Conditional Image Generation Technology Based on Diffusion Models

Chenmingzhe Zhuo *

Department of Computing and Mathematical Sciences, Fujian University of Technology, Fujian, 350118, China

* Corresponding Author Email: 3241311126@smail.fjut.edu.cn

Abstract. Conditional Image Generation (CIG) refers to the process of learning and sampling image distributions that satisfy given explicit condition variables using generative models, thereby synthesising high-quality images with specified semantic, structural, or stylistic attributes. In recent years, diffusion models have demonstrated significant advantages in conditional image generation tasks, driving a paradigm shift from "random generation" to "controllable creation." This paper provides a systematic review of research on conditional image generation based on diffusion models: it illustrates the fundamental principles and methods of diffusion models, and introduces the current mainstream model development trends from several perspectives, including semantic precise control, spatial structure constraints, style variability, heterogeneous modality fusion, and dynamic temporal generation. It summarises the latest results from benchmark datasets, such as MS-COCO, DrawBench, and T2I-CompBench, as well as evaluation metrics like FID and CLIP Score. It also discusses future challenges such as large-scale unified models, physical consistency, privacy protection, and edge deployment, and looks forward to potential breakthroughs in content creation, autonomous driving, medical imaging, and virtual reality scenarios. This review aims to provide researchers with a comprehensive technical roadmap, promoting continuous innovation in the theory and applications of conditional image generation.

Keywords: Conditional Image Generation; Diffusion Models; Semantic Alignment and Spatial Control; Multimodal Fusion; Style Transfer and Dynamic Generation.

1. Introduction

In the field of conditional image generation, traditional generative models, such as variational autoencoders (VAEs) and generative adversarial networks (GANs), have made significant progress in tasks like semantic control and style transfer. However, they still face issues such as insufficient generation quality, missing details, mode collapse, and unstable training when dealing with complex, multi-dimensional conditional constraints. These limitations restrict the models' performance in fine-grained control, high-resolution generation, and multimodal fusion in practical applications, potentially leading to generated results that do not match the conditional semantics, exhibit geometric distortions, or result in inaccurate style control, thereby weakening the models' applicability in industrial design, content creation, and scientific computing [1].

To address these issues, in 2020, Jonathan Ho proposed the denoising diffusion probabilistic model (DDPM), which formalised the diffusion process as a Markov chain and introduced noise prediction loss (simplified evidence lower bound, or ELBO), significantly enhancing the detail fidelity and diversity of generated images [2]. Diffusion models can incorporate external conditional information into the denoising process through classifier guidance, classifier-free guidance, and cross-attention mechanisms, achieving high-precision semantic alignment and controllable generation in conditional generation tasks. Compared to VAE and GAN, diffusion models not only excel in generation quality but also offer flexible architecture, stable training, and a preference for high-frequency details, promoting a paradigm shift from "random generation" to "controllable creation" in conditional image generation.

In recent years, diffusion models have continued to optimise in multiple directions, such as sampling acceleration (e.g., non-Markovian sampling in DDIM), lightweight and edge deployment (e.g., latent space inference in Stable Diffusion), data generalisation enhancement, and multi-modal

fusion. These advancements have enabled diffusion models to excel in semantic precision control, spatial structure constraints, style variability, heterogeneous modality fusion, and dynamic temporal generation, gradually expanding their applications to content creation, medical imaging, autonomous driving simulation, and virtual reality. Meanwhile, large-scale unified models, physical consistency modelling, privacy protection, and efficient deployment in resource-constrained environments remain essential research directions [3-4].

This paper aims to explore conditional image generation technology based on diffusion models. First, it reviews the development of traditional generative models and their limitations. Then, it analyses the implementation mechanisms, performance, and applicable scenarios of various methods from five core technical paths: semantic alignment, spatial structure control, style attribute regulation, multi-modal conditional fusion, and dynamic temporal generation. It summarises the comparative results of mainstream methods on different benchmark datasets and evaluation metrics, revealing their strengths and weaknesses. Finally, it addresses current challenges and outlines potential future directions, offering references and insights for subsequent research and applications.

