 1$ corresponds to pure noise distribution and $t = 0$ corresponds to the distribution of clean images.

Sparse autoencoders (SAEs) – Sparse autoencoders are one of the most popular mechanistic interpretability techniques, and have been demonstrated to find interpretable features at scale [6, 11]. The core assumption underpinning SAEs is the *superposition hypothesis*, the idea that models encode far more concepts than the available dimensions in their activation space by using a combination of sparse and linear representations [39]. SAEs unpack these features in an *over-complete* basis of sparsely activated concepts in their latent space, as opposed to the compressed latent space of autoencoders commonly used in representation learning. Training autoencoders with both low reconstruction error and sparsely activated latents is not an easy feat. An initial approach [2] towards this goal uses ReLU as the activation function and ℓ_1 loss as a regularizer to induce sparsity. However, additional tricks are necessary, such as the initialization of encoder and decoder weights, to ensure that training is stable. Moreover, auxiliary loss terms may be necessary to ensure there are no dead neurons/concepts. Recent work [26, 11, 3] proposes using TopK activation instead of the ReLU function, which enables the precise control of the sparsity level without ℓ_1 loss and results in improved downstream task performance over ReLU baselines.

Interpreting diffusion models – There has been significant effort towards interpreting diffusion models. Authors in Tang et al. [44] find that the cross-attention layers in diffusion models with a U-Net backbone – such as SDXL [32] and Stable Diffusion [34] – can be used to generate saliency maps corresponding to textual concepts. Another line of work focuses on finding interpretable editing directions in diffusion U-Nets to control the image generation process. In particular, Kwon et al. [23] and Haas et al. [13] focus on manipulating bottleneck features, Park et al. [31] finds edit directions based on the SVD of the Jacobian between the input and bottleneck layer of the U-Net, Chen et al. [5] considers the Jacobian between the input and the posterior mean estimate rather than the bottleneck, Orgad et al. [30], Gandikota et al. [10] modify the *key* and *value* projection matrices, and Epstein et al. [9], Chen et al. [4] seek to control object position, size, shape directly by thresholding attention maps.

In recent work [43], authors train SAEs on the activations of a distilled, single-step diffusion model [37] (SDXL Turbo). In particular, they target certain cross-attention transformer blocks in the U-Net and train SAEs based on the residual update made by the transformer block. The features in the latent space of SAEs are found to be highly interpretable. Our work differs from theirs in two important ways. First, we analyze the *time-evolution* of interpretable concepts during the generative process, a key component in understanding and controlling the diffusion process, which is not captured by a single-step model. Second, they leverage vision-language foundation models to extract the semantics of SAE features by reasoning about and summarizing the commonalities between groups of images that activate specific features. This technique, however, is difficult to scale to large number of images due to the limited context of such models and is afflicted by the reasoning limitations, biases and hallucinations of the foundation model. In contrast, we propose a simple, scalable pipeline to extract interpretations for SAE features in the form of a flexible concept dictionary, leveraging open-set object detectors and segmentation models. Concurrent work [7] demonstrates the potential of SAEs in machine unlearning for diffusion models. Even though, similar to our work, they study a non-distilled diffusion model, their analysis focuses on identifying and removing particular concepts from generated images, and not on understanding the time-evolution of internal representations. In fact, they train a single SAE jointly for all time steps, whereas we perform a more granular analysis and train separate SAEs specialized to each time step. Recent work [20] leverages the SAE framework for controlled text-to-image generation. Different from our work, they train SAEs on the activations of the separate text-encoder that guides the diffusion model, and thus they do not investigate the visual representations of the diffusion model itself.

3 Method

3.1 SAE Architecture and Loss

In this section, we discuss the design choices behind our SAE model. We opt for k-sparse autoencoders (with TopK activation) given their success with GPT-4 [11] and SDXL Turbo [43]. In particular, let $\boldsymbol{x} \in \mathbb{R}^d$ denote the input activation to the autoencoder that we want to decompose into a sparse combination of features. Then, we obtain the latent $\boldsymbol{z} \in \mathbb{R}^{n_f}$ by encoding $\boldsymbol{x}$ as

$$\boldsymbol{z} = \mathcal{E}(\boldsymbol{x}) = \text{TopK}(\text{ReLU}(\boldsymbol{W}_{enc}(\boldsymbol{x} - \boldsymbol{b}))),$$

where $\boldsymbol{W}_{enc} \in \mathbb{R}^{n_f \times d}$ denotes the learnable weights of the encoder, $\boldsymbol{b} \in \mathbb{R}^d$ is a learnable bias term, and TopK function keeps the top k highest activations and sets the remaining ones to 0. Note, that due to the superposition hypothesis, we wish the encoding to be expansive and therefore $n_f \gg d$. Then, a decoder is trained to reconstruct the input from the latent $\boldsymbol{z}$ in the form

$$\hat{\boldsymbol{x}} = \mathcal{D}(\boldsymbol{z}) = \boldsymbol{W}_{dec}\boldsymbol{z} + \boldsymbol{b},$$

where $\boldsymbol{W}_{dec} \in \mathbb{R}^{d \times n_f}$ represents the learnable weights of the decoder. Note, that the bias term is shared between the encoder and decoder. We refer to $\boldsymbol{f}_i = \boldsymbol{W}_{dec}[:, i]$ columns of $\boldsymbol{W}_{dec}$ as *concept vectors*. We obtain the learnable parameters by optimizing the reconstruction error

$$\mathcal{L}_{rec}(\boldsymbol{W}_{enc}, \boldsymbol{W}_{dec}, \boldsymbol{b}) = \mathcal{L}_{rec}(\boldsymbol{\theta}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2.$$

In practice, training only on the reconstruction error is insufficient due to the emergence of dead features. Dead features are defined as directions in the latent space that are not activated for some specified number of training iterations resulting in wasted model capacity and compute. To resolve this issue, Gao et al. [11] proposes an auxiliary loss AuxK that models the reconstruction error of the SAE using the top-k_{aux} feature directions that have been inactive for the longest. To be specific, define the reconstruction error as $\boldsymbol{e} = \boldsymbol{x} - \hat{\boldsymbol{x}}$, then the auxiliary loss takes the form

$$\mathcal{L}_{aux}(\boldsymbol{\theta}) = \|\boldsymbol{e} - \hat{\boldsymbol{e}}\|_2^2,$$

where $\hat{\boldsymbol{e}}$ is the approximation of the reconstruction error using the top-k_{aux} dead latents. The combined loss for the SAE training becomes

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{rec}(\boldsymbol{\theta}) + \alpha\mathcal{L}_{aux}(\boldsymbol{\theta}),$$

where α is a hyperparameter.

