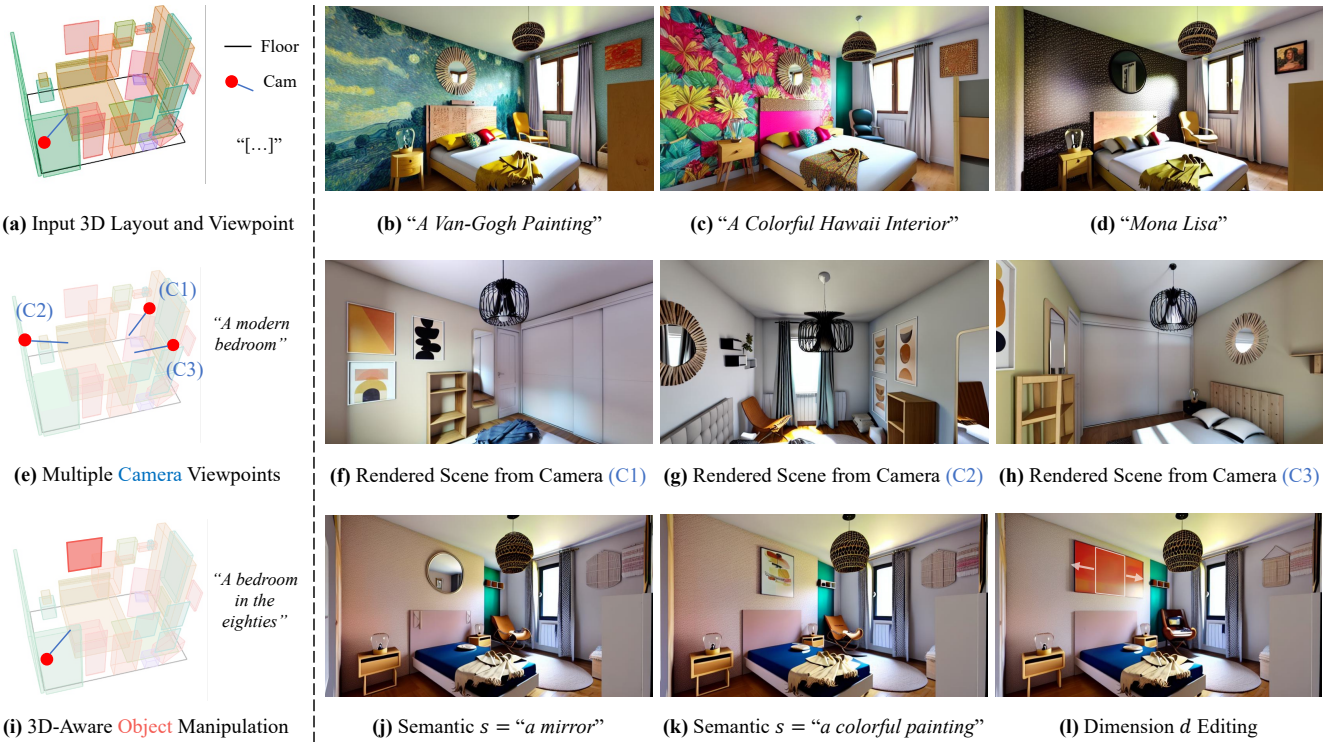
LACONIC: A 3D Layout Adapter for Controllable Image Creation

Léopold Maillard[1,2] Tom Durand[2] Adrien Ramanana Rahary* Maks Ovsjanikov[1]

[1]LIX, École Polytechnique, IP Paris [2]Dassault Systèmes

{maillard,maks}@lix.polytechnique.fr {firstname.lastname}@3ds.com



(a) Input 3D Layout and Viewpoint **(b)** "*A Van-Gogh Painting*" **(c)** "*A Colorful Hawaii Interior*" **(d)** "*Mona Lisa*"

(e) Multiple Camera Viewpoints **(f)** Rendered Scene from Camera (C1) **(g)** Rendered Scene from Camera (C2) **(h)** Rendered Scene from Camera (C3)

(i) 3D-Aware Object Manipulation **(j)** Semantic s = "*a mirror*" **(k)** Semantic s = "*a colorful painting*" **(l)** Dimension d Editing

Figure 1. **Overview of LACONIC Capabilities and Applications**. Our model generates realistic renderings from an input semantic 3D layout and target viewpoint (a), while leveraging the comprehensive knowledge of a text-to-image prior (b–d). A given 3D scene can be rendered from multiple camera poses (e) while maintaining a consistent 3D structure across views (f–h). Finally, objects can be individually manipulated (i) by editing their open-vocabulary text caption, or position and dimension attributes in the 3D space (j–l).

Abstract

Existing generative approaches for guided image synthesis of multi-object scenes typically rely on 2D controls in the image or text space. As a result, these methods struggle to maintain and respect consistent three-dimensional geometric structure, underlying the scene. In this paper, we propose a novel conditioning approach, training method and adapter network that can be plugged into pretrained text-to-image diffusion models. Our approach provides a way to endow such models with 3D-awareness, while leveraging their rich prior knowledge. Our method supports camera control, conditioning on explicit 3D geometries and, for the first time, accounts for the entire context of a scene, i.e.*, both on and off-screen items, to synthesize plausible and semantically rich images. Despite its multi-modal nature, our model is lightweight, requires a reasonable number of data for supervised learning and shows remarkable generalization power. We also introduce methods for intuitive and consistent image editing and restyling, e.g., by positioning, rotating or resizing individual objects in a scene. Our method integrates well within various image creation workflows and enables a richer set of applications compared to previous approaches.*

*Work done during an internship at Dassault Systèmes.

