A Comprehensive Survey on Layout-Guided Controllable Image Synthesis

SurveyForge

Abstract— Layout-guided controllable image synthesis has emerged as a transformative paradigm in generative modeling, enabling precise spatial and contextual control over synthesized imagery. This survey provides a comprehensive analysis of foundational layouts, representation formats, and generative architectures, discussing their applicability across diverse domains and use cases. By employing structured layouts such as bounding boxes, segmentation masks, scene graphs, and hierarchical configurations, these methods enable adherence to spatial constraints while supporting semantic and multi-modal integration. Generative frameworks, including Generative Adversarial Networks (GANs), diffusion models, and transformers, are critically examined for their suitability in maintaining alignment with input layouts, fidelity, and contextual reasoning. Applications, spanning healthcare, urban design, scientific visualization, and creative industries, underscore the utility of these techniques in real-world scenarios. However, significant challenges remain, including scalability to high-resolution, multi-object scenes, generalization across complex layouts, and bias mitigation in synthetic outputs. Emerging research directions, such as 3D layout synthesis, multi-modal fusion, adaptive hierarchical control, and memory-efficient architectures, offer promising pathways to address these limitations. The survey highlights ethical implications and emphasizes the importance of standardized evaluation protocols to foster fairness, inclusivity, and accountability, ensuring layout-guided synthesis evolves into a robust and impactful generative framework.

Index Terms—layout-guided synthesis, generative architectures evaluation, multi-modal integration

1 Introduction

Layout-guided controllable image synthesis is an emerging paradigm in generative modeling, combining advanced machine learning techniques with structured representations to provide greater control over image creation. At its core, this approach allows for the generation of images conditioned on input layouts—spatially explicit configurations such as bounding boxes, segmentation masks, or scene graphs—enabling detailed control over object positioning, spacing, and scene composition. This capability addresses a critical challenge in image synthesis: reconciling the high-level creative flexibility of text-based generative models with the precise spatial and structural constraints required in real-world applications. By integrating explicit layout guidance, this field bridges the gap between artistic freedom and geometric precision, catalyzing advancements across diverse domains such as content creation, scientific visualization, and autonomous systems.

The motivation behind layout-guided synthesis stems from the limitations of unconstrained generative models, which frequently struggle to adhere to desired spatial arrangements, particularly in scenarios involving densely populated or semantically intricate scenes. Early developments in layout-driven synthesis were heavily influenced by rule-based graphic design systems, which employed deterministic algorithms for arranging elements but lacked both adaptability and scalability in diverse visual contexts. Modern approaches address these shortcomings by leveraging deep generative models, such as Generative Adversarial Networks (GANs) and diffusion models, to map rich layout information to photorealistic imagery. For example, LayoutGAN [1] demonstrated the potential of aligning explicit spatial inputs with layout optimization, while Layout2Im [2] further expanded this design, combining object categories with bounding boxes to achieve precise semantic alignment.

Over time, the shift toward probabilistic and attention-based models has enhanced the quality, realism, and compositional adherence of generated images. Transformers, incorporating self- and cross-attention mechanisms, have become a dominant architectural choice in encoding spatial configurations [3], allowing for hierarchical layout comprehension and scalable representations. Complementary innovations, such as LayoutDiffusion [4], build on the successes of diffusion models by introducing modules that specifically ensure spatial fidelity and object-aware conditioning, pushing the boundaries toward more complex, contextually rich scene generation.

The real-world implications of layout-guided synthesis are vast. In media and entertainment, these systems enable the efficient generation of storyboards, game environments, and advertising compositions, ensuring adherence to aesthetic principles while reducing manual design efforts [5]. In scientific and medical imaging, precise spatial constraints provided by layouts allow for the creation of anatomically valid synthetic datasets, addressing issues of data scarcity in high-stakes environments [6]. Furthermore, interactive and user-centric applications, including graphic design and virtual interface creation, can benefit from dynamic layout adjustment technologies, as systems like CoLay [7] enable designers to customize layouts in real-time.

Despite remarkable progress, significant challenges remain. Generating semantically coherent and spatially accurate images from increasingly complex layouts requires robust multi-modal integration and reasoning across textual,

visual, and spatial modalities [8]. Current models must contend with scalability to high-resolution scenarios, disparities in cross-domain generalizability, and ethical concerns, such as bias in layout data that may propagate through generated outputs [9]. Moreover, seamless real-time adaptability and interpretable synthesis processes are essential to democratizing access to these tools for non-expert users.

The objectives of this survey are multi-fold: to systematically categorize existing work, analyze trade-offs across representation and modeling paradigms, and identify emergent trends, such as hierarchical and 3D-guided layout synthesis. By consolidating advancements in layout-guided synthesis, the field moves closer to realizing generative systems that are not only creative and high-fidelity but also controllable, interpretable, and aligned with real-world constraints. Future directions include a deeper exploration of foundational models incorporating self-supervised learning for diverse layouts [10], as well as integrating ethical frameworks to ensure fairness and inclusivity in synthesized outputs. This survey endeavors to serve as a cornerstone for researchers and practitioners aiming to innovate along the multifaceted dimensions of layout-guided controllable image synthesis.

2 Representations and Encoding of Layouts

2.1 Layout Representation Types and Formats

Layout representations serve as the foundational input for layout-guided image synthesis, encapsulating the spatial and semantic structure of a scene. Different representation types offer varying levels of abstraction, detail, and controllability, playing a crucial role in the performance of downstream synthesis pipelines. This subsection explores key layout formats—bounding boxes, segmentation masks, keypoints, and scene graphs—evaluating their encoding efficiency, flexibility, and applicability across diverse tasks, alongside discussing emerging trends and open challenges in the field.

Bounding boxes represent arguably the simplest layout format, defining spatial regions by their rectangular boundaries (minimum and maximum x, y coordinates). This format is computationally efficient and easy to annotate, making it one of the most widely used representations for tasks requiring coarse-grained structural guidance. Bounding boxes effectively constrain object placement and size, supporting synthesis methods that generate semantically realistic scenes in accordance with spatial constraints. For instance, Layout2Im [2] achieves significant gains in controllability by leveraging bounding boxes and encoding object categories alongside spatial information. However, this representation's lack of granularity limits its suitability for tasks demanding precise spatial delineation or shape information, such as detailed layouts for urban planning or scientific visualization.

Segmentation masks extend bounding box-based layouts by encoding pixel-wise object boundaries, with each pixel assigned a unique label corresponding to the object class. This dense representation provides finer-grained control in image synthesis, enabling precise spatial realism and boundary alignment. Techniques such as the layout-to-mask-to-image paradigm [11] utilize segmentation masks to enhance the transition from structured layouts to high-quality synthesized images. However, the computational cost associated with processing dense segmentation maps, particularly in complex and multilayered scenes, can be prohibitive. This trade-off makes segmentation masks most relevant for domains such as medical imaging [6], where spatial precision and structural accuracy are paramount.

Keypoints provide an alternative sparse representation, encoding structural control through a set of landmarks or skeletal points. These representations are often utilized in tasks involving skeleton-based image synthesis, such as human pose generation, where only salient structural cues are required to guide the generative process. For instance, methods incorporating keypoint-layout synthesis, as explored in [12], demonstrate the power of sparse representations to disentangle spatial information from visual appearance, facilitating transferability across domains. While computationally efficient, keypoints are inherently limited in expressiveness, lacking the capability to handle dense, multi-object scenes or capture contextual relationships.

Scene graphs represent a more relational approach to layout representation, modeling entities as nodes and their spatial or semantic relationships as edges. This hierarchical format excels at capturing contextual dependencies, such as "object A is to the right of object B," which enhances semantic coherence in synthesis tasks. The integration of scene graphs with generative pipelines, as demonstrated in [13], allows systematic encoding of object arrangements and interactions. Yet, scene graphs face scalability challenges, particularly when modeling dense, highly interconnected scenes. Additionally, transforming scene graphs into computationally usable representations for processing by neural architectures, such as graph convolutional networks, demands non-trivial computational overhead [14].

The selection of a specific layout format often involves critical trade-offs between computational efficiency, representational capacity, and domain requirements. Emerging trends aim to alleviate these trade-offs through hybrid approaches, which integrate multiple layout types to combine their strengths. Methods such as LayoutGPT [8] propose using large language models to generate multi-modal layouts that incorporate bounding boxes, semantic attributes, and even textual descriptions in a unified representation. Furthermore, hierarchical representations, such as multi-scale layouts [15], offer promising avenues for aligning coarse-to-fine-grained structures to support both high-level planning and pixel-level synthesis.

Despite these advances, several challenges remain. Computational scalability remains a bottleneck for dense formats like segmentation masks and scene graphs. Furthermore, most layout representations are tailored to 2D contexts, limiting their applicability in emerging 3D synthesis tasks [16]. Addressing these challenges will likely require innovations in spatially immersive encoding methodologies, such as 3D layouts and depth-aware representations [16], paired with efficient layout fusion mechanisms to balance detail and efficiency.

In summary, the diversity of layout representation types supports a rich spectrum of controllable image synthesis applications, each with unique trade-offs and constraints. Advances in hybrid layout approaches and integration with

emerging modalities promise to push the boundaries of spatial and semantic control, paving the way for increasingly versatile and scalable generative systems.

2.2 Layout Construction Approaches

The construction of layouts, which serve as structural blueprints for controllable image synthesis, is fundamental to the generative process, enabling spatial organization and coherence across objects. Building on the discussion of layout representations, this subsection delves into methodologies for generating layouts, including manual annotation, semi-automated systems, data-driven techniques, and emerging paradigms leveraging retrieval frameworks and large language models (LLMs). Through a critical analysis, these methods' strengths, limitations, and trade-offs are examined, demonstrating their impact on the synthesis pipeline and potential for advancing the field.

Manual annotation remains one of the earliest and most intuitive methods for layout generation, with human experts arranging objects to meet domain-specific requirements. This approach ensures precision and high fidelity, making it particularly useful in applications with stringent accuracy demands, such as medical imaging and architectural visualization. Despite the diversity and reliability offered by manual design, the method is inherently laborious and time-intensive, presenting clear scalability challenges [1]. Enhancements in user interfaces, such as drag-and-drop tools, have improved accessibility, yet the lack of adaptability to unseen scenarios or the generation of diverse variations limits its broader applicability and efficiency.

Moving toward automation, data-driven approaches capitalize on large datasets to derive coherent layouts, fueled by advancements in machine learning. Probabilistic graphical models, such as scene graph generators, capture essential semantic and spatial relationships among objects, translating these into structured layouts [17], [18]. More sophisticated models like LayoutGAN utilize adversarial learning to refine raw inputs into geometrically plausible arrangements informed by dataset priors [1]. Meanwhile, variational frameworks such as LayoutVAE introduce stochasticity to synthesize diverse layouts under given constraints, adeptly addressing real-world variability [13]. Despite these strides, such methods remain dependent on the quality and diversity of training datasets, exposing them to biases and difficulties in generalizing to novel or unseen distributions.

Automated systems have further propelled layout construction through sophisticated architectures and attention mechanisms, enabling nuanced modeling of multi-object interactions. Transformers, exemplified by LayoutTransformer, leverage self-attention to effectively model contextual relationships, generating complex layouts with innovative spatial compositions [3]. Constraint-aware approaches, such as LayoutFormer++, refine this capability by limiting decoding spaces to ensure adherence to geometric and semantic constraints, thereby preventing infeasible outputs [19]. However, such systems often struggle with tasks requiring domain-specific expertise or human-level intuition, posing challenges in specialized application contexts.

Retrieval-based strategies offer an alternative perspective on construction by reusing and adapting pre-annotated layouts from extensive databases. This approach leverages existing scene prototypes to meet new input requirements, significantly reducing annotation burdens and streamlining the synthesis of coherent layouts [20]. By fostering interpretability and computational efficiency, retrieval-based methods have shown promise; however, they are limited by the diversity and quality of source databases, which can constrain adaptability to novel contexts.

A significant shift in layout construction has emerged with the integration of LLMs, which bridge high-level textual semantics with low-level spatial configurations. For instance, LayoutGPT utilizes LLMs for parsing text descriptions into visually coherent spatial layouts, enabling an unprecedented level of flexibility for generating scene structures [8]. Through this paradigm, LLMs alleviate the constraints of manual annotation and facilitate cross-domain adaptability. Nevertheless, these models face challenges in scaling to complex multi-modal scenarios, where ensuring alignment between textual inputs and generated layouts remains a formidable task.

Emerging trends point to hierarchical and compositional layout methodologies to further advance generative capabilities. Techniques like LayoutDiffusion incorporate diffusion processes to enhance object-aware spatial modeling, allowing for precise and adaptable layouts [4]. Additionally, hybrid frameworks that combine retrieval systems with generative models aim to balance interpretability and diversity dynamically, fostering semantically robust layouts that can accommodate wide-ranging applications.

While substantial progress has been made, pivotal challenges persist. Achieving scalability across diverse domains, mitigating biases inherent in training datasets or retrieved layouts, and effectively combining manual inputs with automated systems represent critical frontiers. Future directions should converge toward integrating retrieval-based pipelines with LLM-driven tools, blending their complementary strengths to maximize efficiency, flexibility, and scalability. By unifying human intuition, data-driven insights, and algorithmic innovation, layout construction methodologies hold immense potential to unlock new horizons in controllable image synthesis, meeting the demands of increasingly sophisticated generative frameworks.

2.3 Encoding and Processing Layouts

Encoding and processing layouts is a critical step in transforming structured spatial information into neural representations suitable for generative image synthesis. Effective encoding ensures that essential spatial, semantic, and relational cues inherent in layouts are preserved, facilitating seamless integration into generative pipelines. This subsection delves into key strategies, including convolutional methods, graph-based representations, attention mechanisms, and hybrid frameworks, assessing their benefits, limitations, and emerging trends.

Convolutional Neural Networks (CNNs) have historically been a cornerstone of layout encoding, especially for dense inputs like segmentation masks or bounding box heatmaps. By leveraging spatial correlations through hierarchical feature extraction, CNN-based encoders excel in capturing fine-grained spatial configurations. For instance,

convolutional backbones are utilized in LayoutGAN [1] to refine geometric relations and render layouts coherent in terms of alignment. However, the locality of convolutional filters imposes a limitation when longer-range dependencies or hierarchical spatial relationships are essential. This weakness becomes pronounced in applications requiring holistic understanding of complex multi-object layouts, as seen in datasets like COCO-Stuff or Visual Genome [2].

Graph-based encoding methods have emerged as a natural choice for layouts characterized by object relationships and hierarchical structures, such as scene graphs. These methods leverage Graph Convolutional Networks (GCNs) or Transformers with graph encoders to capture object-to-object interactions via nodes (objects) and edges (relationships). Ashual et al. [21] demonstrate the utility of scene graph embeddings, where spatial and semantic interactions are embedded directly into the layout representation. Similarly, LayoutTransformer [3] encodes layouts as sequences of primitives, enabling contextual understanding across spatially interdependent elements. One significant challenge for graph-based approaches lies in their scalability to dense, high-resolution layouts, where computational efficiency can degrade due to the quadratic complexity of self-attention or graph aggregation mechanisms.

Attention-based models, particularly those incorporating Transformer architectures, have recently shown great promise in processing layouts with long-range dependencies and diverse compositional constraints. By employing self- and cross-attention mechanisms, such models effectively capture relationships across entire layouts while maintaining interpretability. For example, LayoutVAE [22] couples Variational Autoencoders with self-attention layers to model both local and global relationships, achieving state-of-the-art diversity in layout synthesis. Meanwhile, LayoutDiffusion [23] incorporates cross-attention modules, allowing object-aware and position-sensitive encoding that is pivotal for faithfully mapping layouts to images. Despite these advantages, the computational demands and memory overheads of attention mechanisms present barriers for complex scenarios requiring real-time processing or high-resolution layouts.

Hybrid approaches, combining the strengths of convolutional, graph-based, and attention-driven encoders, represent an emerging paradigm to address encoding challenges across diverse layout types. For example, LostGAN [24] elegantly integrates convolutional feature extraction with layout-aware normalization strategies, enabling both spatial precision and stylistic control. Similarly, LayoutDiffusion [4] incorporates graph-encoded relational information into a Transformer-based diffusion model, achieving robust performance under various constraints. Hybrid methods, however, necessitate careful architectural tuning to balance complexity with performance.

Emerging trends point toward hierarchical and multi-modal encoding frameworks to handle evolving requirements in layout-guided synthesis. Hierarchical models, such as Scene Graph to Image Generation with Contextualized Object Layout Refinement [25], iteratively refine layout representations from coarse to fine granularity, preserving semantic consistency and spatial detail throughout the pipeline. Further advancements, like LayoutGPT [8], leverage large language models to generate layouts based on textual instructions, bridging textual and spatial modalities. These approaches open new frontiers for integrating contextual cues and domain-specific constraints into encoding pipelines, though challenges related to computational scalability and interpretability persist.

In summary, the encoding and processing of layouts are pivotal in preserving semantic and spatial cues for effective image synthesis. While CNNs remain reliable for dense representations, graph-based models and attention mechanisms provide flexibility and relational understanding for complex layouts. Hybrid and hierarchical approaches, alongside multi-modal integrations, represent exciting directions for future research, with the potential to unify layout representations across diverse domains and applications. Addressing challenges in scalability, real-time efficiency, and interpretability will remain critical in advancing this field.

2.4 Representation Trade-offs and Design Choices

In layout-guided image synthesis, choosing the most suitable representation format and encoding model involves critical trade-offs that directly shape performance, scalability, and domain-specific applicability. This subsection delves into these trade-offs by analyzing prominent layout representations and their design considerations, elucidating their strengths, limitations, and adaptability under various constraints.

Layout representations, such as bounding boxes, segmentation masks, keypoints, and scene graphs, differ significantly in their granularity, influencing computational efficiency and expressive power. For instance, bounding boxes offer a lightweight and coarse-grained encoding of an object's location and scale, making them ideal for real-time applications or scenarios with stringent computational limitations [2]. However, due to their simplicity, bounding boxes often fail to capture fine-grained details or intricate inter-object relationships, both of which are critical for high-fidelity outputs. In contrast, segmentation masks enable pixel-precise delineation of object boundaries and spatial context, supporting the synthesis of complex, high-quality visual scenes. These representations are widely adopted for tasks requiring photorealistic outputs, as demonstrated in [26]. However, their higher computational and memory demands pose scalability challenges when applied to dense layouts or real-time synthesis workflows.