3.2 Collecting Model Activations

In this work, we use Stable Diffusion v1.4 (SDv1.4) [34] as our diffusion model due to its widespread use. Inspired by Surkov et al. [43], we use 1.5M training prompts from the LAION-COCO dataset [38] and store $\Delta_{\ell,t} \in \mathbb{R}^{H_\ell \times W_\ell \times d_\ell}$, the difference between the output and input of the ℓth cross-attention transformer block at diffusion time t (i.e. the update to the residual stream). We train our SAE to reconstruct features individually along the spatial dimension. That is the input to the SAE is $\Delta_{\ell,t}[i, j, :]$ for different spatial locations (i, j) whereas ℓ and t are fixed and to be specified next.

To capture the time-evolution of concepts, we collect activations at timesteps corresponding to $t \in [0.0, 0.5, 1.0]$ and analyze *final* ($t = 0.0$, close to final generated image), *middle*, and *early* ($t = 1.0$, close to pure noise) diffusion dynamics respectively. For each timestep t, we target 3 different cross-attention blocks in the denoising model of SDv1.4: `down_blocks.2.attentions.1`, `mid_block.attentions.0`, `up_blocks.1.attentions.0`. We refer to these as `down_block`, `mid_block`, `up_block` for brevity. We specifically include the `mid_block` or the bottleneck layer of the U-Net since earlier work found interpretable editing directions here [23]. Other blocks are chosen to be the closest to the bottleneck layer in the downsampling and upsampling paths of the U-Net. The performance of text guidance is improved through Classifier-Free Guidance (CFG) [14]. The model output is modified as $\tilde{\varepsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c}) = \varepsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c}) + \omega(\varepsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c}) - \varepsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \varnothing))$, where ω denotes the guidance scale, $\boldsymbol{c}$ is the conditioning input and $\varnothing$ is the null-text prompt. At each timestep we collect both the text-conditioned diffusion features (called `cond`) and null-text-conditioned features (denoted by `uncond`).

To provide a granular and in-depth analysis, we train separate SAEs for different `block`, `conditioning` and `timestep` combinations. Training results are in Appendix A. In this work, we focus on `cond` features, as we hypothesize that they may be more aligned with human-interpretable concepts due to the direct influence of language guidance through cross-attention (more on this in Appendix C).

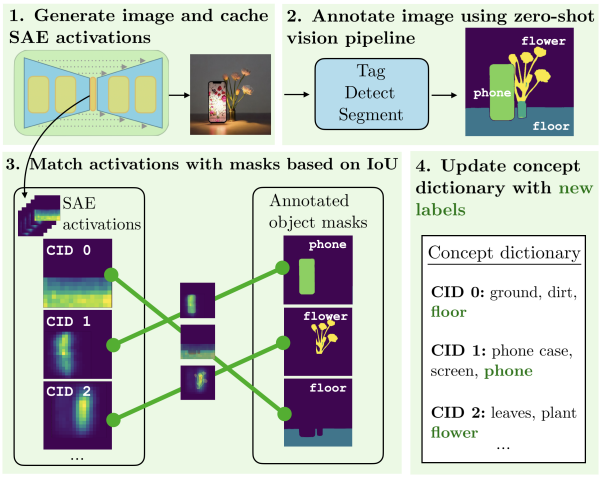


Figure 2: Curating the concept dictionary: 1) We cache SAE activations for various time steps and blocks during image generation. 2) We leverage a pipeline of image tagging, open-set object detection and promptable segmentation to annotate the generated image with segmentation masks and corresponding object labels. 3) We find SAE activations that sufficiently overlap with the object masks. 4) We add the overlapping object's label to the concept dictionary under the matching SAE activation's CID.

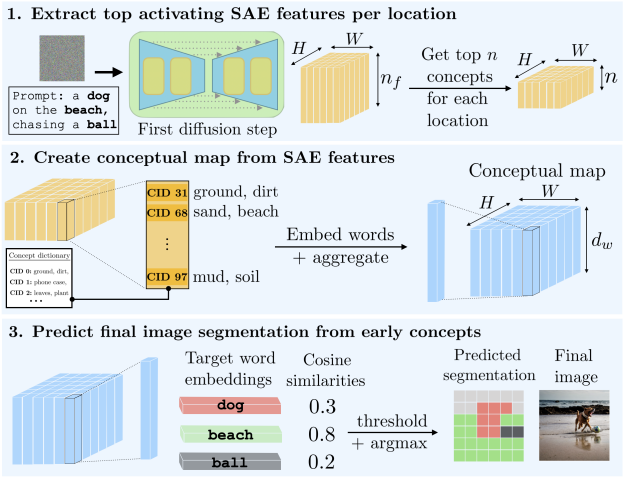


Figure 3: Predicting image composition: 1) We cache SAE activations during the *very first* diffusion step (or other time step of interest) and extract top activated concepts for each spatial location. 2) For each spatial location, we fetch the associated objects from the concept dictionary and produce a conceptual embedding via Word2Vec. 3) We compare the conceptual embedding at each location to the target word embeddings from the input prompt and predict a segmentation map based on cosine similarity.

3.3 Extracting interpretations from SAE features

Multiple work on automatic labeling of SAE features resort to LLM pipelines where the captions corresponding to top activating dataset examples are collected and the LLM is prompted to summarize them. However, these approaches come with severe shortcomings. First, they may incorporate the biases and limitations of the language model into the concept labels, including failures in spatial reasoning [19], object counting, identifying structural characteristics and appearance [45] and object hallucinations [24]. Second, they are sensitive to the prompt format and phrasing, and the instructions may bias or limit the extracted concept labels. Last but not least, it is computationally infeasible to scale LLM-based concept summarization to a large number of images, limiting the reliability of extracted concepts. For instance, Surkov et al. [43] only leverages a few dozens of images to define each concept. Therefore, we opt for designing a scalable approach that obviates the need for LLM-based labeling and instead use a vision-based pipeline to label our extracted SAE features.

