SeeThrough3D: Occlusion Aware 3D Control in Text-to-Image Generation

Vaibhav Agrawal[1] Rishubh Parihar[2] Pradhaan S Bhat[2]
Ravi Kiran Sarvadevabhatla[1†] Venkatesh Babu Radhakrishnan[2†]
[1]IIIT Hyderabad [2]IISc Bengaluru

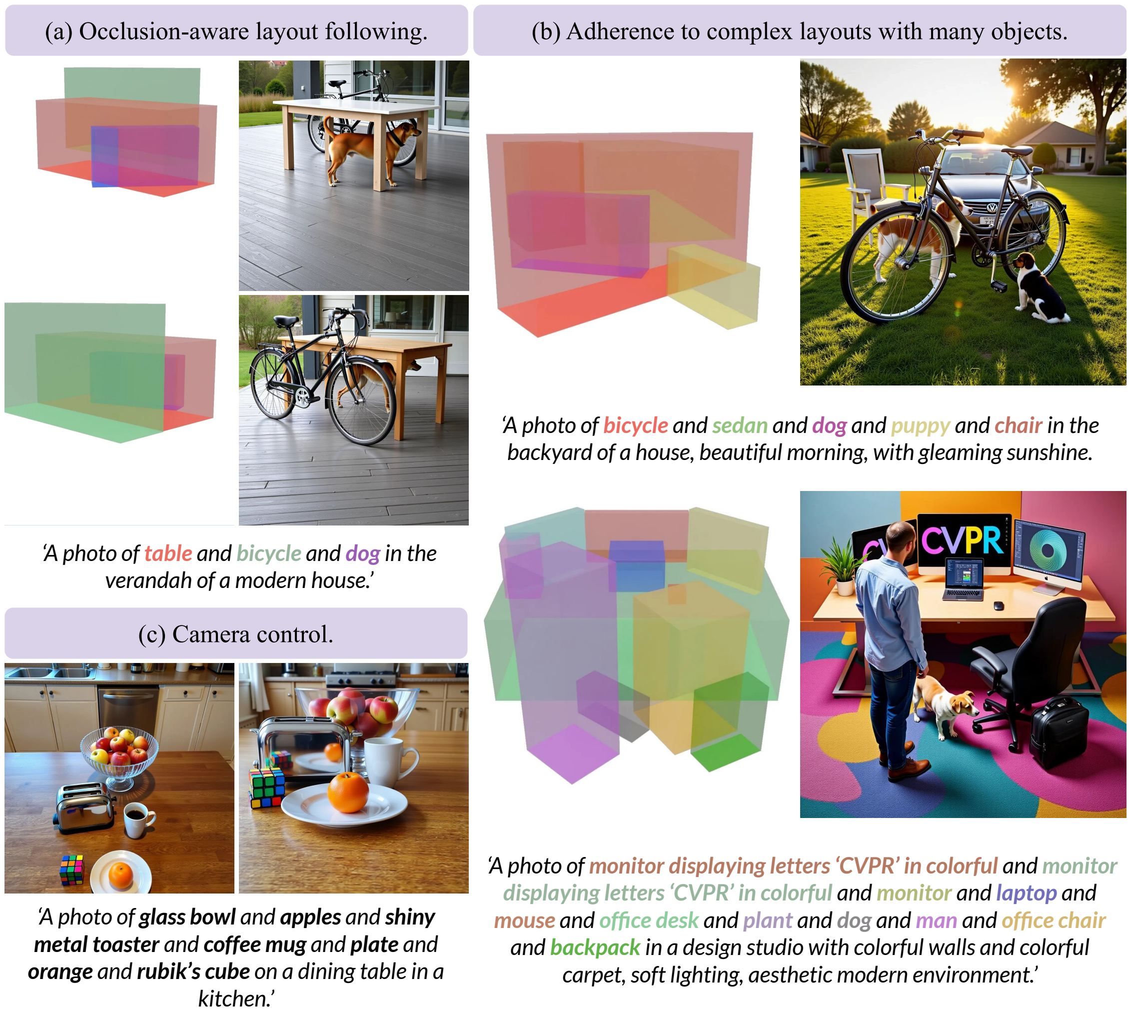


Figure 1. We propose **SeeThrough3D**, a method for occlusion aware 3D scene control in text-to-image generation. Our method enables (a) occlusion-aware 3D object placement in generated images, and (b) adheres well to complex layouts featuring many objects. Additionally, our method allows for (c) control over the camera viewpoint in the generated image.

Abstract

*We identify occlusion reasoning as a fundamental yet overlooked aspect for 3D layout–conditioned generation. It is essential for synthesizing partially occluded objects with depth-consistent geometry and scale. While existing methods can generate realistic scenes that follow input layouts, they often fail to model precise inter-object occlusions. We propose **SeeThrough3D**, a model for 3D layout conditioned generation that explicitly models occlusions. We introduce an occlusion-aware 3D scene representation (OSCR), where objects are depicted as translucent 3D boxes placed within a virtual environment and rendered from desired camera viewpoint. The transparency encodes hidden object regions, enabling the model to reason about*

†Equal advising

occlusions, while the rendered viewpoint provides explicit camera control during generation. We condition a pre-trained flow based text-to-image image generation model by introducing a set of visual tokens derived from our rendered 3D representation. Furthermore, we apply masked self-attention to accurately bind each object bounding box to its corresponding textual description, enabling accurate generation of multiple objects without object attribute mixing. To train the model, we construct a synthetic dataset with diverse multi-object scenes with strong inter-object occlusions. SeeThrough3D generalizes effectively to unseen object categories and enables precise 3D layout control with realistic occlusions and consistent camera control. Project page: `https://seethrough3d.github.io`

1. Introduction

Recent work has introduced various forms of controllability in text-to-image generation, but most methods remain limited to 2D spatial controls, such as bounding boxes or segmentation maps [23, 44, 48, 61, 62, 77, 79]. While effective for coarse control over the scene content, they offer limited control over inherently 3D scene properties, including object arrangement and camera viewpoint. Yet many practical content-creation domains such as design, gaming, and architectural visualization require precise 3D layout control, where object size, orientation, and placement must be explicitly specified. Critically, a truly 3D-aware generative model must also reason about occlusions, generate partially hidden objects with depth-consistent scale and perspective; a fundamental capability that 2D controls cannot provide.

Despite being fundamental to accurate 3D-aware generation, occlusion has been largely overlooked in recent 3D layout based methods. Existing approaches condition the generative model on depth maps derived from 3D bounding-box layouts [4, 65] or on explicit 3D attributes such as object or camera poses [7, 9, 43, 49, 55]. These methods succeed in generating simple scenes with few objects and minimal occlusion, but fail to model significant inter-object occlusions in multi-object layouts (Fig. 3(a)). A related direction represents scenes as a stack of 2D object layers [37, 76] to approximate occlusion, but this collapses the inherently 3D structure of the scene into flat planes (Fig. 3(c)), leading to generating object occlusion that violate true 3D geometry and perspective.

