Comprehensive survey on Diffusion Models: From Image generation to 3D NeRF Generation

Naman Garg

Northwestern University

namangarg2025@u.northwestern.edu

March 1, 2024

Abstract

This survey paper provides an in-depth exploration of diffusion models, tracing their evolution from theoretical concepts in thermodynamics to their essential role in modern machine learning and artificial intelligence. It highlights key technological advancements that have established diffusion models as leading tools for generating high-quality, diverse images, videos, and more, while also adapting to various types of conditioning data. The paper covers a broad range of applications, from traditional image and video processing to innovative uses in 3D model creation and document layout generation. It addresses the significant challenges faced by diffusion models , such as computational demands and the generation accuracy relative to complex text prompts, and proposes potential solutions and future research directions to overcome

these obstacles. By offering a comprehensive review and insightful analysis, this survey aims to serve as a valuable resource for both researchers and practitioners, guiding further advancements and innovation in the field of diffusion models.

1 Introduction

Image generation has long been at the forefront of research in computer vision and artificial intelligence. From the initial forays with Variational Autoencoders (VAEs) [1] and Generative Adversarial Networks (GANs) [2] to the recent breakthroughs in photorealistic image synthesis, the field has seen rapid evolution. VAEs, which are neural networks designed to encode and decode images, laid the groundwork for understanding complex data distributions. GANs, comprising two competing networks, further advanced the capability to generate new, high-quality images by learning to mimic the distribution of real images.

Diffusion models have disrupte the longstanding dominance of GANs [2] and VAEs [1] in the realm of image generation. These novel generative models are currently receiving significant attention. Originating from the domain of thermodynamics, the concept of diffusion models introduces a fresh approach to image generation [3]. The underlying idea posits that gradually introducing noise into an image until it is completely transformed into noise—and then reversing this process—mirrors the core principles of generating images. This reversal process, if successfully learned by deep learning techniques, could revolutionize image generation. Further, the mathematical foundation of diffusion models was solidified with the use of Markov chains in Denoising Diffusion Probabilistic Models (DDPM) [4], providing a robust theoretical framework that enhances our understanding and capabilities in generating complex images.

The advent of open-source initiatives like Stability.ai's Stable Diffusion marked a significant leap forward. These models have democratized access to advanced image generation technologies, catalyzing both academic research and product innovation. Unlike their predecessors, diffusion models generate images through a process of gradually refining random noise, aligning more closely with how humans might imagine and create images from scratch [5].

Recently, these models have been enhanced by training on image-caption pairs, leveraging techniques such as those introduced by CLIP (Contrastive Language–Image Pre-training) to align images with text prompts more effectively [6]. To navigate away from the adversarial nature of GANs, Classifier-Free Guidance (CFG) emerged as a method to improve the fidelity and diversity of generated images without the need for a discriminator network [7].

Despite these advancements, the quest for models that can faithfully adhere to all elements of a text prompt remains ongoing. The gap between the complexity of human language and the interpretative capability of AI models poses a significant challenge. Innovations like ControlNet offer a promising direction by incorporating additional inputs such as depth maps and pose maps to provide nuanced control over the image generation process [8]. However, achieving perfect alignment between text prompts and generated images is still a complex challenge that the field continues to grapple with.

Moreover, these image generation models still struggle with finer details, such as accurately generating text within images or getting the right number of fingers on a hand. These seemingly minor inaccuracies highlight the ongoing challenges in creating truly lifelike and accurate images.

The application of diffusion models extends beyond mere image creation; they have been

successfully applied to image editing, video generation and editing, 3D model generation from text or images, and even document layout generation. This survey paper aims to delve into the multifaceted world of diffusion models, exploring their components, evolutionary trajectory, limitations, and the ingenious solutions devised to overcome these hurdles.

A comprehensive mindmap detailing the development and evolution of diffusion models is included in Figure 1. This visual representation provides an overview of the key milestones, technologies, and innovations that have shaped the current landscape of diffusion models.

As we chart the course of diffusion models from their inception to their current state, and look towards the future, it becomes evident that while significant strides have been made, the journey is far from over. The potential for further breakthroughs and applications is vast, promising a dynamic and exciting future for image generation technology.

2 PRELIMINARIES OF DIFFUSION MODELS

Image Diffusion Models were first introduced by SohlDickstein et al. [3, 9] and have been recently applied to image generation [10]. The Latent Diffusion Models (LDM) [5] performs the diffusion steps in the latent image space [19], which reduces the computation cost. Text-toimage diffusion models achieve state-of-the-art image generation results by encoding text inputs into latent vectors via pretrained language models like CLIP [6]. Glide [11] is a text-guided diffusion model supporting image generation and editing. Disco Diffusion [5] processes text prompts with clip guidance. Stable Diffusion [12] is a large-scale implementation of latent diffusion [5]. Imagen [13] directly diffuses pixels using a pyramid structure without using latent images. Commercial products include DALL-E2 [14] and Midjourney [15].