1. Introduction

The integration of user-prompted conditioning signals from diverse modalities has been a major factor in the recent success of image generation models [21, 34, 47, 49], enabling precise control, expanding creative possibilities, and narrowing the gap between human intent and generated content. Yet, there is still room to make controllable image synthesis of multi-object scenes more intuitive and better integrated with conventional design and composition workflows. In a standard pipeline for photorealistic rendering, designers generally follow a bottom-up approach: first defining the three-dimensional structure of the environment, then introducing objects and their appearance before rendering from different viewpoints and, optionally, refining details through local edits or relighting. However, existing guiding mechanisms for generative image models predominantly rely on text [34, 47, 49] or image [33, 47, 66, 69] representations, which do not inherently preserve consistent, unambiguous, and well-defined 3D structures.

The dominant approach of generative image modeling through text guidance, while convenient, also makes it difficult to convey complex compositional structures [38], as it can be cumbersome to capture nuanced spatial and geometric relations through text alone. Such relations are especially prominent in multi-object environments such as indoor scenes. Consequently, approaches leveraging structural controls, such as bounding box layouts, keypoints or semantic maps [24, 41, 47], have emerged to provide more explicit control over composition. Key to their adoption has been their efficient integration into pretrained diffusion models via the training of lightweight *adapter* modules [33, 66, 69], that preserve prior generative capabilities, while enabling richer and more intuitive controls. Attempts aimed at providing 3D-aware control for image synthesis and editing have gained popularity [7, 9, 36, 51, 62, 64], but typically rely on panorama images, video datasets or conditioning through depth maps that are hard to acquire and offer limited flexibility.

Overall, and as illustrated in Figure 2, the nature of these inputs is inherently associated with critical limitations, as they discard crucial spatial information such as viewpoint, object orientations, or occlusions in nested arrangements. As a result, existing representations struggle to accurately account for cases where objects are placed within others—such as books stored inside a shelf—despite such configurations being highly prevalent in real-world scenes. They also encode geometric information in a highly viewpoint-dependent manner, causing inconsistencies in how spatial structures are perceived across views. Finally, conditioning should ideally encode 3D context in a comprehensive manner. This means, for instance, capturing the influence of off-screen elements, such as lighting from a window outside the frame, but also maintaining stylistic and functional coherence across the entire scene.

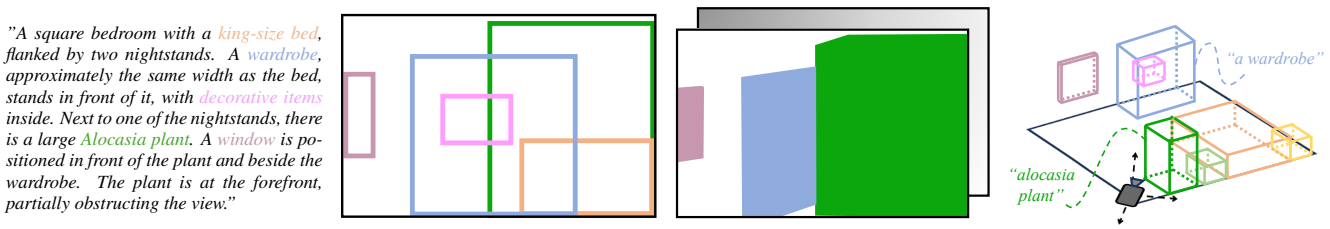
In this context, we propose to inject simple, explicit 3D geometric information as guidance for single-view generative models. Associated challenges are numerous in light of the limited availability of image data with accurate 3D semantic annotations [10, 46, 67], and the lack of established solution to encode 3D inputs as additional control to a pretrained diffusion backbone. From a technical perspective, incorporating conditioning in a form that is different from text or image inputs is indeed highly non-trivial. As a result, our work aims to introduce the representational and architectural components as well as the training dynamics to *augment* pretrained text-to-image models with guidance capabilities from parametric 3D semantic layouts. More precisely, our main contributions are threefold:

1. A parametric conditioning representation, based on semantic bounding boxes, which, for the first time, allows for 3D-informed image synthesis and editing, maintaining consistent structures across views and without relying on depth estimators, multi-view or panorama images.
2. An adapter architecture that establishes a tight relation between inputs in the 3D domain and the image target, while being compatible with pre-existing conditioning modalities.
3. An efficient training framework that enables camera control as well as geometric and free-form semantic object-level guidance in 3D.

We demonstrate the effectiveness of our method in a wide range of experimental settings, highlighting multiple advantages over current approaches. We additionally introduce a novel evaluation methodology that allows to evaluate the adherence to object-level conditioning, and use it to assess different image synthesis methods.

2. Related Work

3D-Aware Content Creation Recent advances in deep generative modeling [16, 20, 25, 40, 54] have facilitated the emergence of powerful methods for user-driven content synthesis. In particular, score-based models have been employed to create realistic images [34, 47, 49], videos [17, 42, 53] or 3D assets [27, 28] of unprecedented quality. Despite these capabilities, achieving *3D-aware* image synthesis and editing, *i.e.*, that naturally incorporate the three-dimensional structure of the underlying scene, remains challenging. This task is typically tackled by leveraging multi-view [4, 11, 12, 56, 60] or multi-frame [32, 62] image datasets, or by using depth maps, which are hard to acquire and manipulate in real-life settings, as additional conditioning inputs [36, 58, 63]. 3D-aware approaches have notably demonstrated advantages over those relying on traditional text-to-image models for 3D scene generation [9, 18, 51]. However, these typically rely on costly



(a) Text Description (b) Typed 2D Bounding Boxes (c) Semantic & Depth Maps (d) **Ours**

Figure 2. **Comparison of high-level conditioning input representations for describing a 3D scene**. Relying solely on text descriptions (a) can make it difficult to convey complex spatial relations. Conditioning via 2D bounding boxes (b) can lead to ambiguity in perspective and does not account for out-of-bound objects. Semantic and depth maps rendered from 3D bounding boxes (c), as introduced in recent work [64], cannot always handle occluded and nested items while being limited to objects typed from a fixed set of categories. In contrast, our layout representation encodes a comprehensive 3D scene structure that is consistent across camera views and supports object-level semantic captioning and direct manipulation of position and orientation in 3D space.

score distillation [43] from the 2D prior [6, 52, 64], limiting their applicability.