In this paper, we propose *SeeThrough3D* - an image generation model that takes 3D layout and text prompt as input and generates scenes with 3D consistent occlusions (Fig. 1). We introduce an efficient and expressive 3D scene representation, termed **O**cclusion-Aware 3D **Sc**ene **R**epresentation (**OSCR**), which jointly encodes object arrangements and camera viewpoint (Fig. 2). In OSCR, each object is modeled as a translucent 3D bounding box, where transparency reveals occluded regions, enabling explicit reasoning about inter-object occlusions. Faces of each box are further color-coded according to a predefined mapping to capture 3D object orientation. The final OSCR representation is obtained by rendering this layout from a specified camera viewpoint.

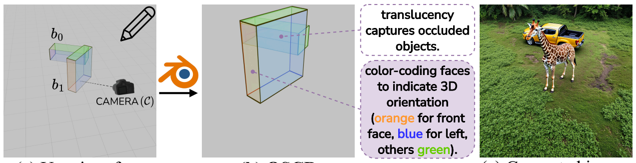
We build on FLUX [35] image generator, conditioning it on our OSCR scene representation. Following the success of recent works [35, 61] on controlling the diffusion transformer (DiT) [21, 54] using condition image tokens, we condition the model with tokens derived from our rendered scene representation. However, spatial conditioning alone fails to associate textual object descriptions with their corresponding box regions. To address this, we apply attention masking to bind each object to its corresponding box, ensuring accurate bounding box adherence for individual objects. Further, we extend this framework to allow 3D control of personalized objects, by conditioning on an image of the object, and binding its appearance to specific box in the OSCR representation.

To train SeeThrough3D, we create a synthetic dataset of scenes by placing diverse 3D assets in a virtual environment [12] and rendering scenes from multiple camera views. Object placement and camera parameters are controlled to induce strong inter-object occlusions in the rendered images. Despite being trained on synthetic data, SeeThrough3D generalizes well to unseen objects, backgrounds and complex scene layouts (see Fig. 1), evaluated qualitatively and through metrics, as well as a user study.

2. Related work

3D control in text-to-image generation: Previous works on 3D control in image generation trains specialized generative models conditioned on various 3D representations [2, 28, 29, 45, 46, 64, 71]. Interestingly, recent works have shown that there is inherent 3D understanding in large text-to-image diffusion models [17, 18, 74]. Several works leverage this insight for enabling precise 3D aware control in generated images [3, 7, 7, 10, 16, 20, 24, 34, 38, 57]. One line of works enable 3D aware editing [47, 57, 66] using scene depth as additional input, but they are limited to manipulation of a single object at a time. Further, a recent work [51] decomposes a scene into depth-based layers, enabling depth-aware editing and scene composition. Others train implicit 3D representations such as radiance fields [33, 53, 73] or 3D Gaussian splats [8, 31, 40, 68, 78] in diffusion feature space to enable 3D aware image editing.

3D layout conditioned generation: Apart from editing, controlling the 3D layout of a scene during generation is an active research area. A recent work for layout-conditioned generation, LooseControl [4] conditions a text-to-image model using depth maps of 3D bounding boxes; however it fails to generate complex scenes with diverse objects. A follow-up work, Build-A-Scene [19] generates the scene using multiple generation-inversion cycles, each iteration



(a) User interface **(b) OSCR** (c) Generated image

Figure 2. **OSCR:** We propose Occlusion-Aware Scene Representation (OSCR) for 3D layout control in text-to-image generation. OSCR describes objects as translucent 3D boxes, which exposes occluded regions, enabling the generative model to reason about occlusions. Further, each box face is color-coded with a mapping to encode its 3D orientation. (a) A user specifies the object bounding boxes (b_0 and b_1) and sets desired viewpoint $\mathcal{C}$ in an interactive graphic environment. (b) These boxes are rendered to obtain our OSCR representation, (c) which is used to condition the generation for occlusion aware 3D control.

adding a new object. However, this leads to inversion artifacts and incoherence in generated images. Another set of works provide partial control over individual 3D properties, such as object orientation [9, 43, 49], but they are limited in their extent to precisely control object placement or camera viewpoint. Another promising direction for 3D layout control is to represent the object bounding box as a set and condition the generative model using a learnable adapter [41, 50, 70]. However, they are limited to a single data domain, e.g. road scenes or indoor scenes, and are less effective than spatial conditioning approaches [55].

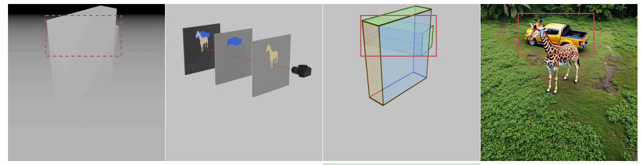
Occlusion awareness: Inter-object occlusions present a significant challenge in perception [22, 27, 32, 36, 42, 60, 75] and generation [37, 39, 49, 69, 76] tasks. Occlusions are particularly important for 3D aware image generation. However, it has received little attention in existing works [4, 19]. Some works model occlusions by decomposing images into flat 2D object layers [14, 37, 76], but they lack 3D awareness, resulting in geometrically inconsistent occlusions. To bridge the gap in existing works, we propose SeeThrough3D, a model that enables generalized occlusion-aware 3D layout control.

3. Method

Our goal is to generate an image conditioned on a text prompt and a scene layout consisting of 3D bounding boxes. We build on a pretrained text-to-image flow model [35] and condition on the proposed Occlusion-Aware 3D Scene Representation (OSCR) (see Fig. 2).

3.1. OSCR

Existing methods for 3D layout–conditioned generation represent scene layouts either by computing depth maps of 3D bounding boxes (see Fig. 3(a)) or by simplifying the scene into a finite set of 2D object layers (Fig. 3(b)). These representations, however, fail to capture true 3D structure of the scene, resulting in inaccurate occlusion modeling and



(a) **Layout depth maps** ✗ Occlusion ✓ 3D control | (b) **Object-based layering** ✓ Occlusion ✗ 3D control | (c) ***Ours:* Translucent boxes** ✓ Occlusion ✓ 3D control | (d) Generated Image

Figure 3. **Towards occlusion aware 3D scene layouts:** existing methods represent scenes as (a) 3D layout depth maps [4, 19, 65], which fail to represent occluded objects (see dashed red box), or (b) object layers [37, 76], which are not 3D aware, hence fail to capture camera viewpoint and perspective. (c) Therefore, we propose OSCR, where objects are described using translucent 3D bounding boxes. The transparency exposes occluded regions (red box), providing cues for occlusion reasoning, while enabling 3D layout control.

limited orientation control. To overcome this, we design OSCR, an efficient yet effective representation that encodes 3D layouts in an occlusion-aware manner.