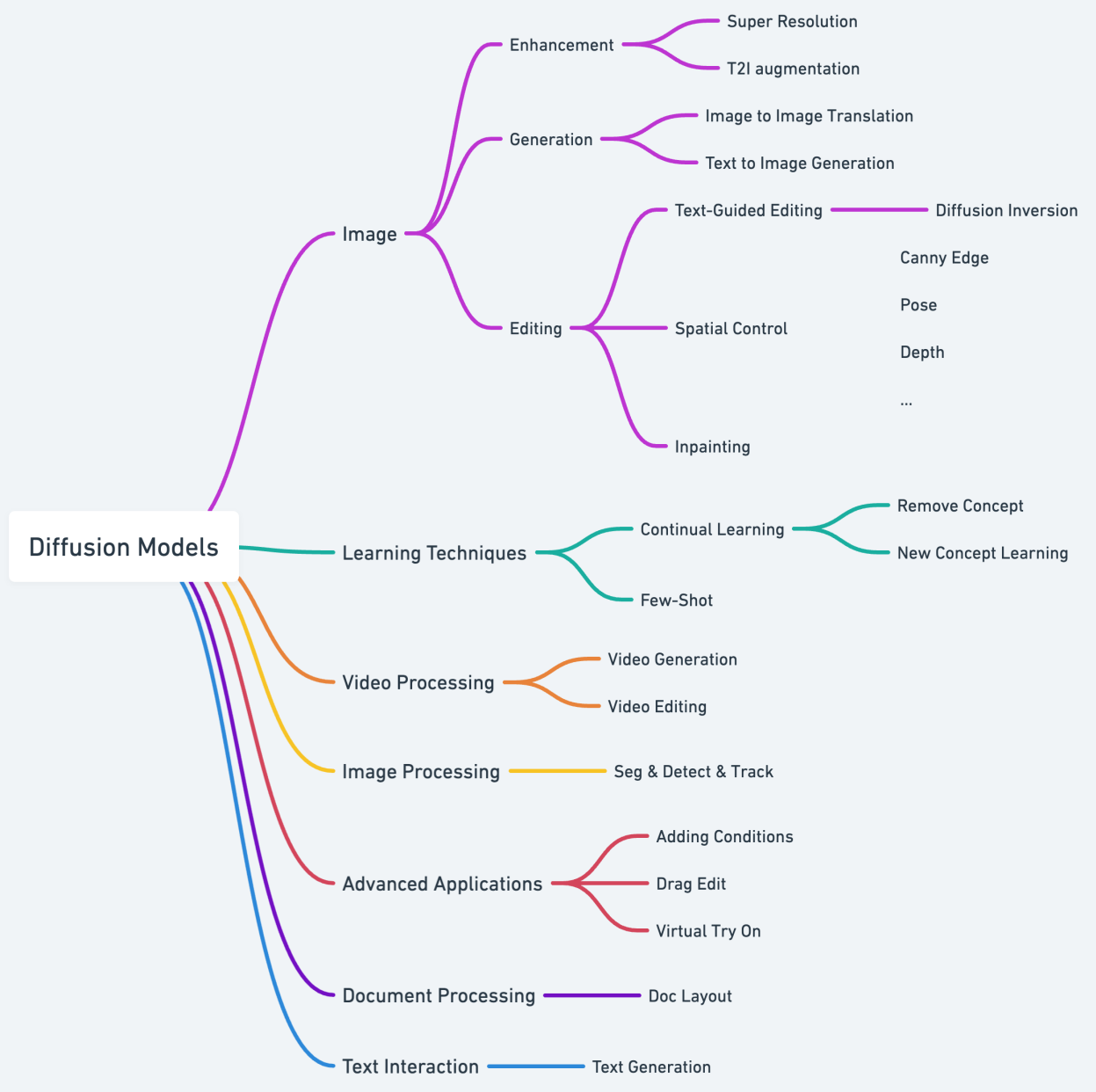

Figure 1: Mind Map of diffusion models

2.1 Exploration of Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) [4] introduce an innovative framework for generative models that emphasize the reverse engineering of diffusion processes. These models utilize parameterized Markov chains to methodically transform noise into organized patterns over multiple iterations.

The diffusion phase commences with an initial distribution $x_0 \sim q(x_0)$, methodically incorporating Gaussian noise over T timesteps. At each step t, the noising process is characterized by:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{1}$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{2}$$

where β_t represents the stepwise noise variance parameters.

In the denoising phase, DDPMs aim to incrementally cleanse the data, effectively reversing the diffusion sequence. This begins from a noisy state x_T and progressively works towards the initial data distribution $q(x_0)$. The model specifies the reverse transition $p_\theta(x_{t-1}|x_t)$ with:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

Deep learning models, notably those based on UNet architectures, parameterize $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$. These models input the noised data x_t and timestep t, predicting the normal distribution's parameters to identify the noise ϵ_θ needed for reversing the diffusion. Generating new data instances x_0 involves starting with a noise vector $x_T \sim p(x_T)$ and sequentially

sampling from $p_\theta(x_{t-1}|x_t)$ until $t = 1$, completing the reverse diffusion pathway.

The methodologies underlying the training and sampling of DDPMs are elucidated through pseudocode in the referenced Algorithm 1 and Algorithm 2. These algorithms detail the procedural steps for both learning the model parameters and generating new samples, providing a comprehensive understanding of the operational framework of DDPMs.

Algorithm 1 DDPM Training

repeat
  $x_0 \sim q(x_0)$
  $t \sim \text{Uniform}\{1, \ldots, T\}$
  $\varepsilon \sim \mathcal{N}(0, I)$
  Take gradient descent step on
  $\nabla_\theta \left\| \varepsilon - \varepsilon_\theta(\sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\varepsilon, t) \right\|^2$
until converged

Algorithm 2 DDPM Sampling

$x_T \sim \mathcal{N}(0, I)$
for $t = T, \ldots, 1$ **do**
  $z \sim \begin{cases} \mathcal{N}(0, I) & \text{if } t > 1 \\ 0 & \text{else} \end{cases}$
  $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \varepsilon_\theta(x_t, t) \right) + \sigma_t z$
end for
return x_0

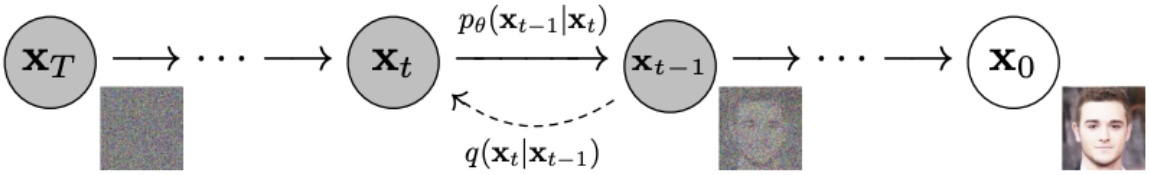

Figure 2: Mind Map of diffusion models [4]

3 Guiderails for Constrained Generation

Diffusion models have emerged as a powerful class of generative models, enabling the generation of high-quality, diverse samples across various domains, including images, text, and au-

dio. To direct the generation process towards desired outcomes, various guiderail mechanisms have been developed, including explicit conditioning, classifier guidance, and classifier-free guidance. These mechanisms constrain the generative process, ensuring that the output adheres to specific criteria or characteristics, thereby enhancing the model's utility for practical applications.

3.1 Explicit Conditioning

Explicit conditioning involves providing the diffusion model with additional context or information, guiding the generation process towards a specific outcome. This can be in the form of textual descriptions, labels, or any form of metadata that describes the desired output characteristics. For instance, in image generation, a text description can serve as a condition to generate images that match the described content. The effectiveness of explicit conditioning lies in its ability to leverage the conditional distribution learned by the model to produce outputs that closely align with the provided context or information.

3.2 Classifier Guidance

Classifier guidance integrates a separate classifier model to steer the generation process of the diffusion model. The classifier is trained to distinguish between desirable and undesirable outputs based on predefined criteria. During generation, the gradient signals from the classifier are used to adjust the diffusion process, pushing the generated samples towards the characteristics identified as desirable. This method allows for fine-grained control over the generation process, enabling the production of outputs that meet specific quality or content standards.

3.3 Classifier-Free Guidance

Classifier-free guidance is a technique that eliminates the need for a separate classifier model. Instead, it leverages the inherent capability of the diffusion model to differentiate between various outcomes. This is achieved by intermittently conditioning the model on a null input (i.e., providing no specific guidance) and comparing the output against those generated with explicit conditioning. The difference in outputs guides the adjustment of the generation process, encouraging the model to produce samples that are more closely aligned with the desired conditions, even in the absence of a dedicated classifier. This approach simplifies the generation process, reducing the computational overhead and complexity associated with maintaining a separate classifier model. [7]

3.4 Some other methods

Each of these guidance mechanisms serves to constrain and direct the generation process of diffusion models, enabling the creation of outputs that meet specific criteria or exhibit desired characteristics. The choice of mechanism depends on the specific requirements of the application, including the level of control needed, the availability of conditional information, and computational constraints.

The ability to steer image diffusion models enhances personalization, customization, and task-specific image generation. Direct manipulation of the image diffusion process enables adjustments in color variations [16] and facilitates inpainting tasks [17]. Text-guided controls extend these capabilities through prompt modification, CLIP feature manipulation, and cross-attention adjustments [18, 19, 20, 21, 22, 23]. Techniques like encoding segmentation masks into tokens for image generation control in MakeAScene [18], or mapping these masks

into localized token embeddings in SpaText [19], showcase the diversity of approaches. GLIGEN [20] introduces parameter adjustments in attention layers for more grounded generation processes. Personalization is further achieved through methods like Textual Inversion [21] and DreamBooth [24], which fine-tune diffusion models using a collection of user-provided example images. Prompt-based editing methods [23] offer practical solutions for image manipulation. Furthermore, optimization techniques that align the diffusion process with sketches have been proposed [25], alongside studies like MultiDiffusion [26], which explore various control strategies over diffusion models.

3.5 ControlNet

[8] presents ControlNet, a novel neural network architecture designed to incorporate spatial conditioning controls into large pre-trained text-to-image diffusion models. ControlNet leverages the depth and robustness of the encoding layers from these pre-trained models, providing a powerful backbone for learning a variety of conditional controls. By employing "zero convolutions," which are convolution layers initialized with zeros, ControlNet ensures a seamless and noise-free integration of conditions into the diffusion process. This architecture allows for the manipulation of images using various conditions, such as edges, depth, segmentation, and human pose, with or without accompanying text prompts. The experiments demonstrate that ControlNet can effectively handle single or multiple conditions and is robust across different dataset sizes. It offers significant advancements in controlled image generation, potentially broadening the application scope of diffusion models in image editing, content creation, and beyond.