Layout-Guided Image Synthesis A prolific line of work has been incorporating structured spatial information to control image generation. Notably, *compositional* synthesis can be achieved from the guidance of 2D semantic bounding boxes [5, 24, 47, 65, 70] or segmentation maps [47, 59]. Closely related to our work, SceneCraft2D [64] renders 3D indoor layouts from any camera viewpoint by projecting object bounding boxes to depth and semantic maps, that are used as conditions to train respective ControlNet [69] modules. Other methods like ControlRoom3D [51] or Ctrl-Room [9] further rely on panorama images to render 360° consistent views of rooms. Recently, Build-A-Scene [7] proposes a training-free approach based on attention guidance [5] to perform 3D layout control from a depth-conditioned prior [2]. While convenient, this method does not scale to *complex* 3D layouts featuring more than a few objects. One common limitation of these approaches is that their conditioning input ultimately lies in the 2D space, making it difficult to disentangle objects in complex arrangements or to account for unseen *contextual* items from the 3D environment.

Adapters for Diffusion Priors *Adapters* are lightweight learnable modules to be plugged into pretrained diffusion models. Notably, *low-rank* adapters [19] have been widely used [30] to efficiently perform customized domain adaptation, while *structure* adapters [33, 69] have enabled to control image generation with various 2D conditions that are spatially aligned with the content of the target image. More related to our work in its methodology, IP-Adapter [66] augments text-to-image models with image guidance, allowing to finely control the style and appearance of the synthesized content. This is done by learning additional cross-attention [57] weights, linking image embeddings extracted by pretrained foundation models [35, 45] with visual features from the diffusion backbone. In contrast, we propose to learn our 3D semantic layout encoder and attention-based adaptation module jointly.

Encoding Semantic 3D Layouts Previous lines of work have proposed a range of representations based on graphs [23, 37, 68] to embed 3D arrangements in a way that can capture the relations between interacting elements. Concurrently, methods defining layouts as unordered sequences of objects, to be encoded by attention-based architectures, have shown compelling results in the context of 3D scene synthesis [31, 39, 55, 61].

Our method broadens the scope of existing control mechanisms and is the first to enable comprehensive control of both object semantics but also the underlying 3D geometric structure in a single unified framework.

3. Method

In this section, we introduce our framework, which is summarized in Figure 3. Our ultimate goal is to provide 3D layout control to pretrained text-to-image (T2I) models. Unlike previous work [64], we use an explicit, view-independent 3D representation and jointly learn feature extractors and adapter modules in a single training experiment.

3.1. Preliminaries

Our framework leverages text-to-image diffusion models (DM). We introduce the key associated generative and conditioning mechanisms that are relevant for our approach, and which we build upon below.

T2I Diffusion Models T2I diffusion models are trained to denoise data samples x, perturbed by Gaussian noise ϵ across multiple timesteps t, given their associated text caption c and, optionally, a task-specific condition y. This is done by parameterizing a neural network ϵ_θ to predict the

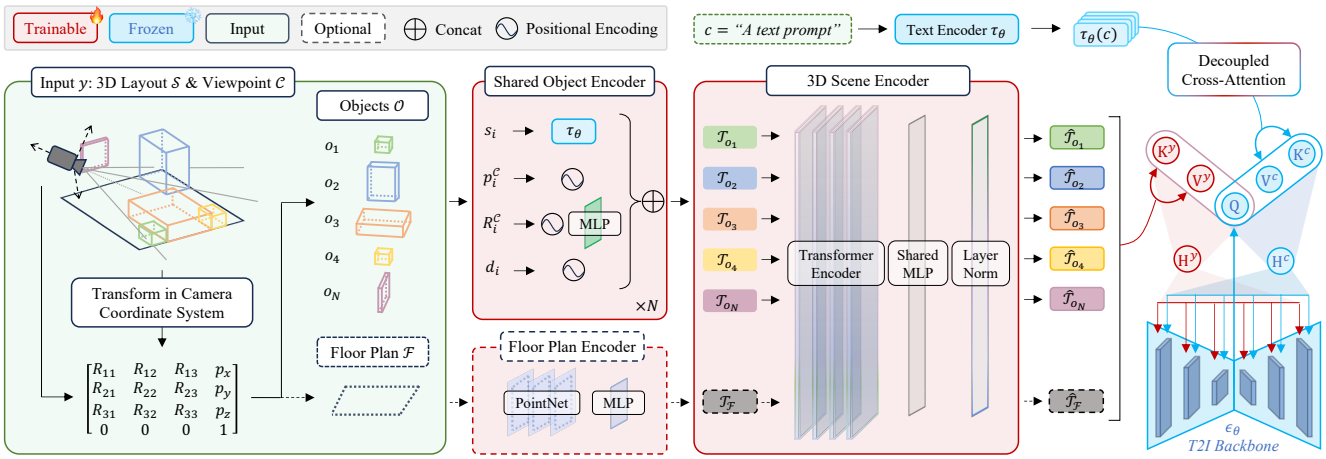


Figure 3. **LACONIC Architecture & Pipeline Overview**. From an input 3D layout $\mathcal{S}$ and camera pose $\mathcal{C}$, trainable modules embeds geometric and semantic properties of individual objects to guide a text-to-image diffusion prior in denoising a target rendering. Camera control is enabled by expressing spatial features (p, R) from the input object 3D bounding boxes $\mathcal{O}$ in the coordinate system defined by the target viewpoint. Resulting objects and, optionally, a room's floor plan $\mathcal{F}$, are embedded by dedicated modules, yielding a set of representations $\mathcal{T}$ processed by a transformer-based encoder. Output sequence of embeddings is subsequently integrated as additional conditioning input to the pretrained T2I backbone via decoupled cross-attention [66].