2. Diffusion model

In the field of conditional image generation, diffusion models hold a critical position and are among the most promising generative models currently. They are a class of deep generative models based on non-equilibrium thermodynamics, learning the data distribution \(p(x) \) through defining a forward diffusion process and an inverse generation process. At its core, it involves state evolution driven by a Markov chain. Unlike other latent variable models, the forward process, or the diffusion process, is fixed as a Markov chain, gradually adding noise to the original image over T time steps until pure noise is obtained. Through learning to denoise, the pure noise is slowly restored to the real image [5].

In the original diffusion model (DDPM), there is no conditional control over the generative capability because only noise is influential. By incorporating classifier guidance during sampling, it became one of the leading image generation models. Diffusion models have the following advantages: they preserve high-frequency details through progressive denoising, avoiding artefacts, which results in high-quality generated images. Additionally, the architecture is relatively flexible and scalable, and during training, it remains highly stable, thereby avoiding issues such as gradient explosion [6].

2.1. Semantic fit

Cross-attention mechanisms are powerful in achieving semantic precision, with the core being that text is first encoded into a dense vector sequence, which then interacts with the feature representations generated by the diffusion model when processing images. The model calculates the attention of image features towards text features, dynamically adjusting the weights of image features based on the importance of the text, thereby effectively integrating text semantics into the image generation process. The advantage of this mechanism lies in its ability to capture complex and abstract semantic concepts, driving the model to generate corresponding global visual content.

To significantly reduce computational costs and make high-quality image generation possible on consumer-grade GPUs, the Stable Diffusion model was developed. It encodes text through the powerful CLIP model. It injects the resulting text embeddings into multiple layers of the cross-attention layers in the U-Net diffusion network (especially after downsample blocks), shifting the entire diffusion process to the latent space of a pre-trained VAE [7].

In contrast, Imagen takes a different approach, utilising a massive T5 language model to encode text for enhanced semantic understanding. It converts text embeddings into a gating signal to modulate the channel importance of feature maps in the diffusion U-Net, thereby more finely controlling the impact of text conditions on the generation process.

2.2. Precise Spatial Structure Control

In today's applications, there is a need for precise control over the spatial structure and geometric layout of images, such as maintaining a specific pose for generated characters or ensuring buildings follow design sketches exactly. Spatial structure conditions are generated based on geometric constraints, with inputs typically being edge maps, depth maps, human pose keypoints, or segmentation masks that clearly define the shape and position of objects. The core challenge lies in effectively injecting these new, often pixel-level, spatial constraint signals without compromising the rich visual knowledge already learned by pre-trained diffusion models. The key mechanism involves introducing a bypass control network. The architecture and weights of the main diffusion model (typically a pre-trained text-to-image model) are preserved mainly [8]. At the same time, geometric condition signals are processed through an additional, trainable neural network branch. The features output by this branch are then fused into the corresponding feature layers of the main model in a residual manner. A key design is zero initialisation: the weights of the convolution layers in the control branch are set to zero when connecting to the main model's convolution layers. This ensures that the control branch's output is zero during the initial training phase, thereby not interfering with the main model's existing capabilities. As training progresses, the control branch gradually learns to interpret geometric conditions and generate effective regulatory signals. ControlNet is a pioneering work in this field. It constructs a bypass network comprising multiple convolutional layers to process input control maps, with its output connected to the encoder intermediate layers of the main U-Net through zero-initialised convolutional layers for feature fusion. This approach successfully protects the knowledge of the pre-trained model, achieving precise pixel-level control, such as generating images that match specified complex human poses. GLIGEN, on the other hand, provides a spatial control method based on attention [9]. It converts user-provided object bounding box coordinates into a spatial attention mask map. Within the attention computation in the diffusion model, this mask map is added to the query-key similarity matrix, guiding the model to "focus" and generate target objects within the specified area, achieving object-level precise localisation. These methods have significant application value in industrial design (such as converting CAD sketches to realistic renderings), film and television special effects (such as converting shot scripts to scenes), and other fields.

2.3. Style attributes are variable

The visual style and specific attributes of an image (such as material, lighting, and colour atmosphere) are also essential control dimensions. The goal is to decouple the image's content and style, allowing users to freely replace or adjust style attributes while keeping the main content unchanged. The input is typically a reference-style image or a set of vectors that describe the attributes. The challenge lies in defining and separating the intertwined visual information into "style." In 2022, the StyleDiffusion model applied the idea of adaptive instance normalisation (AdaIN) to diffusion models. The principle is that style information is encoded into a vector containing scaling and offset factors; the image feature maps are first normalized (subtracted by their own mean and divided by their own standard deviation, removing their own style information), then scaled by the scaling factor of the style vector and offset by the offset factor, thus applying the target style to the content features. The Stylediffusion model utilises a style encoder to extract a style vector from the reference image, and then directly modulates (scales and offsets) the weight parameters of the convolutional layers within the diffusion U-Net itself. This method of stylising the model weights is believed to achieve excellent and diverse style control [10].