In particular, we represent each concept by an associated list of objects, constituting a *concept dictionary*. The keys are unique concept identifiers (CIDs) assigned to each of the concept vectors of the SAE. The values correspond to objects that commonly occur in areas where the concept is activated. To build the concept dictionary (Figure 2), we first sample a set of text prompts, generate the corresponding images using a diffusion model and extract the SAE activations for each CID during generation. We obtain ground truth annotations for each generated image using a pre-trained vision pipeline, that combines image tagging, object detection and semantic segmentation, resulting in a mask and label for each object in generated images. Finally, we evaluate the alignment between our ground truth masks and the SAE activations for each CID, and assign the corresponding label to the CID only if there is sufficient overlap.

The concept dictionary represents each concept with a list of objects. In order to provide a more concise summary that incorporates semantic information, we assign an embedding vector to each concept. In general, we could use any model that provides robust natural language embeddings, such as an LLM, however we opt for a simple approach by assigning the mean Word2Vec embedding of object names activating the given concept.

3.4 Predicting image composition from SAE features

Leveraging the concept dictionary, we predict the final image composition based on SAE features at any time step (Figure 3), allowing us to gain invaluable insight into the evolution of image representations in diffusion models. Suppose that we would like to predict the location of a particular object in the final generated image, but before the

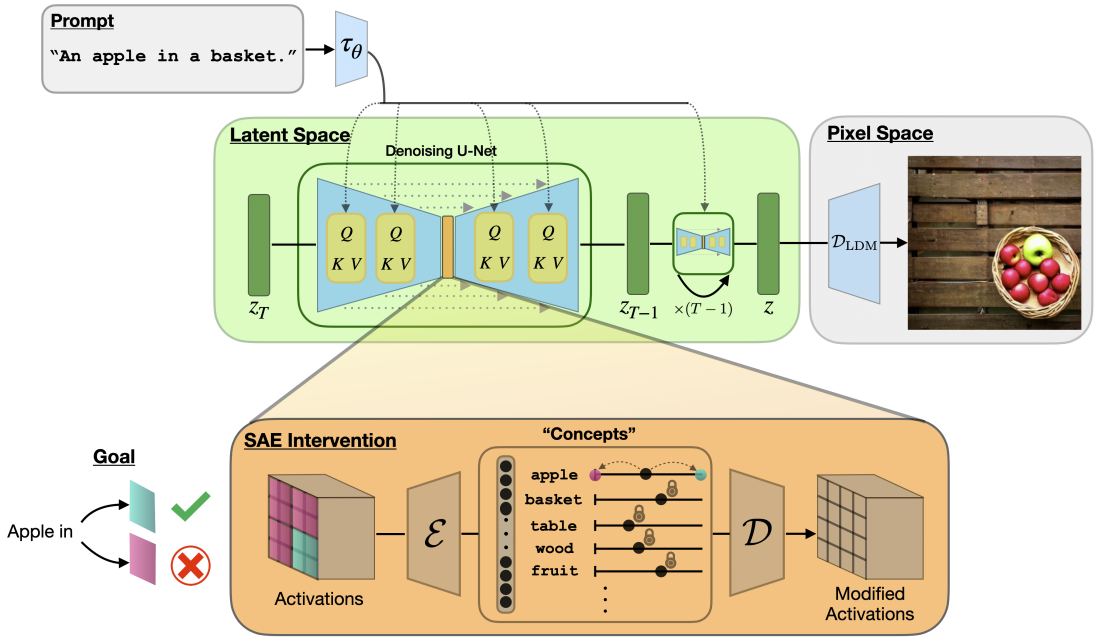


Figure 4: An overview of our SAE intervention technique. The prompt `"An apple in a basket"` specifies the necessary concepts but is vague in terms of spatial composition. We intercept activations of the denoising model and edit the latents after encoding them with the SAE. For the features that are spatially located in the bottom-right quadrant, we increase the coefficient corresponding to `"apple"` concept. For the remaining features, latents corresponding to `"apple"` concept are set to 0. After the intervention, generated image satisfies the specified layout where all the apples are located in the `bottom-right` quadrant.

reverse diffusion process is completed. First, given SAE features from a given intermediate time step, we extract the top activating concepts for each spatial location. Next, we create a *conceptual map* of the image by assigning a word embedding to each spatial location based on our curated concept dictionary. This conceptual map shows how image semantics, described by localized word embeddings, vary spatially across the image. Given a concept we would like to localize, such as an object from the input prompt, we produce a target word embedding and compare its similarity to each spatial location in the conceptual map. To produce a predicted segmentation map, we assign the target concept to spatial locations with high similarity, based on a pre-defined threshold value. This technique can be applied to each object present in the input prompt (or to any concepts of interest) to predict the composition of the final generated image.

3.5 Causal intervention techniques

Analyzing top activating dataset examples and semantic segmentation predictions only establish *correlational* relationship between concepts and the output image. In order to probe *causal* effects, we consider two categories of interventions: *spatially targeted interventions* designed to guide scene layout and *global interventions* directed towards manipulating image style.

