DreamScene: 3D Gaussian-based Text-to-3D Scene Generation via Formation Pattern Sampling Supplementary Materials

Haoran Li[1,2], Haolin Shi[1,2], Wenli Zhang[1,2], Wenjun Wu[1,2], Yong Liao[1,2]*, Lin Wang[3], Lik-Hang Lee[4], and Peng Yuan Zhou[5]*

[1] University of Science and Technology of China
[2] CCCD Key Lab of Ministry of Culture and Tourism
{lhr123, mar, zwl384369, wu_wen_jun}@mail.ustc.edu.cn, yliao@ustc.edu.cn
[3] AI Thrust, HKUST(GZ) and Dept. of Computer Science Eng., HKUST
linwang@ust.hk
[4] The Hong Kong Polytechnic University
lik-hang.lee@polyu.edu.hk
[5] Aarhus University
pengyuan.zhou@ece.au.dk

1 Discussions of Current Methods

Inpainting-based methods employ text-to-image inpainting for scene generation [2, 4, 7]. They first initialize an image and then partially mask it to represent a view from an alternate angle. Leveraging pretrained image inpainting models like Stable Diffusion [5], alongside depth estimation, they reconstruct the occluded image segments and deduce their depths, iteratively composing the entire scene through depth and image alignment. While such methods can achieve good visual effects at the camera positions (e.g., the center of the scene) during the generation process, they encounter significant limitations in visible range. Venturing beyond the predefined camera areas used during generation precipitates scene degradation, as exemplified in Fig. 4a of the paper, underscoring a deficiency in scene-wide 3D consistency. Concurrently, this method's tendency to replicate certain environmental objects, like the proliferation of sofas in a living room scenario, highlights issues with logical scene composition.

Combination-based methods also employ a combination approach to construct scenes [1, 8]. However, they face challenges including low generation quality and slow training speeds. Moreover, [8] utilizes various 3D representations (such as NeRF+DMTet) to integrate objects and scenes, increasing the complexity of scene representation and thus limiting the number of objects that can be placed within the scene (2-3 objects), impacting their applicability. In contrast, DreamScene's Formation Pattern Sampling (FPS) can generate high-quality 3D content in a very short time, using a single 3D representation to compose the entire scene, allowing for more than 20 objects to be placed within the scene. This underscores DreamScene's remarkable superiority.

* Corresponding authors

Objects combination methods do not take the environmental context into account, focusing solely on whether the combination of objects is logical [3, 6, 9]. They generate a simple assembly of objects rather than a complete scene. We believe that the approach to scene composition should be more diverse and offer greater controllability.

DreamScene demonstrates a significant advantage by efficiently, consistently, and flexibly generating 3D scenes, showcasing a substantial superiority over the aforementioned methods.

2 Additional Implementation Details

The overall generation process of DreamScene is shown in Algorithm 1.

2.1 Rendering and Training Settings

We render images with a size of 512×512 for optimization. During the optimization process, we do not reset the opacity, to maintain the consistency of the optimization during the training process and avoid gradient disappearance due to opacity reset.

2.2 Camera Sampling Strategy

This section outlines a three-stage camera sampling strategy for crafting both outdoor and indoor scenes. The process is as follows:

Outdoor

- In the first stage, we freeze the parameters of the ground and objects, focusing solely on optimizing the surrounding environments without rendering the objects. During this phase, we sample cameras near the center of the scene, the camera's pitch angle is set between 80 to 110 degree. After reaching 70% of the iterations of this stage, we select four camera poses for multi-camera sampling at each later iteration. These four cameras, all directed towards the same direction, are positioned on either side of the scene center at distances of either 1/4 or 1/2 of the radius, ensuring the environments achieve satisfactory visual effects across various distances.
- In the second stage, we freeze the parameters of the surrounding environments and objects, and focus solely on optimizing the ground without rendering objects. We sample four camera poses at each iteration, akin to the sampling strategy in the later part of the first stage. We adjust the pitch angle range to 85∼95 degree, which can reduce the occurrence of a singular ground or environment in the rendered images, thereby enhancing the overall scene generation outcome.