It is also feasible not to modify the model directly. The DiffusionCLIP generation model proposed by Shicheng Xu leverages the powerful alignment capability of the CLIP model to achieve zero-shot style transfer. In the latent space (the intermediate representation space where the diffusion process occurs), a direction vector is sought. This direction represents the path from the current image to the target style image. By moving the latent variables along this direction during the diffusion sampling process, the input image can be transferred to the target style without retraining the model. Such methods have great potential in areas like fashion design games [11].

2.4. Multimodal combination conditions

Real-world demands are often complex and multi-dimensional, with AI model users hoping to combine multiple types of conditions to generate images. The challenge lies in effectively integrating heterogeneous signals and making reasonable decisions when they potentially conflict. Composer is a model representative for achieving multimodal condition inputs. It designs an independent "integrator" module. Text, layout (such as bounding boxes), style, and other types of condition signals are processed by their respective encoders and then fed into this integrator. The integrator learns how to weigh the importance of these conditions, especially when they conflict, and outputs a combined condition representation to guide the diffusion model. UniDiffuser, on the other hand, pursues a more unified and general architecture. It utilises a Transformer model as its core encoder, capable of receiving and processing inputs from various modalities, including text, image blocks, and audio spectrograms, and encoding them into a unified representation space. This design inherently enables it to handle combinations of multiple modalities. Such methods pave the way for building more general and flexible generation systems, applicable to complex scenarios such as virtual reality (voice commands + gestures generating 3D scenes) and medical imaging (CT scans + diagnostic reports generating annotated images).

2.5. Dynamic Time-Series Generation

As the models evolve, the demand for image generation extends beyond static images, with the generation of dynamic, temporally coherent videos becoming a more advanced requirement. The input can be a single starting frame with a description of the motion or a text that describes the entire video content. The greatest challenge lies in ensuring visual coherence and motion consistency across each frame in the generated video sequence, avoiding flickering, jittering, or violations of physical laws. Typical approaches involve explicitly modelling the temporal dependencies during the diffusion process. Typically, modules capable of capturing temporal information, such as 3D convolutions or recurrent connections, are introduced into the model architecture, or regularisation terms are added to the loss function to constrain the changes between adjacent frames to be as expected (for example, using optical flow information). Gen-1 (RunwayML) is a significant advancement in video generation. Its architecture first uses a shared encoder to independently process each frame of the input video, extracting features from each frame. These time-ordered feature sequences are then fed into a series of 3D convolutional layers, which are specifically designed to learn the spatiotemporal dependencies between frames, capturing the continuity of motion. Finally, the processed spatiotemporal features are input into a diffusion decoder to generate future frame sequences. Its design ensures the effective integration of temporal information. Video LDM opts to operate in the latent space. It uses optical flow estimation techniques to calculate the expected motion (deformation field) between adjacent frames' latent variables, then adds a regularisation term to the training objective of the diffusion model to constrain the generated latent variable sequence to follow this optical flow prediction as closely as possible, thereby enforcing coherent motion at the latent space level. These methods are driving the development of applications, such as short video creation (converting text scripts to animation) and autonomous driving simulation (generating dynamic scenes from road condition descriptions).

3. Evaluation metrics and model comparison

Currently, there are no comprehensive metrics to evaluate generative image models, and most metrics can only showcase a part of the model's features. The evaluation of conditional generation methods for diffusion models should closely revolve around their control objectives, with different technical approaches reflected in their specific quantification metrics, which deeply reveal the core problems they address and their performance boundaries. The CLIP Score is the core standard for measuring semantic alignment between text and images. This metric maps generated images and conditional texts to a shared embedding space using the CLIP model, calculating their vector cosine similarity (ranging from 0 to 1), which directly reflects the accuracy of semantic matching. For

example, DALL·E 2 achieves a CLIP Score of 0.35 on the MS-COCO dataset, significantly higher than the traditional GAN's 0.28, validating the effectiveness of its cascaded conditional injection mechanism. The FID (Fréchet Inception Distance) metric extracts the feature space characteristics of generated images and real data from the Inception-v3 network, quantifying fidelity and diversity by calculating the Fréchet distance between two multivariate Gaussian distributions (considering both mean difference and covariance dispersion). Imagen achieves an FID of 7.27 in ImageNet 256×256 generation, about 40% lower than that of contemporaneous models.

