한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 31, No. 3, P. 119~126

사용자 상호작용 가능성을 고려한 Diffusion 기반 3D 실내 장면 생성

김미송[1,O] 정승재[3] 황효석[3,*] 강형엽[2,*1]

[1]경희대학교 인공지능학과, [2]고려대학교 컴퓨터학과, [3]경희대학교 소프트웨어융합학과

misong023@khu.ac.kr, teclados078@khu.ac.kr, hyoseck@khu.ac.kr, siamiz_hkang@korea.ac.kr

3D Indoor Scene Generation via Diffusion with User Interaction Awareness

Misong Kim[1,O] SeoungJae Jeong[3] Hyoseok Hwang[3,*] HyeongYeop Kang[2,*]

[1]Department of Artificial Intelligence, Kyung Hee University, [2]Department of Computer Science and Engineering, Korea University, [3]Department of Software Convergence, Kyung Hee University

요 약

3D 실내 장면 생성은 가상현실, 로봇 내비게이션, 인간-환경 상호작용 시뮬레이션 등 다양한 응용에서 중요한 역할을 하며, 객체 간의 구조적 정합성과 시각적 자연스러움이 핵심 요소로 여겨진다. 최근 Diffusion 모델 기반 접근이 복잡한 객체 배치를 안정적으로 생성하는 데 강점을 보이며 주목받고 있지만, 사용자의 손이 닿기 어려운 위치에 객체가 배치되는 등 실질적인 상호작용 가능성은 반영되지 않았다. 본 연구는 이러한 한계를 보완하고자, 인간의 손 도달 가능성을 정량화한 접근 가능성 맵을 구축하고, 이를 Diffusion 모델의 샘플링 과정에 보상 신호로 통합하는 새로운 장면 생성 방식을 제안한다. VPoser 와 SMPL-X 기반의 자세 샘플링을 통해 다양한 상호작용 맥락을 반영하였으며, 제안된 방식은 사람이 실제 상호작용 가능한 위치에 객체가 배치되도록 유도한다. 실험을 통해 제안 기법이 기존 모델보다 상호작용 가능성과 공간적 일관성 측면에서 우수한 성능을 보임을 확인하였다.

Abstract

3D indoor scene synthesis plays a crucial role in a wide range of applications, including virtual reality, robot navigation, and human-environment interaction simulations. While recent diffusion-based approaches have demonstrated strong capabilities in generating complex object arrangements with structural coherence and visual plausibility, they often fail to consider the user's actual ability to interact with the environment. Specifically, objects may be placed in locations that are difficult for a user to physically reach or manipulate. To address this limitation, we propose a novel scene generation framework that integrates human reachability into the diffusion sampling process. We construct a reachability map by quantitatively modeling the spatial accessibility of human hands using pose sampling based on VPoser and SMPL-X. This reachability information is then incorporated as a reward signal during sampling, guiding the model to generate object layouts that are more functionally usable. Experimental results demonstrate that our method outperforms prior diffusion-based models in terms of both interaction feasibility and spatial consistency, enabling the generation of 3D indoor scenes better aligned with real-world human use.

키워드: 3D 실내 장면 생성, 조건 기반 Diffusion 모델, 인간-환경 상호작용

Keywords: 3D Indoor Scene Generation, Conditional Diffusion Model, Human-Scene Interaction

*corresponding author: Hyoseok Hwang / Kyung Hee University (hyoseok@khu.ac.kr), HyeongYeop Kang / Korea University (siamiz_hkang@korea.ac.kr)

Received : 2025.06.13./ Review completed : 1st 2025.06.30. 2nd 2025.07.10./ Accepted : 2025.07.14.
DOI : 10.15701/kcgs.2025.31.3.119
ISSN : 1975-7883(Print)/2383-529X(Online)

1. 서론

3D 실내 장면 생성은 가상현실, 로봇 내비게이션, 인간-환경 상호작용 시뮬레이션 등 다양한 분야에서 중요한 역할을 한다. 기존 연구들은 시각적으로 자연스러운 장면이나 기하학적으로 타당한 배치를 생성하는 데에 집중해 왔다. 그러나 이러한 접근은 사람이 실제로 공간 내에서 어떻게 상호작용할 수 있는지에 대한 고려가 부족하다는 한계가 있다. 즉, 사람이 접근하거나 조작할 수 없는 위치에 객체가 배치되는 등, 실사용 가능성과의 괴리가 발생하는 경우가 많다.