Our input is a set of 3D bounding boxes b_i, each representing an object, arranged in a 3D virtual environment (see Fig. 2(a)). To encode object orientation, we define a canonical color mapping across box faces, where each face is assigned a predefined color (see Fig. 2(b)). This mapping provides an explicit and interpretable encoding of 3D orientation directly in image space. To make the representation aware of spatial ordering and occlusions, we render the boxes as translucent, allowing occluded objects to remain partially visible. This simple yet expressive design compactly captures both orientation and occlusion cues (see Fig. 2(b)). Notably, occlusion may alter the apparent colors of some faces, causing them to deviate from the predefined mapping. However, the relative color differences between faces remain discernible, preserving reliable orientation cues. Finally, we render the composed scene from a specified camera view $\mathcal{C}$ using Blender [12]. The rendered image inherently embeds camera pose information, enabling precise viewpoint control in generation. The rendered image r is used as 'OSCR condition' to the generative model (see Fig. 4).

3.2. SeeThrough3D

We build on FLUX [35], a DiT-based text-to-image model. FLUX comprises a series of multimodal DiT blocks that jointly process text and image tokens through self-attention and feed-forward layers (see Fig. 4). This architecture facilitates rich information exchange between text and image tokens, resulting in strong image-text alignment during generation. Further, this design naturally supports an effective way to condition the model on a new modality by adding condition tokens [61, 62, 79]. Leveraging this, we condi-

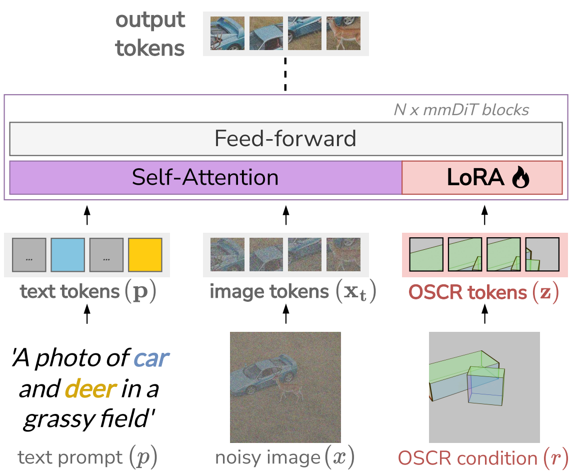


Figure 4. **SeeThrough3D:** We encode the rendered OSCR condition map r using the VAE to obtain OSCR tokens. These are concatenated with text prompt tokens $\mathbf{p}$ and noisy image tokens $\mathbf{x}_t$. The concatenated result is passed through the DiT based text-to-image model where they are jointly processed using self attention modules. We inject LoRA [25] onto the attention projections corresponding to OSCR tokens; this enables control while preserving prior of the base model [61, 62, 79].

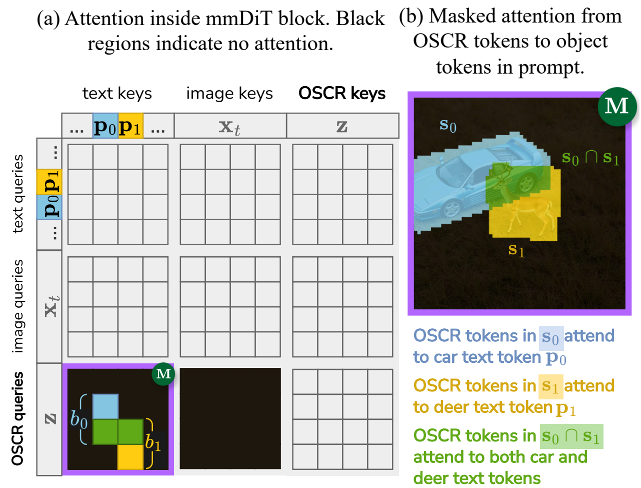


Figure 5. (a) Inside the mmDiT block, text tokens $\mathbf{p}$, image tokens $\mathbf{x}_t$ and OSCR tokens $\mathbf{z}$ are jointly processed using self attention, conditioning the generation on our OSCR representation. To bind objects to corresponding boxes, we mask the attention to enable OSCR tokens within each box $\{b_i\}$ to attend to corresponding object tokens $\{\mathbf{p}_i\}$ using a mask (M) (b) For this, we require spatial extent for each object box b_i, which we obtain we use its amodal segmentation mask $\mathbf{s}_i$. When multiple boxes overlap, their region of intersection (green) attends to multiple objects.

tion the model on the rendered OSCR layout representation r (see Fig. 4). Specifically, we first encode r using the VAE to obtain OSCR tokens $\mathbf{z}$, which are concatenated with text prompt tokens $\mathbf{p}$ and the noisy image tokens $\mathbf{x_t}$. The OSCR tokens $\mathbf{z}$ are assigned the same positional encodings as the noisy image tokens $\mathbf{x}_t$, establishing spatial correspondence between them. The combined token sequence is then processed by mmDiT blocks. To adapt the model to OSCR condition while preserving its text-to-image prior, we train a LoRA [25] only on the projection matrices associated with the newly added tokens (see Fig. 4). In line with recent work [79], we also block attention from OSCR tokens $\mathbf{z}$ to the image tokens $\mathbf{x}_t$ (see Fig. 5).

3.3. Object binding with attention masking

While the conditioning mechanism described above ensures spatial alignment with the given layout, it does not explicitly associate 3D bounding boxes with their corresponding object identities. This ambiguity arises because OSCR encodes geometric arrangements of objects but lacks semantic information about them, which can lead to mismatched object placements during generation. A straightforward solution would be to encode object classes as colors within the boxes, similar to semantic segmentation. However, this approach constrains the model to a fixed set of predefined categories and limits generalization. Instead, we utilize the attention mechanism to enrich OSCR tokens with corresponding object semantics. Specifically, we mask the attention so that OSCR tokens $\mathbf{z}$ within each bounding box only attend to corresponding object noun tokens $\mathbf{p}_i$ in the text prompt, (see Fig. 5(a) (M)), thus enriching the spatial OSCR tokens with corresponding object semantics. For this, we require the spatial extents for each box b_i, which we obtain using its rendered segmentation mask , (see Fig. 5(b)) using Blender.