Keypoints, meanwhile, represent objects as sparse sets of control landmarks, frequently used in domains like human pose generation, where skeletal interactions dominate. Keypoints are computationally efficient and effective in such specialized contexts but lack semantic richness, limiting their utility in broader, more generic synthesis tasks [21]. Conversely, scene graphs encode objects as graph nodes and relationships as edges, delivering an enriched understanding of contextual and hierarchical relationships between entities. Their utility is particularly evident in multi-object scenarios, where relational knowledge is crucial. Models leveraging scene graphs [17], [27] show strong performance in generating context-sensitive imagery. However, they face scalability hurdles as graph complexity increases nonlinearly in cluttered or densely populated layouts.

Encoding strategies play a pivotal role in optimizing representation utility, requiring a careful balance between precision, computational efficiency, and compatibility with downstream generative models. Convolutional Neural Networks (CNNs) dominate the encoding of dense formats like segmentation masks or bounding box heatmaps due to their hierarchical feature extraction and spatial locality [2]. On the other hand, graph-specific models, such as Graph Convolutional Networks (GCNs), align closely with the structural properties of scene graphs, encoding object relationships and dependencies effectively [17]. Recent progress in attention-based architectures, including transformers [3], [19], has introduced new opportunities for encoding flexible, hierarchical layouts with enhanced global contextual understanding. While these models provide remarkable adaptability across diverse spatial compositions, their high computational demands remain a constraint for practical implementations.

Hybrid representation formats offer a promising route to mitigate individual shortcomings by combining complementary strengths. For example, Layout-to-Mask-to-Image pipelines [11] demonstrate how coarse-grained bounding boxes can lay the structural groundwork for synthesis, which segmentation masks subsequently refine to achieve greater spatial precision and detail. This multi-stage strategy balances computational overhead and visual fidelity but introduces architectural complexity, reflecting inherent trade-offs in such hybrid approaches.

Domain-specific requirements further influence representation selection. For medical imaging, where semantic accuracy and structural precision are paramount, segmentation masks remain the preferred choice [28]. Similarly, in urban planning and architectural design, the interplay of spatial dependencies and scalability favors hybrid approaches incorporating scene graphs [29]. These domain-driven preferences underscore that representation trade-offs invariably depend on the prioritization of domain-specific metrics, whether they emphasize precision, relational comprehension, or computational economy.

Despite these advancements, the quest to unify scalability, flexibility, and adaptability across representations remains ongoing. Emerging trends in hierarchical representations, which iteratively refine coarse-to-fine layouts [30], and the expansion of layouts into 3D spaces [31], offer forward-looking pathways for improved synthesis. Furthermore, incorporating richer multimodal inputs like textual descriptions or depth maps [8] paves the way for seamlessly enriching spatial and semantic conditioning while addressing the challenge of scalability.

In conclusion, optimizing representation and encoding models to navigate these trade-offs remains crucial for advancing layout-guided synthesis across diverse domains. Lightweight architectures, modular designs, and hybrid frameworks will continue to shape the evolution of adaptive and robust synthesis pipelines, addressing the complex demands of real-world applications.

2.5 Emerging Trends in Layout Representations

The evolution of layout representations in controllable image synthesis has expanded its focus into new frontiers driven by multi-modal, hierarchical, and three-dimensional (3D) layouts. These emerging trends aim to significantly enhance the capabilities, flexibility, and contextual richness of layout-guided generation while addressing challenges in scalability, domain adaptability, and multimodal reasoning.

One of the most transformative directions is the integration of multi-modal contexts into layouts. Early methods relied exclusively on spatial structures like bounding boxes or segmentation masks to represent layouts, but recent advancements incorporate auxiliary modalities such as textual descriptions, appearance attributes, and depth maps to provide richer conditioning. LayoutGPT [8] exemplifies this trend by enabling Large Language Models (LLMs) to infer layouts from natural language prompts, which allows for more intuitive user control while achieving coherence between textual semantics and spatial configurations. Similarly, SceneComposer [15] explores the synthesis of images from layouts whose precision levels range from pure text to precise segmentation maps, merging textual and spatial constraints during generation. A crucial strength of these frameworks is their ability to transform textual inputs into spatial priors, but they are limited when layout precision is critical, such as in industrial design applications. Addressing this limitation, methods like BoxDiff [32] introduce box-constrained diffusion techniques that simultaneously fuse textual constraints with bounding box guidance, exemplifying robust adaptability without extensive additional training.

Hierarchical representations are another rapidly advancing area. Traditional layout paradigms often fail to capture the contextual relationships across different levels of granularity, especially in complex scenes with nested or interrelated objects. Hierarchical frameworks decompose layouts into coarse-to-fine structures, enabling synesthetic multi-level reasoning. Hierarchical methods such as those presented by Inferring Semantic Layout [30] leverage stepwise object bounding box refinement to support semantically controlled synthesis, preserving spatial alignment even in densely populated scenes. Similarly, LayoutDiffusion models [33] and GALA3D [34] extend diffusion methodologies to hierarchical settings, modeling complex inter-object relationships without sacrificing fidelity. These approaches, however, face challenges when scaling to scenes with extreme object hierarchies or when applied to domains requiring strict geometric coherence, such as architectural visualization.

The expansion into 3D layout encoding represents an essential step toward immersive and spatially aware generative systems. While 2D layouts remain foundational, they are inherently limited for applications like virtual reality (VR), augmented reality (AR), and robotics. Techniques like CC3D [31] and RoomDreamer [35] extend generative pipelines into three dimensions, enabling compositional scene generation with structured depth-aware representations extracted from layouts. By employing 3D Gaussian Splatting [34] or volumetric neural fields, such frameworks efficiently synthesize detailed spatial representations while maintaining global coherence. RoomDreamer [35] further integrates 3D scene geometry with text-input-driven stylistic refinement, merging surface texture and spatial-object interaction optimization. These 3D methods provide

unprecedented capabilities, but reconciling computational demands with the complexities of 3D layout structures remains a persistent challenge.

Another crucial trend is the focus on low-resource and transfer learning approaches for layout representations. Large-scale annotated layout datasets are costly to curate, limiting the feasibility of data-intensive methods. Methods like LayoutDM [23] explore task-agnostic diffusion models with modular plug-and-play architectures that achieve strong generalization across layout generation tasks. Additionally, the training-free paradigm explored in models like Training-Free Composite Scene Generation [36] reduces dependency on extensive annotations by leveraging adaptive attention-based guidance and semantic consistency constraints during inference. These approaches have opened exciting possibilities for few-shot or zero-shot learning, enabling robustness in low-resource settings, though achieving versatility across highly diverse domains remains an open research challenge.

Looking ahead, future research must address core limitations, including the efficient unification of hierarchies, modalities, and dimensions into a single representation framework. Additionally, the blending of 2D and 3D representations into hybrid spatial layouts will play a significant role in addressing mixed-reality applications. Emerging multimodal systems that combine text, image, and spatial data into cohesive semantic embeddings—for example, those structured by adaptive layout-semantic fusion [37]—present promising strategies for holistic layout encoding. By integrating hierarchical reasoning, multimodal conditioning, and domain-specific architectural advancements, the field of layout-guided synthesis is poised to unlock greater levels of control, fidelity, and scalability in the years to come.

3 Generative Models and Methods for Layout-Guided Synthesis

3.1 Generative Adversarial Network (GAN)-Based Approaches

Generative Adversarial Networks (GANs) have emerged as foundational models for layout-guided image synthesis, offering a flexible and powerful framework for generating imagery that adheres to predefined spatial and semantic layouts. By exploiting the adversarial training paradigm, GANs can effectively learn mappings from complex input layouts, such as bounding boxes, segmentation masks, and scene graphs, to detailed and semantically coherent visual representations. In this subsection, we analyze the development and characteristics of GAN-based techniques for layout-guided synthesis, discussing key advancements, challenges, and opportunities for future research.

At the core of layout-guided GAN approaches is the objective of ensuring strict adherence to spatial constraints while maintaining photorealistic image fidelity. Methods such as Layout2Im [2] pioneered this space by introducing a GAN architecture that combines object-specific embeddings with convolutional LSTM modules to process spatial layouts encoded as bounding boxes. By disentangling the representation of object categories (certain information) from their appearance (uncertain information), this approach facilitates diversity in appearance while maintaining geometric alignment with the layout. Similarly, LayoutGAN [1] advanced architectural design by proposing a wireframe rendering layer to optimize geometric relationships, producing layouts with highly accurate spatial alignment. Additionally, region-aware discriminators employed in LayoutGAN enhance the fidelity of generated outputs by supervising spatially localized regions individually, which is particularly critical for complex, multi-object scenes.

Another critical direction in GAN-based synthesis is attribute conditioning to refine control over generated content. Attribute-conditioned methods, such as the Attribute-conditioned Layout GAN [38], integrate specific element attributes (e.g., size, aspect ratio, and reading order) directly into the adversarial framework. By combining these conditions with novel loss functions tailored for layout optimization, these models ensure that generated outputs align with intricate design principles. Such mechanisms allow for flexible user-centric generation and adaptation to application-specific needs, e.g., in graphic design and advertisement generation. However, these approaches often trade computational efficiency for enhanced control, as the incorporation of multiple conditioning factors increases system complexity.

Despite their success, GAN-based layout synthesis methods face notable challenges. A pervasive issue is mode collapse, wherein the generator produces limited variations of outputs, particularly for complex layouts with high object diversity. Techniques such as noise injection and latent vector manipulation have been employed to combat this limitation [2]. For instance, adversarial methods designed for layout-to-mask-to-image generation extend the synthesis pipeline by incorporating intermediate mask representations, enhancing both fidelity and layout compliance [11].

Moreover, semantic and compositional consistency remains a key concern in GAN-based frameworks. Object-aware GANs, including those employed in Layout2Im [2] and Reconfigurable GANs [11], utilize convolutional layers with positional encodings to explicitly model object interactions and spatial dependencies. However, these approaches often struggle to balance global coherence with localized accuracy, particularly in dense scenes. To address this, region-specific discriminators and compositional encodings have been explored, offering improved adherence to layout constraints but at the cost of increased training complexity.

Another emerging challenge lies in generalizing GANs to diverse layout representations, such as hierarchical scene graphs or keypoint-based skeletons. For example, models incorporating scene graph constraints [27] leverage graph convolutional networks to process relational information, generating images with enhanced contextual realism. However, scaling such approaches to large, dense layouts remains computationally intensive due to the quadratic complexity of graph-based operations.

Despite their limitations, GAN-based methods have demonstrated remarkable adaptability across an array of applications. For instance, models such as ControlGAN [39] have successfully extended layout-guided synthesis to multi-modal contexts, integrating text and spatial inputs to enable fine-grained manipulation of generated imagery. Similarly, hybrid approaches combining GANs with other

architectures, such as variational autoencoders (VAEs) and transformers, have shown promise for addressing inherent challenges like mode collapse and semantic misalignment [11]. These integrations point toward a future where GAN-based frameworks are augmented with the strengths of alternative generative paradigms.

In conclusion, GAN-based approaches remain pivotal to layout-guided image synthesis research, providing a flexible yet powerful framework for integrating spatial layouts into generative pipelines. However, challenges such as achieving semantic consistency, handling complex multi-object scenes, and addressing computational inefficiencies continue to necessitate creative architectural innovations. Future research could explore hybrid frameworks that fuse adversarial training with transformer-based or diffusion models [4], as well as scalable solutions for diverse layout types such as scene graphs and 3D layouts. By addressing these challenges, GAN-based techniques will continue to play a transformative role in advancing layout-aware generative modeling.

3.2 Diffusion-Based Models for Layout-Guided Synthesis

Diffusion models have emerged as a compelling alternative in layout-guided synthesis, capitalizing on their iterative denoising processes to achieve fine-grained control over both spatial alignment and generative fidelity. Unlike GANs, which often contend with issues like mode collapse and unstable optimization, diffusion models leverage probabilistic modeling frameworks to generate high-quality images that closely adhere to input layouts. This subsection delves into the methodologies, innovations, and challenges of diffusion-based approaches, focusing on strategies for spatial conditioning, trade-offs between fidelity and controllability, and advancements in multimodal integration, while situating these models within the broader landscape of generative methods.

The iterative noise-to-sample refinement characteristic of diffusion models provides a robust means of preserving spatial structure through learned conditional priors. For example, LayoutDiffusion [4] introduces a Layout Fusion Module (LFM) and Object-aware Cross Attention (OaCA) to explicitly model object relationships and ensure adherence to predefined spatial arrangements. The LFM integrates layout representations during intermediate denoising steps, while OaCA enhances contextual coherence with position-sensitive attention mechanisms. This sophisticated conditioning enables LayoutDiffusion to achieve state-of-the-art performance in spatial fidelity and visual realism, as demonstrated on datasets such as COCO-Stuff and Visual Genome.

The flexibility of diffusion models becomes apparent in composable frameworks such as FreestyleNet [40], which leverages rectified cross-attention mechanisms to incorporate unseen or abstract semantic elements into layouts. By facilitating the inclusion of textual prompts to introduce novel components, FreestyleNet extends the utility of diffusion-based synthesis beyond static layouts, highlighting their versatility in handling open-vocabulary scenarios. This adaptability is particularly relevant in use cases requiring dynamic or interactive image synthesis, underscoring the expanding potential of these models.

Achieving fine-grained spatial conditioning in diffusion-based frameworks often requires specialized techniques to balance spatial alignment with visual fidelity. For instance, LAW-Diffusion [41] employs a spatial dependency parser to embed layout-aware semantic features, along with an adaptive guidance schedule that systematically modulates conditioning strength throughout the denoising pipeline. This strategy prevents overfitting to rigid layouts while ensuring high levels of texture and detail consistency in the synthesized output. Such innovations serve as a pivotal step toward reconciling the inherent trade-offs in precision and visual quality that diffusion models must manage.

Nevertheless, diffusion-based approaches are not without limitations. The computational demands of their iterative refinement process pose significant challenges for real-time synthesis applications [42]. Furthermore, ensuring robustness against irregular or sparsely annotated layout configurations remains an open problem. Developing lightweight inference mechanisms and employing self-supervised pretraining techniques are critical avenues for improving the scalability and efficiency of these models, particularly in resource-constrained or data-scarce environments.

Hybrid approaches blending diffusion models with other generative paradigms have shown promise in addressing some of these challenges. Build-A-Scene [16], for instance, incorporates 3D layout conditioning into diffusion pipelines, offering interactive spatial modifications and bridging the gap between 2D and 3D layout representations. By incorporating dynamic self-attention modules and context-preserving mechanisms, these systems expand the range of user-controlled applications while enhancing cross-domain utility.

Looking forward, diffusion models present substantial opportunities for integrating auxiliary modalities—such as depth maps, scene graphs, or natural language prompts—to further enhance multimodal layout synthesis. Advances such as hierarchical or coarse-to-fine diffusion frameworks [15] hold the potential to scale these models to high-resolution and complex layouts without compromising efficiency. Such developments would significantly extend their utility across diverse application fields, from creative design to simulation.

In summary, diffusion-based models represent a transformative approach to layout-guided synthesis, offering unparalleled control over spatial adherence and generative fidelity. Their inherent flexibility in conditioning, coupled with the potential for seamless multimodal integration, positions them as key enablers of progress in layout-aware generative modeling. To fully unlock their potential, future research must address computational efficiency, improve robustness to incomplete or unconventional layouts, and explore innovative ways of synergizing them with other generative frameworks, as seen in hybrid approaches and multimodal pipelines. With continued advancements, diffusion models are poised to become foundational tools in tackling the challenges of layout-structured image synthesis while complementing GANs, transformers, and hybrid architectures within the broader generative ecosystem.

3.3 Transformer-Based Methods and Attention Mechanisms

Transformer-based methods and attention mechanisms have emerged as pivotal components in layout-guided image synthesis, owing to their exceptional capability to model long-range dependencies, hierarchical spatial structures, and interactions among multiple scene elements. By leveraging attention to capture both local and global relationships, these methods provide a robust framework for spatially and semantically consistent image generation, particularly for scenarios involving complex layouts.

Central to the success of transformer models in this domain is the self-attention mechanism, which effectively encodes pairwise dependencies among elements by dynamically weighing their influence on one another. This capability has been instrumental in generating realistic layouts and guiding conditional synthesis tasks. For instance, LayoutTransformer [3] utilizes self-attention layers to model contextual relationships between layout elements such as bounding boxes, enabling the synthesis of document layouts, mobile interfaces, and object arrangements in natural scenes. The transformer-based architecture ensures scalability when handling diverse layout types and supports both unconditional layout generation and layout completion from partial inputs, demonstrating its adaptability across various application domains.

Token-based representations have further elevated the potential of transformers for fine-grained manipulation in layout-guided synthesis. In Variational Transformer Networks [22], layout elements are tokenized, facilitating the modeling of global design rules such as alignment margins and hierarchical spatial constraints. This integration of token-level representations with attention has been shown to outperform traditional convolutional approaches, particularly in generating diverse, high-quality layouts with inter-element dependencies. Similarly, the Hierarchical Layout Model [3] enriches generation by capturing spatial dependencies at multiple levels of granularity, such as object-object interactions and their alignment within larger scene structures. This hierarchical encoding enables flexible layout synthesis across diverse visual domains, from structured documents to creative compositions.

Another transformative development is the application of cross-attention mechanisms for introducing conditional controls. Cross-attention maps inputs (e.g., layouts) to generated outputs (e.g., images), ensuring alignment between spatial configurations and visual realism. LayoutGPT [8] demonstrates the power of transformer-based models in integrating multi-modal inputs such as textual descriptions with layouts, leveraging cross-attention modules to translate linguistic expressions into coherent image compositions. Additionally, solutions like LayoutDiffusion [4] embed object-aware cross-attention modules that model spatial relationships between objects, allowing precise integration of fine-grained details and layout constraints.