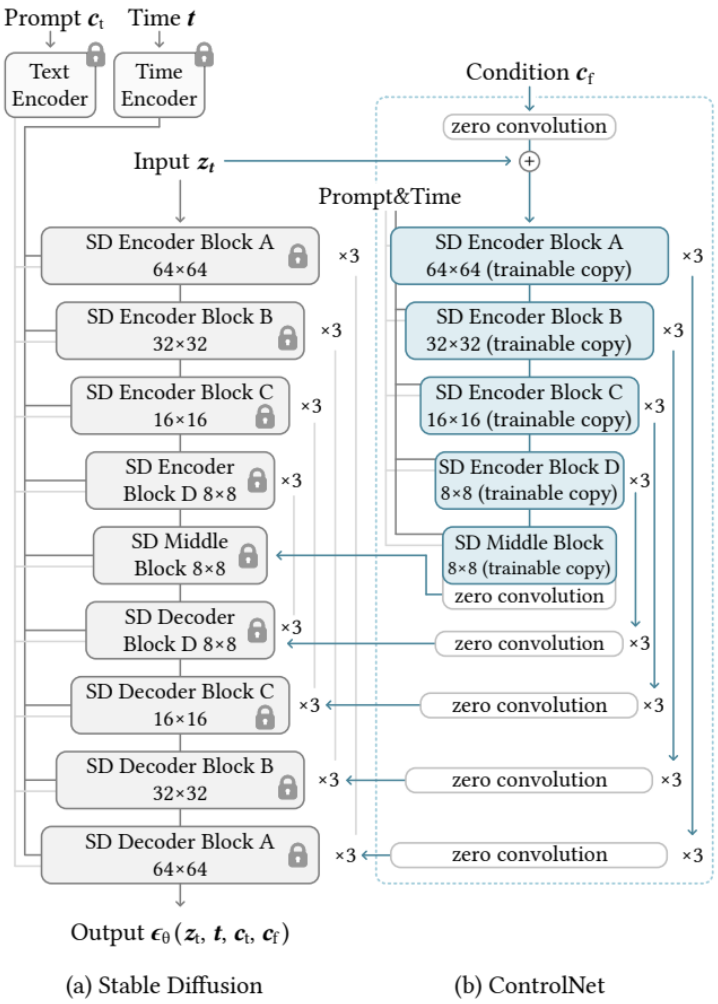

Figure 3: ControlNet: Guiding model through Canny, Depth Map, Pose Map, etc. [8]

Figure 4: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), etc., to control the image generation of large pretrained diffusion models. The default results use the prompt "a high-quality, detailed, and professional image". Users can optionally give prompts like the "chef in kitchen". [8]

4 Learning New Concept

Models such as Stable Diffusion are developed by training on extensive datasets, enabling them to grasp intricate details about a wide range of concepts. Introducing a new concept, such as incorporating images of your dog, to generate further images in various styles, is an intriguing and essential capability for numerous tasks. This includes editing images or transferring the style of your image to create a new one. The ability to seamlessly add personal or unique elements into the model's database not only enhances its versatility but also opens up new avenues for creativity and customization. This process of enriching the model with additional, specific data points allows for a more tailored and personalized user experience, making it invaluable for tasks ranging from personal projects to professional

endeavors that require a high degree of customization and artistic flair.

4.1 Learning

Incorporating new concepts into generative models, particularly diffusion models like Stable Diffusion, represents a significant stride toward enhancing the versatility and personalization capabilities of these models in image generation. This advancement is primarily facilitated through two distinct methodologies: embedding-based learning and full model training.

4.2 Embedding-based Learning

Embedding-based learning focuses on integrating new concepts into the model through specialized embeddings, which are pivotal in teaching the model to recognize and generate images based on newly incorporated ideas. Textual Inversion, for instance, is at the forefront of this approach, specializing in learning textual embeddings for a new concept. This technique enables the model to understand and produce images that embody unique or personal subjects, essentially customizing the model to grasp novel ideas beyond its initial training data [27]. Complementing this, ReVersion shifts the emphasis toward learning embeddings that capture the relational dynamics between concepts. It delves into understanding the interconnections between new and existing concepts within the model, thus facilitating a more nuanced and context-aware integration of new ideas [28]. This embedding-centric approach not only enhances the model's generative capabilities but also ensures that the integration of new concepts is done in a manner that respects the nuanced relationships between different ideas, enabling more cohesive and contextually relevant outputs.

Figure 5: Textual Inversion Results [21]

4.3 Full Model Training

On the other side of the spectrum, full model training approaches such as Dreambooth push the boundaries by employing autogenous, class-specific prior preservation. This technique fine-tunes the entire model to adeptly generate images of a specific subject in various contexts, thereby broadening the model's versatility and adaptive capabilities [24]. Similarly, Custom Diffusion updates the model by fine-tuning a selective set of weights, specifically targeting the key and value mappings in the cross-attention layers. This method underscores a targeted and efficient means of updating the model, ensuring that new concepts are seamlessly woven into the model's fabric without extensive retraining [29]. Furthermore, SINE reimagines model modification through classifier-free guidance (CFG) adjustments, enhancing the model's fidelity to desired outputs in the absence of explicit class labels, thus offering a robust and flexible framework for concept introduction [30]. Lastly, Break-A-Scene introduces a novel paradigm by enabling the learning of multiple concepts from a singular image, deviating from the conventional one-concept-per-multiple-images approach.

This strategy showcases an efficient pathway to enrich the model's comprehension and creative scope [31]. Through these full model training techniques, the landscape of generative modeling is witnessing a paradigm shift towards more dynamic, adaptive, and personalized image generation, promising a future where models can more accurately reflect the diversity and specificity of human imagination.

The convergence of these methodologies underlines a pivotal evolution in the development of diffusion models, marking a significant leap toward creating more adaptive, personalized, and context-aware generative models.

$$\mathbb{E}_{x,c,e,\varepsilon,t}\left[w_t \left\|\mathcal{X}_\theta(\alpha_t X + \sigma_t \varepsilon, c) - X\right\|^2\right] + \lambda w'_t \left\|\mathcal{X}_\theta(\alpha'_t X_{pr} + \sigma'_t \varepsilon', c_{pr}) - X_{pr}\right\|^2, \tag{4}$$

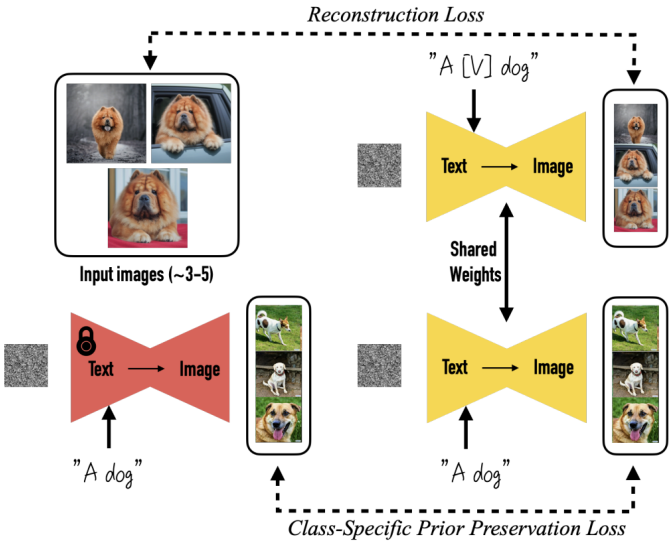

Figure 6: Finetuning Dreambooth: Fine-tuning. Given $\sim 3-5$ images of a subject we finetune a text-to-image diffusion model with the input images paired with a text prompt containing a unique identifier and the name of the class the subject belongs to (e.g., "A [V] dog"), in parallel, we apply a class-specific prior preservation loss, which leverages the semantic prior that the model has on the class and encourages it to generate diverse instances belong to the subject's class using the class name in a text prompt (e.g., "A dog"). [24]