For spatial structure control models (ControlNet, T2I-Adapter, GLIGEN), the evaluation focuses on pixel-level alignment with geometric constraints. PCKh is a key standard for assessing the generation of human poses, statistically measuring the percentage of keypoint prediction errors less than half the size of the head (ranging from 0% to 100%). ControlNet achieves 89% PCKh on the COCO-val dataset, confirming the robust control of the zero-convolution residual branch over complex poses. In general object localisation tasks, IoU (Intersection over Union) calculates the ratio of the overlapping area between the generated target region and the conditional control image (such as an edge map) (ranging from 0 to 1), quantifying spatial accuracy. GLIGEN achieves an IoU of 0.78 in ADE20K scene generation, representing an 86% improvement over the baseline, which demonstrates the precision of its spatial attention mask mechanism.

In the evaluation of style attribute control models (StyleDiffusion, DiffusionCLIP), LPIPS (Learned Perceptual Image Patch Similarity) is commonly used to assess the fidelity of style transfer. This metric utilises a pre-trained VGG network to extract multi-level features, calculating the weighted distance between the generated image and the style reference image in the feature space (ranging from 0 to 1, with lower values indicating a closer style). DiffusionCLIP achieves an LPIPS of 0.31, significantly better than StyleGAN-NADA's 0.41, demonstrating the advantage of zero-shot transfer guided by CLIP. Multi-modal combination models (Composer) require evaluating their conflict resolution capabilities, which is introduced through the multi-condition satisfaction rate (i.e., the proportion of generated results that meet all input conditions simultaneously). Composer achieves an 89% satisfaction rate under complex condition combinations, a 32% improvement over single-condition models, validating the effectiveness of its learnable weighted fusion mechanism.

As shown in Table 1, dynamic temporal models (Gen-1, Video LDM) and FVD (Fréchet Video Distance) evaluate the overall realism of the videos. This metric extends FID to the video domain by extracting spatiotemporal features with an I3D network and calculating distribution distance (lower values are better) [12,13]. Video LDM achieves an FVD of 256 on the UCF-101 action dataset, setting a new SOTA. Meanwhile, t-SSIM (temporal Structural Similarity) quantifies frame coherence (with a value range of 0-1), with Gen-1 reaching 0.92, approaching the 0.95 of real videos, reflecting the superiority of its 3D convolutional temporal modelling [14].

Table 1. Evaluation Metrics for Image Generation Models

Evaluation metrics	Model	Indicator value	Dataset
CLIP Score	DALL·E 2 GAN	0.35 0.28	MS-COCO
FID	Imagen	7.27	ImageNet 256×256
PCKh	ControlNet	89%	COCO-val
IoU	GLIGEN	0.78	ADE20K
LPIPS	DiffusionCLIP StyleGAN-NADA	0.31 0.41	-
multi-condition fulfilment rate	Composer	89%	-
FVD	Video LDM	256	UCF-101
t-SSIM	Gen-1	0.92	-

4. Conclusion

Conditional image generation based on diffusion models is a significant research direction in contemporary computer vision. In the context of rapid advancements in artificial intelligence technology, generative models have garnered considerable attention due to their wide-ranging applications in content creation and data augmentation. As a new generation paradigm, diffusion models demonstrate notable advantages in generation quality and diversity through their unique mechanisms of forward noise addition and reverse denoising. Their training objective, based on variational lower bounds, ensures the stability of the training process. In contrast, the multi-step iterative generation process guarantees the richness and diversity of the output samples, making them one of the most promising generative models.