Despite their undeniable strengths, transformer-based approaches also present challenges and trade-offs. Primarily, their quadratic complexity with respect to the sequence length imposes significant computational and memory overhead, especially for high-resolution layouts or dense multi-object scenes. Techniques such as sparse attention and memory-efficient transformers offer potential solutions but may compromise modeling fidelity. Moreover, transformers often require extensive training on large datasets to generalize effectively across diverse layout styles, which can be limiting in specialized domains with scarce annotated data.

Emerging trends highlight promising directions for addressing these limitations. Hybrid paradigms, combining transformers with diffusion models [43], have begun to surface as highly effective solutions for balancing computational efficiency against generation quality. Additionally, advancements in pre-trained foundational models expand transfer-learning opportunities, which may mitigate the dependency on large-scale layout-specific training datasets. Further, the synergistic integration of transformers with large language models, as exemplified in LayoutGPT [8], provides a compelling pathway for multi-modal and instruction-driven layout synthesis, opening unexplored avenues for interactive and user-adaptive applications.

In summary, transformers and attention mechanisms redefine the paradigm of layout-guided image synthesis by enabling comprehensive spatial reasoning, multi-object interactivity, and robust hierarchical control. While challenges persist, ongoing innovations in model architecture, computational efficiency, and multimodal integration position transformers as key enablers in advancing the frontiers of generative modeling for layout-constrained applications.

3.4 Hybrid Generative Strategies for Layout Integration

Hybrid generative strategies for layout-to-image synthesis combine the complementary strengths of diverse generative architectures—including Generative Adversarial Networks (GANs), diffusion models, and transformers—into cohesive frameworks that address the individual limitations of each approach, thereby enhancing synthesis quality, controllability, and adherence to layout constraints. By integrating these modalities, hybrid models aim to achieve high levels of realism, spatial fidelity, and adaptability, making them well-suited for a range of synthesis tasks.

GAN-diffusion hybrids exemplify the synergy between adversarial training's ability to produce globally coherent images and the iterative refinement strengths of diffusion processes. GANs, bolstered by adversarially trained discriminators, excel at generating visually plausible images, yet often encounter challenges such as mode collapse and difficulties in maintaining precise alignment with input layouts [44]. Diffusion models counterbalance these shortcomings by iteratively refining outputs through probabilistic denoising, which captures spatial relationships and details effectively [4]. In hybrid implementations, GANs establish the foundational global scene structure while diffusion models fine-tune the layout alignment and enhance detail. The LayoutDiffusion framework [4], for instance, leverages object-aware cross-attention modules in conjunction with diffusion-based refinement to ensure both global and local synthesis accuracy. However, the computational costs of such hybrids, particularly the repeated forward passes characteristic of diffusion stages, pose significant efficiency challenges.

Similarly, integrating transformers into GAN architectures facilitates the modeling of long-range spatial relation-

ships while mitigating GANs' limitations in capturing complex layouts. Through their multi-head self-attention mechanisms, transformers effectively encode global dependencies, making them especially advantageous for layouts featuring multiple interacting entities [3], [22]. By embedding transformers within GAN pipelines, hybrid models such as LayoutFormer++ [19] improve adherence to spatial constraints, employing constraint serialization and decoding-space restriction to reduce layout violations while maintaining visual integrity. This approach, however, introduces additional computational demands due to the quadratic complexity of transformers' attention mechanisms, making scalability to high-resolution images a pressing concern.

Multi-modal hybrids incorporate auxiliary inputs—such as textual descriptions, depth maps, or stylistic embeddings—alongside layout information, enriching the generative process and enhancing adaptability. For example, LayoutGPT [45] extends layout-to-image pipelines by aligning textual and numeric semantic layout plans with visual synthesis, enabling more conceptually coherent outputs than traditional text-to-image systems. Similarly, LayoutDiffuse [33] uses neural layout adaptors and task-specific prompts to adapt foundational diffusion models for layout-conditioned tasks, streamlining multi-modal conditioning into a seamless pipeline. While these approaches significantly expand the usability of layout-guided synthesis, they necessitate advanced integration mechanisms to handle potential redundancies and inconsistencies across modalities.

Hierarchical hybrids represent another promising direction, uniting coarse-to-fine hierarchical generative models to simultaneously enforce global structural constraints and refine local attributes. Semantic Hierarchy Emerges in Deep Generative Representations for Scene Synthesis [46] highlights the merits of hierarchical latent spaces, wherein early model layers encode global layout structures and later layers refine object attributes and textures. Similarly, approaches like Layout-to-Image Generation with Localized Descriptions [47] employ hierarchical modulation at varying granularities to balance structural consistency with detail accuracy. Despite their strengths, these approaches often contend with increased architectural complexity and potential training instability, particularly when managing gradients across multiple stages.

While hybrid generative models bring significant advantages in terms of realism, precision, and multi-modal adaptability, they also entail trade-offs in computational efficiency, scalability, and interpretability. The interplay among architectural components can lead to increased training time and resource demands, with challenges exacerbated in transformer- or diffusion-based hybrids. Additionally, ensuring consistent multi-modal conditioning and seamlessly integrating hierarchical representations remain unresolved hurdles.

Nonetheless, the advances achieved by hybrid frameworks demonstrate their critical role in advancing layout-guided synthesis across various domains, from urban simulations and creative design to scientific visualization. Future efforts should focus on further optimizing hybrid architectures by reducing computational overhead through lightweight model designs, exploring self-supervised pretraining paradigms, and leveraging modular training pipelines for improved adaptability. These strategies stand to unlock new possibilities for achieving unparalleled fidelity, diversity, and user control in layout-to-image synthesis, paving the way for broader applications across diverse and complex contexts.

3.5 Specialized Training Techniques and Optimizations

Training layout-guided generative models presents a dynamic set of challenges, necessitating strategies that respect the structural coherence and diversity of layouts while enhancing semantic alignment with synthesized outputs. This subsection delves into specialized training techniques and optimization paradigms that address these requirements, focusing on methods that ensure adherence to layout constraints, augment generalization across layout styles, and improve computational efficiency.

A prominent focus in training layout-guided models is the design of loss functions tailored to enforce spatial and semantic fidelity. Conditioning losses such as Intersection over Union (IoU)-based penalties and position-sensitive constraints are commonly utilized to ensure adherence to the input layouts [18]. For instance, differentiable wireframe rendering layers optimize layout fidelity by directly allowing object alignments to propagate as constraints during adversarial training [1]. Moreover, layout-aware auxiliary losses like segmentation-based or hierarchical object-alignment functions can robustly integrate global coherence with local realism in complex inputs [48].

To further improve model robustness, layout-augmented pretraining has emerged as a foundational paradigm. Pretraining on large-scale synthetic layout datasets ensures coverage of diverse and often underrepresented configurations, fostering improved downstream performance. Layout2Im [2] demonstrated the efficacy of such approaches by leveraging coarse-to-dense layout augmentation pipelines that disentangle semantic encoding (e.g., classes, orientations) from instance-specific details. Similarly, LayoutDiffusion [4] integrates synthetic data generation during model initialization and employs noise schedules tailored to ensure functional proximity between corrupted and corresponding clean layouts.

Domain-specific fine-tuning has further propelled advancements in specialized synthesis tasks. This optimization paradigm targets the transfer of pretrained layout-aware embeddings to context-sensitive domains, such as medical imaging [49], urban planning [50], or creative design [51]. Fine-tuning retains domain-agnostic layout priors while adapting models to handle domain-specific constraints like anatomical accuracy or structural plausibility. Generative frameworks like LayoutGPT [8] leverage transfer learning to refine contextual embeddings, aligning text descriptions to syntactically enriched layouts.

Another promising avenue lies in modular training pipelines. Modular frameworks decouple the layout understanding components (e.g., scene graph parsing, layout-to-representation mapping) from the synthesis components to enhance scalability. Notable examples include LayoutTransformers [3], which adopt element-wise modular attention to improve context aggregation selectively. Similarly, Layout Diffusion Models, such as LayoutDM [43], utilize plug-and-play modules for augmenting layouts with element-specific

constraints in a non-retraining fashion. This approach results in streamlined workflows, reducing computational demands.

From an optimization standpoint, reinforcement learning (RL) has facilitated the iterative refinement of generated layouts. Methods such as SceneComposer [15] employ RL punishments for spatial misalignments and rewards for adherence to global constraints. This synergy ensures layouts with improved semantic interpretability while maintaining configurational realism. Furthermore, parameter-efficient training paradigms like low-rank adapters have emerged to mitigate training overhead in diffusion-based layout models. By directly conditioning diffusion steps with layout constraints, these approaches reduce dependence on annotated data while achieving competitive results [36].

Despite these advancements, persistent challenges remain. Models still struggle to generalize efficiently across multi-modal layouts or configurations involving hierarchical, temporal, or 3D complexities [8], [34]. Another unresolved issue is the scalability of training pipelines for larger, more intricate layouts without sacrificing synthesis fidelity. Emerging trends, such as leveraging foundational models in layout contexts or employing self-supervised embeddings for low-resource domains, offer significant potential for resolving these gaps.

In conclusion, training techniques for layout-guided generative synthesis are rapidly evolving toward broader generalization and richer contextual grounding. Continued focus on modularity, data-driven augmentation, and efficient training pipelines will be key to addressing existing limitations and broadening applicability across domains and modalities.

3.6 Challenges and Future Directions in Generative Layout Synthesis

Generative layout synthesis has achieved remarkable milestones, yet several technical and practical challenges remain, underscoring critical areas for future exploration. To ensure continued progress in the field, it is essential to address these unresolved challenges while drawing connections to recent advancements and emerging opportunities.

One enduring challenge lies in **generalizability across diverse layout types and application domains**. While many generative models excel on constrained datasets such as COCO-Stuff or Visual Genome, their performance degrades significantly when faced with unseen or complex layouts. For example, LayoutGAN [1], which is well-suited to structured 2D layouts in domains like document or graphic design, struggles to adapt to hierarchical or multi-modal settings. Similarly, Layout2Im [2], known for its strengths in generating images from bounding-box layouts, is restricted by its inability to handle open-world semantics or freestyle configurations [40]. Overcoming these limitations requires the development of architectures that leverage robust, foundational multimodal models or fine-tuned diffusion frameworks, enabling adaptation to heterogeneous data distributions without necessitating extensive domain-specific retraining [52], [53].

The **computational demands of real-time layout-guided generation** constitute another key bottleneck. While diffusion-based methods such as LayoutDiffusion [4] and LayoutDiffuse [52] achieve high-quality results, their iterative refinement processes are computationally intensive, rendering them less feasible for real-time or interactive applications. Transformer-based techniques like LayoutTransformer [3] improve efficiency through parallelization, but their memory overhead and latency continue to hamper scalability for industrial and real-world scenarios, such as interactive content design or large-scale simulations. Addressing this issue may require hybrid pipelines that combine the strengths of autoregressive architectures with lightweight diffusion or GAN variants, striking a balance between fidelity and efficiency [48], [54].

A further critical obstacle is the presence of **semantic and spatial inconsistencies during synthesis**. While cross-modal systems like LayoutGPT [8] and SceneComposer [15] integrate textual descriptions into layout generation, they frequently face challenges in ensuring precise spatial alignment and maintaining stylistic or semantic coherence across objects. These inconsistencies often arise from limitations inherent to layout-aware representation learning. For instance, transformers that rely on standard cross-attention mechanisms have been found lacking in object-level fidelity [3]. Novel solutions, such as Object-aware Cross Attention (OaCA) [4] and geometry alignment modules [49], show promise for addressing these challenges but have yet to generalize effectively in high-density or multi-object scenes.

Ethical and sociotechnical considerations also pose significant challenges for generative layout synthesis. Many state-of-the-art models are trained on datasets with limited diversity, which risks perpetuating stereotypes or failing to represent inclusive layouts [38]. Incorporating fairness-aware strategies, such as bias detection during training [55], into generative systems is a critical research direction. Additionally, applications in domains like medical imaging or urban planning amplify concerns about privacy. Developing synthetic-layout frameworks that include robust privacy guarantees will be crucial to mitigate such risks [20].

Exciting progress in areas such as **3D layouts and multi-modal controllability** is unlocking new possibilities but also introduces its own set of complexities. Models like Build-A-Scene [16] demonstrate advancements in integrating depth-conditioned frameworks to enable 3D-aware layout guidance. However, achieving fine-grained scene-level coherence and control in 3D environments remains a major challenge. Similarly, multi-modal synthesis methods that incorporate audio, text, and visual layouts, such as VideoComposer [56], encounter difficulties in preserving cross-modal consistency while sustaining high resolution and fidelity.

Future research should prioritize the development of scalable, efficient, and adaptable frameworks to address these challenges. Leveraging foundational models pre-trained on large-scale data, combined with modular, task-specific fine-tuning approaches [43], offers promising pathways toward addressing domain generalization issues. Additionally, efforts to unify layout synthesis with explainable AI principles can enhance interpretability and user-centric application design, enabling broader industry adoption. By confronting these limitations and harnessing cutting-edge techniques, generative layout synthesis stands poised to

redefine creative design, simulation, and education across diverse domains.

4 Evaluation Frameworks and Metrics

4.1 Metrics for Spatial Fidelity and Layout Alignment

Evaluating spatial fidelity and adherence to prescribed layouts remains a pivotal focus in layout-guided image synthesis. Accurate metrics play a fundamental role in quantifying the consistency between generated outputs and input layouts, thus ensuring usability across diverse applications such as graphic design, medical imaging, and architectural visualization. This subsection delves into prominent metrics, their strengths and limitations, and insights for future directions.

One of the most widely adopted metrics for spatial fidelity is the Intersection over Union (IoU), which measures the overlap between the spatial regions of objects in the generated image and their corresponding regions in the input layout. Formally, given a predicted region R_p and a ground-truth region R_g, IoU is calculated as:

$$\text{IoU} = \frac{|R_p \cap R_g|}{|R_p \cup R_g|}$$

IoU is particularly advantageous for evaluating the alignment of bounding boxes and segmentation masks due to its simplicity, scalability, and robustness across varying scene complexities. Methods like Layout2Im [2] and LayoutDiffusion [4] have extensively employed IoU to validate alignment precision. However, a limitation of IoU arises in complex scenes with overlapping or occluded objects, where it may fail to capture nuanced spatial misalignments or hierarchical relationships between objects.

Building on IoU, structural IoU variants—tailored for hierarchical scene graphs or layouts involving complex shapes—offer more context-sensitive evaluations. For instance, scene graph-guided methods [27], [55] leverage topological consistency checks to account for relational misalignments, enabling the evaluation of scene-level spatial arrangements. These approaches, however, often involve higher computational costs and require well-labeled hierarchical data, which may not always be available.

Another line of metrics focuses on pixel-wise evaluation, offering finer granularity to assess misalignments at the subregion level. Metrics such as Chamfer Distance, which computes the ℓ_2 distance between nearest points on predicted and ground-truth spatial contours, have been employed in conditional synthesis paradigms [1]. Additionally, metrics like position-sensitive pixel accuracy, which assess the correctness of pixel-location predictions within segmented regions, have emerged as effective complements. Despite their precision, pixel-based metrics are computationally expensive and sensitive to minor perturbations, making them less robust for high-level layout evaluation.

For alignment tasks that require integration of global geometric consistency, pre-trained vision models like CLIP have been leveraged in an embedding space setup to assess contextual alignment between layouts and images. For instance, LayoutGPT [8] leverages similarity scores from CLIP to measure alignment between generated outputs and layout descriptions. This embedding-based evaluation excels in scenarios requiring semantic validation, though it may overlook precise geometric alignment due to its reliance on latent feature representations.

Moreover, metrics addressing object-specific spatial coherence are gaining importance. Semantic Object Accuracy (SOA), introduced in models like LayoutVAE [13] and LayoutFormer++ [19], evaluates whether the generated image respects the object types, sizes, and positions as prescribed by the layout. These approaches ensure compatibility between synthesis outputs and input layouts in domains like document design or medical imaging.

Despite these advances, critical challenges persist in defining universal evaluation protocols. For instance, the ambiguity in metric selection—depending on the complexity of the input layout or the target domain—hinders the generalizability of proposed methods. Furthermore, emerging needs, such as generative synthesis involving dynamic or interactive layouts, highlight gaps in existing methodologies. While traditional metrics like IoU or Chamfer Distance excel in static configurations, they struggle to fully capture layout evolution in interactive scenarios, as noted in interactive systems like Build-A-Scene [16].

Future directions should explore hybrid evaluation paradigms that combine geometric, semantic, and user-centered metrics. For instance, a multi-faceted framework could integrate pixel-level metrics (e.g., Chamfer Distance) with semantic relevance checks (e.g., CLIP-based scores) and domain-specific utility scores, such as diagnostic accuracy in medical imaging [6]. Simultaneously, incorporating user studies to validate perceptual alignment in subjective domains such as graphic design could bridge existing gaps between automated metrics and real-world usability. In advancing these directions, research must also address the computational and interpretability trade-offs inherent in multi-metric integration.

Ultimately, as layout-guided image synthesis continues to evolve, robust, scalable, and domain-flexible metrics will remain vital to ensuring the consistency and quality required for deployment in high-stakes applications.

4.2 Visual Realism and Perceptual Quality Metrics

Evaluating the visual realism and perceptual quality of layout-guided controllable image synthesis is crucial for assessing the overall quality and practical effectiveness of generative models. This subsection explores both automated and human-centric metrics, examining their comparative methodologies, recent advancements, and inherent challenges in the context of layout-guided synthesis.

Automated evaluation metrics play a central role in measuring visual realism, with Fréchet Inception Distance (FID) [2], [4], [57] being one of the most widely adopted. FID quantifies the statistical similarity between feature distributions of generated and real images by computing the Wasserstein-2 distance between multivariate Gaussian distributions parameterized by the means and covariances of embeddings in an Inception model. While highly robust for assessing dataset-level realism, FID falls short when tasked with evaluating layout-driven tasks that require fine-grained spatial realism and adherence to specified object arrangements. To address such limitations, extensions like

SceneFID [18] have been developed. By incorporating object-centric embeddings, SceneFID adapts FID for multi-object scene synthesis, enabling a dual focus on compositional consistency and visual fidelity. These advancements underscore the shift toward automating more layout-aware quality evaluations to cater to the unique requirements of layout-guided synthesis.