기존 딥러닝 기반 3D 장면 생성 기법들은 주로 VAE [1, 2], GAN [10], Transformer [3, 4] 모델 등을 활용하여 장면 내 객체의 구성 요소들을 예측하는 방식으로 발전해왔다. 최근에는 Diffusion 모델의 강력한 표현력을 활용하여, 복잡한 배치 구조를 동시에 모델링할 수 있는 DiffuScene [11] 과 같은 방법이 제안되었다. 하지만 이들 대부분은 사람의 위치나 행동 가능성을 고려하지 않기 때문에, 실제로 사람이 사용할 수 있는 장면을 생성하는 데에는 제약이 존재한다.

본 연구는 장면의 시각적/기하학적 일관성뿐만 아니라, 사용자의 실제 상호작용 가능성까지 반영된 실내 3D 장면을 생성하는 것을 목표로 한다. 이를 위해 Diffusion 기반 장면 생성 모델을 기반으로 사람의 손이 도달할 수 있는 공간을 정량화 한 접근 가능성 맵을 구축하고, 이를 샘플링 과정에 guidance 로 활용하는 방식을 제안한다. 이때 접근 가능성 맵은 VPoser [12] 기반의 다양한 인간 자세로부터 도출되며, 상호작용 상황은 실제 데이터에서 자주 공존하는 객체 조합을 클러스터링하여 추론한다.

본 논문의 주요 기여는 다음과 같다:

· 사용자의 실제 동작을 반영할 수 있도록, VPoser [12] 기반 자세 샘플링을 통해 손 도달 가능 영역을 계산하고, 이를 기반으로 3D 접근 가능성 맵을 구축하였다.

· 접근 가능성 맵을 바탕으로, 사람이 손으로 도달할 수 있는 공간을 정량적으로 평가하고, 이를 기반으로 샘플링 단계에서 객체 배치를 유도하는 방식을 설계하여, 실제 사용자가 상호작용할 수 있는 위치에 객체가 배치되도록 하였다.

· 기존 DiffuScene [11] 기반 구조에 최소한의 추가 조건만으로, 실용적인 상호작용 가능 장면 생성을 가능하게 하는 확장 프레임워크를 제안하였다.

2. 관련 논문

2.1 기존 딥러닝 기반 3D 장면 생성 모델

3D 실내 장면 생성을 위한 초기 연구들은 주로 VAE [1, 2], GAN [10], Transformer [3, 4] 모델 등 다양한 딥러닝 기반 생성 구조를 활용해 왔다. 예를 들어 3D-FRONT [5] 와 같은 대규모 데이터셋을 활용하여 장면 내 객체의 위치, 크기, 방향을 예측하는 방식으로 전체 레이아웃을 생성하는 접근이 제안되었다.

VAE 기반 모델들은 전체 장면의 잠재 공간을 학습함으로써 다양한 형태의 장면 구성이 가능하다는 장점을 갖는다 [1, 2]. 하지만 복잡한 상호작용 패턴이나 정확한 공간 제약을 반영하는 데에는 한계가 있다.

Transformer 기반 모델들 [3, 4]은 객체 간 관계성을 더 정교하게 포착할 수 있도록 설계되었으나, 대부분 순차적 구조를 기반으로 하여 복잡한 배치 조합을 동시에 고려하기에는 제약이 있다.

이와 같은 생성 모델들은 장면 내 객체 간의 복합적인 관계 및 상호작용 가능성을 효과적으로 모델링하지 못한다는 한계가 존재한다. 본 연구는 이러한 한계를 극복하고자, diffusion 기반 모델의 표현력과 안정성을 활용하여 더욱 자연스러운 장면 구성을 도모하며, 추가적으로 사용자의 실제 상호작용 가능성을 반영할 수 있는 조건을 통합하고자 한다.

2.2 Diffusion 모델 기반 3D 장면 생성

최근에는 Diffusion 모델을 기반으로 한 3D 장면 생성 기법들이 주목받고 있다. Diffusion 모델은 고차원 공간에서의 안정적인 샘플링이 가능하며, 반복적인 노이즈 제거 과정을 통해 복잡한 데이터 분포를 정밀하게 학습할 수 있다는 장점이 있다. 이러한 특성은 다양한 객체 간의 상호작용과 공간적 배치를 포함하는 3D 장면 합성에 적합하다.