noise residual in x_t following the simple objective:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x,c,y,\epsilon\sim\mathcal{N}(0,I),t}\Big[\|\epsilon - \epsilon_\theta(x_t, t, c, y)\|_2^2\Big]. \quad (1)$$

Central to our approach, this learning objective can be used to incorporate diverse conditioning controls y through fine-tuning experiments, leveraging a pretrained, frozen T2I backbone alongside additional trainable components.

Cross-Attention Conditioning Input signals to control text-to-image generative models are learned through the use of cross-attention layers within the architecture. We introduce here the established mechanisms and notations that intertwine with our own conditioning approach, described in Section 3.3. In particular, given a tokenized text caption $c = [c_1, c_2, \ldots, c_M]$, a pretrained foundation model like CLIP [45] computes a sequence representation $\tau_\theta(c) \in \mathbb{R}^{M\times d_\tau}$ where d_τ is the text embedding size. Resulting *key* K^c and *value* V^c, of hidden dimension d, are obtained from respective learnable linear projections. Similarly, the intermediate $h \times w$ feature map with d_{img} channels from the image backbone, that is typically implemented as a UNet [48] with residual blocks, is flattened and projected to the *query* Q. New hidden state $H^c \in \mathbb{R}^{hw\times d}$ incorporating the text condition c is finally computed via dot-product attention [57]. In practice, this mechanism is applied in a multi-headed fashion and at various feature resolutions within the conditional image denoiser ϵ_θ.

3.2. Semantic Layout Representation

At the heart of our method lies the additional conditioning input y that we define as an intuitive and explicit proxy for furnished 3D layouts. As described in this section, individual objects are encoded from their high-level semantic and spatial properties using a parametric bounding box representation in the 3D space, as described below.

3D Scene Parameterization We represent a 3D scene $\mathcal{S}$ as an unordered set of N objects $\mathcal{O} = \{o_1, \ldots, o_N\}$ and, optionally, an indoor floor plan modeled by a point cloud $\mathcal{F} \in \mathbb{R}^{P\times 3}$, where P points are sampled along its boundary contours. We make $\mathcal{F}$ optional because (i) 3D layout datasets often lack detailed structural data, like floors, ceilings and walls, and (ii) we empirically found that 3D objects alone can implicitly reflect the scene's structure via their spatial arrangements, given that they are commonly positioned on floors or aligned against walls. We represent each object by its "semantic 3D bounding box": $o_i = (p_i, d_i, R_i, s_i)$. Here $p_i \in \mathbb{R}^3$ is the object's center position in the world coordinate system, $d_i \in \mathbb{R}^3$ is the size along each dimension, and $R_i \in \mathbb{R}^{3\times 3}$ is the rotation matrix. Similar to the global scene caption c, the object-level semantic description $s_i = [s_i^1, s_i^2, \ldots, s_i^M]$ is processed to M tokens from natural language. Those design choices are motivated by the recent success of similar lightweight representations, which have demonstrated their expressiveness in related 3D scene generative tasks [31, 39, 55, 61].

Camera Viewpoint The camera $\mathcal{C}$ from which to render the 3D layout is represented by its extrinsic parameters, *i.e.*, position $p_\mathcal{C} \in \mathbb{R}^3$ and rotation $R_\mathcal{C} \in \mathbb{R}^{3\times 3}$. Note that, for the purpose of our general methodology, we assume that the image samples x from the data distribution are rendered using consistent camera intrinsics.

3.3. Adapter Architecture

In this section, we introduce the trainable modules and mechanisms designed to (i) capture meaningful representations from the scene conditioning input and (ii) incorporate them to the T2I backbone, which, as described in Section 3.1, is conditioned through cross-attention modules and remains frozen to fully retain its original capabilities.

3D Layout Encoder The part of our architecture that is responsible for encoding the input 3D layout representation is inspired by previous work [31, 39, 61]. Notably, individual objects o_i are handled by a common module that embeds scalar spatial features through sinusoidal positional encoding [57] and dense layers, while encoding the semantic caption using the pretrained text encoder backbone τ_θ. Concatenation of the resulting attributes produces per-object tokens $\mathcal{T}_{o_i}$. The optional indoor floor plan is encoded by a PointNet [44] module, leading to an individual token $\mathcal{T}_\mathcal{F}$. The sequence defined by the embedded tokens is passed to a transformer encoder computing new representations. Subsequently and following previous work [66], a dense unit shared between tokens outputs representations of dimension d_τ, which is consistent with the global text embeddings, and is followed by Layer Normalization [1].