Spatially targeted interventions – To assess layout controllability using the discovered concepts, we propose a simple task: enforce a specific object to appear only in a designated quadrant (e.g., top-left) of the image. To achieve this, we intercept activations and edit features in the SAE latent space by amplifying the desired concept in the target region and setting it to 0 otherwise. Recall, that the contribution of the ℓth transformer block at time t is given by $\Delta_{\ell,t}$. Let $\boldsymbol{Z}_{\ell,t}$ denote the latents after encoding the activations with the SAE encoder $\mathcal{E}$. Let S denote the set of coordinates to which we would like to restrict the object. Let C_o be the set of CIDs that are relevant to object o. We wish to modifty the latents as follows:

$$\forall c \in C_o, \quad \tilde{\boldsymbol{Z}}_{\ell,t}[i,j,c] = \begin{cases} \beta, & \text{if } (i,j) \in S \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where β is our *intervention strength*. However, decoding the modified latents directly is suboptimal as the SAE cannot reconstruct the input perfectly. Instead, we modify the activations directly using the concept vectors. The modification

(a) $t = 1.0$ (b) $t = 0.0$

Figure 5: Concept dictionary and visualization of the activation maps for the top 5 activating concepts extracted from `up_blocks.1.attentions.0` for the (a) first and (b) last diffusion steps. Sample ID: 2000031

.

in Eq. (1) can be equivalently written as:

$$\tilde{\Delta}_{\ell,t}[i,j] = \begin{cases} \Delta_{\ell,t}[i,j] + \beta \sum_{c \in C_o} \boldsymbol{f}_c & \text{if } (i,j) \in S \\ \Delta_{\ell,t}[i,j] - \sum_{c \in C_o} \boldsymbol{f}_c, & \text{otherwise} \end{cases}. \tag{2}$$

An overview of this intervention can be seen in Eq. (4). In prior experiments, we observe that the same intervention strength β does not work well across different objects o. To solve this, we introduce a normalization where the intervention at a spatial coordinate (i, j) is proportional to the norm of the latent at that coordinate $\|\boldsymbol{Z}_{\ell,t}[i,j]\|$. Therefore, the effective intervention strength is $\beta_{ij} = \beta \, \|\boldsymbol{Z}_{\ell,t}[i,j]\|$.

Global interventions – Beyond image composition, we investigate whether image style can be manipulated through our discovered concepts. To this end, given a CID c related to the style of interest, as image style is a global property we modify the activation at each spatial location as follows:

$$\tilde{\Delta}_{\ell,t}[i,j] = \Delta_{\ell,t}[i,j] + \beta \boldsymbol{f}_c. \tag{3}$$

Similar to spatially targeted interventions, we find that normalization is necessary for β to work well across different choices of style. We let β to be adaptive to spatial locations and modify them as $\tilde{\beta}_{ij} = \frac{\|\boldsymbol{Z}_{\ell,t}[i,j]\|}{\sum_{i,j} \|\boldsymbol{Z}_{\ell,t}[i,j]\|} \beta$.

4 Experiments

We perform extensive experiments on SD v1.4 aimed at understanding how internal representations emerge and evolve through the generative process.

4.1 Building the concept dictionary

We sample $40k$ prompts from the LAION-COCO dataset from a split that has not been used to train the SAEs. We build the concept dictionary following our technique introduced in Section 3.3. For annotating generated images, we leverage RAM [47] for image tagging, Grounding DINO [25] for open-set object detection and SAM [21] for segmentation, following the pipeline in Ren et al. [33]. We assign a label to a specific CID if the IoU between the corresponding annotated mask and activation is greater than 0.5. We binarize the activation map for the IoU calculation by first normalizing to $[0, 1]$ range, then thresholding at 0.1. We visualize the top 5 activating concepts and the corresponding concept dictionary entries in Figure 5 for a generated sample. More samples can be found in Appendix E.

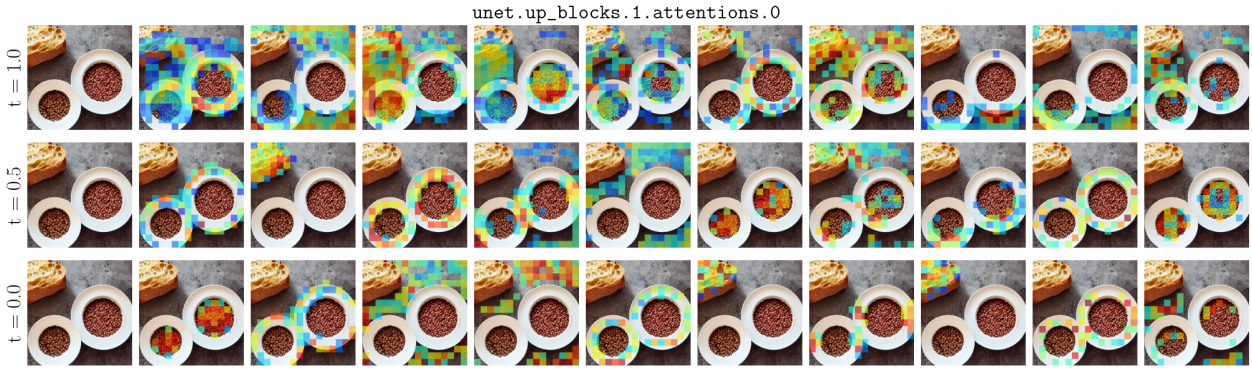


Figure 6: Visualization of top activating concepts in a generated sample. Concepts are sorted by mean activation across spatial locations and top 10 activation maps are shown. Each row depicts a different snapshot along the reverse diffusion trajectory starting from pure noise ($t = 1.0$) and terminating with the generated final image ($t = 0.0$). Note that each row within the same column may belong to a different concept, as concepts are not directly comparable across different diffusion time indices (separate SAE is trained for each individual timestep). Sample ID: 2000035.

4.2 Qualitative analysis of concept activations

We visualize the activation maps for top 10 (in terms of mean activation across the spatial dimensions) activating concepts in Figure 6 across time steps. Based on our empirical observations, the activations can be grouped into the following categories.

Local semantics – Most concepts fire in semantically homogeneous regions, producing a semantic segmentation mask for a particular concept. Examples include the segmentation of the plate, food items and background in Figure 6. We observe that these semantic concepts can be redundant in the sense that multiple concepts often fire in the same region (e.g. see Fig. 6, second row with multiple concepts focused on the food in the bowl). We hypothesize that these duplicates may add different conceptual layers to the same region (e.g. *food* and *round* in the previous example). In terms of diffusion time, we observe that the segmentation masks are increasingly more accurate with respect to the final generated image, which is expected as the final scene progressively stabilizes during the diffusion process. This observation is more thoroughly verified in Section 4.3 and Figure 7a. In terms of different U-Net blocks, we observe that `up_block` provides the most accurate segmentation of the final scene, especially at earlier time steps.