Complementing FID, other metrics such as Multi-Scale Structural Similarity Index (MS-SSIM) [58] provide perceptually grounded measures of visual similarity. MS-SSIM assesses structural congruence over multiple resolutions, rendering it particularly effective for tasks involving adherence to semantic layouts. However, its limitations become apparent in tasks where high-level semantic accuracy or aesthetic nuances are critical to perceptual quality. Recent strategies have incorporated masked or region-specific evaluation metrics, such as those used in LayoutDiffusion [4], to more directly assess spatial fidelity by focusing on local adherence to specified layout regions using object masks. While such metrics improve granularity, they often require task-specific implementations, which can hinder generalizability.

Human-centric approaches, on the other hand, capture subjective aspects of visual realism and perceptual quality that automated metrics may overlook. Human preference studies, where individuals rank synthesized outputs on criteria like realism, coherence, and alignment, remain the gold standard for evaluating perceptual quality [26], [59]. These evaluations inherently factor in subtle elements of visual plausibility and semantic coherence, making them indispensable in domains like art or design. Nonetheless, human-centric evaluations are resource-intensive and exhibit variability due to inter-rater differences. Combining controlled user studies with scalable crowd-sourced experiments has been effective in mitigating these drawbacks, providing broader insights into user satisfaction and applicability across diverse audiences.

Emerging methodologies leveraging multimodal systems like CLIP have bridged automated and semantic metrics by introducing contextual evaluations of textual and visual alignment [45], [60]. CLIP's ability to compute alignment scores between text/image embeddings facilitates a more holistic assessment of how well generated imagery matches input specifications, considering both spatial configurations and visual semantics. However, these embedding-based techniques are sensitive to the pre-trained biases inherent in the underlying models, often prioritizing style consistency over precise spatial alignment unless fine-tuned for layout-specific objectives. This trade-off reflects the limitations of relying solely on embedding models in the context of complex, layout-driven synthesis.

The trade-offs between automated and human-centric methodologies emphasize the need for hybrid frameworks to ensure balanced and comprehensive evaluations. While metrics like FID and MS-SSIM are efficient and reproducible, they miss contextual subtleties that affect how outputs are perceived in real-world applications. Conversely, human evaluations, while providing nuanced insights, are inherently subjective and resource-demanding. Hybrid approaches, integrating objective metrics with perceptual feedback, could address these gaps, enhancing both accuracy and interpretability in assessment processes.

Another critical challenge is the absence of standardized evaluation protocols for layout-guided systems, particularly those generating outputs from diverse spatial configurations, such as 3D-aware layouts [31], [61]. For instance, tasks involving 3D perspective consistency and viewpoint adaptations require evaluation metrics that go beyond traditional 2D assessments to account for depth, orientation, and three-dimensional spatial realism. Similarly, for multimodal tasks integrating text and layout, comprehensive evaluation methodologies must assess both modality-specific alignment and the coherence of interactions across modalities [15], [45]. Addressing these gaps is crucial for creating benchmarks that can assess performance across increasingly complex and multimodal generation tasks.

In conclusion, while significant progress has been made in the evaluation of visual realism and perceptual quality, challenges remain in fully capturing the multifaceted requirements of layout-guided synthesis. The development of adaptive, context-aware metrics capable of balancing computational efficiency with perceptual relevance will be essential for advancing this field. Future research efforts should also focus on the creation of standardized benchmarks that incorporate diverse datasets spanning multiple domains, accompanied by interactive, feedback-driven metrics for real-world usability. By addressing these challenges, the field can progress toward a more robust, holistic evaluation paradigm that aligns automated metrics with subjective user expectations, ultimately bridging the gap between research methodologies and practical deployments in layout-guided image synthesis.

4.3 Semantic and Contextual Coherence Metrics

Semantic and contextual coherence are fundamental metrics for evaluating whether generated images align with the semantic relationships, spatial configurations, and hierarchical structures specified by input layouts. This coherence ensures that synthesized images not only appear realistic but are also meaningful in relation to their prescribed context. In this subsection, we explore approaches designed to assess semantic and contextual coherence, analyzing their methodologies, advantages, limitations, and emerging trends.

One essential metric for evaluating coherence is **Semantic Object Accuracy (SOA)**, which quantifies whether the categories and attributes of objects in the generated image match those defined in the input layout or textual description. Many methods, such as those based on bounding box-conditioned synthesis [2], integrate these labels directly into the generative process, making SOA a useful starting point for evaluation. However, this metric alone may fail to capture relational and spatial mismatches among objects, as it focuses on object correctness in isolation.

To extend beyond per-object evaluation, **Scene Graph Consistency** metrics have emerged as compelling tools. These evaluate whether the hierarchical and relational dependencies encoded in scene graphs, such as "object A next to object B" or "object C above object D," are respected in the generated images. For example, models like those proposed in [17], [25] emphasize adherence to these graphs, leveraging structural scene components to enhance relational fidelity. While scene graph consistency captures object

interactions effectively, it proves computationally expensive for complex graphs and can struggle with ambiguous or underspecified relational input.

Another critical approach involves the use of **pre-trained vision-language models**, such as CLIP, to compute contextual alignment scores. This involves measuring the cosine similarity between visual embeddings extracted from the generated image and text embeddings derived from the input description or layout. By aligning multimodal representations, this evaluation provides a holistic measure of contextual coherence, enabling quantification of both spatial relationships and semantic agreement [15], [40]. However, issues such as model bias and sensitivity to dataset-specific training distributions can hinder generalizability. Furthermore, CLIP-based metrics may poorly capture nuances in hierarchical semantics, where object arrangement or complex relationships play a critical role.

For hierarchical settings, methods like **context-aware localization loss** add a layer of granularity by penalizing deviations from spatial or relational hierarchies defined in input semantics. For instance, LayoutDiffusion [4] incorporates position-sensitive modules to ensure alignment of generated regions with predefined layouts. This hierarchical view enables evaluation of both global and local contextual adherence, addressing shortcomings of simpler object-centric or pixel-level metrics.

A significant challenge in the evaluation of semantic and contextual coherence lies in defining ground truth for tasks with underspecified or abstract instructions. For example, synthetic augmentation frameworks using unconstrained compositional prompts [23] often introduce layout configurations that are novel or unstructured. This necessitates the development of adaptive metrics capable of dynamically approximating coherence without reliance on exhaustive ground truth annotations. Proposed solutions include self-supervised correspondence techniques, where generated images are re-encoded to recover their corresponding input layouts, thereby validating alignment indirectly [33].

Emerging trends in this domain involve multi-modal and multi-resolution evaluative pipelines. For example, adaptive frameworks like those introduced in [37] integrate fine-grained alignment constraints through multi-scale attention, achieving more robust evaluations of layout fidelity across object granularities. Another avenue includes leveraging large language models (LLMs), such as LayoutGPT [8], to automate layout validation by generating paraphrased descriptions of synthesized scenes for reverse mapping and coherence scoring.

In conclusion, while significant progress has been made in developing semantic and contextual coherence metrics, gaps remain in evaluating increasingly complex and hierarchical tasks. Future directions could explore metric fusion strategies, combining object-centric, relational, and global-contextual measures into unified scoring systems. Additionally, adaptive evaluation frameworks capable of learning coherence definitions in novel, open-domain scenarios present a promising frontier for ensuring robust and comprehensive metric adoption in layout-guided image synthesis.

4.4 Utility-Driven and Domain-Specific Metrics

Utility-driven and domain-specific metrics constitute a pivotal dimension of evaluating layout-guided controllable image synthesis, especially in use cases demanding bespoke objectives and functionality. Distinguishing themselves from generic metrics focused on visual fidelity or spatial alignment, these specialized metrics aim to measure how well the generated images fulfill practical and industry-specific requirements. By aligning evaluation frameworks with the distinct needs of domains like healthcare, urban planning, and graphic design, researchers can better benchmark the real-world applicability of synthesis systems and align them with specific functional goals.

In healthcare, for instance, the evaluation of synthetic medical imagery prioritizes measures reflecting clinical relevance and diagnostic integrity. Synthetic images must not only demonstrate high anatomical realism but also contribute meaningfully to tasks such as disease detection or treatment planning. Metrics like Dice coefficient and Intersection over Union (IoU) are often employed to compare synthetic segmentation masks against real counterparts in tasks such as pathology generation. Furthermore, utility-oriented evaluations extend to gauging the improvement of downstream tasks, such as classification accuracy when augmented with synthetic data. Another critical consideration in this domain is the creation of privacy-preserving datasets, wherein metrics must balance fidelity with obfuscation rigor to ensure preserved anonymity alongside diagnostic validity. However, achieving sufficient diversity in synthetic datasets while retaining diagnostic usability remains a persistent challenge, underscoring the demand for more nuanced evaluation techniques.

In urban planning and architectural design, domain-specific metrics assess the practical utility of synthesized layouts by evaluating their capacity to support functional simulations and spatial optimization. Key performance indicators include spatial efficiency, traffic flow quality, and compliance with zoning regulations. For example, models like CityGen [29] introduce systems for generating 3D urban layouts, requiring metrics that measure not just spatial precision but also urban functionality at scale. Similarly, methods like LayoutGAN [1] target architectural and graphic layout generation, with metrics spanning structural integrity, blueprint alignment, and adherence to design constraints. These metrics highlight the importance of coupling automated evaluation with human feedback, allowing real-time design adjustments in dynamic layout-to-image pipelines. Despite significant advancements, the field still necessitates developing more context-sensitive evaluative techniques, capable of capturing nuanced trade-offs between functional accuracy and contextual demands.

For creative design and advertising, utility-driven metrics often balance aesthetic, functional, and user-centric considerations. For example, evaluating an automatically generated poster layout may involve combining compositional measures like spatial symmetry and alignment with subjective metrics such as user satisfaction or crowd-sourced aesthetic preferences. Specialized systems like Desigen [62] align layouts with domain-specific design principles, such as typographic harmony and appropriate utilization of neg-

ative space, refining layouts to cater to guidelines while maintaining creative variability. However, challenges frequently arise in balancing creativity with functional clarity, reflecting the broader need for adaptable evaluation protocols that address overlapping goals across different design domains.

Emerging trends also emphasize utility-driven metrics for augmenting machine learning workflows. Synthetic datasets derived from layout-guided synthesis pipelines are increasingly leveraged to improve model training, particularly in scenarios with limited real-world data. Metrics in these cases measure performance gains, such as error reduction in tasks like object detection or improved data heterogeneity to address class imbalance. However, the utility of these datasets depends on carefully aligned data distributions; inconsistencies in synthetic layouts may introduce biases or propagate errors, complicating their contribution to generalizable model performance.

The inherent domain specificity of utility-driven metrics poses a significant challenge to standardization, necessitating contextualized frameworks for evaluation. Future directions should emphasize integrating multi-domain metrics into unified frameworks while preserving the granularity required for domain-specific objectives. Incorporating user-centered evaluation methodologies, such as real-world usability testing in creative design or diagnostic acceptability in healthcare, will be key to bridging automated and manual assessment gaps. The emergence of foundational models like LayoutGPT [8] provides additional opportunities to broaden utility benchmarks by supporting cross-modal, multi-domain synthesis applications. As evaluation practices evolve, a focus on flexible, inclusive, and practical validation pipelines will be essential to advancing utility-driven metrics and ensuring their alignment with the diverse landscape of real-world applications.

4.5 Benchmark Datasets and Protocols

The evaluation of layout-guided image synthesis methods relies on robust benchmark datasets and standardized protocols to ensure fair, comprehensive, and reproducible comparisons. This subsection provides an analysis of the existing datasets, evaluates their suitability for diverse synthesis tasks, and examines standardization criteria for effective evaluation.

A cornerstone for benchmarking layout-guided synthesis is the availability of datasets with rich annotations that link spatial (layout) characteristics to semantic content. Datasets such as COCO-Stuff and Visual Genome remain widely used for tasks like layout-to-image generation due to the detailed annotations they provide for object categories, bounding boxes, spatial relationships, and pixel-level segmentation [2], [17]. COCO-Stuff, with its broad coverage of real-world scenes, allows testing models under diverse layout configurations, while Visual Genome's densely annotated scene graphs facilitate the evaluation of more structured, relational synthesis approaches. However, the scalability and compositional complexity of these datasets often pose challenges for models designed to generalize across multi-object, high-density scenes [2], [4].

Specialized datasets, including RICO (user interfaces) and PubLayNet (documents), have enabled domain-specific evaluations, particularly for design and structured layout applications [19], [22]. For example, RICO's fine-grained element annotations improve benchmarking in generating UI layouts, whereas PubLayNet's hierarchical structure proves critical for testing document-layout generation models. However, such datasets are limited by their domain specificity, which may restrict their applicability to broader layout synthesis tasks, necessitating the creation of more generalized benchmarks.

Recently, novel datasets have been introduced to address compositional challenges inherent in complex layouts. T2I-CompBench, for instance, enables the evaluation of text-to-image models under compositional constraints across attributes, relationships, and multi-object scenarios [63]. Its focus on compositional instructions, such as spatial relationships and hierarchical associations, fills a critical gap in benchmarking alignment with input layouts. Another notable addition is the Structured3D dataset, which extends benchmarks into 3D layout conditioning by providing realistic layouts and corresponding ground-truth meshes for room-scale scenes [64]. Such datasets open avenues for bridging 2D layout control with emerging demands in 3D and immersive environments, a frontier increasingly explored by studies like LayoutDiffusion and GALA3D [4], [65].

Standardized evaluation protocols are paramount for ensuring fair comparative analysis. Layout-guided image synthesis entails multiple performance dimensions, including spatial alignment, semantic consistency, visual realism, and diversity. Metrics like Intersection over Union (IoU) are widely employed to quantify spatial conformity with input layouts, often supplemented by pixel-wise accuracy or the Chamfer distance for finer geometric evaluations [4], [32]. Additionally, Fréchet Inception Distance (FID) persists as a primary metric for visual fidelity and realism, although it overlooks layout-specific factors and compositional alignment, leading to the introduction of measures like SceneFID, which emphasize multi-object adherence and categorical accuracy [18].

Emerging protocol designs are increasingly aimed at multi-modal settings, where text, layout, or other modalities (like depth or scene graphs) jointly guide synthesis. A significant advancement is the use of pre-defined prompts and hierarchical layouts for controlled comparisons. For example, LayoutGPT provides a standardized pipeline for evaluating reasoning and fidelity in text-to-layout translation, achieving more interpretable results within conditioned synthesis benchmarks [45]. Similarly, layout-aware conditional sampling strategies, as proposed in studies such as LayoutDM, demonstrate the importance of masking-based protocols in achieving precise spatial control during layout generation [45].

Despite progress, key challenges persist in dataset diversity and protocol standardization. First, existing datasets often fail to capture the complexity of real-world layouts, particularly in niche domains such as scientific visualization or augmented reality. Secondly, standardized pipelines that integrate IoU-like metrics with user-centric assessments remain underdeveloped. Comprehensive benchmarks that combine synthetic and real-world data, spanning 2D, 3D, and hierarchical layouts, are critical for advancing the field

further. Future directions might also involve adaptive protocols tailored to dynamically evolving layouts or interactive synthesis settings, allowing iterative improvements during evaluation phases.

In sum, while benchmark datasets like COCO-Stuff, Visual Genome, and emerging contributions (e.g., T2I-CompBench, Structured3D) have significantly advanced the systematic evaluation of layout-guided synthesis, domain-specific gaps and the limited coverage of complex relationships and compositional hierarchies still warrant further attention. Advances in evaluation protocols, particularly multi-modal and interactive setups, will play a crucial role in pushing layout-guided synthesis toward applications that demand higher levels of spatial precision, semantic coherence, and adaptability.

4.6 Limitations, Emerging Challenges, and Future Directions in Evaluation

Existing evaluation frameworks for layout-guided image synthesis face considerable limitations in addressing the expanding complexity and diversity of generative modeling tasks. These challenges stem from the inadequacy of current protocols to fully encapsulate multi-modal interdependencies, hierarchical spatial relationships, semantic coherence, and real-world user-centric requirements. This subsection examines these gaps across different evaluation dimensions, highlights recent advancements, and explores emerging avenues for designing more robust and holistic evaluation paradigms.

One core limitation is the absence of universally applicable metrics to assess alignment between input layouts and generated images across diverse formats such as bounding boxes, segmentation masks, or scene graphs. While metrics like Intersection over Union (IoU) and Chamfer Distance provide effective geometric alignment measurements for structured elements, they often fall short in capturing higher-level spatial semantics or intricate relational interactions, particularly in complex scenes [2], [45]. For instance, a bounding box or segmentation mask that is geometrically accurate might fail to reflect relational inconsistencies, such as incorrect layering, occlusions, or size hierarchy conflicts. Attempts to address these issues through semantic alignment metrics, such as Structural Similarity (SSIM) or CLIP-based scores, have shown potential but remain constrained by their reliance on pre-trained visual-textual models that inadequately capture detailed spatial dependencies [4], [66].

Another critical gap is the evaluation of hierarchical and multi-resolution layouts. Certain domains — like urban planning or medical imaging — demand not only strict pixel-level spatial coherence but also structural congruency across various granular levels. However, current benchmarks often focus either on global layout accuracy or fine-grained segmentation, without accommodating the need for multi-scale hierarchical evaluations that traverse these intermediate resolutions [4], [22]. Moreover, the evaluation of semantic hierarchies, such as nested relationships within scene graphs, remains largely underexplored in mainstream frameworks [27].

Multi-modal synthesis workflows, where layouts are integrated with auxiliary inputs—such as text, depth maps, or temporal layouts—introduce further evaluation complexities. Existing metrics are rarely equipped to assess coherence across multiple conditioning modalities. For example, widely adopted perceptual quality metrics like Fréchet Inception Distance (FID) fail to evaluate the mutual alignment of spatial constraints with textual or temporal semantics [67], [68]. Although innovative frameworks like LayoutGPT's compositional evaluation pipeline have attempted to address these challenges, they still require refinement for robustly capturing latent interdependencies between modality-conditioned inputs and generation outputs [45].

