Placing Any Object at Any 3D Position

Junhao Zhang[1], **Ming Kong**[1], **Zhanbin Hu**[2], **Hao Qin**[1], **Zhijie Xu**[3], **Xiaojun Zhu**[1], **Qiang Zhu**[1, *]

[1]Zhejiang University
[2]Beijing Information Science and Technology University
[3]University of Michigan - Ann Arbor
[1]{junhao.zhang, zjukongming, 12321248, zhuq, Zhuxiaojun}@zju.edu.cn, [2]zhanbinhu@bistu.edu.cn, [3]zhijiexu@umich.edu

Abstract

In this work, we propose a diffusion-based method for 3D-aware image composition. Previous approaches have focused on 2D-view image composition, which limits their handling of complex 3D spatial relationships. Consequently, they are not well-suited for applications requiring precise 3D object control and iterative refinement, including interior design visualization, visual effects prototyping, and virtual reality scene construction. In contrast, our method extracts 3D bounding boxes for all objects in the scene image. Users can then specify a new 3D bounding box based on existing spatial context and provide an image of the target object. Leveraging a fine-tuned diffusion model, our approach enables high-fidelity image composition while preserving the underlying 3D structure of the scene.

Video — https://youtu.be/5sURhdPQBds

Sliders — https://github.com/Orzjh/Placing-Any-Object-at-Any-3D-Position

Introduction

In recent years, the development of diffusion models has significantly advanced visual content creation (Ho, Jain, and Abbeel 2020; Rombach et al. 2022), achieving impressive results in image generation and editing (Podell et al. 2023; Brooks, Holynski, and Efros 2023). Recent approaches have enhanced controllability by incorporating conditional signals such as text prompts, depth maps, and sketches (Zhang, Rao, and Agrawala 2023). However, existing image composition methods are fundamentally limited by their 2D nature. They often treat the task as a blending problem. Although some recent approaches have incorporated geometric cues like depth maps to improve coherence, most existing methods still lack an explicit, object-level understanding of the 3D layout. This prevents precise control over an object's 3D position and orientation, leading to failures in complex occlusion scenarios and making it difficult for users to iteratively adjust and refine the object's placement until satisfied. The lack of explicit 3D reasoning in prior methods (Zhao 2024; Ye et al. 2023; Mao et al. 2025) often leads to unrealistic composites with distorted objects, incorrect depth ordering, and sticker-like artifacts, as shown in Figure 1. Our work directly addresses this core limitation by introducing a 3D-aware framework. By conditioning the synthesis on a user-defined 3D bounding box, we provide explicit spatial constraints that ensure that the final result is not only visually harmonious but also geometrically coherent.

We propose a 3D-aware image composition framework that leverages 3D object layouts without requiring explicit 3D modeling. Specifically, our method extracts 3D bounding boxes of all objects in the scene, allows users to specify the new object's 3D position, orientation and scale, and uses a fine-tuned diffusion model to generate high-fidelity composites guided by both image and depth. This design improves spatial realism, supports interactive placement control, and is suitable for AI-assisted editing and virtual scene construction.

Figure 1: Comparison of our 3D-aware composition with previous methods. Using a user-defined 3D bounding box, our approach preserves depth and occlusion and avoids sticker-like artifacts.

Our Method

As illustrated in Figure 2, our method follows a two-stage pipeline: 3D-aware object detection and 3D-aware object placement.

Stage 1: 3D-aware Object Detection. In this stage, we first use a vision-language model (VLM) built on InternVL2.5 (Chen et al. 2024) to perform semantic parsing on the input scene and extract candidate object categories. Guided by these category priors, a monocular 3D detector (Yao et al. 2024) fuses cues from Grounding DINO (Liu et al. 2024), DepthPro (Bochkovskii et al. 2024), and SAM (Kirillov et al. 2023) to back-project into 3D, yielding object-level 3D bounding boxes for all visible objects and the scene's initial geometry.

Stage 2: 3D-aware Object Placement. In this stage, the user-specified 3D location for the target object is merged