Figure 7: Comparisons with Textual Inversion [20] Given 4 input images (top row), we compare: DreamBooth Imagen (2nd row), DreamBooth Stable Diffusion (3rd row), Textual Inversion (bottom row). Output images were created with the following prompts (left to right): "a [V] vase in the snow", "a [V] vase on the beach", "a [V] vase in the jungle", "a [V] vase with the Eiffel Tower in the background". DreamBooth is stronger in both subject and prompt fidelity. [24]

Name	**Tags**	**Text**	**Citations**
Textual Inversion	Embeddings	Learns textual embedding for new concept	
Dreambooth	Full model new loss	Trains full model using autogenous, class-specific prior preservation loss	[24]
Custom Diffusion	Learns Attention KV	A small subset of model weights, namely the key and value mapping from text to latent features in the cross-attention layers. Fine-tuning these is sufficient to update the model with the new concept	[29]
ReVersion:	Embeddings	Learns embedding to learn relation instead of feature of image	[28]
SINE	Modifies CFG	Model-based classifier-free guidance	[30]
Break-A-Scene	Learning multiple concept	Learning multiple concepts using single image instead of learning one concept using multiple images	

Table 1: Contribution of different concept learning methods

5 Removing Concept

The endeavor to remove specific concepts from diffusion models highlights a critical and multifaceted challenge in the realm of generative artificial intelligence, merging technical sophistication with ethical considerations. As these models increasingly permeate various sectors, the generation of undesired or controversial content raises significant concerns, necessitating interventions that are not only technologically adept but also cognizant of broader societal implications and individual privacy concerns. This complex interplay demands solutions that adeptly navigate the technical intricacies of altering model outputs without compromising on accuracy or creativity, while also ensuring adherence to ethical standards, legal requirements, and cultural sensitivities. As such, the pursuit of concept removal from diffusion models transcends mere algorithmic adjustments, embodying a broader commitment to responsible AI development that respects human values and rights, thereby reinforcing the importance of aligning technological advancements with ethical imperatives in the development and deployment of generative models.

5.1 Dataset Curation and Post-Hoc Modifications

Methods to mitigate undesirable image generation have primarily followed two paths. The first approach involves censoring or selectively curating the training dataset to exclude specific classes of images that are considered undesirable, such as removing all images of people or more narrowly targeting specific undesired content [32, 33, 34, 35]. While straightforward, this method is notably resource-intensive due to the significant computational demands of retraining large models. Furthermore, it risks unintended consequences stemming from large-scale censorship [36].

The alternative, post-hoc strategy modifies the output after training. This can be achieved through the use of classifiers to filter out undesired content [37, 38, 39] or by incorporating guidance into the inference process [40]. Although these methods are more cost-efficient and quicker to implement, they suffer from vulnerability; knowledgeable users can bypass these safeguards by manipulating model parameters [41].

In light of these limitations, recent developments have seen the emergence of novel techniques. For example, Stable Diffusion 2.0 represents an effort to retrain models on censored datasets [42], while Safe Latent Diffusion introduces state-of-the-art guidance-based approaches [40]. Our investigation builds upon these foundations, proposing a third methodology that refines model parameters through a guidance-based model-editing technique, offering both rapid deployment and resilience against circumvention efforts.

5.2 Image Cloaking

Image cloaking emerges as a proactive measure enabling content creators to shield their works from being learned by generative models. By introducing adversarial perturbations, artists can obscure their images, effectively disguising them from machine learning algorithms either during training or inference, without impacting human perception significantly [43, 44]. This approach, however, diverges from the primary concern of concept removal from the perspective of model creators.

5.3 Model Editing

As an extension of concept removal, model editing seeks to adjust the generative capabilities of models without extensive retraining. Techniques vary from modifying specific neurons or layers [45, 46] to employing hypernetworks [47, 48] for text generators, and analogous methods for image synthesis, such as using textual or sketch inputs, gestures, or direct editing of features [49, 50, 51, 52]. The evolution towards model editing underscores the industry's shift towards more efficient and nuanced control over generative models, aligning with our proposed method of parameter tuning for concept removal. Among these advancements, the "Forget-Me-Not" [53] approach represents a significant stride, utilizing a guidance-based model-editing technique that refines model parameters to selectively forget or correct specific concepts. This method enhances the flexibility and efficiency of concept manipulation in diffusion models, offering a solution that is both quick to deploy and robust against attempts to circumvent concept removal safeguards. The evolution towards model editing, as exemplified by "Forget-Me-Not," underscores the industry's shift towards more efficient and nuanced control over generative models, aligning with our proposed method of parameter tuning for concept removal.

Methods	**Performance**	**Integrity**	**Generality**	**Flexibility**
Token Blacklisting	No forgetting	Inevitably affects other concepts sharing overlapping prompts	Within the vocabulary of the tokenizer	Tokenizer required
Naive Finetuning	Successfully removes concept	Removes unrelated concept by fault	Applies to any concepts with sufficient data.	Applies to any models
Forget-Me-Not	Successfully removes concept	Maintains most of the model's integraity.	Applies to any concepts with few data samples	Only applies to models with cross attention

Table 2: This table compares pros (green) and cons (red) on the four major aspects of concept forgetting between baselines and the proposed Forget-Me-Not. If an approach can handle an aspect to some extent, the corresponding explanation is marked in yellow. [53]

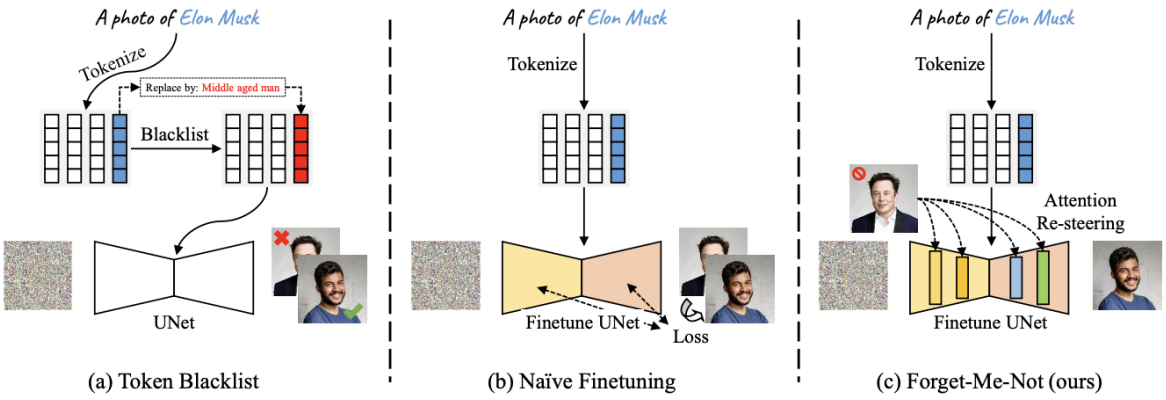

Figure 8: This figure shows two baseline forgetting methods and our proposed Forget-Me-Not. The target concept to forget is Elon Musk. One baseline is (a) Token Blacklist that simply replaces the target token with a different one. The other baseline is (b) Naive Fintuning in which instead of replacing tokens, it finetunes model weights so that the new weights generate outputs containing unrelated concepts. Our method (c) Forget-Me-Not utilizes Attention Re-steering in which we finetune only UNet to minimize each of the intermediate attention maps associated with the target concepts to forget. [53]

6 Applications

In the rapidly evolving landscape of artificial intelligence, diffusion models have emerged as a groundbreaking approach, offering transformative capabilities across a wide array of applications. These models, through a process of iterative refinement, have unlocked unprecedented potential in creative and analytical tasks alike. Among their most notable applications are Text to Image generation, where models synthesize visually compelling images from descriptive text, blending creativity with precision to bridge the gap between textual concepts and their visual representations. In Image to Image translation, they adeptly modify or transform one image into another style or context, showcasing their versatility in reimagining visual content. Image editing with diffusion models goes a step further, allowing for nuanced adjustments and enhancements that redefine the boundaries of digital artistry. In

the realm of Video, these models extend their prowess to dynamic sequences, enabling the generation and modification of video content with an eye for detail and continuity. The exploration into 3D brings a spatial dimension to their capabilities, where diffusion models generate and manipulate three-dimensional objects and environments, opening new frontiers in virtual and augmented reality. Lastly, their application in Document Layout emphasizes their utility beyond the visual arts, automating the design of document structures in a way that balances aesthetics with information delivery. This diverse spectrum of applications not only highlights the versatility of diffusion models but also sets the stage for their future developments, promising to redefine the intersection of technology and creativity.