Global semantics (style) – We find concepts that activate more or less uniformly in the image. We hypothesize that these concepts capture global information about the image, such as artistic style, setting or ambiance. We observe such concepts across all studied diffusion steps and architectural blocks.

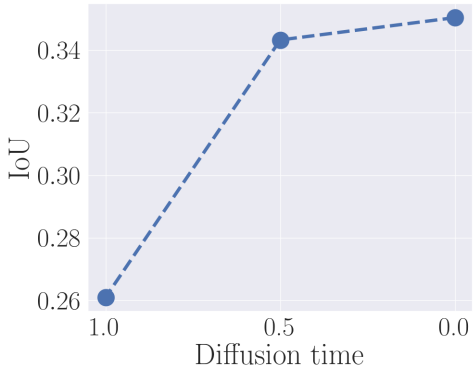
Context-free – We observe that some concepts fire exclusively in specific, structured regions of the image, such as particular corners or bordering edges of the image, irrespective of semantics. We hypothesize that these concepts may be a result of optimization artifacts, and are leveraged as semantic-independent knobs for the SAE to reduce reconstruction error. Visual examples and further discussion can be found in Appendix F.

More visualized concept activations for multiple blocks, time steps and samples can be found in Appendix D.

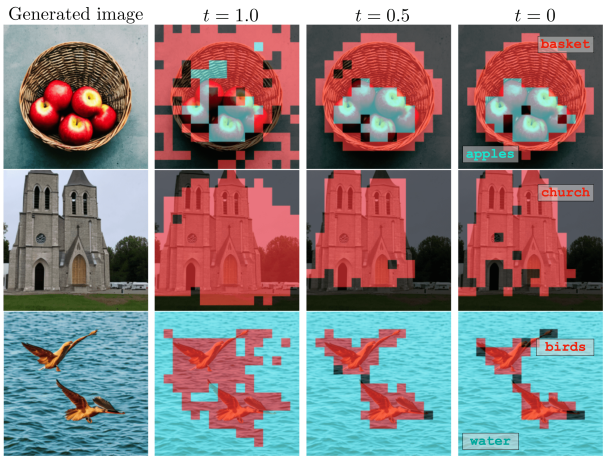
4.3 Emergence of image composition

Next, we investigate how image composition emerges and evolves in the internal representations of the diffusion model. We sample $5k$ LAION-COCO test prompts that have not been used for SAE training or to build the concept dictionary, and generate corresponding images with SDv1.4. Then, we follow the methodology described in Section 3.4 to predict a segmentation mask for every noun in the input prompt using SAE features at various stages of diffusion. We filter out nouns that are not in Word2Vec and those not detected in the generated image by our zero-shot labeling pipeline. We evaluate the mean IoU between the predicted masks and the ground truth annotations from our labeling pipeline for the first generation step ($t = 1.0$), the middle step ($t = 0.5$) and final diffusion step ($t = 0.0$). Numerical results are summarized in Figure 7a.

First, we surprisingly find that the image composition emerges during the very first reverse diffusion step (even before the first complete forward pass!), as we are able to predict the rough layout of the final scene with $IoU \approx 0.26$ from `mid_block` SAE activations. As Figure 1 demonstrates, the general location of objects from the input prompt is already



(a) Evolution of predicted image composition accuracy (in terms of IoU) over the reverse diffusion process (`mid_block`).



(b) Visualization of segmentation maps predicted from extracted concepts across reverse diffusion steps (`up_block`).

Figure 7: Evolution of predicted image composition during the reverse diffusion process, shown through segmentation accuracy (left) and visualizations (right). Features from later time steps become progressively more accurate at predicting the final layout of the image. However, the general image composition emerges as early as the first time step.

determined at this stage, even though the model output (posterior mean prediction) does not contain any visual clues about the final generated scene yet. More examples can be seen in the second column of Figure 7b.

Second, we observe that the image composition and layout is mostly finalized by the middle of the reverse diffusion process ($t = 0.5$), which is supported by the saturation in the accuracy of predicted masks. Visually, predicted masks for $t = 0.5$ and $t = 0.0$ look similar, however we see indications of increasing semantic granularity in represented concepts. For instance, the second row in Figure 7b depicts predicted segmentation masks for the noun *church*. Even though the masks for $t = 0.5$ and $t = 0.0$ are overall similar, the mask in the final time step excludes doors and windows on the building, suggesting that those regions are assigned more specific concepts, such as *door* and *window*. Moreover, we would like to emphasize that the segmentation IoU is evaluated with respect to our zero-shot annotations, which are often *less accurate* than our predicted masks for $t = 0.0$, and thus the reported IoU is bottlenecked by the quality of our annotations.

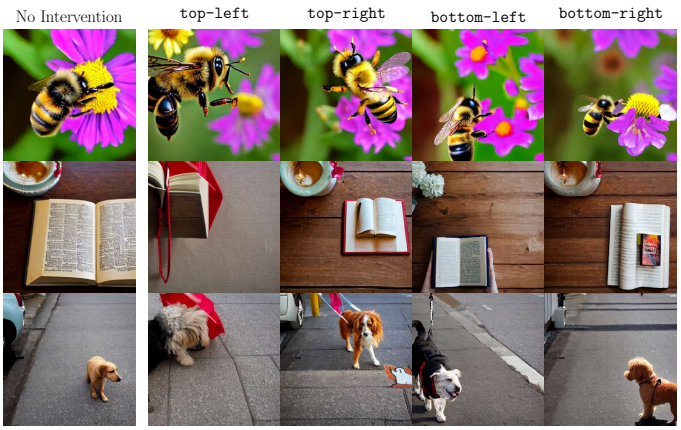
Finally, we find that image composition can be extracted from any of the investigated blocks, and thus we do not observe strong specialization between these layers for composition-related information. However, `up_block` provides generally more accurate segmentations than `down_block`, and `mid_block` provides the lowest due to the lower spatial resolution. We also find that `cond` features result in more accurate prediction of image composition than `uncond` features, likely due to more semantic information as an indirect result of text conditioning. Results for all block and conditioning combinations can be found in Appendix C.