This paper systematically introduces the basic principles of diffusion models, elucidating their mechanism of achieving high-quality image generation through a gradual denoising process. In terms of conditional control, it focuses on the current mainstream technical routes: utilizing cross-attention mechanisms to achieve semantic alignment between text and images, making the generated content more consistent with the text description; introducing bypass network structures to achieve precise control of geometric constraints, enhancing the accuracy of spatial layout; and using adaptive normalization techniques to decouple style attributes, achieving fine-grained style control. These methods expand the conditional control capabilities of diffusion models in various dimensions.

Despite significant progress, diffusion models still face numerous challenges. One of these is the lack of physical consistency, where generated results often lack detailed representations that conform to physical laws, such as lighting effects and material textures. Another challenge is the incomplete evaluation system, as existing metrics (such as FID and CLIP Score) struggle to comprehensively measure the conditional conformity and visual authenticity of generated images. Additionally, computational efficiency remains a bottleneck, with the multi-step sampling process resulting in slower generation speeds, which makes it challenging to meet real-time application needs.

Future research can delve into several directions: first, developing physically enhanced generative architectures by incorporating physical engines or constraints to improve the rationality of generated results; second, constructing a more comprehensive evaluation system, establishing a multi-faceted evaluation standard that considers semantic alignment, spatial accuracy, and visual quality; and third, optimizing sampling efficiency through techniques such as knowledge distillation and fast solvers to achieve real-time generation. In terms of application, diffusion models are expected to play a significant role in fields such as autonomous driving simulation, industrial design, and medical imaging, particularly in scenarios that require high-precision conditional control. As the technology continues to mature, diffusion models are expected to become a crucial bridge between visual perception and content creation, driving the development of generative artificial intelligence to a higher level.

References

[1] Rombach Robin, Blattmann Andreas, Lorenz Dominik, et al. High-resolution image synthesis with latent diffusion models. IEEE Trans Pattern Anal Mach Intell, 2023, 45(1): 10674–10685.

[2] Saharia Chitwan, Chan William, Saxena Saurabh, et al. Photorealistic text-to-image diffusion models with deep language understanding. Nature Machine Intelligence, 2023, 5(6): 545–556.

[3] Zhang Lvmin, Zhang Yujun, Zhang Bo, et al. Adding conditional control to text-to-image diffusion models. IEEE Trans Image Process, 2024, 33(3): 3813–3824.

[4] Mou Chong, Qiu Tao, Wang Yu, et al. T2I-Adapter: Learning adapters for controllable text-to-image generation. Pattern Recognition, 2024, 150: 110220.

[5] Li Yuheng, Qi Chenyang, Liu Xueyan, et al. GLIGEN: Open-set grounded text-to-image generation. IEEE Trans Multimedia, 2024, 26(2): 22511–22521.

[6] Wang Hongyu. Estimates of the Kobayashi metric and Gromov hyperbolicity on convex domains of finite type. J Lond Math Soc, 2022, 110(1): 1–22.

[7] Xu Shicheng, Yin Wenpeng, Liu Jingjing, et al. Match-Prompt: Prompt learning for multi-task generalisation in neural text matching. ACM Trans Inf Syst, 2023, 41(2): 1–25.

[8] Huang Lianghua, Zhang Haoye, Wu Tianshui, et al. Composer: Creative and controllable image synthesis with composable conditions. IEEE Trans Multimedia, 2024, 26(4): 1–12.

[9] Zhou Hao, Wang Guangyi, Jiang Yue, et al. Temporal grounding with diverse label learning under weak supervision. IEEE Trans Pattern Anal Mach Intell, 2024, 46(5): 1–13.

[10] Bilitewski Thomas, Rey Ana Maria. Manipulating correlation propagation in dipolar multilayers: From pair production to bosonic Kitaev models. Phys Rev Lett, 2023, 131(5): 053001.

[11] Song Jiaming, Meng Chenlin, Ermon Stefano. Denoising diffusion implicit models for efficient sampling. IEEE Trans Pattern Anal Mach Intell, 2022, 44(11): 1321–1333.

[12] Charpentier Bertrand, Hasenclever Leonard, Paige Brooks, et al. Deterministic uncertainty quantification via architectural and prior design. J Mach Learn Res, 2023, 24(109): 1–27.

[13] Nichol Alex, Dhariwal Prafulla. Improved denoising diffusion probabilistic models. Adv Neural Inf Process Syst, 2021, 34: 12447–12458.

[14] Ho Jonathan, Jain Ajay, Abbeel Pieter. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst, 2020, 33: 6840–6851.