A particularly pivotal shortcoming across evaluation paradigms is the lack of user-centric metrics. Practical applications often demand the incorporation of subjective preferences, such as positional accuracy, compositional aesthetics, or usability for specific tasks. However, most automated evaluations overlook these subjective factors, resulting in a gap between technical performance metrics and real-world applicability [38], [69]. Emerging interactive evaluation approaches that include real-time feedback mechanisms—potentially through adversarial learning or reinforcement frameworks—are in their infancy but show promise in enhancing user relevance and task-specific adaptability [48], [70].

Emerging trends indicate promising solutions to these challenges. First, the development of multi-scale contextual alignment metrics that account for layout fidelity across hierarchical resolutions and multi-modal inputs could establish a unified and flexible evaluation standard. Recent methods, such as Layout Fusion modules and object-aware attention mechanisms, demonstrate progress in encoding these relationships during synthesis but lack complementary evaluation methodologies [4], [49]. Additionally, leveraging pre-trained foundational models specialized for diverse domains, such as 3D datasets or depth-aware tasks, could standardize evaluation for immersive applications like augmented and virtual reality [31], [71].

Looking forward, the need for standardized, adaptable evaluation pipelines remains critical, particularly those that integrate quantitative metrics with user-focused qualitative feedback. Such frameworks should prioritize accessibility for non-experts while ensuring applicability across diverse domains and cultural contexts. Furthermore, ethical considerations, such as addressing potential biases within layout datasets or disparities in subjective user experiences, must be a core focus to ensure fairness and inclusivity in generative evaluation protocols [54], [72].

By systematically addressing these gaps, the field can bridge the divide between technical evaluation benchmarks and practical user applicability, ultimately supporting both research advancements and real-world deployment of layout-guided image synthesis technologies.

5 Applications and Use Cases

5.1 Media and Entertainment Applications

The media and entertainment industries have witnessed transformative advancements driven by layout-guided controllable image synthesis, enabling content creation processes that are both highly efficient and creatively versatile.

By providing precise spatial control, these methods streamline the development of personalized, high-quality visuals across domains such as advertising, animation, and digital art. Recent innovations in generative modeling and layout-based controls highlight the critical role of such technologies in enhancing creativity while maintaining economic and artistic feasibility.

In advertising and branding, layout-guided image synthesis models have enabled the creation of visuals with customizable compositions tailored to meet specific brand guidelines and target demographics. Systems such as LayoutGAN [1] use self-attention mechanisms and differentiable wireframe representations to generate layouts with precise alignment and visual appeal. This capability ensures that advertising visuals maintain structural consistency while adhering to design principles like balance, symmetry, and aspect ratio. For instance, advertisers can create responsive banner layouts across multiple formats with minimal manual intervention, thereby reducing labor costs and accelerating time-to-market. Additionally, methods like ControlGAN [39] can generate region-specific visuals conditioned by text, offering fine-grained control for localized advertising campaigns. However, while such systems provide reliable alignment between layouts and output images, challenges remain in handling extreme variability in object attributes and brand-specific stylistic elements, as explored in attribute-conditioned approaches [38].

The integration of layout guidance into animation preproduction, particularly for storyboard generation and visual pre-visualization, has significantly streamlined workflows. Traditional approaches to pre-visualization rely on manual design, which is labor-intensive and iterative. By adopting hierarchical text-to-layout frameworks [30], studios can semi-automatically convert narrative descriptions into scene-specific semantic layouts, reducing the cognitive and physical effort required from designers. These layouts can then be further refined into coherent visuals using advanced systems like Layout2Im [2], which disentangle layout attributes (e.g., object positions and categories) from uncertain visual properties such as appearance. Additionally, bidirectional layout transformations [5] allow iterative editing of layouts during the pre-visualization phase, empowering animators with dynamic control. However, scalability in generating rich, high-fidelity visuals for complex movie scenes remains a major limitation, as the interplay of multiple characters, props, and environmental details often challenges even the most robust layout-conditioned pipelines.

Virtual content creation, particularly for gaming and immersive XR (extended reality) environments, also benefits significantly from layout-driven synthesis. For example, models like LayoutDiffusion [4] enhance multi-object generation by leveraging object-aware cross-attention mechanisms, ensuring precise spatial relationships among virtual assets. These systems can create expansive game scenes with highly controlled compositions, allowing developers to integrate themed objects or architectural elements in specified regions while preserving global coherence. LayoutGPT [8] has further shown promise by incorporating large language models to convert freeform textual descriptions into structured layouts, enabling intuitive guidance even for non-specialist users. Nonetheless, limitations in adapting these models to real-time applications, particularly for large-scale multiplayer game environments, indicate an area ripe for further exploration.

In creative arts, layout-guided systems have opened unprecedented possibilities for interactive and generative art forms. Artists can experiment with compositional frameworks dynamically using tools like SceneComposer [15], which interpolate between text-only inputs and precise shape-based layouts. These systems allow artists to manipulate a layout's precision based on their creative intent and level of expertise, facilitating seamless collaboration between machine and human creativity. Additionally, freestyle layout-to-image synthesis models [40] promote artistic expression by enabling the generation of unconventional, non-predefined object-semantic pairs, thus expanding the boundaries of traditional generative modeling. Despite these advancements, the practical application of these creative tools depends on resolving issues such as computational overhead during high-resolution image generation and ensuring that outputs reflect nuanced artistic intent without repeated trial-and-error adjustments.

Looking forward, emerging research seeks to enhance layout-guided synthesis in media and entertainment through better multi-modal integration and adaptive spatial refinement. For instance, multimodal frameworks like Uni-Control [10] aim to incorporate language, style, and layout inputs within a unified synthesis model. Meanwhile, the ongoing exploration of 3D scene layouts [31] promises to revolutionize virtual reality design and cinematic storytelling by providing fully immersive, spatially coherent environments. However, scalability, real-time performance, and user accessibility remain crucial challenges for future advancements. By addressing these obstacles, layout-guided controllable synthesis holds immense potential to revolutionize the creative landscape of media and entertainment.

5.2 Healthcare and Medical Imaging

Layout-guided controllable image synthesis is redefining the landscape of healthcare and medical imaging, addressing crucial challenges in data availability, privacy, and computational efficiency. By employing spatial control mechanisms, this paradigm enables the generation of synthetic medical data tailored for diagnostic support, medical training, and research—all while adhering to stringent ethical regulations. The convergence of generative modeling and spatial layouts has provided a pathway for significant advancements in these domains, ensuring functionality and utility in highly sensitive medical contexts.

One of the primary applications of layout-guided synthesis in healthcare is the augmentation of limited medical datasets. Diagnostic models often encounter limitations stemming from data imbalance or scarcity, particularly for rare pathological conditions. By generating medical images from predefined anatomical layouts—such as segmentation masks or bounding boxes—researchers can create datasets that reflect accurate real-world variations while ensuring spatial coherence. For instance, works like [2] and [4] demonstrate how bounding boxes can map organ locations while segmentation masks define intricate anatomical

boundaries, facilitating refined control over the synthesis process. These capabilities not only address the issue of data scarcity but also improve the training of diagnostic models while simultaneously ensuring strict privacy compliance.

Another critical use case lies in synthetic pathology generation, where layout-guided approaches enable the controlled simulation of pathological structures, such as lesions or abnormalities, based on specific spatial configurations. This has profound implications for the study of rare diseases, where real-world imaging data is often insufficient. Techniques like those explored in [30] adapt hierarchical frameworks to the medical domain, bridging the gap between semantic annotations (e.g., "a lesion in the upper left quadrant of the liver") and spatially precise visual outputs. Such advances allow for the generation of diverse and representative pathological scenarios, facilitating research and training in underrepresented medical conditions.

Layout-guided synthesis also addresses critical privacy concerns prevalent in medical imaging. Data-sharing restrictions dictated by regulations like HIPAA often limit access to patient records for research purposes. By employing methods such as [20], which blend non-parametric retrieval with generative networks, researchers can produce realistic yet anonymized medical images that uphold privacy standards. These synthetic images ensure that no identifiable patient data is embedded, enabling ethically sound collaboration across institutions while mitigating concerns around potential privacy breaches.

Despite these advancements, challenges persist in achieving high visual fidelity and precise semantic accuracy, both of which are critical in medical imaging. Computationally intensive techniques, such as [26], deliver strong alignment between spatial inputs and generated outputs but often exhibit scalability limitations in real-time medical environments. Similarly, graph-based methods like those in [17] effectively model relational dependencies between anatomical structures but might struggle with the complexity of dense medical imaging datasets, where highly interconnected anatomical regions introduce additional processing demands.

Emerging innovations are pushing the boundaries of layout-guided synthesis to further enhance medical imaging. Hybrid models that incorporate multimodal inputs, such as textual prompts and clinical metadata, are beginning to address these challenges. For instance, frameworks like [45] combine spatial layout synthesis with language-driven spatial planning, showcasing the potential to create medical images guided not only by spatial constraints but also by textual medical annotations or reports. Moreover, ventures into 3D-aware synthesis, as suggested in [12], are unlocking the potential for volumetric data generation in MRI or CT imaging, granting richer spatial representations and deeper contextual relevance.

Beyond synthetic data generation, layout-guided synthesis holds promise for real-world clinical applications. Automated report generation systems could integrate synthesized images and their spatial layouts to assist radiologists, offering plausible diagnostic scenarios during consultations. Additionally, interactive tools like those derived from [3] enable iterative refinement of medical images, allowing clinicians to collaboratively and dynamically explore various diagnostic possibilities—aided by the adaptability of spatial guidance frameworks.

In summary, layout-guided controllable image synthesis offers groundbreaking potential to transform healthcare and medical imaging. By addressing challenges related to privacy, data scarcity, and diagnostic precision, this technology paves the way for more robust, scalable, and ethically compliant frameworks. Future research will need to focus on integrating multimodal inputs, improving scalability, and ensuring interpretability to fully realize its potential in key medical applications. Combining generative innovation with clinical workflows promises to advance personalized medicine, accelerate medical research, and uphold rigorous ethical standards, thereby reshaping the field of healthcare for years to come.

5.3 Urban and Architectural Design

Layout-guided controllable image synthesis has become a transformative tool in urban planning and architectural design, reshaping workflows through automated generation, simulation, and optimization of functional spaces with high spatial fidelity. By leveraging spatially explicit layouts, these methods enable efficient generation of urban environments, buildings, and interior designs that adhere to predefined spatial constraints and aesthetic goals. This subsection explores the strengths, limitations, and emerging directions of layout-guided image synthesis approaches in addressing challenges faced by urban and architectural design.

The inherent complexity of urban planning and architectural workflows necessitates tools that bridge the gap between functional requirements and aesthetic demands. Layout-based methodologies address these dual needs by encoding spatial relationships explicitly, such as road networks, building placements, or interior layouts, into generative processes. For instance, models like CityGen [29] employ multi-scale diffusion methods to generate diverse and controllable city layouts, offering scalability and the ability to tailor urban designs to specific planning objectives. Similarly, approaches like LayoutGPT [8] extend the capabilities of language-based planning, converting text-based natural language descriptions into structured urban layouts, demonstrating the synergy between large language models and layout synthesis techniques.

In urban contexts, the ability to generate realistic road networks, open spaces, and mixed-use environments is paramount. Models like LayoutDiffusion [33] and LayoutDM [43] excel in synthesizing large-scale layouts by capturing critical geometric and relational features of urban spaces. These models leverage transformer-based architectures or diffusion probabilistic frameworks to balance realism, diversity, and adherence to spatial constraints. One noteworthy strength of these approaches is their ability to synthesize infinite city layouts (as demonstrated in CityGen [29]) while allowing user input to refine specific areas interactively.

Architectural design workflows benefit from layout-guided synthesis by enabling rapid prototyping of building floorplans, interior designs, and structural arrangements. For instance, LayoutGAN [1] demonstrates the effective use of wireframe discriminators to optimize layouts in image

space, particularly for interior building designs, ensuring that spatial alignments and aesthetic coherence are preserved. Similarly, methods like LayoutVAE [13] introduce stochasticity into the generation pipelines, enabling variations in floorplans and room layouts that meet predefined functional requirements. These advancements reduce manual design effort while fostering creativity by offering multiple viable design solutions.

While layout-guided systems have shown promise, challenges persist, especially in achieving high-resolution synthesis for densely populated urban scenes. Transformer-based methods such as LayoutFormer++ [19] have made strides in improving the semantic consistency and visual quality of generated outputs. However, fine control over real-world alignment, especially in complex multi-object environments, remains limited. Furthermore, the scalability of these methods to incorporate environmental factors, such as terrain constraints or green spaces, lags behind their generative capabilities, presenting an opportunity for future work.

Emerging trends in this domain underscore the importance of integrating multi-modal and hierarchical representations. For example, advances such as GALA3D [34] expand the scope of layout-based planning to three-dimensional urban and architectural design, allowing designers to transition seamlessly between 2D layouts and detailed 3D environments. Similarly, tools like Ctrl-Room [50] emphasize the importance of layered workflows, in which coarse layouts are iteratively refined through object-level manipulation and interactive adjustments.

Trade-offs between computational efficiency and spatial fidelity are central to the design of recent models. Diffusion-based frameworks such as LayoutDiffusion [33] achieve state-of-the-art performance in aligning layout attributes, but often require longer inference times, which can limit scalability in real-time applications. Meanwhile, advances in conditional synthesis, such as those introduced with LayoutGPT [8] and LayoutFormer++ [19], provide essential insights into reducing user burden by offering flexible input modalities, from text instructions to pre-defined spatial constraints.

In summary, layout-guided image synthesis has become an indispensable tool for urban and architectural design, driving substantial efficiency gains, fostering creativity, and enabling better alignment with planning objectives. Future advancements will likely converge on hybrid methodologies that integrate text, image, and generative planning frameworks while addressing computational scalability, domain-specific customization, and user-centric interaction. Building on the strengths of current state-of-the-art models, such as CityGen [29], LayoutGPT [8], and LayoutGAN [1], the field is poised to deliver increasingly powerful and accessible solutions for shaping the built environment.

5.4 Scientific Visualization and Education

The field of scientific visualization and education has significantly advanced through the application of layout-guided controllable image synthesis, leveraging its spatial and semantic precision to enhance the understanding and communication of complex phenomena. These techniques enable the seamless translation of abstract data into intuitive visual representations, bridging the gap between conceptual knowledge and visual comprehension. As a result, researchers and educators are empowered to create high-quality simulations, visualizations, and educational materials tailored to specific spatial and semantic constraints, fostering clearer insights and deeper engagement.

One prominent application of layout-guided synthesis is the generation of detailed scientific illustrations and simulations that adhere to domain-specific spatial requirements. For instance, in molecular biology, the accurate depiction of protein structures or molecular pathways benefits from the spatial control provided by these methods, which ensure proper alignment and interaction among components [73]. Similarly, in astronomy, layout-guided synthesis assists in visualizing celestial formations, such as planetary systems or star clusters, elucidating spatial relationships at both macroscopic and microscopic scales [74]. These capabilities provide invaluable tools for interpreting complex systems across diverse scientific domains.

Educational visualization represents another key area of impact, wherein such techniques have proven instrumental in creating illustrations and interactive teaching materials that align closely with curriculum objectives. For example, the generation of anatomical diagrams benefits from the high spatial fidelity enabled by methods like LayoutDiffusion [33], which can produce detailed and dynamic visualizations of body systems in accordance with predefined spatial hierarchies. Dynamic capabilities like zooming into specific sections or toggling between layers further enhance interactive learning experiences. Similarly, in STEM education, layout-based synthesis facilitates the creation of compelling visual examples for complex concepts, such as quantum mechanics or fluid dynamics, where accuracy in representing intricate spatial patterns is essential [15].

In the realm of scientific simulation, layout-guided techniques have been transformative for generating synthetic datasets to support machine learning model training. In geophysics and environmental science, for example, these methods enable the creation of layouts depicting terrain, vegetation, and water bodies to train algorithms for segmentation, object detection, or disaster risk assessment. Approaches incorporating hierarchical 3D modeling, such as CC3D [31], add further granularity, simulating changes across spatial and temporal dimensions to better approximate real-world dynamics. These efforts not only enhance model performance but also expand the repertoire of scenarios that can be effectively analyzed through machine learning.

Layout-guided synthesis is also crucial for addressing data scarcity in specialized domains. In medical research, for example, techniques like Layout-to-Image frameworks [26] enable the creation of labeled synthetic datasets for rare conditions, such as tumor pathologies mapped within specific anatomical contexts. These spatially accurate syntheses empower researchers to explore high-resolution outputs while adhering to privacy and ethical constraints. Such applications are particularly impactful in domains where collecting real-world datasets is challenging, costly, or invasive.

Despite these advances, there remain challenges in advancing the generalizability and usability of layout-guided synthesis for scientific visualization. Current models often

fall short when tasked with creating multi-domain datasets that meet both physics-based realism and domain-specific requirements. While methods like LayoutDiffusion and LayoutDM [43] excel in layout-to-image fidelity, their potential for integration in multi-modal and hierarchical scientific workflows remains underexplored. Furthermore, enhancing interactivity for educational applications—such as enabling real-time modifications aligned with adaptive lesson plans—remains a critical area for improvement. Innovations like LayoutGPT [8], which extend functionalities to text-based spatial planning, suggest promising pathways for developing systems that integrate natural language input with scientific content generation.

Looking forward, emerging technologies such as 3D layout-based synthesis hold particular promise for immersive educational environments when paired with virtual and augmented reality [16]. These systems could offer learners hands-on interaction with scientific content in three-dimensional space, deepening engagement across fields such as structural engineering and planetary science. Furthermore, the integration of layout-guided synthesis with large-scale foundational models like those explored in LayoutGPT illustrates how such advancements may democratize access to scientific visualization, simplifying content creation for non-experts. Overcoming challenges related to computational scalability, the ethical generation of synthetic data, and the development of collaborative tools will be essential for realizing the full potential of layout-guided synthesis in supporting both scientific discovery and education in the years ahead.