Decoupled Cross-Attention The unordered sequence of scene representations $\hat{\mathcal{T}} = \{\hat{\mathcal{T}}_\mathcal{F}, \hat{\mathcal{T}}_{o_i}, \dots, \hat{\mathcal{T}}_{o_N}\}$ output by the 3D layout encoder is intuitively ideal to establish the individual visual contributions of each object within the image backbone through cross-attention conditioning. It is projected accordingly to *key* K^y and *value* V^y both in $\mathbb{R}^{(N+1)\times d}$ by additional trainable linear projections. The hidden state associated to the input 3D layout is computed similarly to the one H^c from the text condition, via dot-product attention with the same *query* Q obtained from image feature maps:

$$H^y = \text{softmax}\left(\frac{Q(K^y)^\top}{\sqrt{d}}\right) \cdot V^y, \quad H^y \in \mathbb{R}^{hw\times d} \tag{2}$$

Following the decoupled cross-attention methodology [66], the final hidden state incorporating both the global text c and 3D layout y conditions is obtained via a weighted sum:

$$H = H^c + \gamma H^y, \quad H \in \mathbb{R}^{hw\times d} \tag{3}$$

where γ is a scalar controlling the strength of the scene control with respect to the global caption one. It is finally projected back to be added to image feature maps of the diffusion backbone's residual blocks.

3.4. Supervised Training

At training time, we assume to be given a collection of scenes, from which to extract structural and semantic 3D bounding box representations, as described in Section 3.2. Each scene $\mathcal{S}$ is associated to a rendering image x_0 and corresponding camera pose $\mathcal{C}$. In this section, we describe the training dynamics adopted to efficiently map our additional guidance signal to the target scene image, enabling free camera control and object-level, open-vocabulary semantic conditioning.

Camera Viewpoint Transformation One important challenge associated to our method is that our conditioning input is defined in a three-dimensional world coordinate system, while the generation target lies in the image domain. As a result, we introduce an explicit mechanism to handle the crucial translation between the 3D input representation and underlying output image. Importantly, the information of the input viewpoint from which $\mathcal{S}$ is rendered is directly integrated into the object spatial representations, by expressing them in the 3D coordinate system defined by the camera $\mathcal{C}$. More precisely, we apply the following series of transformations:

$$p_i^\mathcal{C} = R_\mathcal{C}^\top (p_i - p_\mathcal{C}), \quad R_i^\mathcal{C} = R_\mathcal{C}^\top R_i, \quad \forall o_i \in \mathcal{O} \tag{4}$$

We found this reframing mechanism of the 3D input signal to the 2D image domain to be key in order to efficiently leverage the pretrained diffusion model. It can be performed in closed form on the fly, without requiring the network to learn such complex mapping between domains. When it is provided, a similar transformation is performed with the floor point cloud input $\mathcal{F}$. In practice, object rotation features $R_i^\mathcal{C}$ are further mapped to a continuous representation, following [71]. The resulting scene serves as the additional conditioning input $y = \mathcal{S}^\mathcal{C}$ to guide the pretrained T2I model ϵ_θ in denoising the target image x_0 consistent with the 3D layout structure and camera view, following the training objective from Equation (1).

Conditioning Dynamics We adopt classifier-free guidance [15] training by randomly dropping the 3D layout input y, which, at each iteration, is set to $\emptyset$ with probability $\mathbf{p}_{\text{drop}}$, so that the denoiser network models both the layout-conditional and unconditional image densities. Additionally, image targets x do not need to be associated to a global textual description and, as a result, c is always obtained from an empty prompt during training. Unlike previous work [64], which relies on *one-hot* category representations, our method conditions individual objects with free-form text captions s_i. If not directly part of the dataset annotations, caption supervision can be derived from a vision-language model [22, 26] applied to an image showcasing the object of interest.

3.5. Application Scenarios

As illustrated in Figure 1, we provide a broad overview of the range of generative and editing capabilities enabled by