6.1 Text to Image

Diffusion models have demonstrated remarkable achievements in the domain of text-to-image synthesis, showcasing their ability to creatively merge distinct concepts such as objects, shapes, and textures, thereby producing innovative imagery. This assertion was validated through the application of Stable Diffusion [54] for crafting images from diverse textual descriptions, as illustrated in Figure 2.

The Imagen model, introduced by Saharia et al. [13], represents a method for text-to-image synthesis, incorporating a text encoder and a series of diffusion models for crafting high-resolution images, conditioned on text embeddings produced by the encoder. Additionally, a novel benchmark for text-to-image evaluation named DrawBench was proposed. In terms of architecture, the authors developed Efficient U-Net to enhance efficiency, applying this design within their text-to-image synthesis experiments.

Gu et al. [55] put forward the VQ-Diffusion model, a novel approach for text-to-image

synthesis that eliminates the unidirectional bias found in earlier methods. Through its unique masking strategy, it prevents error accumulation during inference. The model operates in two phases: initially utilizing a VQ-VAE to represent images as discrete tokens, followed by a discrete diffusion model acting on the VQ-VAE's latent space, guided by caption embeddings. This process is inspired by masked language modeling, replacing some tokens with a [mask] token.

Avrahami et al. [56] developed a text-conditional diffusion model reliant on CLIP [57] embeddings for both image and text. Employing a dual-stage approach, the first generates the image embedding, while the second, acting as a decoder, synthesizes the final image based on both the image embedding and the text caption. To create image embeddings, a diffusion model in the latent space was utilized, with a subjective human evaluation assessing the generative outcomes.

Addressing the slow sampling issue inherent in diffusion models, Zhang et al. [**zhang2022fast**] introduced a novel discretization strategy that minimizes error and allows for larger step sizes, thereby reducing the number of required sampling steps. Utilizing high-order polynomial extrapolations for the score function and an Exponential Integrator to solve the reverse SDE, they significantly decreased the number of network evaluations needed without compromising the models' generative capabilities.

Shi et al. [58] merged a VQ-VAE [59] with a diffusion model for image generation. Initiating with the VQ-VAE for encoding, they substituted the decoder with a diffusion model, applying the U-Net architecture [60] and incorporating image tokens into its mid-block.

Expanding on the concepts in Blattmann et al. [61], Rombach et al. [62] introduced a

modification for crafting artistic images by extracting nearest neighbors in the CLIP [57] latent space from a dataset, then guiding the reverse denoising process with these embeddings. Given the shared CLIP latent space for text and images, diffusion can also be text-guided. At inference, an artistic image database substitutes the original, steering the model to generate images reflective of the new dataset's style.

Jiang et al. [63] unveiled a framework to generate full-body human images with detailed clothing representation from three inputs: a human pose, and text descriptions for both the clothing's shape and texture. The initial phase encodes the shape description into an embedding vector, infusing it into a generative encoder-decoder module for shape mapping. Subsequently, a diffusion-based transformer leverages multiple texture-specific, multi-level codebooks for sampling the texture description's embedded representation, as suggested in VQ-VAE [59]. Initially, coarse-level codebook indices are sampled, with finer levels predicted via a feed-forward network, utilizing Sentence-BERT [64] for text encoding.

6.2 Image To Image

Saharia and colleagues [65] introduce a unified framework for converting images across different contexts using diffusion models, concentrating on tasks such as colorization, inpainting, extension of image borders, and enhancement of compressed images. This framework remains consistent across these tasks, avoiding the necessity for individual modifications per task. They start by evaluating the effectiveness of L1 versus L2 losses, advocating for the latter due to its ability to increase the variety of generated samples. They also underscore the pivotal role of self-attention mechanisms in generating conditional images.

In an effort to facilitate image translation without corresponding pairs, Sasaki and team

[66] develop a strategy that employs two diffusion models trained in tandem. Each model, through the process of reverse denoising, incorporates feedback from the other model's ongoing outputs. Additionally, they introduce a cycle-consistency loss to refine the training of these models.

Zhao et al. [67] focus on enhancing the efficacy of image translation models based on diffusion by equally valuing source domain data. They use an energy-based approach that operates on both the source and target domains to guide the stochastic differential equation (SDE) solver. This method produces images that maintain universal features while accurately transferring domain-specific attributes. The model utilizes dual feature extractors, each tailored to a particular domain.

Wang and colleagues [68] leverage the pretrained GLIDE model to create a semantically rich latent space for image generation. By modifying the model's architecture to suit various conditions and then fine-tuning it for specific tasks, they achieve notable improvements. This process involves an initial focus on training a new encoder while keeping the decoder static, followed by concurrent training of both components. Adversarial training techniques and normalization strategies are also applied to further refine the quality of generated images.

Li et al. [69] propose a novel image translation diffusion model that combines the concepts of Brownian bridges and Generative Adversarial Networks (GANs). Their approach starts with encoding images via a Vector Quantized GAN, followed by a diffusion process that acts as a Brownian bridge in the quantized latent space, facilitating the transition between source and target domains. Subsequently, another VQ-GAN decodes these quantized representations to generate the translated image. Each GAN is trained independently within its respective domain.

Building on their previous research, Wolleb and associates [70] enhance their diffusion model by integrating a task-specific model in lieu of the standard classifier. This model enriches the sampling process by incorporating gradients from a network designed for a specific application, demonstrated through either a regression or segmentation task. This approach benefits from the existing diffusion model frameworks, eliminating the necessity for comprehensive retraining except for the component tailored to the specific task.

6.3 Image editing

Meng et al. [71] explored the application of diffusion models in tasks such as stroke painting, stroke-based editing, and image composition, starting with images that offer some form of guidance. These models preserve the original images' characteristics while smoothing out deformations through a forward diffusion process, and subsequently, a reverse process is applied to denoise these images to produce realistic outputs according to the provided guidance, effectively solving the reverse stochastic differential equation (SDE) without necessitating specialized datasets or modifications in training.

An earlier method for modifying specific image regions based on natural language descriptions was introduced [56], where users define regions for editing via a mask. This approach uses CLIP guidance to generate images according to text inputs. However, it was observed that merging the generated output with the original image did not always result in a globally coherent image. To overcome this, a modified denoising process was applied that incorporates the masked latent image with a noisy version of the original image during each iteration.

Advancing this work, Avrahami et al. applied latent diffusion models for localized image

editing through text. This method encodes both the image and a dynamically adjustable mask into the latent space, where a diffusion process guided by textual descriptions within the target area takes place [72]. Inspired by Blended Diffusion [56], this technique uniquely combines the masked region in latent space with the contemporaneously noised image before decoding, resulting in enhanced performance and efficiency.