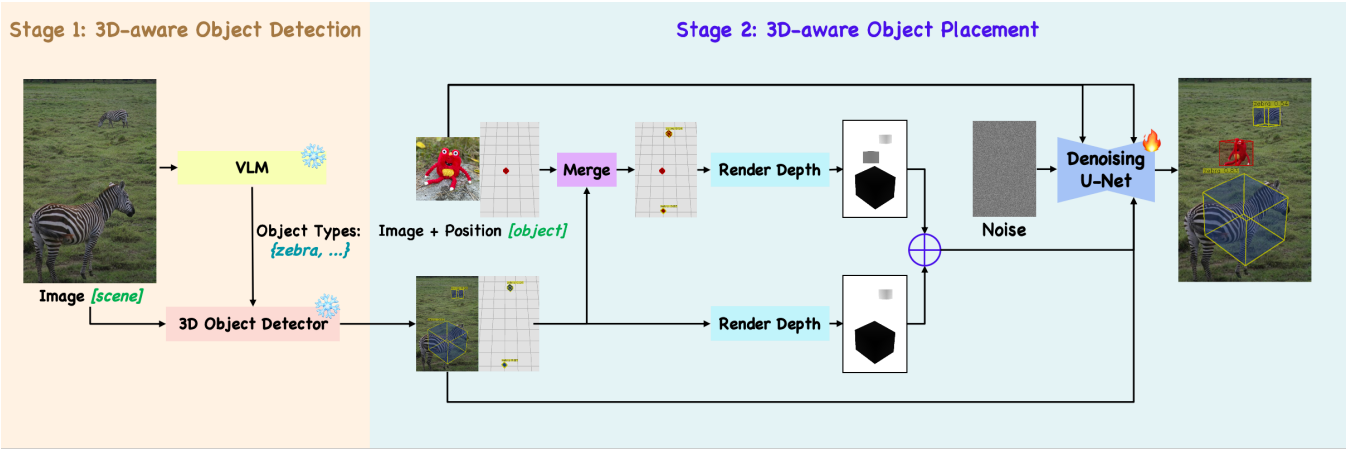

Figure 2: **Overview of the proposed 3D-aware image composition process.** In stage 1, the system performs 3D-aware object detection using a vision-language model for semantic parsing, followed by monocular 3D object detection. In stage 2, the system focuses on 3D-aware object placement, where the target object is positioned in 3D space. By using the scene image, the object image and its 3D position, the model can generate a high-quality composite image.

with the previously detected scene-level 3D bounding boxes to form an updated layout. The system then renders both the original and updated layouts into two depth maps, which are used as geometric constraints to guide the subsequent image synthesis process. We extract features from the rendered depth maps, the input scene image, and the reference object image, and inject them into a fine-tuned diffusion model to generate the final composition, ensuring spatial coherence and seamless foreground-background integration. Specifically, we first synthesize the target object's desired viewpoint using Zero123 (Liu et al. 2023). An ID extractor (DINOv2 backbone with a fully connected layer) then processes this view for object-specific ID tokens, while a high-pass filter captures its fine-grained texture. These texture details, concatenated with the scene image, are fed to a ControlNet-style detail extractor. Similarly, a ControlNet-style depth extractor processes the concatenated rendered depth maps. Finally, a fine-tuned denoising U-Net is conditioned on these ID tokens, detail features, and depth features to iteratively synthesize the composite image, ensuring 3D consistency, object identity, and scene coherence.

Interactive Workflow

As illustrated in Figure 3, our interactive workflow involves two main stages: **3D-aware Object Detection** and **3D-aware Object Placement**. In the first stage, 3D-aware object detection, the user simply provides a single scene image as input. The system automatically analyzes the image, detects all existing objects in the scene, and visualizes their corresponding 3D bounding boxes along with their spatial positions in 3D space. The workflow then proceeds to the second stage, 3D-aware object placement. Here, the user can upload an image of a new object to be inserted into the scene, and specify its placement by adjusting the parameters of a 3D bounding box—such as position, rotation, and scale. Finally, based on the user-defined object and its placement, the system generates a high-quality composite image in which the inserted object is seamlessly integrated into the scene.

While our method is broadly applicable to indoor design visualization, VFX prototyping, and VR scene construction, its final output quality is naturally dependent on the accuracy of the underlying monocular 3D detector. In the future, we plan to package it as plugins for image editing software to enhance usability and explore incorporating generative relighting to better match the object with the scene's specific illumination conditions.

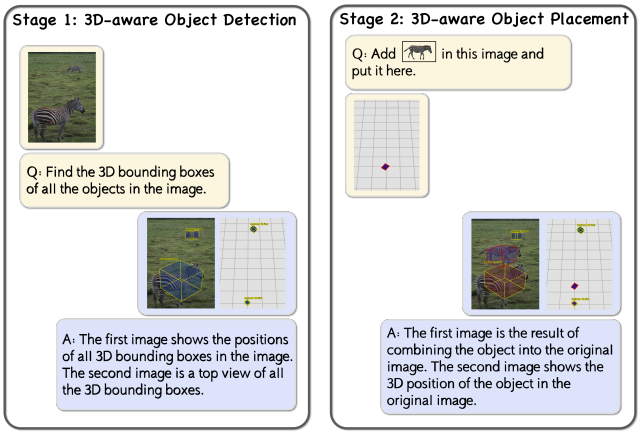

Figure 3: The 3D-aware image composition process mainly comprises two stages: 3D-aware object detection and 3D-aware object placement. In the first stage, the model detects all objects in the scene and provides the corresponding 3D bounding boxes. In the second stage, the model places the target object according to the user's specified 3D position and generates the composite image.

References

Bochkovskii, A.; Delaunoy, A.; Germain, H.; Santos, M.; Zhou, Y.; Richter, S. R.; and Koltun, V. 2024. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9298–9309.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55. Springer.

Mao, C.; Zhang, J.; Pan, Y.; Jiang, Z.; Han, Z.; Liu, Y.; and Zhou, J. 2025. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Yao, J.; Gu, H.; Chen, X.; Wang, J.; and Cheng, Z. 2024. Open Vocabulary Monocular 3D Object Detection. *arXiv preprint arXiv:2411.16833*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhao, H. 2024. AnyDoor: Zero-shot Object-level Image Customization. *Computer Vision and Pattern Recognition (CVPR), 2024 (17/06/2024-21/06/2024, Seattle, USA)*.