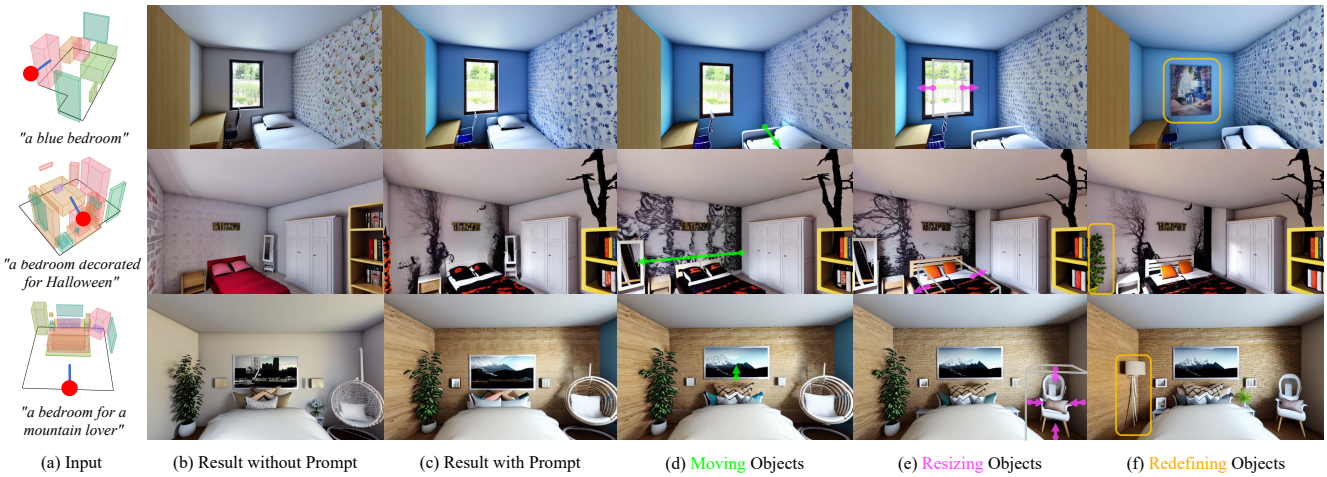


Figure 4. **Iterative Scene Editing Results**. From left to right: given an input semantic 3D layout and camera viewpoint (a), we render the scene both without (b) and with (c) global text prompt conditioning. Then, individual objects are subsequently moved (d), resized (e) and re-captioned (f). Results highlight the ability of our approach to additively perform local manipulations and generalize to out-of-distribution concepts through the use of the global text conditioning, while maintaining a consistent 3D structure. Remarkably, we observe that semantic concepts from the global text prompt are represented on *relevant* objects in the generated image, *e.g.*, patterns on wallpapers or bedsheets, art in wall frames, and do not *leak* to *e.g.*, ceilings or floors. Interestingly, removing a window has an impact on the global illumination of the scene, which would have not been possible by editing local objects via *e.g.*, 2D image inpainting [29].

our trained model at test time. Here, we consider an initial input scene $\mathcal{S}_0$, which can be conveniently user-provided by disposing, sizing and describing a collection of objects.

Generating Structurally-Consistent Views Although our method is designed to be trained on single-view datasets, the comprehensive 3D context provided by our input representation allows synthesizing multiple views of a given scene, sharing a consistent 3D structure. More formally, from a set of target camera viewpoints $\{\mathcal{C}_1, \ldots, \mathcal{C}_C\}$, the corresponding inputs $\mathcal{S}_0^{\mathcal{C}_i}$ conditioning the trained model can be obtained by applying the space transformation logic from Equation (4).

Text-Driven Scene Restyling Remarkably, while our model has not been trained with global text caption c supervision, its adapter design allows to fully leverage the prior knowledge from the T2I backbone, benefiting from strong out-of-distribution generalization to a wide range of semantic concepts. Notably, as described in Section 4.2, strict adherence to the input scene and viewpoint can be relaxed by lowering γ in Equation (3), thus interpolating structural variations matching the global input description.

Object Attribute-Level Scene Editing The initial input scene $\mathcal{S}_0$ can be iteratively edited based on user preferences, by modifying parts of the objects in $\mathcal{O}_0$ or adjusting its items set through additions or removals. Importantly, individual objects can be manipulated with per-attribute granularity *i.e.*, by individually and independently adjusting their size, positioning or semantic features. This shows practical advantages over previous approaches [36] whose attempts to solve 3D-aware image editing required to edit conditioning depth maps in the pixel space.

4. Experiments

In this section, we showcase our method's capabilities at generating realistic images that incorporate the desired semantic and geometric input information. To this end, we conduct various experiments comparing it to baseline approaches on a range of standard metrics. We also introduce a novel metric that captures the alignment between the 3D layout prompt and the rendered image on the object level.

4.1. 3D Layout-Guided Image Synthesis

Datasets We follow previous work [64] and use HyperSim [46] indoor scene dataset to extract semantic 3D bounding box layouts as well as camera poses with their associated photorealistic renderings. As HyperSim only features 326 unique 3D layouts, paired to a total of 24,383 images, methods taking as input the comprehensive 3D layout associated to each image sample, such as ours, may tend to memorize individual fixed scenes from the dataset, making it difficult to perform the complex out-of-distribution manipulations of individual objects that our framework enables. In response, and in order to provide additional qualitative results showcasing the full range of our model's capabilities, we also gather a custom dataset featuring 72,000 *bedroom* 3D scene layouts, each being paired to a single,

(a) 3D Scene (b) SceneCraft (c) DM-FS (d) **ours** (e) SceneCraft w/prompt (f) **ours w/prompt** (g) SceneCraft w/prompt (h) **ours w/prompt**

Figure 5. **3D layout-guided image synthesis baseline comparisons**. Our method produces more detailed and natural images compared to baseline approaches. Methods leveraging our 3D layout encoder (DM-FS & ours) better represent the guiding layout, while our adapter-based approach additionally shows advantages at producing higher quality images. We also observe in text-driven scenarios (e–h) that compared to SceneCraft [64], LACONIC better blends conditioning inputs coming from different sources, adhering both to elements from the 3D layout and the semantic text caption conditionings.

high-quality image and viewpoint. See supplementary for additional details and statistics on the used datasets.

Baselines We ensure a relevant and fair comparison by competing against SceneCraft [64], a recent baseline which, in the like of our method, (i) proposes to synthesize 2D images from 3D layout controls, (ii) is based on supervised training on single-view, non-panoramic images and (iii) is an adapter-based approach, enabling text-driven synthesis at test time. Additionally, we evaluate the impact of using our adapter approach instead of training a layout-conditioned *Diffusion Model From Scratch* (DM-FS) on layout-image pairs by introducing a dedicated baseline that retains the 3D layout conditioning encoder from our methodology.

Implementation Throughout our experiments, we employ Stable Diffusion v1.5 [47] as the pretrained text-to-image diffusion backbone upon which our adapter network is trained. We train the DM-FS baseline using a similar but downsized UNet architecture, to account for the scale of the dataset. SceneCraft [64] is used with its pretrained weights and default parameters from the official implementation. All the compared methods in quantitative evaluations and side-by-side qualitative comparisons are trained on the same HyperSim [46] subset. We provide comprehensive implementation details, training and inference settings, and an ablation study in the supplementary material.

Metrics We evaluate generation quality and diversity with respect to the image data distribution by reporting the Fréchet Inception Distance (FID) [14] and Kernel Inception Distance (KID $\times 1{,}000$) [3]. Rendering realism and variety is further assessed using the Inception Score (IS) [50]. For methods supporting text-driven synthesis, we additionally report the CLIP Score (CS) [13] as a measure of how well the synthesized images align with a global caption describing the target scene.