대표적인 연구로는 DiffuScene [11] 이 있으며, 이 모델은 3D-FRONT [5] 데이터셋을 기반으로 학습한 후, 객체의 위치, 크기, 회전, 클래스, 형상 임베딩을 포함한 속성 벡터들을 생성하는 구조를 갖는다. DiffuScene [11] 은 장면 내 객체들을 순서 없는 집합으로 간주하여, 전체 장면 구성을 동시에 모델링할 수 있다는 특징이 있다. 이러한 접근은 특히 복잡한 레이아웃 구조나 다수의 객체가 포함된 환경에서 높은 유연성과 표현력을 보인다.

또한 SceneDiffuser [14]와 같이 Diffusion 모델에 물리 기반 조건을 통합하여 생성 결과의 물리적 타당성을 높이려는 시도도 이어지고 있다.

이러한 연구들은 Diffusion 모델이 기존 생성 모델의 한계를 보완하며, 더욱 사실적이고 상호작용 가능한 장면 생성을

가능하게 함을 보여준다. 본 연구 또한 이러한 Diffusion 모델의 구조를 바탕으로 하여, 실제 사용자의 동작 가능성을 반영하는 접근 가능성 기반 샘플링 전략을 통합하였다.

2.3 상호작용

최근 3D 장면 생성 연구에서는 단순한 기하학적 배치나 시각적 자연스러움을 넘어서, 사용자가 실제로 공간 내 객체들과 상호작용할 수 있는지를 고려하려는 시도가 나타나고 있다. 예를 들어, PhyScene [6]는 관절형 객체의 동작 가능성을 판단하기 위해 물리적 충돌 여부를 평가하는 제약 조건을 도입하였다. 그러나 이러한 접근은 상호작용의 일부 측면에 국한되며, 사람이 장면 내에서 어떤 자세로 어떤 객체와 상호작용할 수 있는지를 포괄적으로 다루지는 않는다.
본 연구는 사용자의 행동 가능성과 공간 내 사용성이라는 관점에서, 인간 중심의 상호작용 조건을 정량적으로 모델링하고 이를 장면 생성과 과정에 통합함으로써, 실질적으로 사용 가능한 배치를 유도하는 것을 목표로 한다.

3. 방법

본 장에서는 본 연구에서 제안하는 접근 가능성 기반 3D 장면 생성 기법의 전체 구조를 설명한다. 먼저, 기반이 되는 Diffusion 모델 구조로서 DiffuScene [11] 의 학습 방식과 객체 표현 방법을 소개하고(3.1 절), 이어서 사용자의 실제 상호작용 가능성을 반영하기 위한 자세 샘플링 절차(3.2 절)와 접근 가능성 맵 구축 과정(3.3 절)을 자세히 설명한다. 마지막으로, 구축된 접근 가능성 맵을 활용하여 샘플링 과정에서 객체 배치를 유도하는 접근 가능성 기반 생성 방식(3.4 절)을 제안한다. 이러한 흐름을 통해 본 연구는 기존의 시각적/기하학적 일관성 위주의 장면 생성 방식에서 나아가, 사용자의 상호작용 가능성까지 반영한 실용적 장면 생성 기법을 제안한다.

3.1 실내 장면을 위한 Diffusion 모델

본 연구는 실내 3D 장면 생성을 위해 DiffuScene [11]구조를 기반으로 한 Diffusion 모델을 사용하였다. DiffuScene [11] 은 3D-FRONT [5] 데이터셋을 기반으로 학습되며, 각 장면을 구성하는 객체들의 위치, 크기, 회전, 클래스, 형상 임베딩 정보를 포함하는 속성 벡터들을 샘플링하는 구조로 이루어진다. 각 장면의 객체들은 순서 없는 집합으로 표현되며, 학습 과정에서는 이 벡터들을 점진적으로 복원하는 denoising diffusion probabilistic model (DDPM) [9] 방식이 사용된다.