4.4 Effectiveness of interventions across diffusion time

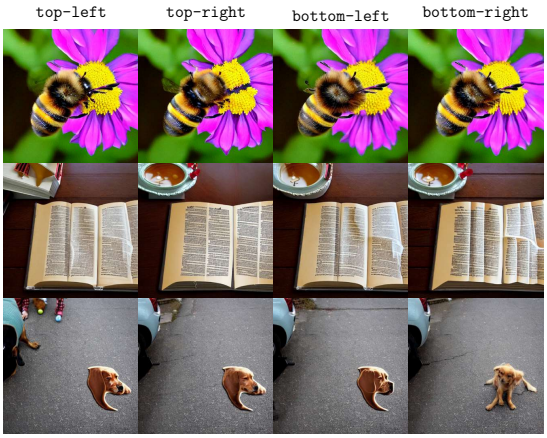
Beyond establishing correlational effects, we analyze how our discovered concepts can be leveraged in causal interventions targeted at manipulating image composition and style. We specifically focus on the effectiveness of these interventions as a function of diffusion time, split into 3 stages: *early* for $t \in [0.6, 1.0]$, *middle* for $t \in [0.2, 0.6]$ and *final* for $t \in [0, 0.2]$. Motivated by the success of bottleneck intervention techniques [23, 13, 31], we target `mid_block` in our experiments.

Spatially targeted interventions– We consider *bee*, *book*, and *dog* as the objects of interest and attempt to restrict them to four different quadrants: `top-left`, `top-right`, `bottom-left`, and `bottom-right`. In order to find the CIDs to be intervened on, we sweep the concept dictionary of the given time step and collect all the CIDs where the word of interest appears. Results are summarized in Figure 8.

Global interventions– Through our concept dictionary and visual inspection of top dataset examples at $t = 0.5$, we select the following CIDs: #1722 that controls the *cartoon* look of the image, #524 appears mostly with beach images where *sea* and *sand* are visible together, and #2137 activates the most on *paintings* (top activating images can be found in Appendix G). We find matching concepts for other time steps by picking the CIDs with the highest Word2Vec embedding similarity to the above target CIDs. An overview of results is depicted in Figure 9.



(a) Early-stage intervention.



(b) Middle-stage intervention

(c) Final-stage intervention

Figure 8: Effect of spatially targeted interventions at different stages of diffusion, aimed at manipulating image layout. We can efficiently restrict objects to the specified quadrant of the image when intervened in early stages of diffusion. However, in middle and final stages our interventions are unsuccessful.

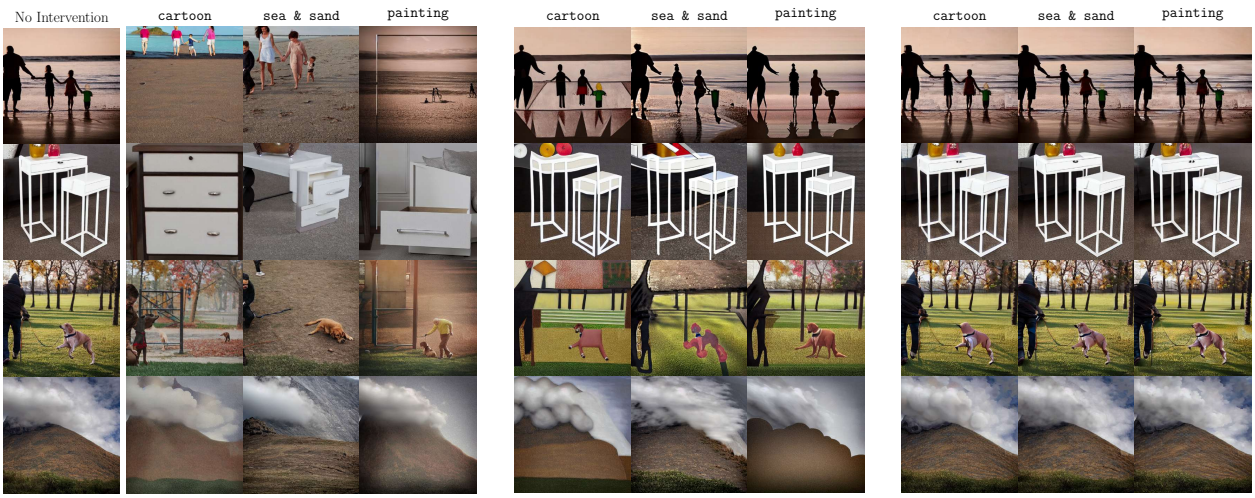
4.4.1 Early-stage interventions

First, we apply spatially targeted interventions according to Eq. 2 using an SAE trained on `cond` activations of `mid_block` at $t = 1.0$. We observe that a large intervention strength β is needed to successfully control the spatial composition consistently. We hypothesize that the skip connections in the U-Net architecture and the features from the `null-text` conditioning in classifier-free guidance reduce the effect of our interventions, as they provide paths that bypass the intervention. Thus, a larger value of intervention strength is needed to mask the leakage effects. In Eq. 8a, we observe that the objects of interest are successfully guided to their respective locations. Moreover, the concepts that we do not intervene on, such as the flower in the first row are preserved.

Next, we perform global interventions according to Eq. (3) aimed at manipulating image style. Interestingly, as depicted in Figure 9a, we find that instead of controlling image style, these global interventions broadly modify the composition of the image, without imbuing it with a particular style. As depicted in Figure 10a, this phenomenon holds for a wide range of β. As we vary the intervention strength, we obtain images with various compositions, but without the target style. This observation is consistent with our hypothesis that early stages of diffusion are responsible for shaping the image composition, whereas more abstract and high-level concepts, such as those related to consistent artistic styles emerge later.