Handling overlapping objects: A challenging case for the proposed object binding arises when the rendered regions of two boxes significantly overlap. In this scenario, the OSCR tokens in the intersection region attend to multiple object tokens (see Fig. 5(b)). At first glance, it appears that attending to multiple objects would lead to semantic blending or visual artifacts at object boundaries. To investigate this, we condition our model on a complex layout with heavy occlusion (see Fig. 6(a)), and observe that the output contains precise occlusion boundaries (see Fig. 6(b)). To understand this further, we visualize attention from image tokens $\mathbf{x}_t$ to object tokens $\{\mathbf{p}_i\}$ in Fig. 6(c,d) Interestingly, the attention maps themselves reveal occlusion boundaries: inside the empty regions of the bicycle structure, attention on the van remains visible, accurately reflecting its presence behind the bicycle. This indicates that object-specific features remain distinct in the model's latent space, and that the text-to-image model encodes necessary priors for occlusion reasoning. Our OSCR representation (see Fig. 6(a)) leverages these priors for precise control over scene layout, in an occlusion-aware manner. Further analysis of attention is provided in appendix Sec. D.

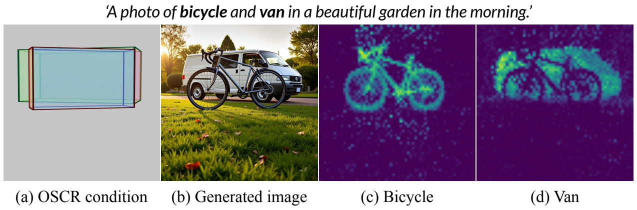


Figure 6. **Visualizing object disentanglement in latent space:** Given a layout with heavy occlusion like (a), our model's outputs show precise occlusion boundaries (b). To understand this, we visualize attention from image-tokens to object tokens in prompt (bicycle and van). Interestingly, the attention maps themselves reveal occlusion boundaries: inside the empty regions of the bicycle structure, attention on the van remains visible, accurately reflecting its presence behind the bicycle. This suggests that object-specific features remain distinct in the model's latent space, indicating strong priors for occlusion reasoning.

3.4. Personalization

The proposed method naturally supports layout-conditioned generation with personalized objects. Given a reference object image v, a text prompt p, and OSCR layout r, the goal is to generate the object adhering to a specific 3D bounding box b_i in the layout r. We first encode object appearance by passing the reference image v through the VAE encoder, resulting in 'appearance tokens' $\mathbf{v}$. These are concatenated with text tokens $\mathbf{p}$, target image tokens $\mathbf{x}_t$, and OSCR tokens $\mathbf{z}$ before passing through the mmDiT blocks. To bind the object's appearance to its corresponding 3D box b_i, we re-use the attention masking strategy described above. Specifically, we enable OSCR tokens inside the segmentation mask $\mathbf{s}_i$ to attend to appearance tokens $\mathbf{v}$. This enables layout-aware generation of personal objects, and can be extended to multiple objects by adding separate appearance token sets for each reference image (see Fig. 10).

3.5. Dataset

To adapt the model to OSCR representation, we require a dataset of paired images and 3D bounding boxes. While existing 3D object detection datasets [13, 58] could be used, they are often domain specific, lack occlusion scenarios, have minimal viewpoint variation and contain marginal errors in 3D annotations, making them unsuitable for our purposes. Therefore, we create a synthetic dataset using Blender [12]; where we procedurally place 3D assets in controlled configurations on the floor (x-y plane). Next, we render the paired ground truth image and OSCR representation from diverse camera viewpoints. We discard trivial scenes with minimal object overlap or very low visibility of any object, as we find such filtering crucial for maintaining occlusion consistency in the generated results (see Sec. 4.4).

Augmentations: Training solely on rendered images risks overfitting to synthetic backgrounds [9, 49], due to limited realism and lack of diversity in object appearance and backgrounds. Since creating highly varied 3D scenes is an expensive process, we adopt a scalable alternative. We generate realistic augmentations for the rendered images, that follow the same layout but are rich in terms of appearance diversity. For each rendered image, we extract its depth and feed it through a depth-to-image generation pipeline (FLUX.1-Depth-dev) [35] to synthesize realistic images that preserve the same spatial layout. Although this pipeline produces high-quality results, it occasionally misaligns objects with their intended depth regions, causing incorrect placements. We mitigate this by applying object-level CLIP-based filtering [56] to retain only those augmentations that adhere to the original layout. Our final dataset comprises $25K$ rendered images and $25K$ augmentations. Further details about dataset pipeline and dataset statistics are provided in appendix Sec. B.

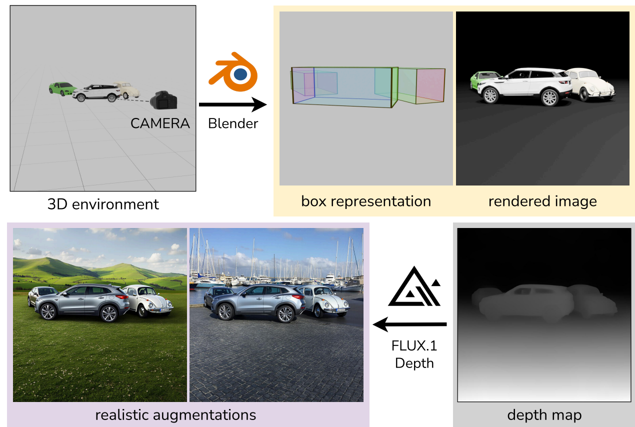


Figure 7. **Dataset creation:** We place 3D assets in controlled configurations in Blender [12]. Object placements and camera viewpoint are controlled to ensure strong occlusions, while ensuring adequate visibility for each object. To generate realistic augmentations, we estimate image depth, and pass it through a depth-to-image model [35] with diverse background prompts.

4. Experiments

4.1. Experimental setup

Implementation details: We use FLUX.1-dev [35] as the text-to-image model. We train for $30K$ steps at a learning rate of 10^{-4}, using a LoRA rank of 128. A detailed implementation report can be found in appendix Sec. E.