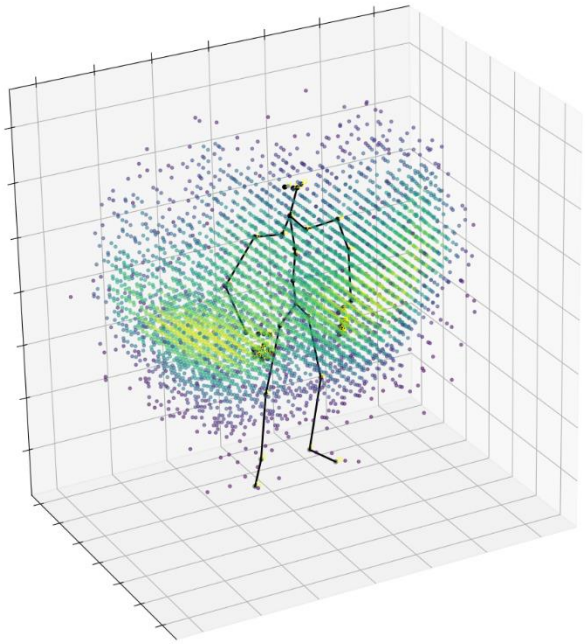

Figure 1. An example visualization of 3D reachable space obtained by accumulating reachable regions of the left and right hands. The standing poses were sampled using VPoser [12] and converted into 3D skeletons using SMPL-X [13]. Yellow regions indicate areas that are frequently reached and easily accessible, while purple regions represent areas that are more difficult to reach.

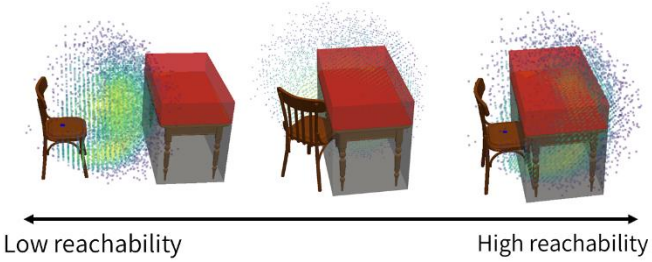

Figure 2. Variation in reachability based on the relative positions of a chair and a table in a seated posture. The arrangements are ordered from low to high reachability, illustrating how changes in the placement of the chair and table affect the feasibility of interaction.

DiffuScene [11] 은 학습된 분포로부터 새로운 장면을 생성할 수 있으며, 본 연구에서는 이 모델을 사전 학습한 뒤, 사람이 실제로 상호작용 가능한 배치를 유도하기 위해 접근 가능성 기반 샘플링 전략을 추가적으로 적용하였다. 이를 위해 먼저 사용자의 자세를 다양하게 생성하고, 도달 가능한 공간을 정량화한 후, 이를 바탕으로 Diffusion 모델의 샘플링을 유도하는 방식을 설계하였다.