4.4.2 Middle-stage interventions

We keep the setting from early-stage experiments, but use an SAE trained on the activations at $t = 0.5$. In contrast with early-stage results, as shown in Figure 8b, we find that our spatially localized intervention fails to manipulate image



(a) Early-stage intervention (b) Middle-stage intervention (c) Final-stage intervention

Figure 9: Effect of global interventions aimed at manipulating image style. Intervening in the early stages of diffusion drastically modifies image composition without imbuing the image with a particular style. In stark contrast, middle-stage global interventions successfully manipulate image style without interfering with image composition. However, in the final stages of diffusion, such global interventions have no effect on style or composition, and only result in minor textural changes.

composition at this stage. This result suggests that the locations of prominent objects in the scene have been finalized by this stage. The interventions cause visual distortions, while maintaining image composition. Interestingly, in some cases we see semantic changes in the targeted regions. For instance, intervening on the *book* concept in the second row of Fig. 8b in the `top-left` quadrant changes the tea cup into a book, instead of moving the large book making up most of the scene.

In an effort to control image style, we perform global interventions in the middle stages. We show results in Figure 9b. We find that the these interventions do not alter image composition as in early stages of diffusion. Instead, we observe local edits more aligned with stylistic changes (cartoon look, sandy texture, smooth straight lines, etc.), while the location of objects in the scene are preserved. Contrasting this with early-stage interventions, we hypothesize that the middle stage of diffusion is responsible for the emergence of more high-level and abstract concepts whereas the image layout is already determined in the earlier time steps (also supported by our semantic segmentation experiments). Moreover, varying the intervention strength impacts the intensity of style transfer in the output image (Figure 10b).

4.4.3 Final-stage interventions

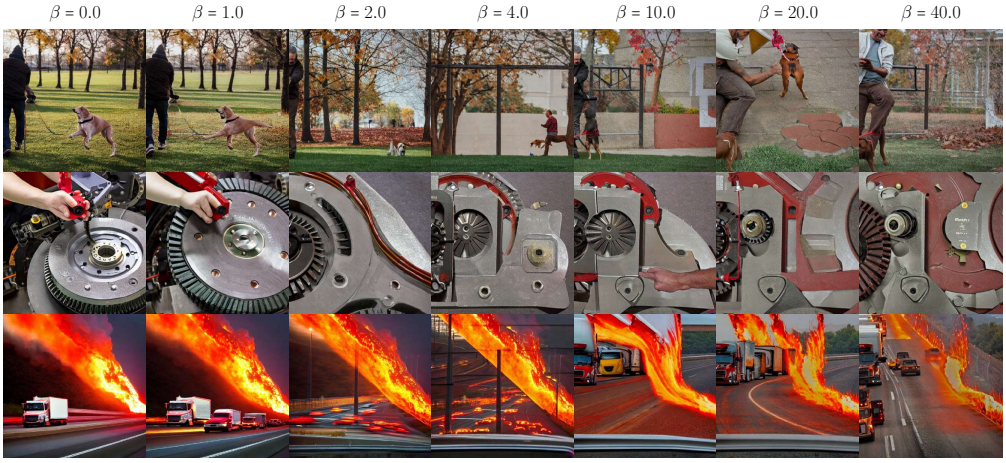
Performing spatially targeted interventions in the final stage of diffusion (Figure 8c) has no effect on image composition and only causes some minor changes in local details. This outcome is expected, as we observe that even by the middle stages of diffusion, image composition is finalized.

Similarly, we find that our global intervention technique is ineffective in manipulating image style in the final stage of diffusion (Figure 9c), as we only observe minor textural changes across a wide range of intervention strengths (Figure 10c).

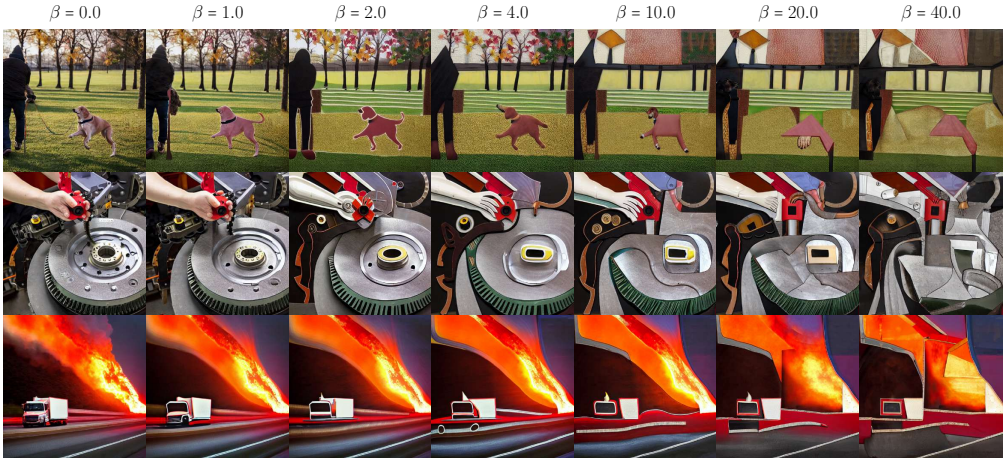
4.5 Summary of observations

Our experimental observations can be summarized as follows:

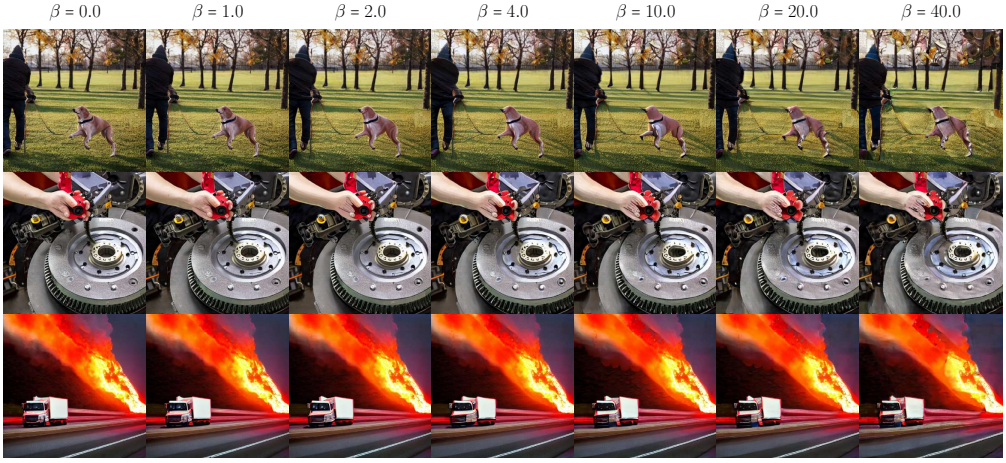
- **Early stage of diffusion:** coarse *image composition* emerges as early as during the very first diffusion step. At this stage, we are able to approximately identify where prominent objects will be placed in the final generated image (Section 4.3 and Figure 1). Moreover, image composition is *still subject to change*: we can manipulate the generated scene (Figure 8a) by spatially targeted interventions that amplify the desired concept in some regions and dampens it in others. However, we are *unable to steer image style* (Figure 9a) at this stage using our global intervention technique. Instead of high-level stylistic edits, these interventions result in major changes in image composition.