Evaluation dataset: Accurate evaluation of occlusion-aware 3D control requires a benchmark of paired images and 3D bounding box annotations that exhibit 1) diverse object configurations 2) challenging occlusion scenarios, and 3) wide range of camera viewpoints. To facilitate this, we introduce **3D** Control with **Oc**clusions benchmark, *3DOcBench*, a dataset with 500 samples of paired 3D bounding-box layouts, rendered images, and scene text prompts. We construct the benchmark in Blender [12] by placing 3D assets on a ground plane and procedurally varying object arrangements and camera poses to produce strong occlu-

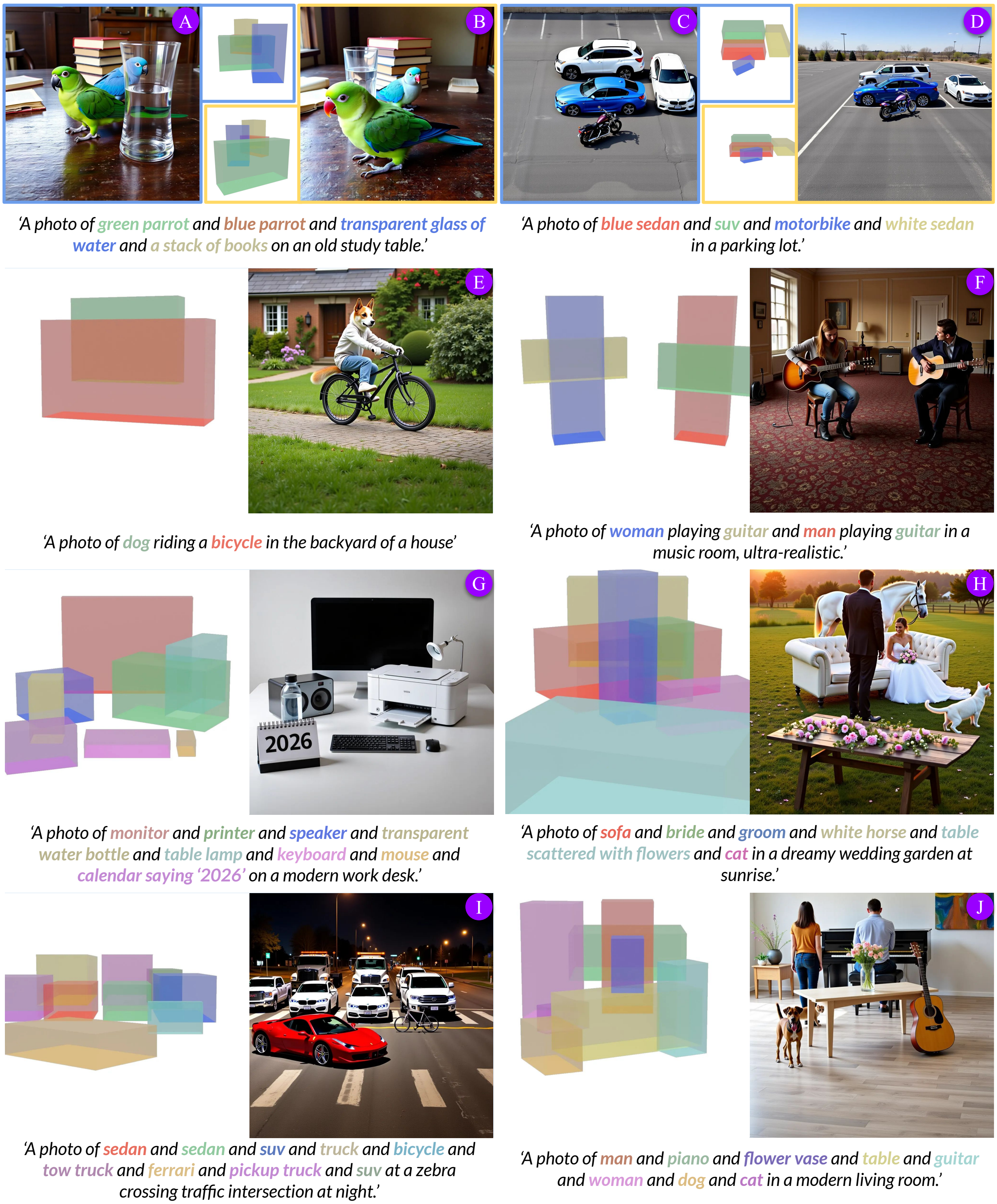


Figure 8. **Qualitative results:** Our method is able to precisely follow 3D scene layouts, with high occlusion consistency. Our approach preserves the prior of text-to-image model, as evident from capabilities like see-through transparent objects (A,B,G,J), text rendering (G) and inter-object interactions (E,F). Additionally, our method enables control over viewpoint of generated image (C,D). Despite being trained on layouts with only upto 4 objects, our method is able to generalize to complex scenes with many objects (G,H,I,J).

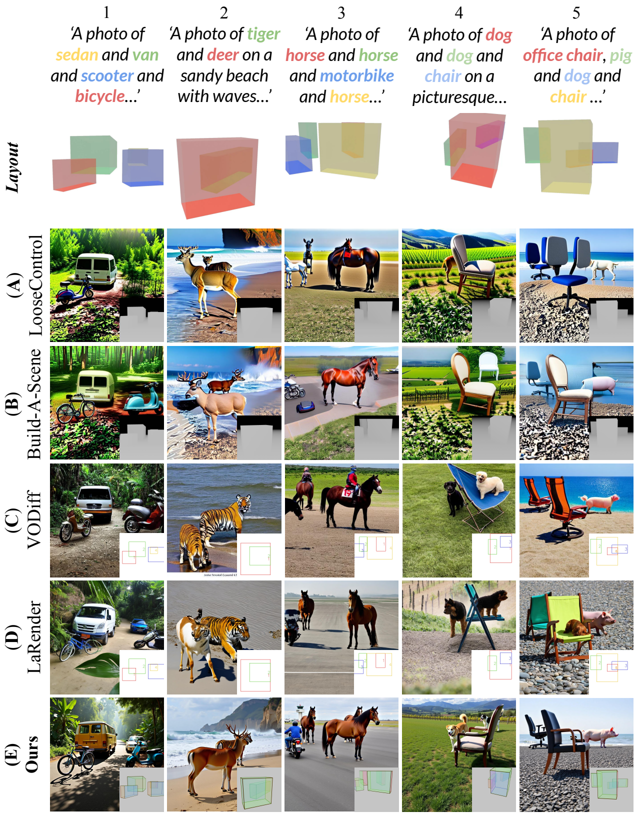


Figure 9. **Qualitative comparison:** We compare against works on **3D layout control**: *LooseControl* [4] and *Build-A-Scene* [19], and on **occlusion control**: *LaRender* [76] and *VODiff* [37].

Baselines	depth ord.↑	obj. score↑	angular err.↓	text align.↑	KID($\times 10^{-3}$) ↓
VODiff [37]	0.68	19.70	92.73	29.51	15.40
LooseControl [4]	0.82	20.02	89.88	28.43	14.32
Build-A-Scene [19]	0.89	21.0	91.62	28.05	20.12
LaRender [76]	1.02	21.83	89.63	30.20	13.46
Ours	**1.46**	**22.86**	**47.92**	**31.87**	**5.43**

Table 1. **Quantitative comparison:** We compute (a) depth ordering, which reflects 3D location and occlusion consistency, (b) CLIP *objectness* score, which indicates layout adherence and object fidelity (c) angular error, which indicates orientation correctness (d) image-text prompt alignment using CLIP [56], and (e) KID [5], which measures image fidelity.

sions while preserving a minimum visible area for each object. We will release the benchmark for future research in occlusion-aware generation. Detailed benchmark statistics are provided in the appendix Sec. B.