5.5 Interactive and User-Centric Systems

Interactive and user-centric systems represent a critical frontier in layout-guided controllable image synthesis, aimed at democratizing technology by empowering non-expert users with intuitive tools for dynamic control over synthesized layouts and imagery. These systems bridge the gap between high-performance generative models and practical usability, fostering accessibility across diverse applications, from creative industries to content personalization.

Central to this paradigm is the development of interactive design tools where users can iteratively define or adjust layouts and receive real-time visual feedback. Methods like Sequential Gallery [75] epitomize this approach, incorporating Bayesian optimization with a gallery-based interface to simplify the search over complex design parameters. This not only reduces user cognitive load but also accelerates workflows by guiding users through 2D subtasks while maintaining high-dimensional design control. Similarly, layout modification systems like LayoutGPT [8] demonstrate the role of large language models (LLMs) in transforming natural language inputs into visually coherent layouts across domains such as 2D images, text-based designs, and 3D environments. These methods showcase flexibility in processing high-level user instructions while enabling granular adjustments, thus enhancing the balance between user agency and automation.

Real-time feedback systems have emerged as a key aspect of user-centric workflows. Techniques such as LayoutDiffusion [4] allow users to iteratively modify spatial configurations during the synthesis process, leveraging object-aware cross-attention mechanisms to maintain fidelity and alignment with the user-specified layouts. This dynamic refinement is crucial in collaborative applications where layouts evolve progressively based on iterative user feedback. Another noteworthy approach is Build-A-Scene [16], which enables iterative 3D layout control through dynamic self-attention modules, allowing for precise object placement and repositioning within multi-stage generation workflows. By addressing the limitations of static 2D layout controls, these systems provide a more natural interface for applications such as interior design and complex scene generation.

Among emerging trends, training-free approaches, such as LooseControl [76], significantly augment user flexibility. By introducing generalized depth conditioning, users can define loose spatial boundaries or specify object positions in 3D environments without requiring annotated depth maps. Such methods achieve significant layout fidelity while empowering users with minimal spatial constraints, paving the way for intuitive and interactive workflows. These innovations align closely with the idea of composable and modular systems, where methods like Editable Image Elements [77] allow users to manipulate spatial elements directly by encoding input images into editable components amenable to spatial adjustments, resizing, and object removal.

Despite rapid advancements, these systems must contend with trade-offs between model complexity and user accessibility. High fidelity and semantic alignment often come at the cost of additional computational resources, limiting real-time interactivity. For example, while adversarial feedback mechanisms, such as those in ALDM [48], enhance layout-image alignment, they introduce significant computational overhead during inference. Similarly, modular frameworks such as LayoutDiffuse [52] amalgamate the strengths of foundational diffusion models and task-aware prompts, balancing layout adherence with high-quality visual output. However, their reliance on extensive training data can restrict scalability.

The need for transparency and inclusion represents another challenge. User-centric systems must incorporate interpretable guidance mechanisms to democratize layout synthesis across varying levels of expertise. Approaches like LayoutGPT [8] offer strides in this direction by allowing users to express constraints in natural language forms, but achieving seamless translation of vague or conflicting user intents into coherent visual outputs remains an ongoing area of exploration.

The future of interactive systems in layout-guided synthesis lies in advancing multi-modal and hierarchical interfaces. Techniques like Hierarchical Semantic Layouts [15] facilitate precise alignment of coarse-to-fine-grained user inputs, addressing diverse use cases from free-form sketching to pixel-perfect designs. Additionally, real-time collaborative platforms leveraging lightweight architectures and decentralized inference pipelines could revolutionize industrial workflows by enabling cloud-based synthesis on mobile and edge devices.

In conclusion, interactive and user-centric systems hold transformative potential in making generative synthesis practical and accessible. As methods evolve, ensuring robust trade-offs between control, scalability, and fidelity will be

paramount in enabling these systems to cater effectively to broad and diverse user domains.

5.6 Ethical Applications and Bias Mitigation

The increasing societal impact of layout-guided controllable image synthesis necessitates a rigorous examination of its ethical applications and mechanisms for bias mitigation. As these systems find application across diverse domains, from urban planning to healthcare, ensuring fairness, inclusivity, and ethical standards becomes paramount. This discussion focuses on how layout-guided approaches can address societal challenges by fostering equitable and accountable outcomes while proposing strategies to combat biases and encourage responsible data generation.

A key ethical concern stems from the bias embedded in the datasets used to train layout-guided generative models. Many commonly used datasets, such as COCO-Stuff and Visual Genome, suffer from demographic, categorical, or contextual imbalances, which can propagate and amplify biases in synthesized outputs. For example, models trained on skewed distributions may over-represent specific demographics in certain settings (e.g., gendered roles in domestic scenes) or fail to reflect underrepresented scenarios. To address this, intentional dataset curation that promotes balanced representation is critical. Techniques like those in LayoutVAE [13], which emphasize diversity in layout synthesis, can help counterbalance these disparities during training. Additionally, data augmentation and adversarial training approaches have demonstrated promise in generating synthetic datasets to bridge these representation gaps.

Equally important is mitigating bias in multi-modal integration, where layout guidance paired with auxiliary modalities, such as text or 3D scene graphs, risks inheriting biases from pre-trained models. Hierarchical models like LayoutTransformer [3] and LayoutGPT [8] highlight advanced semantic representation capabilities for layout planning but remain prone to biases in text-to-layout alignment, inadvertently privileging certain spatial or relational patterns. Techniques like masked input guidance and alignment constraints, as demonstrated in LayoutDiffusion [4], aim to ensure equitable cross-modal alignment and reduce misrepresentations. Grounding these multi-modal synthesis technologies in ethics-aware frameworks, such as Composer [68], enhances their contextual fairness and societal sensitivity.

In accessibility-focused applications, layout-guided synthesis has transformative potential by creating content tailored to diverse user needs. For instance, simplifying layouts for visually impaired users or designing universally accessible urban plans represents a rapidly emerging but crucial area of research. Examples such as BLT [5] illustrate how layout designs can incorporate constraints for readability, simplicity, and inclusivity, fostering universal design principles. By enabling diverse audiences to engage with synthesized outputs, such methods align closely with ethical imperatives of inclusivity and accessibility.

Transparency and accountability further compound the ethical considerations surrounding layout-guided systems, particularly in high-stakes domains like healthcare and urban planning. Ensuring traceability and compliance in generated outputs prevents misuse or untraceable modifications. Approaches such as differentiable rendering and interpretable scene refinement, as seen in LayoutDiffusion [33] and Constrained Graphic Layout Generation [78], integrate mechanisms for control and provenance tracking into the synthesis process. These efforts ensure that models remain faithful to user-defined constraints while safeguarding against unethical manipulations. Similarly, adversarial feedback techniques utilized in ALDM [48] create validation mechanisms that dynamically confirm the alignment between input specifications and generated results, further enhancing accountability.

Despite these advancements, balancing the trade-off between fairness and utility remains an ongoing challenge. Efforts to enrich layout representation diversity may inadvertently raise computational demands or reduce domain-specific synthesis fidelity. Techniques like End-to-End Optimization of Scene Layout [27], which employ optimization-driven refinements, showcase potential solutions by mediating between diversity, efficiency, and context-specific requirements. Additionally, developing domain-agnostic debiasing strategies, as explored in LayoutVAE [13] and LayoutLLM-T2I [79], signals progress toward scalable, fair generative pipelines across diverse applications.

Emerging trends emphasize the importance of establishing ethical frameworks that extend beyond technical bias mitigation. Future research should prioritize participatory design practices, incorporating input from diverse stakeholders to refine layout-semantics alignment in ways that reflect societal values. Furthermore, as foundational models for layout-guided synthesis advance, explicit integration of ethical guidelines into their pretraining and fine-tuning objectives is essential, ensuring fairness and accountability at every stage. Through collaborative innovation, layout-guided synthesis has the potential not only to mitigate biases but also to exemplify responsible and ethical AI deployment on a broader scale.

6 Challenges, Open Problems, and Future Directions

6.1 Scalability and Real-Time Adaptability

Scalability and real-time adaptability in layout-guided controllable image synthesis present significant challenges given the computational demands of high-quality generation, diverse dataset requirements, and latency constraints in real-world applications. The ability to extend existing architectures to handle large-scale, complex layouts while supporting real-time interaction is pivotal for advancing this domain.

One critical bottleneck in scaling generative models lies in their computational inefficiency when tasked with maintaining high visual fidelity across increasingly complex layouts featuring numerous objects with intricate spatial relationships. Models such as LayoutDiffusion [33] and LayoutDM [43] have demonstrated advancements in conditional layout generation by leveraging diffusion-based frameworks capable of fine-grained control over spatial details. However, such approaches exhibit high computational

costs during inference due to iterative denoising or transformation processes, as evidenced in LayoutDM's transformer-based architecture, which, while effective for fidelity, suffers from prolonged generation times for high-resolution layouts. The cost of scaling demands solutions that either optimize existing architectures or reimagine model design to prioritize computational efficiency without compromising quality.

Another prominent scalability challenge stems from the necessity of accommodating growing datasets to improve generative models' generalization. Many state-of-the-art systems rely on carefully curated datasets, such as COCO-Stuff or Visual Genome, to ensure semantic precision and realistic spatial arrangements. However, both the storage and training time requirements grow exponentially with dataset expansion. Solutions such as large-scale, self-supervised pretraining introduced in LayoutGPT [8] highlight an emerging trend towards leveraging generalized pretrained transformers, followed by fine-tuning for specific layout constraints. This approach reduces the need for fully annotated datasets while maintaining robust alignment to input layouts. Still, fine-tuning remains resource-intensive, particularly for transfer to novel domains, necessitating further exploration into lightweight, domain-adaptive frameworks.

Real-time adaptability introduces additional complexity as it requires rapid inference to support interactive systems, such as those employed in graphic design or urban planning. Interventions like BLT's bidirectional transformers [5], which learn highly efficient non-autoregressive generation processes, seem promising for reducing overall latency. Similarly, Zero-Painter [80] proposes training-free approaches with localized cross-attention mechanisms to bypass traditional compute-heavy pipelines. Despite these innovations, achieving sub-second response times for high-resolution, multi-object scenes remains a challenge, particularly when incorporating complex multi-modal inputs, as demonstrated by techniques involving spatial-depth integration in Build-A-Scene [16]. The trade-off between control granularity, as seen in these methods, and speed underscores an unresolved tension in model optimization.

Memory-efficiency is another key consideration when scaling models to handle dense or hierarchical layouts, particularly in applications like 3D synthesis or multi-domain generation. Methods such as UniControl [10] exemplify promising unified architectures that consolidate multiple layout conditions while maintaining efficiency. However, even these systems must carefully manage memory trade-offs during fine-tuned diffusion processes to avoid degradation in task-specific control. Hybrid approaches that combine memory-efficient latent diffusion with task-aware adaptations, as seen in PLay [7], have shown improvement in meeting scalability requirements, though further research is required to ensure these optimizations translate effectively across larger datasets.

A potential pathway forward involves exploring modular architectures and distributed computation strategies. By decomposing synthesis pipelines into modular subprocesses, such as layout encoding, object-specific refinement, and final image synthesis, models can achieve greater task specialization, reducing redundant computation across tasks [3]. Furthermore, the integration of techniques such as task-specific adapting modules, as proposed by Composer [68], could better harness domain knowledge while distributing compute-intensive operations across cloud or edge computing frameworks.

Future advances will likely emerge from innovations in foundational model design, such as sparse attention mechanisms to reduce computational overhead for high-resolution layouts and hierarchical planning schemes that enable multi-scale spatial reasoning. The evolution of real-time optimizations, including forward-looking approximate methods, as mentioned in LayoutDiffusion [33], coupled with resource-aware techniques such as dynamic cache utilization during latent optimization, will be critical in democratizing layout-guided synthesis for interactive, real-world applications. Addressing scalability holistically—through algorithmic efficiency, hardware-aware implementation, and principled trade-offs between quality and speed—remains an open frontier in this rapidly evolving field.

6.2 Generalization and Robustness Across Domains

Achieving generalization and robustness across diverse domains and input layouts remains one of the most formidable challenges in layout-guided controllable image synthesis. The inherent variability in domain-specific data distributions, layout configurations, and semantic representations often leads to performance degradation when models are exposed to novel datasets or layouts not encountered during training. This subsection critically examines methods and challenges associated with enhancing domain generalization and layout variability, while providing insights into potential research directions to overcome these limitations.

A fundamental challenge lies in the tendency of generative models to overfit to specific training distributions, limiting their flexibility and adaptability to unseen layouts or content variations. Many layout-guided synthesis models are overly reliant on domain-specific priors, which hinders performance in broader contexts or when encountering novel spatial arrangements. For example, models such as Layout2Im [2] demonstrate strong results on datasets like COCO-Stuff and Visual Genome but often falter in maintaining fidelity when tackling complex multi-object scenes outside these curated domains. Similarly, hierarchical generation pipelines, such as those proposed in [30], perform well under dataset-specific hierarchies but encounter difficulties in understanding diverse domains or resolving structural ambiguities in novel configurations.

Domain transfer and adaptation techniques offer a potential pathway to expand generalization capabilities. Semi-parametric approaches that combine learned priors with external memory banks, such as [20], showcase improved robustness by drawing from memory-based references of training data to interpolate novel configurations. Yet, these methods are heavily dependent on the diversity and quality of the external memory bank, raising questions about their scalability in domains with limited or minimally annotated data. Diffusion-based methods, including LayoutDiffusion [4], have introduced iterative refinement frameworks to facilitate smoother transitions across layouts, but these approaches often contend with trade-offs between preserving spatial alignment and maintaining visual fi-

delity—especially in domains involving abstract spatial configurations.

The challenge exacerbates when addressing extreme layout variability, including layouts with novel aspect ratios, sparse or densely packed components, or incomplete inputs. Scene graph-based methods, such as [17] and [18], provide graph-encoded priors to capture relational dependencies, enabling these models to operate effectively in scenarios where spatial relationships are critical for semantic consistency. However, the computational overhead and potential for noisy graph encodings limit scalability as layouts grow more complex with dense inter-object interactions. Probabilistic generative techniques, such as LayoutVAE [13], attempt to address variability through stochastic sampling of plausible spatial arrangements, yet this approach lacks the deterministic control required for domains like medical imaging or urban design where spatial and semantic precision are paramount.

Mitigating the effects of domain shifts—where there are discrepancies between training and testing data distributions—is another focal point. Transfer learning approaches and pretraining on large foundational models provide promising results, as demonstrated by LayoutGPT [8], which uses large-scale language model priors to achieve zero-shot synthesis across diverse domains. Furthermore, advances in multimodal conditioning techniques, like those in [60], enrich semantic guidance by integrating complementary contextual information. However, effective alignment between spatial and semantic subtleties across diverse domains remains a difficult task, necessitating innovations in cross-modality fusion strategies to ensure cohesive synthesis outcomes.

Looking ahead, research must strive to balance specificity and generality within layout-guided synthesis pipelines. One promising direction involves the development of self-supervised learning techniques to capture universal spatial priors without the need for exhaustive annotations, drawing inspiration from advancements in unsupervised 3D understanding [12]. Another avenue lies in hybrid approaches, such as combining GAN and diffusion-based architectures, as explored in LayoutDiffusion [4], to leverage the strengths of multiple paradigms, enhancing fidelity and control while achieving domain robustness. Additionally, continuous adaptation mechanisms during inference could address distributional shifts and ensure adaptable performance in real-time use cases.

In summary, achieving strong generalization and robustness across diverse domains and layouts requires an intricate interplay of architectural innovation, adaptive frameworks, and multimodal integrations. Addressing these challenges effectively will enable layout-guided image synthesis systems to extend their usability beyond current benchmarks, unlocking transformative applications across fields such as digital content creation, urban planning, and scientific visualization. Furthermore, ensuring seamless integration with scalable, efficient architectures—aligned with the demands of real-time and multimodal contexts—positions this field for substantial advancements in usability and impact across an array of industries.

6.3 Enhancing Multi-Modal and Hierarchical Control

The integration of multi-modal inputs such as text, sound, and depth maps, alongside the adoption of hierarchical layout specifications, represents an exciting frontier for advancing control in layout-guided image synthesis. Multi-modal and hierarchical paradigms are particularly appealing as they address not only the need for spatial precision and contextual alignment but also the complexity and diversity of real-world synthesis tasks, ranging from semantic-rich content generation to spatially immersive 3D scene creation.

Multi-modal synthesis leverages complementary inputs to enhance contextual understanding within generative models. For instance, combining textual prompts with layouts has shown substantial promise in improving semantic consistency and control. Large language models (LLMs) integrated with layout generation pipelines, such as LayoutGPT [8], demonstrate the potential for converting textual descriptions into flexible layouts encompassing complex relationships. This approach eases the cognitive load on users while preserving rich semantic alignment. Moreover, hybrid methods like LayoutLLM-T2I [79] further interweave textual prompts with layouts to enhance spatial reasoning capabilities, addressing misalignment challenges often associated with textual-to-visual mappings. However, a notable limitation of multi-modal approaches stems from the mismatched granularity of inputs. For example, while textual inputs are inherently abstract, layouts often demand spatial and quantitative precision. This disparity underscores the importance of fine-tuning fusion mechanisms such as cross-attention layers with layout constraints [54]. Further advancements, like integrating spatially aware latent initialization techniques, as proposed in [81], mitigate challenges by leveraging reference images to encode spatial relationships robustly.

On the other hand, hierarchical control aims to reconcile coarse-to-fine layout understanding, enabling holistic planning at multiple granularity levels. Hierarchical synthesis frameworks decompose layouts into spatial hierarchies, cascading from broad scene arrangements to finer object-level details [30]. This paradigm aligns well with natural multiscale synthesis processes, exemplified by SceneComposer [15], which supports workflows accommodating varying precision levels, from text-driven generation to segmentation-driven image control. Similarly, layout refinement models such as [25] adopt iterative, dependency-aware approaches for generating layouts that align inter-object relationships incrementally. Hierarchical methods demonstrate specific advantages in handling complex, layered scenes, particularly in applications like urban layouts or indoor planning. Yet, these approaches often face scalability challenges, particularly when translating hierarchical refinements into computationally efficient pipelines.