6.4 Video

In the exploration of video diffusion models, a broad spectrum of video analysis techniques is examined, encompassing the generation, modification, and interpretation of video content. These methods frequently adopt diffusion generation approaches or capitalize on the advanced generative capacities of diffusion models for subsequent applications. This survey prioritizes areas such as the creation of videos from text [73, 74, 75], generating videos without predefined conditions [76, 77, 78], and video modification guided by text descriptions [79, 80, 81], among others.

6.5 Creation of Videos from Text

Creation of Videos from Text is concerned with the automated transformation of textual narratives into videos. This requires the interpretation and conversion of text-described scenes, objects, and actions into a sequence of frames that visually cohere, thus yielding a video that is consistent both logically and visually. This technology finds application across various domains, including the automated production of films [82], animations [83, 84], virtual reality, and educational materials [85].

6.6 Video Generation Without Predefined Conditions

Video Generation Without Predefined Conditions describes the challenge of producing a continuous, visually coherent series of video frames from either stochastic noise or a predetermined starting point, with no reliance on specific input conditions. This mode of video generation distinguishes itself by not necessitating any external data or conditions [76, 77, 78], thus challenging the generative model to independently learn the necessary temporal dynamics and visual elements for realistic and varied video content. This underscores the model's ability to understand content from unlabelled data and its potential for showcasing diversity.

6.7 Text-driven Video Modification

Text-driven Video Modification involves leveraging textual instructions to direct video content alterations. Through this method, a textual input specifies the desired video modifications, and the system employs this description to identify and implement changes based on the text-identified objects, actions, or scenes. This approach enhances editing efficiency and intuitiveness by enabling the use of natural language to describe edits [79, 80, 81], potentially diminishing the reliance on meticulous, frame-by-frame adjustments.

6.8 3D

The DreamFusion [86] method represents a significant milestone in the field of artificial intelligence and computer vision, particularly in the synthesis of 3D models from textual de-

scriptions. This innovative approach leverages a loss based on probability density distillation to utilize a pretrained 2D diffusion model as a prior for the optimization of a parametric image generator. Essentially, it optimizes a randomly-initialized Neural Radiance Field (NeRF) through a process akin to DeepDream, using gradient descent to ensure that the 3D model's 2D renderings from various angles align with the set low-loss criteria. This process enables the creation of 3D models that can be viewed from any angle, lit by any light source, and integrated into any 3D environment without the need for 3D training data or modifications to the existing diffusion model, showcasing the effectiveness of pretrained image diffusion models as priors.

The methodology behind DreamFusion involves initializing a NeRF-like model with random weights and iteratively refining it. This refinement process includes rendering views of the NeRF from random camera positions and angles, calculating a score distillation loss function that integrates with an existing image model like Imagen, and updating the NeRF parameters through an optimizer based on the computed gradients of the score distillation loss. This process is repeated for each text prompt, with each optimization cycle consisting of random sampling of camera and lighting conditions, rendering of the NeRF, and subsequent parameter updates. This iterative optimization results in the gradual refinement of the 3D model to more accurately reflect the textual description it is based on.

The compelling results of DreamFusion's process highlight the model's capability to produce detailed and realistic 3D renderings that closely match the given textual prompts. The provided images demonstrate the versatility of the model in generating a variety of scenes and objects, viewable from multiple perspectives. This underscores the model's potential in creating dynamic and interactive 3D models for a range of applications in 3D modeling, virtual reality, and more. DreamFusion's novel approach sets a new standard in the text-to-3D synthesis domain, offering a highly adaptable and efficient framework for the generation of

intricate 3D models directly from text, thereby opening new avenues for both creative and practical applications in the field.

6.9 Document Layout

The paper LayoutDM [87] by Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi introduces a novel method for generating graphic layouts through a process named LayoutDM. This process leverages discrete state-space diffusion models, offering a significant advancement in controllable layout generation.

Graphic layouts are essential for visual communication, playing a pivotal role in various applications ranging from printed media to user interface design. The core challenge addressed by this work is the generation of plausible arrangements of elements within a layout, considering optional constraints such as the type or position of specific elements. Traditional methods often struggle with this task due to the structured nature of layout data, which necessitates careful consideration of the relationships between elements.

LayoutDM stands out by its use of discrete state-space diffusion models for layout generation. The model corrupts a layout in a modality-wise manner during the forward process and progressively denoises it, considering all elements and modalities in the reverse process. This approach naturally handles structured layout data in a discrete representation, learning to infer a noiseless layout from an initially corrupted input. Moreover, LayoutDM introduces a novel method to handle variable-length layout data by extending the discrete state space with a special PAD token, facilitating the incorporation of complex layout constraints through logit adjustment during inference.

The researchers demonstrate LayoutDM's superior performance over existing methods through extensive experiments on various layout generation tasks, utilizing large-scale datasets like Rico and PubLayNet. Their results showcase not only high-quality layout generation but also significant improvements over both task-specific and task-agnostic baselines.

A notable contribution of LayoutDM is its flexibility in conditional layout generation, addressing limitations of both autoregressive and non-autoregressive models. It enables the generation of variable-length elements without the need for additional training or external models, presenting a solution to the immutable dependency chain issue prevalent in autoregressive models. Through techniques such as masking and logit adjustment, LayoutDM effectively injects layout constraints during inference, demonstrating its ability to solve a wide range of tasks with a single model.

Furthermore, the paper explores various aspects crucial for effective layout generation, including modality-wise diffusion, adaptive quantization, and decoupled positional encoding. These innovations contribute to the model's ability to accurately capture and reproduce the structured and highly variable nature of layout data.

In summary, LayoutDM represents a significant step forward in the field of layout generation, providing a versatile and effective tool for synthesizing graphic layouts with optional constraints. Its approach combines the strengths of diffusion models with innovative techniques for handling structured and variable-length data, offering promising directions for future research in visual communication and design automation.

7 Future research areas

In recent years, diffusion models have made remarkable strides in the field of artificial intelligence, particularly in image, video, and 3D model generation. As outlined in the comprehensive survey, these models have evolved from theoretical concepts in thermodynamics to crucial tools in modern machine learning. While they have demonstrated impressive capabilities, there remains a vast scope for further advancement and refinement. The future research directions in this field are aimed at overcoming existing limitations and unlocking new possibilities, ensuring that diffusion models continue to evolve and remain at the forefront of technological innovation.

7.1 Diverse Training Environments

Exploring the integration of various data sources such as audio, video, and tactile feedback to enhance the training of diffusion models. This diversification can lead to more versatile models capable of handling a broader range of applications, including multimodal interaction and immersive virtual environments.

7.2 Enhanced Latent Representations

Investigating advanced techniques for improving the latent space representations of diffusion models. This includes developing methods for better interpretability, increased efficiency, and enhanced accuracy in handling complex data. Research in this area can significantly improve the models' ability to understand and generate more nuanced and contextually

relevant content.

7.3 Lower Computational Costs

Focusing on developing more efficient algorithms and exploring novel hardware solutions to reduce the computational expense associated with training and operating diffusion models. This effort can make these models more accessible and sustainable, especially for applications requiring real-time processing or deployment in resource-constrained environments.

7.4 Generating Finer Details

Enhancing the capability of diffusion models to accurately generate fine details such as realistic text within images, correct anatomical features (like the correct number of fingers), and facial expressions. This requires advancements in both the models' understanding of intricate details and their ability to replicate these details accurately in the generated content.

7.5 Ethical and Responsible AI

As diffusion models become more powerful and widespread, it's crucial to address the ethical implications of their use. This includes developing guidelines and technologies for preventing misuse (such as deepfakes), ensuring privacy, and promoting fairness and inclusivity in the generated content.

7.6 Interactivity and User Control

Researching ways to enhance user interaction with diffusion models, allowing users to specify detailed preferences and control the generation process more precisely. This can include developing intuitive interfaces and control mechanisms that cater to both expert users and the general public.