Algorithm 1 DreamScene

$Y \rightarrow y_1, y_2, ..., y_N, y_e$;
for $n = [1, 2, ..., N, e]$ **do**
  if n is not e **then**
    Initialize 3D Gaussian of obj_n
  else
    Initialize 3D Gaussian of environment
  end if
  for $iter = [0, 1, ..., max\_iter]$ **do**
    if n is not e **then**
      Sample camera pose c
    else
      Sample camera pose c follow strategy in Sec. 2.2
    end if
    $x_0 = g(\theta, c)$
    $T_{end} = (1 - \frac{iter}{max\_iter})timesteps$
    for $i = [1, 2, ..., m]$ **do**
      $t_i = T_{end} \cdot random(\frac{i-1}{m}, \frac{i}{m})$
      $x_i = DDIM_(x_{i-1}, i)$
      $\epsilon_\phi(x_{t_i}; y_n, t_i)$ =U-Net(x_{t_i}, y_n, t_i)
      $\epsilon_\phi(x_{t_i}; \emptyset, t_i)$ =U-Net$(x_{t_i}, \emptyset, t_i)$
    end for
    $\nabla_\theta \mathcal{L}_{\text{MTS}}(\theta) = \mathbb{E}_{t,\epsilon,c}\left[\sum_{i=1}^{m} w(t_i)(\epsilon_\phi(x_{t_i}; y_n, t_i) - \epsilon_\phi(x_{t_i}; \emptyset, t_i))\frac{\partial g(\theta,c)}{\partial \theta}\right]$
    Update θ
    if $iter\%compress\_iter = 0$ **then**
      $Score_k = \sum_{j=1}^{H\times W\times M} \frac{V(k)}{D(r_j,k)^2 \times maxV(r_j)}$
      $Sort(Score_k)$
      Delete last z 3D Gaussians
    end if
  end for
  if n is e **then**
    Save 3D Gaussian Representation of the Scene
    break
  end if
  Save 3D Gaussian Representation obj_n of text y_n
  $world(x) = r \cdot s \cdot obj_n(x) + t$
  Add obj_n to the Scene
end for

– In the third stage, we optimize both the surrounding environments and the ground, and render objects into the scene to achieve a harmonious and unified effect. We integrate the camera positions used in the previous two stages, ensuring that areas within the scene are evenly and comprehensively covered, thereby attaining a consistent reconstruction result.

Indoor

– In the first stage, we freeze the ground and object parameters and **render** the objects into the scene. We primarily sample camera poses around the center of the scene and set the radius as large as possible to encompass all objects and thereby minimize the multi-head problem. At this stage, the pitch angle range of cameras is set to 75∼115 degree. After the same iterations at outdoor settings, we sample camera poses around the objects to reduce the impact of object occlusion on the environment.
– In the second stage, we freeze the environment parameters and begin optimizing the ground parameters. The indoor camera sampling strategy remains largely unchanged, but we adjust the pitch angle range to 45∼90 degree to ensure coverage of the ground. Additionally, we increase the camera sampling around objects and from the center to the periphery of the scene, thereby enhancing the integration between the ground and the walls.
– In the third stage, we optimize both the surrounding environments and the ground, and render objects into the scene in the same way as outdoor's.

Our indoor and outdoor camera sampling strategies correspond to typical bounded and unbounded scenes, respectively. For bounded scenes, we focus on ensuring consistency across various orientations. Hence we pay more attention to integrating objects with environments to avoid illogical layouts caused by generating excessive objects in the environment. For unbounded scenes, our primary concern lies with maintaining scene-wide consistency across varied distances. Therefore, we employ two different strategies for scene generation.

3 Additional Results

3.1 Qualitative Results

More qualitative results of FPS are shown in Fig. 1. More qualitative results of scene generation are shown in Fig. 2 and Fig. 3.

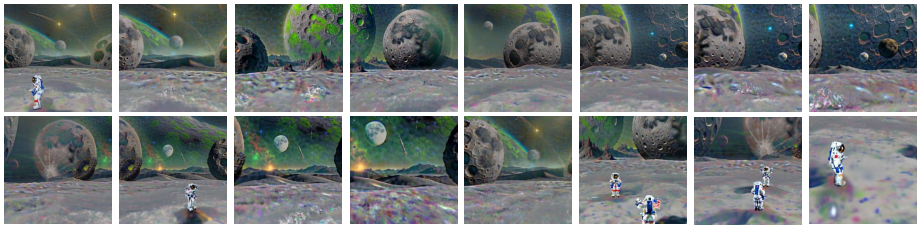
3.2 Ablation Study

Fig. 4(a) depicts a scene generated by randomly sampling cameras within the scene. Due to the difficulty in ensuring consistency of multi-angle views at the same location, the optimization process often tends to collapse. Fig. 4(b) utilizes a strategy that initiates from the center to the surroundings, where the environment and ground are not differentiated. It can be observed that while scene

Fig. 1: More object generation results of DreamScene.