An emerging trend in multi-modal synthesis is the use of auxiliary depth, audio, or temporal modalities to bridge the gap between 2D and 3D spatial reasoning. Depth-aware models, such as Build-A-Scene [16], leverage depth-conditioned diffusion to augment spatial organization with a third-dimensional perspective, allowing intricate control over object location in 3D layouts. Meanwhile, frameworks integrating sound inputs, such as sound-conditioned lay-

out modeling, remain largely unexplored but hold promise in synchronizing auditory-driven spatial cues with visual arrangements. Bridging temporal information, for example, using time-dependent layouts, could also improve performance in dynamic scenes like storytelling, where object persistence and motion trajectories are critical [82].

Despite the progress, several challenges persist. Cross-modal fusion mechanisms often lack robustness, where noisy or incomplete modalities degrade synthesis quality. Hierarchical models may struggle with achieving global coherence across granular refinements, as early-stage decisions in coarse layouts can propagate errors downstream. Methods like compositional priors or incremental refinement modules could help alleviate these issues [14]. Moreover, as models become increasingly complex, computational efficiency remains an obstacle, particularly for real-time or interactive applications. Optimizing hierarchical methods for low latency—such as by employing joint training of coarse and fine submodules or introducing memory-efficient architectures like transformers adapted with layout fusion [19]—offers a pathway forward.

Future directions must also consider the democratization of multi-modal synthesis tools. Collaborative frameworks enabling non-expert users to seamlessly integrate multi-modal inputs with hierarchical control, as suggested in [83], promise significant advancements in accessibility and creative workflows. Additionally, the push toward 3D and spatially immersive applications introduces the need for encoding multi-modal, volumetric, and temporal layouts that extend the boundaries of traditional 2D paradigms.

In conclusion, the convergence of multi-modal and hierarchical control in layout-guided synthesis holds immense potential for enriching the realism, versatility, and usability of generative systems. Future advancements require innovations in multi-modal fusion mechanisms, computational optimizations for hierarchical layouts, and inclusivity-driving frameworks to broaden the range of real-world applications.

6.4 Bias, Fairness, and Ethical Considerations

Bias, fairness, and ethical considerations play a pivotal role in the development of layout-guided controllable image synthesis systems, given their profound societal impact and the potential risks of misuse. Bias often originates from the dependencies on training data or the structural design of generative frameworks, both of which reflect inherent inequalities or limitations. Training datasets, such as COCO-Stuff and Visual Genome, frequently exhibit skewed distributions in spatial layouts, object co-occurrences, or demographic representations, perpetuating stereotypes and exacerbating underrepresentation of certain groups [17]. Similarly, generative frameworks themselves can amplify these biases, introducing additional ethical concerns. This subsection delves into these sources of bias, evaluates current mitigation strategies, and highlights future directions for fostering inclusivity and fairness in layout-guided synthesis systems.

A key source of bias stems from the diversity and balance of the training datasets. Data distributions that fail to adequately represent multicultural, geographic, or demographic contexts constrain the generalizability of generated outputs, often replicating stereotypical spatial relationships or scene configurations. For instance, object co-occurrence patterns can skew models toward replicating normative pairings—such as associating particular activities with specific genders or spaces—which risks entrenching existing stereotypes. Addressing such challenges requires rigorous data curation processes. Frameworks like LayoutGPT have demonstrated initial success in incorporating broader corpora to diversify layout generation, mitigating some data biases during synthesis [8]. Despite these advances, challenges persist in devising unbiased data augmentation strategies that sufficiently balance representation while preserving layout precision and semantic fidelity.

Bias is not only confined to datasets but also systematically embedded in generative architectures. Techniques such as those based on Generative Adversarial Networks (GANs) are particularly susceptible to mode collapse, a phenomenon where outputs disproportionately reflect high-density areas of the input distribution, thereby reducing emphasis on rare or unconventional features [44]. Diffusion-based methods like LayoutDiffusion provide greater robustness through iterative refinement pipelines, enabling better exploration of latent spaces and ensuring broader coverage across configurations [4]. Moreover, fairness-aware architectural adjustments, such as loss functions designed to redistribute emphasis across underrepresented spatial or object relationships, present promising paths forward. Constraint-based systems that regulate spatial prominence or balance object importance [78] offer models the capability to synthesize layouts that more equitably represent diverse scene possibilities.

In multi-modal applications, such as text-to-layout or text-to-image synthesis, the integration of language-driven prompts introduces an added dimension of complexity. Biases ingrained in text-based embeddings, often extracted from pretrained language models, can propagate through the synthesis pipeline, manifesting as prejudiced visual outputs. Cross-modal approaches, while powerful in bridging linguistic and spatial reasoning—for example, using CLIP-guided visual-text alignment—may inadvertently reflect the biases of the source language models [79]. Techniques aimed at decoupling spatial semantics from linguistic semantics or attention-driven mechanisms to amplify underweighted contextual details have proven effective in mitigating these risks [28].

Ethical implications expand beyond data and model biases, addressing broader concerns like consent, privacy, and misuse. In high-stakes domains such as healthcare and urban design, improperly controlled generative models risk reconstructing sensitive or identifiable content, which may inadvertently infringe on individual privacy [22]. Additionally, these systems harbor misuse potential, such as propagating manipulated visuals or reinforcing harmful societal narratives. The development of transparent workflows, explainable model behavior, and accountable outputs is vital for minimizing these risks. Provenance tracking systems for synthesized content offer an essential safeguard, enabling validation and verification of visual outputs and fostering trust in their applications.

The integration of fairness-aware evaluation pipelines plays a crucial role in addressing these issues. Metrics like

SceneFID, adapted to weight diversity-specific alignment in complex scenes, represent meaningful innovations in aligning synthesis outputs with principles of equity [18]. User-centric evaluation paradigms, which incorporate participatory and iterative refinement feedback into multi-modal workflows, open pathways for identifying latent biases that automated systems might overlook [19]. These strategies make it possible to align generative technologies with the nuanced requirements of diverse end-user groups.

Future work in ensuring fairness and inclusivity within layout-guided synthesis systems must adopt a deeply interdisciplinary perspective. This includes designing pretraining regimes based on representative foundational models, implementing domain-agnostic bias-reduction techniques, and integrating participatory frameworks that leverage input from underrepresented communities. By embedding fairness measures and ethical considerations throughout a system's lifecycle—from training and architecture design to evaluation and deployment—the field can establish a foundation for socially responsible applications. Coupled with the adoption of regulatory and accountability frameworks, these strides will pave the way toward inclusive and transformative generative technologies.

6.5 Evaluation Paradigms and Benchmarking

Addressing the challenges of evaluating layout-guided controllable image synthesis, this subsection identifies key gaps in existing evaluation frameworks and proposes directions for future standardization. The field has made considerable progress with task-specific and dataset-dependent metrics, but these approaches often fail to generalize across the diverse applications of layout-guided synthesis. Critical issues include inconsistencies in how frameworks measure spatial fidelity, realism, and semantic alignment, as well as a lack of user-centric benchmarks.

Spatial fidelity metrics are fundamental in assessing the adherence of synthesized images to their prescribed layouts. Intersection over Union (IoU) remains a standard quantitative measure for evaluating the overlap between predicted object regions and input layouts [2]. However, IoU and its derivatives are insufficient for capturing nuanced spatial relationships, such as hierarchical arrangements or complex 3D layouts. Scene Graph Similarity [18] and position-sensitive loss functions [38] offer more context-aware evaluations but often exhibit scalability limitations in multi-object or densely populated scenes. Emerging strategies, like Layout Fusion Modules for contextual spatial processing [4], suggest promising avenues for enhancing spatial validation metrics.

Visual quality metrics such as Fréchet Inception Distance (FID) and Structural Similarity Index Measure (SSIM) are commonly used to evaluate realism [17], [26]. While FID quantifies the perceptual alignment between the distribution of generated and real images, it does not account for layout-specific constraints. Recent advancements propose domain-specific adaptations like SceneFID [18], which integrates object-wise evaluation into the overall image realism assessment. Such contextual refinements are necessary for comparing the perceptual quality of densely compositional scenes or domain-specific outputs like medical or architectural imagery [54].

Semantic alignment metrics focus on the coherence between input layouts and generated outputs, particularly in complex relational scenes. Semantic Object Accuracy (SOA) [21] measures categorical and attribute-level correctness, but it does not consider contextual relationships between objects. Scene Graph-based consistency checks have shown promise in evaluating interaction fidelity between objects, especially using metrics like graph alignment and hierarchy preservation [17]. The incorporation of pretrained multi-modal models like CLIP for semantic evaluation [79] introduces a scalable framework for assessing textual and visual alignment, but its applicability for highly specific layouts, such as those requiring 3D spatial semantics [14], remains unexplored.

Another significant gap is the inadequacy of existing datasets for benchmarking. Popular datasets like COCO-Stuff and Visual Genome [2] provide robust annotations for single-modality scenarios but lack the multimodal and hierarchical complexity required for advanced evaluation. Benchmarks such as T2I-CompBench [63] and custom synthetic datasets designed for controlled experiments [84] play a pivotal role in advancing evaluation protocols. Still, the need for standardized, high-quality datasets featuring diverse layouts, domain coverage, and multi-task scenarios persists.

User-centric evaluation paradigms offer another avenue for standardization. While human studies are often employed to assess perceptual quality and usability, these assessments are resource-intensive and inherently subjective. The introduction of interactive metrics that capture user preferences in real time—such as those evaluating iterative refinement in collaborative systems [45]—marks a shift toward more dynamic evaluation frameworks. Incorporating user-defined constraints and real-time feedback loops into benchmarking protocols will significantly enhance their applicability to practical, real-world scenarios.

Future directions must address these limitations through the development of unified benchmarks and multi-modal evaluation pipelines that integrate spatial, semantic, and user-centric criteria. Metrics should evolve to assess alignment across hierarchical layouts, multimodal inputs, and 3D contexts [35]. Furthermore, adaptive standards leveraging advances in explainability and interpretability [16] could improve reproducibility and foster cross-domain applicability. By aligning evaluation frameworks with the complex demands of layout-guided controllable synthesis, the field can ensure that advancements are rigorously validated and practically relevant across diverse applications.

6.6 Future Trends and Emerging Directions

As the field of layout-guided controllable image synthesis progresses, emerging research directions and trends are redefining both foundational methodologies and applications, building upon the challenges and evaluation advancements discussed earlier. These trends seek to address current limitations while introducing novel paradigms that expand the scope of layout-guided synthesis toward more robust, multi-modal, hierarchical, and resource-efficient systems, thus paving the way for practical, real-world deployments.

One of the most groundbreaking trends involves the integration of 3D spatial reasoning into generative pipelines,

which extends beyond conventional 2D layouts to enable a richer, volumetric understanding of scenes. While current methods are predominantly limited to planar layouts, expanding to 3D layouts is vital for applications such as virtual reality, autonomous systems, and advanced physical simulations. Systems like CC3D [31] and RoomDreamer [35] showcase progress in synthesizing 3D scenes that incorporate geometric and structural awareness via scene graphs, latent shape modeling, and depth priors. However, significant challenges remain in achieving consistent high-resolution outputs while maintaining computational efficiency in large, densely populated scenes. Techniques like Neural Mixtures of Planar Experts (NeurMiPs) [85] highlight promising pathways for bridging explicit scene geometry with implicit appearance models, addressing the challenges of high-dimensional data representation and reducing computational overhead.

Another significant direction involves the adoption of multi-modal and hierarchical generative approaches that align layout synthesis with auxiliary inputs, such as textual descriptions, sketches, audio, or temporal constraints [56], [86]. Multi-modal frameworks like SceneComposer [15] and LayoutGPT [8] exemplify the flexibility to combine diverse input modalities. To unlock their full potential, integrating multi-modal inputs in hierarchical settings that fuse global spatial reasoning with fine-grained local synthesis remains an active challenge. Recent innovations, such as LayoutDiffusion [4] and hierarchical diffusion techniques like MovieDreamer [87], demonstrate models capable of coarse-to-fine spatial guidance but highlight limitations when tackling highly complex, multi-layered scenarios.

Advancements in resource-efficient and decentralized systems form another transformative frontier. As synthesis models grow in complexity, addressing constraints imposed by real-time systems, edge computing, and low-resource environments gains paramount importance. Methods like cross-attention-based modular augmentation [54] and lightweight latent diffusion architectures [88] suggest promising strategies for reducing computational cost without compromising quality or controllability. Moreover, innovative approaches in parameter efficiency and compression, such as those demonstrated in Cones 2 [89], emphasize the increasing priority of sustainability and deployment feasibility in layout synthesis.

Simultaneously, the community must actively contend with challenges related to bias and fairness inherent to layout-guided synthesis. Models often inherit biases from training datasets, which, if unchecked, can propagate into generated outputs. Initial steps to mitigate these biases include attribute-conditioned modeling [38] and domain-agnostic frameworks [45]. Future research must focus on embedding fairness constraints during training and ensuring inclusive, representative outputs, particularly in multi-modal and cross-cultural applications.

Finally, explainability and interpretability have gained heightened importance in fostering trust, especially for sensitive or ethically significant applications like healthcare and urban planning. Techniques emphasizing disentangled representations, such as hierarchical disentanglement [73], offer pathways for increasing transparency and user understanding of generative decisions. Additionally, adversarial strategies like multi-step discriminator unrolling [48] integrate semantic alignment into the training process while exposing model decision-making to scrutiny, fostering trustworthiness in deployment.

Looking ahead, these research directions call for holistic exploration, synthesizing advancements across disciplines. Leveraging large language models, diffusion techniques, and geometrically informed methods could enable adaptive, ethically sound solutions that redefine layout-guided synthesis' capabilities. Through these converging innovations, the field stands poised to revolutionize domains from multimedia design to immersive AI-driven environments, aligning with the multifaceted demands and opportunities outlined in prior evaluative frameworks.

7 Conclusion

This survey has provided an extensive examination of the field of layout-guided controllable image synthesis, a domain that has progressed significantly due to advancements in deep generative models. By synthesizing insights from diverse approaches and frameworks, this subsection highlights the notable contributions, limitations, and future opportunities that collectively define the trajectory of this field.

At its core, layout-guided controllable image synthesis addresses the challenge of generating visually realistic and semantically aligned images while adhering to spatial and compositional constraints. The survey emphasized the fundamental role of different layout representations—bounding boxes, segmentation masks, keypoints, scene graphs, and hierarchical forms—in structuring input guidance. The trade-offs between these options were rigorously discussed, analyzing their computational efficiency, semantic granularity, and suitability across domains such as document layout generation [1], medical imaging enhancement [6], and urban planning [90]. Furthermore, emerging trends, such as multi-modal integration and 3D layout synthesis, demonstrate a shift towards more contextually rich and geometrically nuanced paradigms [12], [16].

Comparing generative models, Generative Adversarial Networks (GANs), diffusion models, and transformer-based architectures each bring unique strengths and limitations to layout-controlled synthesis. Early successes with GANs were heralded for high-resolution outputs but faced challenges in maintaining spatial consistency and aligning to complex layouts [1], [2]. Diffusion models, such as LayoutDiffusion and LayoutDM, represent a transformative advancement by incorporating iterative denoising processes that ensure better spatial fidelity, context coherence, and sample diversity [33], [45]. Additionally, transformer-based models, notably LayoutFormer++ and LayoutTransformer, excel in capturing long-range dependencies and generating semantically faithful compositions through their attention-based mechanisms [3], [19]. Integrating these diverse architectures into unified approaches, as seen in UniControl for multi-condition synthesis, underscores the growing demand for versatility and robustness in generative pipelines [91].

Despite these advancements, significant challenges persist. One notable issue is the generalization of models across diverse and complex layout types. Most current frameworks

achieve strong performance on domain-specific datasets, such as COCO-Stuff or Visual Genome, but struggle with unseen configurations or domains requiring multi-modal reasoning [40], [45]. Furthermore, balancing interdependent objectives—spatial alignment, semantic consistency, and computational efficiency—remains a critical bottleneck. For instance, tools like ControlNet have demonstrated the ability to tightly adhere to localized descriptions during synthesis, yet they often encounter trade-offs between strict localization and visual quality [47]. Addressing these competing demands will require further innovation in architectural design and optimization techniques.

An equally critical dimension of future research lies in ethical and societal considerations. As these models are adopted in real-world applications, biases embedded in layout datasets—such as unequal demographic representation in generative tasks—pose significant risks. Responsible frameworks, such as CoLay, which offer conditional generation while considering inclusive design principles, set a precedent for addressing fairness and transparency [92]. Moreover, privacy-preserving approaches remain pivotal, especially in sensitive domains like healthcare, where synthetic medical imagery can both enhance diagnostic accuracy and mitigate risks of data leakage [6].

Reflecting on the contributions of this survey, it is evident that layout-guided synthesis transcends traditional generative modeling by operationalizing spatial reasoning and user-centric control. The integration of foundational advancements, such as pre-trained large language models for layout planning [45], and innovations in decoding strategies for precise alignment [81], signal a maturity in the field. As future benchmarks evolve to capture real-world complexities, emerging methodologies—spanning 3D layouts [16], [31], multi-modal synthesis [66], and interactive generation tools [15]—promise to expand the frontiers of this domain.

In conclusion, layout-guided controllable image synthesis represents a nexus of technical sophistication and practical applicability. While remarkable progress has been achieved, further efforts towards enhancing scalability, generalization, multi-modal integration, and ethical compliance will be critical in shaping the next generation of generative models. By addressing these challenges, the field has the potential to redefine human-machine collaboration in creative and functional domains, bridging theoretical innovation with transformative applications.