7.7 Integration with Other AI Technologies

Exploring how diffusion models can be effectively combined with other AI technologies like reinforcement learning, symbolic AI, or decision-making algorithms to create more comprehensive AI systems capable of complex tasks like automated storytelling, content creation, or advanced simulations.

7.8 Domain-Specific Applications

Investigating the application of diffusion models in specific domains such as healthcare, education, and environmental modeling. This involves customizing the models to handle domain-specific data and requirements, potentially leading to breakthroughs in these fields.

8 Conclusion

In this comprehensive survey, we traversed the expansive and dynamic terrain of diffusion models, from their theoretical underpinnings in thermodynamics to their revolutionary applications in artificial intelligence, specifically in the domains of image, video, 3D model generation, and beyond. The journey through the evolution of diffusion models revealed their profound impact on the field of machine learning, showcasing their ability to generate high-quality, diverse outputs that push the boundaries of creativity and realism.

Throughout this exploration, we encountered the pivotal advancements that have cemented diffusion models as a cornerstone of modern generative AI. From the early implementations in image synthesis to the intricate mechanisms enabling text-to-image translations, each development has contributed to the models' increasing sophistication and versatility. We delved into the challenges that accompany diffusion models, such as computational efficiency and alignment with complex text prompts, and highlighted the innovative solutions proposed by the research community to address these issues.

The application of diffusion models extends beyond mere content creation; they have proven instrumental in editing, translating, and even removing content, demonstrating an unparalleled flexibility. This versatility, coupled with the continuous advancements in model architecture and training methodologies, signals a future where diffusion models could offer even more personalized, efficient, and ethically conscious solutions to a wide array of challenges in AI and beyond.

As we look to the horizon, the potential of diffusion models appears boundless. The ongoing research into enhancing their efficiency, accuracy, and ethical considerations paints a promising picture for the future. By continuing to refine and expand upon the capabili-

ties of diffusion models, we can anticipate a new era of AI-driven creativity and innovation, where the boundary between human and machine-generated content blurs, leading to unprecedented opportunities for artistic expression, technological advancement, and societal impact.

In conclusion, this survey underscores the significant strides made in the development of diffusion models and their applications. It highlights the model's transformative potential across various domains, advocating for continued exploration and innovation to unlock even greater capabilities. The journey of diffusion models is far from complete; it is a field ripe with opportunities for discovery, improvement, and revolutionary applications that will continue to shape the future of artificial intelligence.

References

[1] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *CoRR* abs/1312.6114 (2013). URL: http://arxiv.org/abs/1312.6114.

[2] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[3] Jascha Sohl-Dickstein et al. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*. Ed. by Francis Bach and David Blei. Lille, France: PMLR, 2015, pp. 2256–2265. URL: https://proceedings.mlr.press/v37/sohl-dickstein15.html.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". Version 2. In: *arXiv preprint arXiv:2006.11239* (2020). DOI: 10.48550/arXiv.2006.11239. URL: https://doi.org/10.48550/arXiv.2006.11239.

[5] Robin Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". Version 2. In: *arXiv preprint arXiv:2112.10752* (2022). DOI: 10.48550/arXiv.2112.10752. URL: https://doi.org/10.48550/arXiv.2112.10752.

[6] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". Version 1. In: *arXiv preprint arXiv:2103.00020* (2021). DOI: 10.48550/arXiv.2103.00020. URL: https://doi.org/10.48550/arXiv.2103.00020.

[7] Jonathan Ho and Tim Salimans. "Classifier-Free Diffusion Guidance". Version 1. In: *arXiv preprint arXiv:2207.12598* (2022). DOI: 10.48550/arXiv.2207.12598. URL: https://doi.org/10.48550/arXiv.2207.12598.

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. "Adding Conditional Control to Text-to-Image Diffusion Models". Version 3. In: *arXiv preprint arXiv:2302.05543* (2023). DOI: 10.48550/arXiv.2302.05543. URL: https://doi.org/10.48550/arXiv.2302.05543.

[9] Diederik Kingma et al. "Variational Diffusion Models". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 21696–21707.

[10] Prafulla Dhariwal and Alexander Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 8780–8794.

[11] Alex Nichol et al. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*. Available at SSRN. 2022.

[12] RunwayML. *Stable Diffusion v1.5 Model Card*. https://huggingface.co/runwayml/stable-diffusion-v1-5. Accessed: 2022-09-30. 2022.

[13] Chitwan Saharia et al. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”. In: *arXiv preprint arXiv:2205.11487* (2022).

[14] OpenAI. *DALL·E 2*. `https://openai.com/product/dall-e-2`. Accessed: 2023-09-30. 2023.

[15] Midjourney. *Midjourney*. `https://www.midjourney.com/`. Accessed: 2023-09-30. 2023.

[16] Chenlin Meng et al. “SDedit: Guided Image Synthesis and Editing with Stochastic Differential Equations”. In: *International Conference on Learning Representations* (2021).

[17] Omri Avrahami, Dani Lischinski, and Ohad Fried. “Blended Diffusion for Text-Driven Editing of Natural Images”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18208–18218.

[18] Oran Gafni et al. “Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors”. In: *European Conference on Computer Vision (ECCV)*. Springer. 2022, pp. 89–106.

[19] Omri Avrahami et al. “Spatext: Spatio-Textual Representation for Controllable Image Generation”. In: *arXiv preprint arXiv:2211.14305* (2022).

[20] Yuheng Li et al. “GLIGEN: Open-Set Grounded Text-to-Image Generation”. In: 2023.

[21] Rinon Gal et al. “An Image is Worth One Word: Personalizing Text-to-Image Generation Using Textual Inversion”. In: *arXiv preprint arXiv:2208.01618* (2022).

[22] Nataniel Ruiz et al. “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation”. In: *arXiv preprint arXiv:2208.12242* (2022).

[23] Tim Brooks, Aleksander Holynski, and Alexei A Efros. “InstructPix2Pix: Learning to Follow Image Editing Instructions”. In: *arXiv preprint arXiv:2211.09800* (2022).

[24] Nataniel Ruiz et al. "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation". Version 2. In: *arXiv preprint arXiv:2208.12242* (2023). DOI: 10.48550/arXiv.2208.12242. URL: https://doi.org/10.48550/arXiv.2208.12242.

[25] Andrey Voynov, Kfir Abernan, and Daniel Cohen-Or. "Sketch-Guided Text-to-Image Diffusion Models". In: 2022.

[26] Omer Bar-Tal et al. "MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation". In: *arXiv preprint arXiv:2302.08113* (2023).

[27] Rinon Gal et al. "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion". Version 1. In: *arXiv preprint arXiv:2208.01618* (2022). DOI: 10.48550/arXiv.2208.01618. URL: https://doi.org/10.48550/arXiv.2208.01618.

[28] Ziqi Huang et al. "ReVersion: Diffusion-Based Relation Inversion from Images". Version 1. In: *arXiv preprint arXiv:2303.13495* (2023). DOI: 10.48550/arXiv.2303.13495. URL: https://doi.org/10.48550/arXiv.2303.13495.

[29] Nupur Kumari et al. "Multi-Concept Customization of Text-to-Image Diffusion". Version 2. In: *arXiv preprint arXiv:2212.04488* (2023). DOI: 10.48550/arXiv.2212.04488. URL: https://doi.org/10.48550/arXiv.2212.04488.

[30] Zhixing Zhang et al. "SINE: SINgle Image Editing with Text-to-Image Diffusion Models". Version 1. In: *arXiv preprint arXiv:2212.04489* (2022). DOI: 10.48550/arXiv.2212.04489. URL: https://doi.org/10.48550/arXiv.2212.04489.