Evaluation metrics: We measure the models' performance for layout adherence, text-to-image alignment, and image quality. For text-to-image alignment, we use CLIP image-text similarity, and for image quality, we use Kernel Inception Distance (KID) [6]. Evaluating 3D layout adherence using a single metric is challenging, as the generated scene may not conform to the metric depth specified by the 3D bounding-box layout. To this end, we compute three metrics that in unison effectively evaluate 3D layout adherence. Specifically, we compute 2D bounding box adherence, relative visibility order and 3D orientation consistency. (1) For evaluating 2D layout adherence, we first obtain object masks by combining 2D layouts with Segment Anything [30]. Next, we compute CLIP similarity between the object masks and textual object descriptions, leading to CLIP *objectness* score. We aggregate this objectness score to evaluate the 2D layout adherence (2) For evaluating relative visibility order, we adopt a similar method as [26]: we estimate per-pixel depth [72] and obtain object depth estimates by averaging the depth within each object mask. Since all objects may not be present in the generated output, we use previously defined objectness score to filter out object masks. Finally, we compare relative depth ordering of each object pair against the ground-truth ordering, assigning a score of 1 if the ordering is correct and 0 otherwise. We aggregate this score over all such pairs 3) For assessing orientation accuracy, we employ OrientAnything [67] to estimate object orientations using filtered object segments, and compute mean absolute error against ground truth.

Baselines: We compare our method with state-of-the-art works in **3D layout control**: *LooseControl* [4] and *Build-A-Scene* [19]. LooseControl uses layout depth maps to condition a diffusion model for scene layout control, while Build-A-Scene is an inference time method that uses pretrained LooseControl checkpoint. For fair evaluation, we train LooseControl on our dataset, and use the checkpoint to evaluate both methods. We also consider works on **orientation control**, *Compass Control* [49] and *ORIGEN* [43], though they do not support 3D object placement, hence not directly relevant. We compare against them in appendix Sec. G. We further evaluate against **occlusion control** methods, *LaRender* [76] and *VODiff* [37]. These methods decompose an image into 2D object layers to manage visibility ordering.

4.2. Results

Qualitative: We present our qualitative results in Fig. 8. Our method is able to generate realistic scenes with intricate inter-object overlaps. It effectively preserves the prior of the base text-to-image model, evident from capabilities like see-through transparent objects (A,B,G,J) and text rendering (G). Additionally, our method enables control over viewpoint of generated image (C,D). Despite being trained on layouts with only upto 4 objects, our method is able to generalize to complex scenes with many objects (G,H,I,J). Even though our synthetic data consists of rigid objects in fixed canonical poses, our method is able to generate diverse poses such as sitting (H,J) and cycling (E). The model generates natural inter-object interactions (dog *riding* bicycle in E, person *playing* guitar in F), even though our synthetic data does not contain such interactions. Further, it generalizes strongly to out-of-domain objects. Notably, our training dataset does not contain any musical instruments (F,J), electronic devices (G), transparent object (A,B,G,J), but our model effectively generalizes to them.

Baseline comparisons: We present results in Tab. 1 and Fig. 9. **3D scene control**: LooseControl [4] fails to handle complex occlusions, as layout depth fails to represent occluded objects (see Fig. 9 A1,3-5). Additionally, the objects are generated in incorrect locations, due to lack of binding (A1,3), also reflected in low objectness-score (see Tab. 1). Build-A-Scene [19] uses multiple generation and inversion cycles to sequentially add objects to the scene. While this improves upon layout adherence and occlusion consistency compared to LooseControl [4], it leads to inversion artifacts (B2-3,5), and hence worse KID value. The sequential generation also leads to lack of coherence in the generated scene (B4), since initial generations are independent of final scene layout. Both the methods fail to provide precise orientation control, since layout depth maps can only encode orientation upto 180° flip, leading to high angular error. In contrast, our method is able to generate coherent images with precise 3D layout and orientation control. **Occlusion control:** LaRender [76] and VODiff [37] rely on 2D layouts as conditioning input, which fail to discertain exact object arrangements. For instance, in Fig. 9 (C4, D4-5), the object is generated on '*top* of the chair', against the intended configuration '*behind* the chair'. In contrast, our OSCR representation is 3D aware, hence offers more precise control than 2D layouts. In case of large overlap between 2D bounding boxes, baselines often fail to generate occluded objects (C1,3-4, D3-4) while SeeThrough3D succeeds in such cases (E).

User study: We conducted an A/B user study where 60 participants were asked to choose between output of our method and a randomly chosen baseline. We evaluate a) image realism, b) layout adherence, and c) text prompt alignment. Results highlight high preference for our method in all evaluation categories (see Appendix Sec. J).

4.3. Personalization

We show personalization results in Fig. 10. We adapt our training dataset for personalization by applying textures to 3D assets, and using this textured object as reference image. Further details and results are in appendix Sec. I.

4.4. Ablations

We study the impact of key design choices, with results shown in Tab. 2 and Fig. 11. Box transparency plays a crucial role in the effectiveness of the OSCR representation, enabling reasoning about occluded objects and relative depth. Color-coding the box faces helps encode orientation and significantly reduces angular error (see Tab. 2). Interestingly, opaque boxes yield the best orientation accuracy due to a clearer color signal. The attention-based binding is essential for layout adherence—without it, objects appear at incorrect locations (see Fig. 11, 1C and 3C), resulting in lower objectness score. Finally, filtering out overly simplistic layouts in data improves performance.

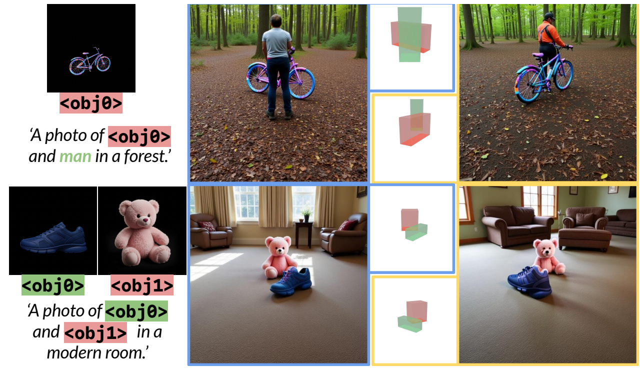


Figure 10. **Personalization:** Our method can be extended for personalized 3D control using reference image of an object.