"A DSLR photo in the open area of the zoo"

"A DSLR photo of an autumn park"



"Gray land at the moon, black tranquil universe in the distance, Sci-fi style"

"A minecraft cubes world with lake and mountains in the far distance and grass cubes in the near distance"

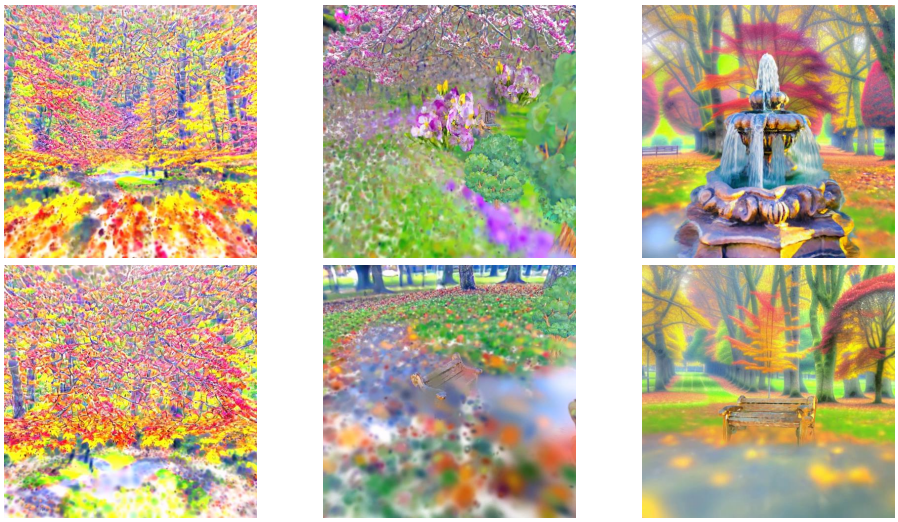
Fig. 2: More outdoor scene generation results of DreamScene.

"A DSLR photo of an European style kitchen"

"A DSLR photo of an Ukiyo-e style bedroom, Ukiyo-e style"

"A DSLR photo of a living room"

Fig. 3: More indoor scene generation results of DreamScene.



(a) Randomly camera sampling

(b) No distinction between environment and ground

(c) DreamScene three-stage camera sampling strategy

Fig. 4: The ablation results of different camera sampling strategies.

consistency improves, the connection between the ground and the scene is poorly generated, and the ground is prone to coarse Gaussian points. Fig. 4(c) employs our three-phase strategy, enhancing the generation quality while ensuring the consistency of the surrounding environment and ground.

References

1. Cohen-Bar, D., Richardson, E., Metzer, G., Giryes, R., Cohen-Or, D.: Set-the-scene: Global-local training for generating controllable nerf scenes. arXiv preprint arXiv:2303.13450 (2023)
2. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. arXiv preprint arXiv:2303.11989 (2023)
3. Lin, Y., Bai, H., Li, S., Lu, H., Lin, X., Xiong, H., Wang, L.: Componerf: Text-guided multi-object compositional nerf with editable 3d scene layout. arXiv preprint arXiv:2303.13843 (2023)
4. Ouyang, H., Heal, K., Lombardi, S., Sun, T.: Text2immersion: Generative immersive scene with 3d gaussians. arXiv preprint arXiv:2312.09242 (2023)
5. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
6. Vilesov, A., Chari, P., Kadambi, A.: Cg3d: Compositional generation for text-to-3d via gaussian splatting. arXiv preprint arXiv:2311.17907 (2023)
7. Zhang, J., Li, X., Wan, Z., Wang, C., Liao, J.: Text2nerf: Text-driven 3d scene generation with neural radiance fields. IEEE Transactions on Visualization and Computer Graphics (2024)

8. Zhang, Q., Wang, C., Siarohin, A., Zhuang, P., Xu, Y., Yang, C., Lin, D., Zhou, B., Tulyakov, S., Lee, H.Y.: Scenewiz3d: Towards text-guided 3d scene composition. arXiv preprint arXiv:2312.08885 (2023)
9. Zhou, X., Ran, X., Xiong, Y., He, J., Lin, Z., Wang, Y., Sun, D., Yang, M.H.: Gala3d: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting. arXiv preprint arXiv:2402.07207 (2024)