(a) Early-stage intervention



(b) Middle-stage intervention



(c) Final-stage intervention

Figure 10: Effect of intervention strength. We perform global intervention on a concept (#1722 for all time steps) corresponding to *cartoon* look in top activating images. Early-stage interventions, at any strength, are unable to modify image style consistently but broadly influence image composition. Interventions in the middle stages imbue the image with the target style with increasing intensity. We only observe minor textural changes in final stages of diffusion, even at high intervention strengths.

- **Middle stage of diffusion**: image composition has been finalized at this stage and we are able to predict the location of various objects in the final generated image with high accuracy (Figure 7). Moreover, our spatially targeted intervention technique fails to meaningfully change image composition at this stage (Figure 8b). On the other hand, through global interventions we *can effectively control image style* (Figure 9b) while preserving image composition, in stark contrast to the early stages.
- **Final stage of diffusion**: Image composition can be predicted from internal representations to very high accuracy (empirically, often higher than our pre-trained segmentation pipeline), however manipulating image composition through our spatially localized interventions fail (Figure 8c). Our global intervention technique only results in minor textural changes without meaningfully changing image style (Figure 9c). These observations are consistent with prior work [46] highlighting the inefficiency of editing in the final, 'refinement' stage of diffusion.

5 Conclusions and limitations

In this paper, we take a step towards demystifying the inner workings of text-to-image diffusion models under the lens of mechanistic interpretability, with an emphasis on understanding how visual representations evolve over the generative process. We show that the semantic layout of the image emerges as early as the first reverse diffusion step and can be predicted surprisingly well from our learned features, even though no coherent visual cues are discernible in the model outputs at this stage yet. As reverse diffusion progresses, the decoded semantic layout becomes progressively more refined, and the image composition is largely finalized by the middle of the reverse trajectory. Furthermore, we conduct in-depth intervention experiments and demonstrate that we can effectively leverage the learned SAE features to control image composition in the early stages and image style in the middle stages of diffusion. Developing editing techniques that adapt to the evolving nature of diffusion representations is a promising direction for future work. A limitation of our method is the leakage effect rooted in the U-Net architecture of the denoiser, which enables information to bypass our interventions through skip connections. We believe that extending our work to diffusion transformers would effectively tackle this challenge.

6 Acknowledgements

We would like to thank Microsoft for an Accelerating Foundation Models Research grant that provided the OpenAI credits enabling this work. This research is also in part supported by AWS credits through an Amazon Faculty research award and a NAIRR Pilot award. M. Soltanolkotabi is also supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, an NSF-CAREER under award #1846369, and NSF-CIF awards #1813877 and #2008443. and NIH DP2LM014564-01.

References

[1] Daniel Beaglehole, Adityanarayanan Radhakrishnan, Enric Boix-Adserà, and Mikhail Belkin. Aggregate and conquer: detecting and steering llm concepts by combining nonlinear predictors over multiple layers, 2025.

[2] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[3] Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders, 2024.

[4] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5343–5353, 2024.

[5] Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspaces in diffusion models for controllable image editing, 2024.

[6] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

[7] Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. *arXiv preprint arXiv:2501.18052*, 2025.

[8] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

[9] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation, 2023.

[10] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models, 2024.

[11] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024.

[12] Daniela Gottesman and Mor Geva. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4019, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[13] René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Graßhof, Sami S. Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models, 2024.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2006.11239*, 2020.

[16] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022.

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[18] Jakub Hoscilowicz, Adam Wiacek, Jan Chojnacki, Adam Cieslak, Leszek Michon, Vitalii Urbanevych, and Artur Janicki. Non-linear inference time intervention: Improving llm truthfulness, 2024.

[19] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023.

[20] Dahye Kim and Deepti Ghadiyaram. Concept steerers: Leveraging k-sparse autoencoders for controllable generations. *arXiv preprint arXiv:2501.19066*, 2025.

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[22] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[23] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space, 2023.

[24] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[26] Alireza Makhzani and Brendan Frey. k-sparse autoencoders, 2014.

[27] Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore, 2023. Association for Computational Linguistics.

[28] Alex Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2102.09672*, 2021.

[29] Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. `https://www.transformer-circuits.pub/2022/mech-interp-essay`, 2022. Accessed: 2025-04-14.

[30] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models, 2023.

[31] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry, 2023.

[32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[33] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[35] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. *arXiv:2104.07636 [cs, eess]*, 2021.

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[37] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024.

[38] Christoph Schuhmann, Andreas Kopf, Richard Vencu, Theo Coombes, Romain Beaumont, and Benjamin Trom. Laion coco: 600m synthetic captions from laion2b-en.

[39] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025.

[40] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[41] Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. *arXiv:1907.05600 [cs, stat]*, 2020.

[42] Yang Song and Stefano Ermon. Improved Techniques for Training Score-Based Generative Models. *arXiv:2006.09011 [cs, stat]*, 2020.

[43] Viacheslav Surkov, Chris Wendler, Mikhail Terekhov, Justin Deschenaux, Robert West, and Caglar Gulcehre. Unpacking sdxl turbo: Interpreting text-to-image models with sparse autoencoders, 2024.

[44] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention, 2022.

[45] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[46] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023.

[47] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024.