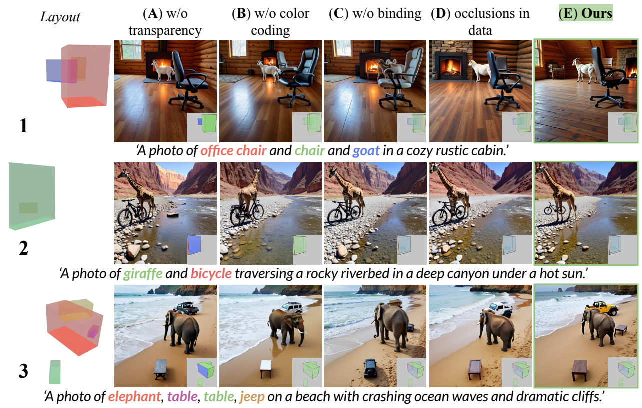


Figure 11. **Ablations:** We ablate upon key aspects of OSCR representation, our binding mechanism and data preparation strategy.

Ablations	depth ord.↑	obj. score↑	angular err.↓	text align.↑	KID($\times10^{-3}$) ↓
w/o transparency	1.20	21.67	**46.15**	31.39	5.90
w/o color-coding	1.36	22.23	88.77	31.57	5.93
w/o binding	0.98	20.45	57.44	31.61	6.35
w/o hard data	1.24	21.89	49.73	31.32	6.34
Ours	**1.46**	**22.86**	47.92	**31.87**	**5.43**

Table 2. **Quantitative results of ablative experiments.**

5. Conclusion

We present SeeThrough3D, a model for occlusion-aware 3D layout control. We introduce OSCR, an occlusion-aware 3D scene representation. We show that our approach can faithfully model heavy occlusion scenarios while preserving the strong text-to-image prior of the model. Despite training on limited synthetic data, it exhibits strong generalization capabilities. We perform evaluations to show that our method outperforms existing baselines, and also ablate upon key design choices, providing useful insights for future research. While effective in layout adherence, our method does not preserve image consistency under layout changes. A future direction is to address this by using editing based techniques.

6. Acknowledgements

We thank Harshavardhan P., Ayan Kashyap, Vansh Garg, Jainit Bafna, Abhinav Raundhal, Varun Gupta, Shivank Saxena, Akshat Sanghvi and Aishwarya Agarwal for helpful discussions and reviewing the manuscript. This work is supported by a grant from Google and Kotak IISc AI-ML Centre (KIAC).

References

[1] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7877–7888, 2025. 6

[2] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3D scene generation. *Advances in Neural Information Processing Systems*, 35:25102–25116, 2022. 2

[3] Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Geoffroy, Xin Sun, and Diego Gutierrez. Precisecam: Precise camera control for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2724–2733, 2025. 2

[4] 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. 2, 3, 7, 8, 5, 6, 12, 13

[5] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd GANs. *arXiv preprint arXiv:1801.01401*, 2018. 7

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

[7] James Burgess, Kuan-Chieh Wang, and Serena Yeung. Viewpoint textual inversion: Unleashing novel view synthesis with pretrained 2d diffusion models. *arXiv preprint arXiv:2309.07986*, 2023. 2

[8] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3D editing with gaussian splatting. corr abs/2311.14521 (2023), 2023. 2

[9] Ta-Ying Cheng, Matheus Gadelha, Thibault Groueix, Matthew Fisher, Radomir Mech, Andrew Markham, and Niki Trigoni. Learning continuous 3D words for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6753–6762, 2024. 2, 3, 5, 1

[10] Yen-Chi Cheng, Krishna Kumar Singh, Jae Shin Yoon, Alexander Schwing, Liang-Yan Gui, Matheus Gadelha, Paul Guerrero, and Nanxuan Zhao. 3D-fixup: Advancing photo editing with 3D priors. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. 2

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[12] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2, 3, 5, 1, 4, 9, 10

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[14] Aneel Damaraju, Dean Hazineh, and Todd Zickler. Cobl: Toward zero-shot ordinal layering without user prompting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8154–8164, 2025. 3

[15] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 1

[16] Ankit Dhiman, Manan Shah, Rishubh Parihar, Yash Bhalgat, Lokesh R Boregowda, and R Venkatesh Babu. Reflecting reality: Enabling diffusion models to produce faithful mirror reflections. *arXiv preprint arXiv:2409.14677*, 2024. 2

[17] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv preprint arXiv:2311.17137*, 2023. 2

[18] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 2

[19] Abdelrahman Eldesokey and Peter Wonka. Build-a-scene: Interactive 3D layout control for diffusion-based image generation. *arXiv preprint arXiv:2408.14819*, 2024. 2, 3, 7, 8, 5, 6, 12, 13

[20] Alejandro Escontrela, Shrinu Kushagra, Sjoerd van Steenkiste, Yulia Rubanova, Aleksander Holynski, Kelsey Allen, Kevin Murphy, and Thomas Kipf. Neural USD: An object-centric framework for iterative editing and control. *arXiv preprint arXiv:2510.23956*, 2025. 2

[21] 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*, 2024. 2

[22] Alhussein Fawzi and Pascal Frossard. Measuring the effect of nuisance variables on classifiers. In *BMVC*, pages 137–1, 2016. 3

[23] Tsu-Jui Fu, Yusu Qian, Chen Chen, Wenze Hu, Zhe Gan, and Yinfei Yang. Univg: A generalist diffusion model for unified image generation and editing. *arXiv preprint arXiv:2503.12652*, 2025. 2

[24] Richard Higgins and David Fouhey. Seldom: Scene editing via latent diffusion with object-centric modifications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 7046–7058, 2025. 2

[25] 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. *arXiv preprint arXiv:2106.09685*, 2021. 4

[26] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 7

[27] Kaleb Kassaw, Francesco Luzi, Leslie M Collins, and Jordan M Malof. Are deep learning models robust to partial object occlusion in visual recognition tasks? *Pattern Recognition*, page 112215, 2025. 3

[28] Kunal Kathare, Ankit Dhiman, K Vikas Gowda, Siddharth Aravindan, Shubham Monga, Basavaraja Shanthappa Vandrotti, and Lokesh R Boregowda. Instructive3D: Editing large reconstruction models with text instructions. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 3246–3256, 2025. 2

[29] Hyunsu Kim, Gayoung Lee, Yunjey Choi, Jin-Hwa Kim, and Jun-Yan Zhu. 3D-aware blending with generative nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22906–22918, 2023. 2

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 7, 3, 6

[31] Eunseo Koh, Sangeek Hyun, MinKyu Lee, Jiwoo Chung, Kangmin Seo, and Jae-Pil Heo. Diffusion feature field for text-based 3D editing with gaussian splatting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[32] Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, and Alan Yuille. Combining compositional models and deep networks for robust object classification under occlusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1333–1341, 2020. 3