Scene Object CLIP score (SOC) We also evaluate different methods on a new metric that we call Scene Object CLIP score (SOC). Our score is motivated by the fact that most established metrics for evaluating image generative models primarily focus on image quality, diversity, and prompt adherence, rather than spatial or semantic alignment at the object level. Instead, SOC explicitly aims at measuring the alignment of individual objects in the conditioning layout with the corresponding locations in the synthesized image.

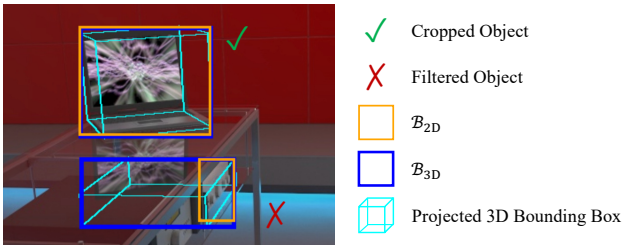


Figure 7. **Object selection for the SOC metric**. Obstructed and out-of-bounds objects are identified from their 2D and 3D bounding box annotations and filtered out from the evaluation set.

Our metric takes as input a synthesized image, a guiding 3D semantic layout and 2D object bounding box $\mathcal{B}_{2D}$ annotations from the ground truth image. To compute SOC, we proceed as follows: for each object in the 3D layout, we (i) project its 3D bounding box on the rendered image using the camera $\mathcal{C}$ extrinsic and intrinsic parameters, (ii) derive the corresponding 2D enclosing bounding box $\mathcal{B}_{3D}$, where $\mathcal{B}_{2D} \subset \mathcal{B}_{3D}$ and (iii) define a minimum threshold on the intersection area ratio $\mathcal{B}_{2D}/\mathcal{B}_{3D}$ for an object to be considered sufficiently visible. Finally, we crop the selected items from their enclosing bounding box $\mathcal{B}_{3D}$, obtaining a collection of per-object images x_{o_i}. These steps are meant to only keep objects in 3D space which are expected to be sufficiently visible in the given 2D image, as illustrated in Figure 7.

(a) Input (b) $\gamma = 0.00$ (c) $\gamma = 0.25$ (d) $\gamma = 0.40$ (e) $\gamma = 0.70$

Figure 6. **Impact of the adapter strength on generated images.** From a text caption c = *"A Retro Synthwave Indoor Bedroom"*, adherence with the input 3D layout improves accordingly with the parameter γ from Equation (3), achieving strict control at higher scales. We can observe that even at lower scales (c) the control enforces the geometric and semantic plausibility of the synthesized image.

Since each visible object o_i is associated with a text description, we can compute a semantic correlation between the object's text caption s_i and its crop image x_{o_i}. For this last step, we follow [13] among others, and use the pretrained CLIP [45] model. We provide details on our Scene Object Clip score (SOC) in the supplementary material.

Table 1. **Quantitative experiment results on 3D layout-guided images synthesis.** We compare our method against other learning-based approaches that we outperform on most evaluation metrics, both with and without providing a global text caption c.

Methods		FID ↓	KID ↓	IS ↑	CS ↑	SOC ↑
SceneCraft [64]	*w/o text prompt*	39.36	28.26	7.72	—	17.59
DM-FS		15.83	7.29	8.69	—	18.22
ours		**9.50**	**3.44**	**9.74**	—	**18.36**
SceneCraft [64]	*w/ text prompt*	27.69	15.21	**14.55**	**19.75**	17.40
ours		**10.12**	**3.91**	10.60	19.74	**18.39**

Results Our main quantitative results are summarized in Table 1. We report results both with and without text prompting. In the latter case our approach achieves state-of-the-art results across all metrics and outperforms the very recent baseline SceneCraft [64] by a significant margin. When using text conditioning (the captions provided in HyperSim [46]), our method still outperforms SceneCraft [64] on most metrics and in several cases by a significant margin. We note that our approach is able to capture the HyperSim data distribution significantly better compared to SceneCraft [64]. However, due to the relative limited scope of 3D structures in that dataset, our approach demonstrates somewhat less variety in text-driven scenarios as indicated by a lower Inception Score. Importantly, our adapter-based approach outperforms the DM-FS approach on all metrics, while enabling text-conditioned synthesis, thus validating the importance of using a rich pretrained T2I model. We provide qualitative comparisons on the 3D layout-guided image synthesis task in Figure 5. Notably, we observe that our approach produces more natural images, featuring fewer visual artifacts compared to baseline methods, while being more faithful to the input 3D layout and viewpoint. These qualitative findings are confirmed by a perceptual study detailed in the supplementary material.

4.2. Additional Qualitative Results

We provide additional generation results, showcasing the intuitive and iterative editing capabilities of our framework as well as its ability to leverage the pretrained text-to-image backbone for out-of-distribution generalization in Figure 1 and Figure 4. Overall, these results highlight that our model can capture nuanced individual object relations within structured scenes. Moreover, our approach enables manipulations in 3D space, and provides wholistic scene modeling, accounting, *e.g.*, for off-screen objects. Finally, results reported in Figure 6 show how introducing our adapter control, even at low scale γ, allows synthesizing 3D-aware images that are more geometrically and semantically sound than base text-to-image outputs, while maintaining adherence with the textual description. Additional qualitative results can be found in the supplementary material, including a comparison against Build-A-Scene [7] and, most notably, a demonstration of our adapter's compatibility with a modern DiT backbone [8].