REFERENCES

[1] J. Li, J. Yang, A. Hertzmann, J. Zhang, and T. Xu, "Layoutgan: Generating graphic layouts with wireframe discriminators," *ArXiv*, vol. abs/1901.06767, 2019. 1, 3, 4, 6, 9, 10, 11, 13, 16, 17, 18, 25

[2] B. Zhao, L. Meng, W. Yin, and L. Sigal, "Image generation from layout," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8576–8585, 2018. 1, 2, 4, 5, 6, 9, 10, 11, 12, 14, 15, 16, 21, 24, 25

[3] K. Gupta, J. Lazarow, A. Achille, L. S. Davis, V. Mahadevan, and A. Shrivastava, "Layouttransformer: Layout generation and completion with self-attention," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 984–994, 2020. 1, 3, 4, 5, 8, 9, 10, 17, 20, 21, 25

[4] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "Layoutdiffusion: Controllable diffusion model for layout-to-image generation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22 490–22 499, 2023. 1, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, 21, 22, 23, 24, 25

[5] X. Kong, L. Jiang, H. Chang, H. Zhang, Y. Hao, H. Gong, and I. Essa, "Blt: Bidirectional layout transformer for controllable layout generation," *ArXiv*, vol. abs/2112.05112, 2021. 1, 16, 20, 21

[6] W. Chen, P. Wang, H. Ren, L. Sun, Q. Li, Y. Yuan, and X. Li, "Medical image synthesis via fine-grained image-text alignment and anatomy-pathology prompting," *ArXiv*, vol. abs/2403.06835, 2024. 1, 2, 11, 25, 26

[7] C.-Y. Cheng, R. Gao, F. Huang, and Y. Li, "Colay: Controllable layout generation through multi-conditional latent diffusion," *ArXiv*, vol. abs/2405.13045, 2024. 1, 21

[8] W. Feng, W. Zhu, T.-J. Fu, V. Jampani, A. R. Akula, X. He, S. Basu, X. Wang, and W. Y. Wang, "Layoutgpt: Compositional visual planning and generation with large language models," *ArXiv*, vol. abs/2305.15393, 2023. 2, 3, 4, 5, 8, 9, 10, 11, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 25

[9] J. Chen, R. Zhang, Y. Zhou, and C. Chen, "Towards aligned layout generation via diffusion model with aesthetic constraints," *ArXiv*, vol. abs/2402.04754, 2024. 2

[10] C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese, S. Ermon, Y. Fu, and R. Xu, "Unicontrol: A unified diffusion model for controllable visual generation in the wild," *ArXiv*, vol. abs/2305.11147, 2023. 2, 16, 21

[11] W. Sun and T. Wu, "Learning layout and style reconfigurable gans for controllable image synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 5070–5087, 2020. 2, 5, 6, 7

[12] Y. Liao, K. Schwarz, L. Mescheder, and A. Geiger, "Towards unsupervised learning of generative models for 3d controllable image synthesis," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5870–5879, 2019. 2, 17, 22, 25

[13] A. A. Jyothi, T. Durand, J. He, L. Sigal, and G. Mori, "Layoutvae: Stochastic scene layout generation from a label set," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9894–9903, 2019. 2, 3, 11, 18, 20, 22

[14] G. Zhai, E. P. Örnek, S. cheng Wu, Y. Di, F. Tombari, N. Navab, and B. Busam, "Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion," 2023. 2, 23, 24

[15] Y. Zeng, Z. Lin, J. Zhang, Q. Liu, J. Collomosse, J. Kuen, and V. M. Patel, "Scenecomposer: Any-level semantic image synthesis," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22 468–22 478, 2022. 2, 5, 7, 10, 12, 13, 16, 18, 19, 22, 25, 26

[16] A. Eldesokey and P. Wonka, "Build-a-scene: Interactive 3d layout control for diffusion-based image generation," *ArXiv*, vol. abs/2408.14819, 2024. 2, 7, 10, 11, 19, 21, 22, 24, 25, 26

[17] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1219–1228, 2018. 3, 4, 5, 12, 14, 17, 22, 23, 24

[18] T. Sylvain, P. Zhang, Y. Bengio, R. D. Hjelm, and S. Sharma, "Object-centric image generation from layouts," in *AAAI Conference on Artificial Intelligence*, 2020, pp. 2647–2655. 3, 9, 12, 14, 22, 24

[19] Z. Jiang, J. Guo, S. Sun, H. Deng, Z. Wu, V. Mijović, Z. Yang, J.-G. Lou, and D. Zhang, "Layoutformer++: Conditional graphic layout generation via constraint serialization and decoding space restriction," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 403–18 412, 2022. 3, 5, 9, 11, 14, 18, 23, 24, 25

[20] X. Qi, Q. Chen, J. Jia, and V. Koltun, "Semi-parametric image synthesis," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8808–8816, 2018. 3, 10, 17, 21

[21] O. Ashual and L. Wolf, "Specifying object attributes and relations in interactive scene generation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4560–4568, 2019. 4, 24

[22] D. M. Arroyo, J. Postels, and F. Tombari, "Variational transformer networks for layout generation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 637–13 647, 2021. 4, 8, 9, 14, 15, 23

[23] N. Inoue, K. Kikuchi, E. Simo-Serra, M. Otani, and K. Yamaguchi, "Layoutdm: Discrete diffusion model for controllable layout generation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 167–10 176, 2023. 4, 6, 13

[24] W. Sun and T. Wu, "Image synthesis from reconfigurable layout and style," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 530–10 539, 2019. 4

[25] M. Ivgi, Y. Benny, A. Ben-David, J. Berant, and L. Wolf, "Scene graph to image generation with contextualized object layout refinement," *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2428–2432, 2021. 4, 12, 22

[26] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2332–2341, 2019. 4, 12, 17, 18, 24

[27] A. Luo, Z. Zhang, J. Wu, and J. Tenenbaum, "End-to-end optimization of scene layout," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3753–3762, 2020. 4, 6, 11, 15, 20

[28] Z. Tan, D. Chen, Q. Chu, M. Chai, J. Liao, M. He, L. Yuan, G. Hua, and N. Yu, "Efficient semantic image synthesis via class-adaptive normalization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 4852–4866, 2020. 5, 23

[29] J. Deng, W. Chai, J. Guo, Q. Huang, W. Hu, J.-N. Hwang, and G. Wang, "Citygen: Infinite and controllable 3d city layout generation," *ArXiv*, vol. abs/2312.01508, 2023. 5, 13, 17, 18

[30] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7986–7994, 2018. 5, 16, 17, 21, 22

[31] S. Bahmani, J. J. Park, D. Paschalidou, X. Yan, G. Wetzstein, L. Guibas, and A. Tagliasacchi, "Cc3d: Layout-conditioned generation of compositional 3d scenes," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7137–7147, 2023. 5, 12, 15, 16, 18, 25, 26

[32] J. Xie, Y. Li, Y. Huang, H. Liu, W. Zhang, Y. Zheng, and M. Z. Shou, "Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7418–7427, 2023. 5, 14

[33] J. Zhang, J. Guo, S. Sun, J.-G. Lou, and D. Zhang, "Layoutdiffusion: Improving graphic layout generation by discrete diffusion probabilistic models," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7192–7202, 2023. 5, 9, 13, 17, 18, 20, 21, 25

[34] X. Zhou, X. Ran, Y. Xiong, J. He, Z. Lin, Y. Wang, D. Sun, and M.-H. Yang, "Gala3d: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting," *ArXiv*, vol. abs/2402.07207, 2024. 5, 10, 18

[35] L. Song, L. Cao, H. Xu, K. Kang, F. Tang, J. Yuan, and Y. Zhao, "Roomdreamer: Text-driven 3d indoor scene synthesis with coherent geometry and texture," *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 5, 24, 25

[36] J. Liu, T. Huang, and C. Xu, "Training-free composite scene generation for layout-to-image synthesis," *ArXiv*, vol. abs/2407.13609, 2024. 6, 10

[37] Z. Lv, Y. Wei, W. Zuo, and K.-Y. K. Wong, "Place: Adaptive layout-semantic fusion for semantic image synthesis," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9264–9274, 2024. 6, 13

[38] J. Li, J. Yang, J. Zhang, C. Liu, C. Wang, and T. Xu, "Attribute-conditioned layout gan for automatic graphic design," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, pp. 4039–4048, 2020. 6, 10, 15, 16, 24, 25

[39] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, "Controllable text-to-image generation," *ArXiv*, vol. abs/1909.07083, 2019. 6, 16

[40] H. Xue, Z. Huang, Q. Sun, L. Song, and W. Zhang, "Freestyle layout-to-image synthesis," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 256–14 266, 2023. 7, 10, 13, 16, 26

[41] B. Yang, Y. Luo, Z. Chen, G. Wang, X. Liang, and L. Lin, "Law-diffusion: Complex scene generation by diffusion with layouts," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22 612–22 622, 2023. 7

[42] H. Lin, S. Peng, Z. Xu, T. Xie, X. H. He, H. Bao, and X. Zhou, "High-fidelity and real-time novel view synthesis for dynamic scenes," *SIGGRAPH Asia 2023 Conference Papers*, 2023. 7

[43] S. Chai, L. Zhuang, and F. Yan, "Layoutdm: Transformer-based diffusion model for layout generation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 349–18 358, 2023. 8, 9, 10, 17, 19, 20

[44] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *ArXiv*, vol. abs/1606.03498, 2016. 8, 23

[45] T. Yang, Y. Luo, Z. Qi, Y. Wu, Y. Shan, and C. W. Chen, "Posterllava: Constructing a unified multi-modal layout generator with llm," *ArXiv*, vol. abs/2406.02884, 2024. 9, 12, 14, 15, 17, 24, 25, 26

[46] C. Yang, Y. Shen, and B. Zhou, "Semantic hierarchy emerges in deep generative representations for scene synthesis," *International Journal of Computer Vision*, vol. 129, pp. 1451 – 1466, 2019. 9

[47] D. Lukovnikov and A. Fischer, "Layout-to-image generation with localized descriptions using controlnet with cross-attention control," *ArXiv*, vol. abs/2402.13404, 2024. 9, 26

[48] Y. Li, M. Keuper, D. Zhang, and A. Khoreva, "Adversarial supervision makes layout-to-image diffusion models thrive," *ArXiv*, vol. abs/2401.08815, 2024. 9, 10, 15, 19, 20, 25

[49] Y. Cao, Y. Ma, M. Zhou, C. Liu, H. Xie, T. Ge, and Y. Jiang, "Geometry aligned variational transformer for image-conditioned layout generation," *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 9, 10, 15

[50] C. Fang, X. Hu, K. Luo, and P. Tan, "Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints," *ArXiv*, vol. abs/2310.03602, 2023. 9, 18

[51] Y. Wang, G. Pu, W. Luo, Y. Wang, P. Xiong, H. Kang, and Z. Lian, "Aesthetic text logo synthesis via content-aware layout inferring," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2426–2435, 2022. 9

[52] J. Cheng, X. Liang, X. Shi, T. He, T. Xiao, and M. Li, "Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation," *ArXiv*, vol. abs/2302.08908, 2023. 10, 19

[53] C. Ham, J. Hays, J. Lu, K. K. Singh, Z. Zhang, and T. Hinz, "Modulating pretrained diffusion models for multimodal image synthesis," *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 10

[54] M. Chen, I. Laina, and A. Vedaldi, "Training-free layout control with cross-attention guidance," *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5331–5341, 2023. 10, 15, 22, 24, 25

[55] W. Para, P. Guerrero, T. Kelly, L. Guibas, and P. Wonka, "Generative layout modeling using constraint graphs," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6670–6680, 2020. 10, 11

[56] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "Videocomposer: Compositional video synthesis with motion controllability," *ArXiv*, vol. abs/2306.02018, 2023. 10, 25

[57] M. Jahn, R. Rombach, and B. Ommer, "High-resolution complex scene synthesis with transformers," *ArXiv*, vol. abs/2105.06458, 2021. 11

[58] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1529, 2017. 12

[59] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Neural Information Processing Systems*, 2018, pp. 1152–1164. 12

[60] J. Liang, W. Pei, and F. Lu, "Layout-bridging text-to-image synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, pp. 7438–7451, 2022. 12, 22

[61] Z. Yang, D. Liu, C. Wang, J. Yang, and D. Tao, "Modeling image composition for complex scene generation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7754–7763, 2022. 12

[62] H. Weng, D. Huang, Y. Qiao, Z. Hu, C.-Y. Lin, T. Zhang, and C. L. P. Chen, "Desigen: A pipeline for controllable design template generation," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 721–12 732, 2024. 13

[63] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," *ArXiv*, vol. abs/2307.06350, 2023. 14, 24

[64] W. Para, P. Guerrero, N. Mitra, and P. Wonka, "Cofs: Controllable furniture layout synthesis," *ACM SIGGRAPH 2023 Conference Proceedings*, 2022. 14

[65] C. Rockwell, D. Fouhey, and J. Johnson, "Pixelsynth: Generating a 3d-consistent experience from a single image," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14 084–14 093, 2021. 14

[66] L. Karacan, Z. Akata, A. Erdem, and E. Erdem, "Learning to generate images of outdoor scenes from attributes and semantic layouts," *ArXiv*, vol. abs/1612.00215, 2016. 15, 26

[67] K. Li, S. Peng, T. Zhang, and J. Malik, "Multimodal image synthesis with conditional implicit maximum likelihood estimation," *International Journal of Computer Vision*, vol. 128, pp. 2607 – 2628, 2020. 15

[68] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," in *International Conference on Machine Learning*, 2023, pp. 13 753–13 773. 15, 20, 21

[69] F. Zhan, Y. Yu, R. Wu, J. Zhang, and S. Lu, "Multimodal image synthesis and editing: The generative ai era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 15 098–15 119, 2021. 15

[70] Y. Zhang, Y. Kang, Z. Zhang, X. Ding, S. Zhao, and X. Yue, "Interactivevideo: User-centric controllable video generation with synergistic multimodal instructions," *ArXiv*, vol. abs/2402.03040, 2024. 15

[71] T. Li, M. W. Ku, C. Wei, and W. Chen, "Dreamedit: Subject-driven image editing," *Trans. Mach. Learn. Res.*, vol. 2023, 2023. 15

[72] M. Hui, Z. Zhang, X. Zhang, W. Xie, Y. Wang, and Y. Lu, "Unifying layout generation with a decoupled diffusion model," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1942–1951, 2023. 15

[73] D. Epstein, B. Poole, B. Mildenhall, A. A. Efros, and A. Holynski, "Disentangled 3d scene generation with layout learning," *ArXiv*, vol. abs/2402.16936, 2024. 18, 25

[74] H. Dhamo, F. Manhardt, N. Navab, and F. Tombari, "Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16 332–16 341, 2021. 18

[75] Y. Koyama, I. Sato, and M. Goto, "Sequential gallery for interactive visual design optimization," *ACM Transactions on Graphics (TOG)*, vol. 39, pp. 88:1 – 88:12, 2020. 19

[76] S. F. Bhat, N. Mitra, and P. Wonka, "Loosecontrol: Lifting controlnet for generalized depth conditioning," *ArXiv*, vol. abs/2312.03079, 2023. 19

[77] J. Mu, M. Gharbi, R. Zhang, E. Shechtman, N. Vasconcelos, X. Wang, and T. Park, "Editable image elements for controllable synthesis," *ArXiv*, vol. abs/2404.16029, 2024. 19

[78] K. Kikuchi, E. Simo-Serra, M. Otani, and K. Yamaguchi, "Constrained graphic layout generation via latent optimization," *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 20, 23

[79] L. Qu, S. Wu, H. Fei, L. Nie, and T. seng Chua, "Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation," *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 20, 22, 23, 24

[80] M. Ohanyan, H. Manukyan, Z. Wang, S. Navasardyan, and H. Shi, "Zero-painter: Training-free layout control for text-to-image synthesis," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8764–8774, 2024. 21

[81] W. Sun, T. Li, Z. Lin, and J. Zhang, "Spatial-aware latent initialization for controllable image generation," *ArXiv*, vol. abs/2401.16157, 2024. 22, 26

[82] W. Wang, C. Zhao, H. Chen, Z. Chen, K. Zheng, and C. Shen, "Autostory: Generating diverse storytelling images with minimal human effort," *ArXiv*, vol. abs/2311.11243, 2023. 23

[83] Z. Wang, A. Li, Z. Li, and X. Liu, "Genartist: Multimodal llm as an agent for unified image generation and editing," *ArXiv*, vol. abs/2407.05600, 2024. 23

[84] C. Jiang, S. Qi, Y. Zhu, S. Huang, J. Lin, L. Yu, D. Terzopoulos, and S.-C. Zhu, "Configurable 3d scene synthesis and 2d image rendering with per-pixel ground truth using stochastic grammars," *International Journal of Computer Vision*, vol. 126, pp. 920 – 941, 2017. 24

[85] Z.-H. Lin, W.-C. Ma, H.-Y. Hsu, Y. Wang, and S. Wang, "Neurmips: Neural mixture of planar experts for view synthesis," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 681–15 691, 2022. 25

[86] L. Han, J. Ren, H.-Y. Lee, F. Barbieri, K. Olszewski, S. Minaee, D. N. Metaxas, and S. Tulyakov, "Show me what and tell me how: Video synthesis via multimodal conditioning," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3605–3615, 2022. 25

[87] C. Zhao, M. Liu, W. Wang, J. wei Yuan, H. Chen, B. Zhang, and C. Shen, "Moviedreamer: Hierarchical generation for coherent long visual sequence," *ArXiv*, vol. abs/2407.16655, 2024. 25

[88] C.-Y. Cheng, F. Huang, G. Li, and Y. Li, "Play: Parametrically conditioned layout generation using latent diffusion," in *International Conference on Machine Learning*, 2023, pp. 5449–5471. 25

[89] Z. Liu, Y. Zhang, Y. Shen, K. Zheng, K. Zhu, R. Feng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao, "Cones 2: Customizable image synthesis with multiple subjects," *ArXiv*, vol. abs/2305.19327, 2023. 25

[90] J. Su, S. Gu, Y. Duan, X. Chen, and J. Luo, "Text2street: Controllable text-to-image generation for street views," *ArXiv*, vol. abs/2402.04504, 2024. 25

[91] Y. Sun, Y. Liu, Y. Tang, W. Pei, and K. Chen, "Anycontrol: Create your artwork with versatile control on text-to-image generation," *ArXiv*, vol. abs/2406.18958, 2024. 25

[92] I. Garcia-Dorado, P. Getreuer, M. Le, R. Debreuil, A. Kauffmann, and P. Milanfar, "Graphic narrative with interactive stylization design," *ArXiv*, vol. abs/1712.06654, 2017. 26