3.2 인간 자세 샘플링

본 연구에서는 사람의 손이 도달 가능한 공간을 정량적으로 모델링하기 위해, 먼저 VPoser [12]를 이용하여 다양한

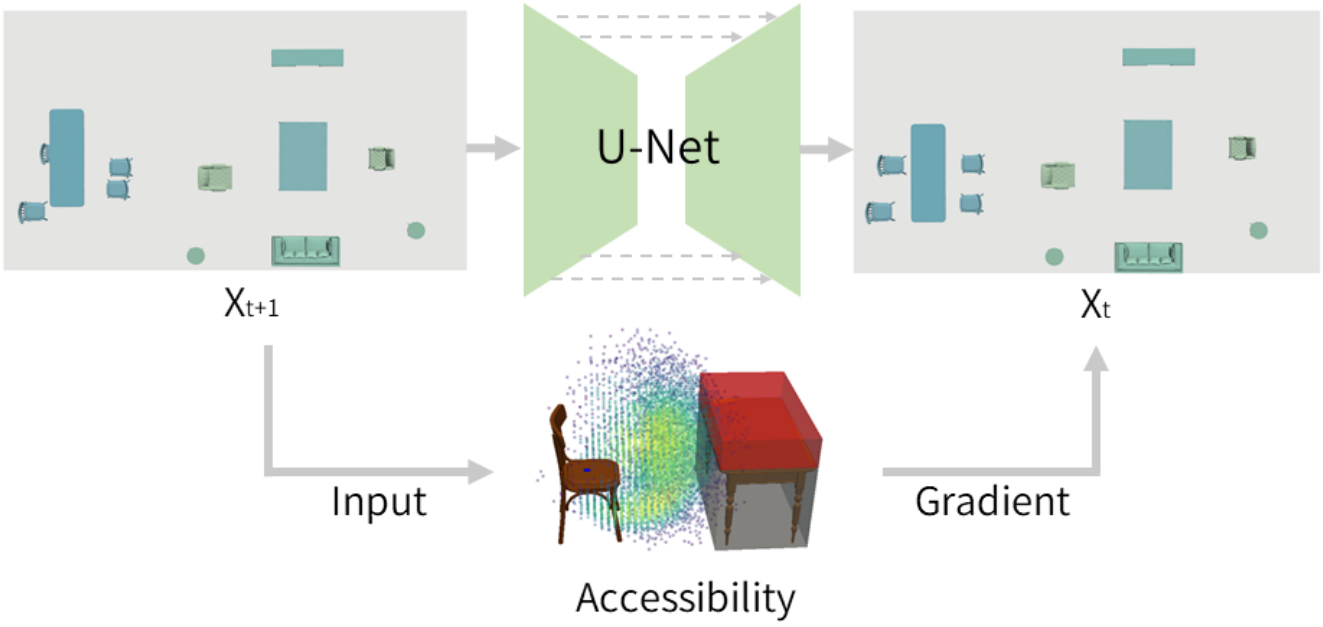

Figure 3. **Overview.** The diffusion model progressively generates object arrangements guided by accessibility scores. Starting from an initial noisy state, the model refines object placements at each step by incorporating reachability constraints as guidance.

자연스러운 자세를 대량으로 샘플링한다. VPoser [12]는 AMASS [15] 데이터셋을 기반으로 학습된 변분 오토인코더로, 인간 자세의 잠재 공간을 학습한 모델이다. 우리는 잠재 벡터 $z \in \mathbb{R}^{32}$ 를 무작위로 샘플링한 뒤, 이를 VPoser [12] 디코더에 통과시켜 21 개 관절의 3D 회전 정보를 포함하는 자세 벡터 $p_{body} \in \mathbb{R}^{63}$를 얻는다.

생성된 자세 벡터는 SMPL-X [13] 모델을 통해 3D 메쉬와 관절 위치로 변환되며, 왼손과 오른손의 위치를 기반으로 도달 가능 영역을 계산한다.

본 연구에서는 두 가지 대표적인 상호작용 자세를 가정하여 데이터를 구성하였다. 첫째, 앉은 상태에서는 골반 관절의 위치를 기준으로 포즈를 정렬하며, 둘째 선 상태에서는 양발 관절의 평균 위치를 기준으로 포즈를 정렬한다. 이러한 두 가지 정렬 방식은 사용자의 상호작용 맥락에 따라 손의 도달 가능 공간이 달라지는 점을 반영하기 위한 것이다.

3.3 접근 가능성 맵 구축

샘플링된 모든 손 위치들을 기반으로, 사람이 도달 가능한 3D 공간을 접근 가능성 맵 형태로 모델링한다. 구체적으로, 사람 주변 공간을 격자 $G \in \mathbb{R}^{K \times K \times K}$ 로 나누고, 각 셀마다 다음 두 가지 정보를 집계한다.

1. 접근 빈도: 손이 해당 위치에 도달한 횟수
2. 포즈 에너지: 해당 포즈를 수행하는 데 필요한 노력을 나타내는 스칼라 값

포즈 에너지는 다음과 같이 정의된다.

$$E = \left\| w \odot p_{body} \right\|_2$$

여기서 $w \in \mathbb{R}^{63}$ 는 관절별 동작에 필요한 에너지 기여도를 조정하기 위한 가중치 벡터로, 손이나 발처럼 말단에 위치한 관절에는 낮은 값을 부여한다. 이는 말단의 움직임보다 몸통을 포함한 중심 관절의 움직임이 더 많은 에너지를 요구한다고 가정한 것이다.

이는 관절 회전량의 크기를 기반으로 자세의 복잡성을 정량화하고, 이를 자세 유지 시 요구되는 생체역학적 부담을 반영하는 지표로 활용한 것이다.

접근 가능성은 각 격자에서 다음과 같이 계산된다:

$$\eta(x,y,z) = \frac{frequency(x,y,z)}{energy\_sum(x,y,z) + \epsilon}$$

여기서 ϵ 는 0 으로 나누는 것을 방지하는 작은 상수이다.

이 정의에 따라, 높은 빈도와 낮은 에너지를 동시에 만족하는 위치일수록 접근 가능성 값이 커진다.