[31] Omri Avrahami et al. "Break-A-Scene: Extracting Multiple Concepts from a Single Image". Version 2. In: *arXiv preprint arXiv:2305.16311* (2023). DOI: 10.48550/arXiv.2305.16311. URL: https://doi.org/10.48550/arXiv.2305.16311.

[32] Alex Nichol et al. "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models". In: *arXiv preprint arXiv:2112.10741* (2021).

[33] Christoph Schuhmann et al. “LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models”. In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022).

[34] OpenAI. “DALL-E 2 Preview - Risks and Limitations”. In: (2022).

[35] Robin Rombach and Patrick Esser. “Stable Diffusion V2 Model Card”. In: (2022).

[36] Ryan O’Connor. “Stable Diffusion 1 vs 2 - What You Need to Know”. In: (2022).

[37] Praneeth Bedapudi. “NudeNet: Neural Nets for Nudity Detection and Censoring”. In: *2022*. 2022.

[38] Gant Laborde. “NSFW Detection Machine Learning Model”. In: (2022).

[39] Javier Rando et al. “Red-Teaming the Stable Diffusion Safety Filter”. In: *arXiv preprint arXiv:2210.04610* (2022).

[40] Patrick Schramowski et al. “Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models”. In: *arXiv preprint arXiv:2211.05105* (2022).

[41] SmithMano. “Tutorial: How to Remove the Safety Filter in 5 Seconds”. In: (2022).

[42] Robin Rombach. “Stable Diffusion 2.0 Release”. In: (2022).

[43] Hadi Salman et al. “Raising the Cost of Malicious AI-Powered Image Editing”. In: *arXiv preprint arXiv:2302.06588* (2023).

[44] Shawn Shan et al. “GLAZE: Protecting Artists from Style Mimicry by Text-to-Image Models”. In: *arXiv preprint arXiv:2302.04222* (2023).

[45] Damai Dai et al. “Knowledge Neurons in Pretrained Transformers”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 8493–8502.

[46] Kevin Meng et al. “Locating and Editing Factual Associations in GPT”. In: *Advances in Neural Information Processing Systems*. 2022.

[47] Nicola De Cao, Wilker Aziz, and Ivan Titov. "Editing Factual Knowledge in Language Models". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 6491–6506.

[48] Eric Mitchell et al. "Fast Model Editing at Scale". In: *International Conference on Learning Representations*. 2021.

[49] David Bau et al. "Rewriting a Deep Generative Model". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020.

[50] Rinon Gal et al. "StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators". In: *ACM Transactions on Graphics (TOG)* 41.4 (2022), pp. 1–13.

[51] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. "Sketch Your Own GAN". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 14050–14060.

[52] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. "Rewriting Geometric Rules of a GAN". In: *ACM Transactions on Graphics (TOG)* 41.4 (2022), pp. 1–16.

[53] Eric Zhang et al. *Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models*. https://doi.org/10.48550/arXiv.2303.17591. 2023. DOI: 10.48550/arXiv.2303.17591. arXiv: 2303.17591 [cs.CV].

[54] Robin Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *CVPR* (2022).

[55] Shuyang Gu et al. "Vector quantized diffusion model for text-to-image synthesis". In: *CVPR*. 2022, pp. 10696–10706.

[56] Omri Avrahami, Dani Lischinski, and Ohad Fried. "Blended diffusion for text-driven editing of natural images". In: *CVPR*. 2022, pp. 18208–18218.

[57] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *ICML* 139 (2021), pp. 8748–8763.

[58] Jiahui Shi et al. "DiVAE: Photorealistic Image Synthesis with Denoising Diffusion Decoder". In: *arXiv preprint arXiv:2206.00386* (2022).

[59] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. "Neural Discrete Representation Learning". In: *NIPS* 30 (2017), pp. 6309–6318.

[60] Alexander Quinn Nichol and Prafulla Dhariwal. "Improved Denoising Diffusion Probabilistic Models". In: *ICML* (2021), pp. 8162–8171.

[61] Andreas Blattmann et al. "Retrieval-Augmented Diffusion Models". In: *arXiv preprint arXiv:2204.11824* (2022).

[62] Robin Rombach, Andreas Blattmann, and Björn Ommer. "Text-Guided Synthesis of Artistic Images with Retrieval-Augmented Diffusion Models". In: *arXiv preprint arXiv:2207.13038* (2022).

[63] Yining Jiang et al. "Text2Human: Text-Driven Controllable Human Image Generation". In: *ACM Transactions on Graphics* 41.4 (2022), pp. 1–11.

[64] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *EMNLP* (2019), pp. 3982–3992.

[65] Chitwan Saharia et al. *Palette: Image-to-image diffusion models*. 2022. arXiv: 2204.02641 [cs.CV].

[66] Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. *UNIT-DDPM: Unpaired Image Translation with Denoising Diffusion Probabilistic Models*. 2021. arXiv: 2104.05358 [cs.CV].

[67] Mengyu Zhao et al. *EGSDE: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations*. 2022. arXiv: 2207.06635 [cs.CV].

[68] Tinghui Wang et al. *Pretraining is All You Need for Image-to-Image Translation*. 2022. arXiv: 2205.12952 [cs.CV].

[69] Baoyang Li et al. *VQBB: Image-to-image Translation with Vector Quantized Brownian Bridge*. 2022. arXiv: 2205.07680 [cs.CV].

[70] Janis Wolleb et al. *The Swiss Army Knife for Image-to-Image Translation: Multi-Task Diffusion Models*. 2022. arXiv: 2204.02641 [cs.CV].

[71] Chong Meng et al. "SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2021.

[72] Omri Avrahami, Ohad Fried, and Dani Lischinski. "Blended Latent Diffusion". In: *arXiv preprint arXiv:2206.02779* (2022).

[73] U. Singer et al. "Make-a-video: Text-to-video generation without text-video data". In: *ICLR*. 2023.

[74] J. Ho et al. "Imagen video: High definition video generation with diffusion models". In: *arXiv:2210.02303* (2022).

[75] Z. Xing et al. "Simda: Simple diffusion adapter for efficient video generation". In: *arXiv:2308.09710* (2023).

[76] J. Ho et al. "Video diffusion models". In: *NeurIPS*. 2022.

[77] Y. Hu, Z. Chen, and C. Luo. "Lamd: Latent motion diffusion for video generation". In: *arXiv:2304.11603* (2023).

[78] K. Mei and V. Patel. "Vidm: Video implicit diffusion models". In: *AAAI*. 2023.

[79] J. Z. Wu et al. "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation". In: *ICCV* (2023).

[80] P. Esser et al. "Structure and content-guided video synthesis with diffusion models". In: *ICCV*. 2023.

[81] E. Molad et al. "Dreamix: Video diffusion models are general video editors". In: *arXiv:2302.01329* (2023).

[82] J. Zhu et al. “Moviefactory: Automatic movie creation from text using large generative models for language and images”. In: *arXiv:2306.07257* (2023).

[83] Y. He et al. “Animate-a-story: Storytelling with retrieval-augmented video generation”. In: *arXiv:2307.06940* (2023).

[84] Y. Guo et al. “Animatediff: Animate your personalized text-to-image diffusion models without specific tuning”. In: *arXiv:2307.04725* (2023).

[85] M. Yang et al. “Probabilistic adaptation of text-to-video models”. In: *arXiv:2306.01872* (2023).

[86] Ben Poole et al. *DreamFusion: Text-to-3D using 2D Diffusion*. Submitted on 29 Sep 2022 (v1), last revised version. 2022. DOI: 10.48550/arXiv.2209.14988. arXiv: 2209.14988 [cs.CV]. URL: https://doi.org/10.48550/arXiv.2209.14988.

[87] Naoto Inoue et al. “LayoutDM: Discrete Diffusion Model for Controllable Layout Generation”. In: *arXiv:2303.08137* (2023). To be published in CVPR2023. URL: https://doi.org/10.48550/arXiv.2303.08137.