[33] Nupur Kumari, Grace Su, Richard Zhang, Taesung Park, Eli Shechtman, and Jun-Yan Zhu. Customizing text-to-image diffusion with camera viewpoint control. *arXiv preprint arXiv:2404.12333*, 2024. 2

[34] Nupur Kumari, Grace Su, Richard Zhang, Taesung Park, Eli Shechtman, and Jun-Yan Zhu. Customizing text-to-image diffusion with camera viewpoint control. *arXiv preprint arXiv:2404.12333*, 2024. 2

[35] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2, 3, 5, 1, 4, 9, 10

[36] Zhenyu Li, Mykola Lavreniuk, Jian Shi, Shariq Farooq Bhat, and Peter Wonka. Amodal depth anything: Amodal depth estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9673–9682, 2025. 3

[37] Dong Liang, Jinyuan Jia, Yuhao Liu, Zhanghan Ke, Hongbo Fu, and Rynson WH Lau. Vodiff: Controlling object visibility order in text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18379–18389, 2025. 2, 3, 7, 8, 6, 12, 13

[38] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[39] Zhengzhe Liu, Qing Liu, Chirui Chang, Jianming Zhang, Daniil Pakhomov, Haitian Zheng, Zhe Lin, Daniel Cohen-Or, and Chi-Wing Fu. Object-level scene deocclusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[40] Guan Luo, Tian-Xing Xu, Ying-Tian Liu, Xiao-Xiong Fan, Fang-Lue Zhang, and Song-Hai Zhang. 3D gaussian editing with a single image. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6627–6636, 2024. 2

[41] Léopold Maillard, Tom Durand, Adrien Ramanana Rahary, and Maks Ovsjanikov. Laconic: A 3D layout adapter for controllable image creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18046–18057, 2025. 3

[42] Rupayan Mallick, Sibo Dong, Nataniel Ruiz, and Sarah Adel Bargal. D-feat occlusions: Diffusion features for robustness to partial visual occlusions in object recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1722–1731, 2025. 3

[43] Yunhong Min, Daehyeon Choi, Kyeongmin Yeo, Jihyun Lee, and Minhyuk Sung. Origen: Zero-shot 3D orientation grounding in text-to-image generation. *arXiv preprint arXiv:2503.22194*, 2025. 2, 3, 7, 5, 6, 9

[44] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 2

[45] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 2

[46] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2

[47] 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*, pages 7695–7704, 2024. 2

[48] Rishubh Parihar, Harsh Gupta, Sachidanand VS, and R Venkatesh Babu. Text2place: Affordance-aware text guided human placement. In *European Conference on Computer Vision*, pages 57–77. Springer, 2024. 2

[49] Rishubh Parihar, Vaibhav Agrawal, Sachidanand VS, and Venkatesh Babu Radhakrishnan. Compass control: Multi object orientation control for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2791–2801, 2025. 2, 3, 5, 7, 1, 6, 9

[50] Rishubh Parihar, Srinjay Sarkar, Sarthak Vora, Jogendra Nath Kundu, and R Venkatesh Babu. Monoplace3D: Learning 3D-aware object placement for 3D monocular detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6531–6541, 2025. 3

[51] Rishubh Parihar, Sachidanand VS, and R Venkatesh Babu. Zero-shot depth aware image editing with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15748–15759, 2025. 2

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 4

[53] Or Patashnik, Rinon Gal, Daniel Cohen-Or, Jun-Yan Zhu, and Fernando De la Torre. Consolidating attention features for multi-view image editing. *arXiv preprint arXiv:2402.14792*, 2024. 2

[54] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[55] Zhenyuan Qin, Xincheng Shuai, and Henghui Ding. Scenedesigner: Controllable multi-object image generation with 9-DoF pose manipulation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2, 3

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 5, 7, 1, 2

[57] Rahul Sajnani, Jeroen Vanbaar, Jie Min, Kapil Katyal, and Srinath Sridhar. Geodiffuser: Geometry-based image editing with diffusion models. *arXiv preprint arXiv:2404.14403*, 2024. 2

[58] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 5

[59] Florian Spiess, Raphael WaltenspÃžl, and Heiko Schuldt. The sketchfab 3D creative commons collection (s3D3c). *arXiv preprint arXiv:2407.17205*, 2024. 1

[60] Wei-En Tai, Yu-Lin Shih, Cheng Sun, Yu-Chiang Frank Wang, and Hwann-Tzong Chen. Segment anything, even occluded. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29385–29394, 2025. 3

[61] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024. 2, 3, 4, 9

[62] Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280*, 2025. 2, 3, 4, 9

[63] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 4

[64] Qian Wang, Yiqun Wang, Michael Birsak, and Peter Wonka. BlobGAN-3D: A spatially-disentangled 3D-aware generative model for indoor scenes. *arXiv preprint arXiv:2303.14706*, 2023. 2

[65] Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3D-aware and controllable framework for cinematic text-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. 2, 3

[66] Ruicheng Wang, Jianfeng Xiang, Jiaolong Yang, and Xin Tong. Diffusion models are geometry critics: Single image 3D editing using pre-trained diffusion priors. *arXiv preprint arXiv:2403.11503*, 2024. 2

[67] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3D models. *arXiv preprint arXiv:2412.18605*, 2024. 7, 5

[68] Minghao Wen, Shengjie Wu, Kangkan Wang, and Dong Liang. Intergsedit: Interactive 3D gaussian splatting editing with 3D geometry-consistent attention prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26136–26145, 2025. 2

[69] Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3D reconstruction from occluded 2d images. *arXiv preprint arXiv:2503.13439*, 2025. 3

[70] 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. *arXiv preprint arXiv:2406.09292*, 2024. 3

[71] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3D-aware generative model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18440–18449, 2022. 2

[72] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 7

[73] Jiraphon Yenphraphai, Xichen Pan, Sainan Liu, Daniele Panozzo, and Saining Xie. Image sculpting: Precise object editing with 3D geometry control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4241–4251, 2024. 2

[74] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. What does stable diffusion know about the 3D scene? *arXiv preprint arXiv:2310.06836*, 2023. 2

[75] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28003–28013, 2024. 3

[76] Xiaohang Zhan and Dingming Liu. Larender: Training-free occlusion control in image generation via latent rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19679–19688, 2025. 2, 3, 7, 8, 6, 12, 13

[77] 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*, pages 3836–3847, 2023. 2

[78] Qihang Zhang, Yinghao Xu, Chaoyang Wang, Hsin-Ying Lee, Gordon Wetzstein, Bolei Zhou, and Ceyuan Yang. 3Ditscene: Editing any scene via language-guided disentangled gaussian splatting. *arXiv preprint arXiv:2405.18424*, 2024. 2

[79] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025. 2, 3, 4, 9