5. Conclusion, Limitations & Future Work

In this paper, we introduced LACONIC, a novel 3D layout-conditional image generative framework that is distinct in both representation and architecture, expanding the creative capabilities in 3D-aware image synthesis and editing. By efficiently mapping the underlying 3D representations to the image domain and integrating them to pretrained text-to-image model, our method introduces fine-grained structural and semantic control, enhancing both user interactivity and the expressiveness and coherence of generated content.

While powerful and capable of generalization to unseen concepts, our model's output domain is still bounded by the training data distribution. As a result, a model trained on bedroom layout data will be unlikely to synthesize *e.g.*, plausible kitchens while allowing fine-grained control. We believe that our model would be ideal to derive continuous 3D scene representations, such as NeRFs, that finely adhere to user-defined structures from the conditioning input. We also believe that exploring ways to couple geometric and *visual* consistency of objects across changes in camera views is an important future research direction.

Acknowledgments

We thank the anonymous reviewers for their insights and suggestions. We also thank Farah Ellouze and Ana Marcusanu for their feedback on the paper draft. This work was supported by Dassault Systèmes SE. The views and conclusions contained in the paper are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the company.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[2] Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[3] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018. 7

[4] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4217–4229, 2023. 2

[5] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5343–5353, 2024. 3

[6] Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Set-the-scene: Global-local training for generating controllable nerf scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2920–2929, 2023. 3

[7] Abdelrahman Eldesokey and Peter Wonka. Build-a-scene: Interactive 3d layout control for diffusion-based image generation. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. 2, 3, 8

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning (ICML)*, 2024. 8

[9] Chuan Fang, Yuan Dong, Kunming Luo, Xiaotao Hu, Rakesh Shrestha, and Ping Tan. Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints. In *International Conference on 3D Vision (3DV) 2025*. 2, 3

[10] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10933–10942, 2021. 2

[11] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3d: Create anything in 3d with multi-view diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 2

[12] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning (ICML)*, pages 11808–11826. PMLR, 2023. 2

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7, 8

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 7

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 2

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:8633–8646, 2022. 2

[18] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7909–7920, 2023. 2

[19] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[20] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:26565–26577, 2022. 2

[21] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in neural information processing systems (NeurIPS)*, 32, 2019. 2

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022. 5

[23] Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao Zhang. Grains: Generative re-

cursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)*, 38(2):1–16, 2019. 3

[24] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22511–22521, 2023. 2, 3

[25] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023. 5

[27] Jiayi Liu, Denys Iliash, Angel X Chang, Manolis Savva, and Ali Mahdavi-Amiri. SINGAPO: Single image controlled generation of articulated parts in objects. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. 2

[28] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023. 2

[29] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, 2022. 6

[30] Michael Luo, Justin Wong, Brandon Trabucco, Yanping Huang, Joseph E. Gonzalez, Zhifeng Chen, Russ Salakhutdinov, and Ion Stoica. Stylus: Automatic adapter selection for diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 3

[31] Léopold Maillard, Nicolas Sereyjol-Garros, Tom Durand, and Maks Ovsjanikov. Debara: Denoising-based 3d room arrangement generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 109202–109232, 2024. 3, 4, 5

[32] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023. 2

[33] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2, 3

[34] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, pages 16784–16804. PMLR, 2022. 2

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. Featured Certification. 3

[36] Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J Mitra. Diffusion handles enabling 3d edits for diffusion models by lifting activations to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7695–7704, 2024. 2, 6

[37] Wamiq Para, Paul Guerrero, Tom Kelly, Leonidas J Guibas, and Peter Wonka. Generative layout modeling using constraint graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6690–6700, 2021. 3

[38] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 2

[39] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 12013–12026, 2021. 3, 4, 5

[40] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 2

[41] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio (CVPR)*, pages 7932–7942, 2024. 2

[42] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2

[43] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 3

[44] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 5

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language super-

vision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 3, 4, 8

[46] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10912–10922, 2021. 2, 6, 7, 8

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3, 7

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 4

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 2

[50] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016. 7

[51] Jonas Schult, Sam Tsai, Lukas Höllein, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, et al. Controlroom3d: Room generation using semantic proxy rooms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6201–6210, 2024. 2, 3

[52] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 3

[53] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2

[54] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[55] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20507–20518, 2024. 3, 4

[56] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 51202–51233, 2023. 2

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3, 4, 5

[58] Ruicheng Wang, Jianfeng Xiang, Jiaolong Yang, and Xin Tong. Diffusion models are geometry critics: Single image 3d editing using pre-trained diffusion priors. In *European Conference on Computer Vision (ECCV)*, pages 441–458. Springer, 2024. 2

[59] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 3

[60] Daniel Watson, William Chan, Ricardo Martin Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2

[61] Qiuhong Anna Wei, Sijie Ding, Jeong Joon Park, Rahul Sajnani, Adrien Poulenard, Srinath Sridhar, and Leonidas Guibas. Lego-net: Learning regular rearrangements of objects in rooms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19037–19047, 2023. 3, 4, 5

[62] Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew A. Hudson, Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste, Kelsey R Allen, and Thomas Kipf. Neural assets: 3d-aware multi-object scene synthesis with image diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 2

[63] Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 3d-aware image generation using 2d diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2383–2393, 2023. 2

[64] Xiuyu Yang, Yunze Man, Junkun Chen, and Yu-Xiong Wang. Scenecraft: Layout-guided 3d scene generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:82060–82084, 2024. 2, 3, 5, 6, 7, 8

[65] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14246–14255, 2023. 3

[66] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 4, 5

[67] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International*

Conference on Computer Vision (ICCV), pages 12–22, 2023. 2

[68] Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graphs. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 3

[69] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2, 3

[70] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22490–22499, 2023. 3

[71] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 5