생성된 접근 가능성 맵은 가우시안 필터를 통해 연속적으로 보정되어, 학습 시 역전파가 전달될 수 있도록 하였다.

3.4 접근 가능성 기반 샘플링 유도

접근 가능성 맵을 활용하여 장면의 사용 가능성을 평가하기 위해서는, 사람이 어떤 자세에서 어떤 객체와 상호작용하는지를 정의할 필요가 있다. 본 연구에서는 이를 위해 학습 데이터에 포함된 객체들의 3 차원 위치 및 크기 정보를 기반으로 HDBSCAN 클러스터링을 수행하였다. 이때 각 객체는 공간적으로 밀집된 위치를 기준으로 클러스터링 되며,

같은 클러스터에 자주 함께 포함되는 객체쌍을 통계적으로 분석하여 상호작용 조합 후보로 정의하였다. 그 결과, 소파-커피테이블, 의자-식탁과 같은 조합이 빈번하게 나타났으며, 본 연구에서는 이러한 두 가지 대표 조합에 대해 사용자가 앉은 자세에서 손으로 상호작용할 수 있는 상황을 가정하여 접근 가능성을 계산하였다.

앞서 구축된 접근 가능성 맵은 장면 내에서 사람이 손으로 도달하기 쉬운 영역을 정량적으로 나타낸다. 본 연구에서는 이 정보를 Diffusion 기반 샘플링 과정에 직접 통합함으로써, 사람이 실제로 상호작용하기 적합한 배치로 객체가 생성되도록 유도한다.

매 샘플링 단계 t에서 중간 샘플 x_t에 대해 다음과 같은 절차를 수행한다. 먼저 예측된 정제 샘플 $\hat{x}_0$에 대해 보상 함수 $r(\hat{x}_0)$를 계산하고, 이를 음의 방향으로 손실로 변화하여 역전파한다. 보상함수 $r(x_0)$는 예측된 샘플 x_0에서 상호작용 대상으로 간주된 객체들, 즉 소파 또는 의자에 앉아 상호작용이 일어날 수 있는 테이블류(예: 커피테이블, 식탁)에 대한 접근 가능성을 기반으로 정의된다.

구체적으로는, 각 소파 또는 의자의 중심 위치를 기준으로 앉은 자세를 배치한 후, 해당 위치 전방에 상호작용이 일어나는 테이블 상단 높이 (예: z=0.2)를 중심으로 한 가상의 영역을 설정한다. 이 가상 박스 내의 3D 격자 위치들에 대해 접근 가능성 맵 $\eta(x_i, y_i, z_i)$ 값을 조회하여 한 값을 보상 함수로 사용하였다:

$$r(x_0) = \left(\frac{1}{M}\right)\sum_{i=1}^{N} \eta(x_i, y_i, z_i)$$

여기서 M 은 장면 내 의자 수를 의미한다.

전체 손실은 다음과 같이 표현된다:

$$\mathcal{L}_{guidance} = -\mathbb{E}_{x_0 \sim x_p}[r(x_0)]$$

여기서 $x_0 \sim x_p$는 DDPM 샘플링 과정 중 x_t가 시간 단계 $t \to 0$으로 점차 정제되며 얻는 예측된 초기 샘플 분포를 의미한다. 즉, x_p는 prior 에 해당하는 분포로, x_0는 그로부터 얻어진 정제 샘플이다.

계산된 gradient 는 정제 샘플 x_t에 적용되어, 접근 가능성이 높은 방향으로 샘플을 이동시킨다. 전체 샘플링 갱신은 다음 식과 같이 정의된다:

$$x_t \leftarrow x_t - \alpha \cdot \frac{\nabla_{x_t}\mathcal{L}_{guidance}}{\left\|\nabla_{x_t}\mathcal{L}_{guidance}\right\| + \epsilon}$$

여기서 α는 갱신 강도를 조절하는 계수로, 본 연구에서는 실험적으로 0.1 을 설정하였다. ϵ 은 정규화를 위한 작은 상수이다.

이러한 방식은 샘플링 중 생성 결과를 접근 가능성 맵 기반으로 직접 평가하고, 이를 통해 사람이 실제로 사용 가능한 객체 배치 구성을 효과적으로 유도할 수 있도록 한다.

4. 실험

본 연구의 실험은 제안한 접근 가능성 기반 샘플링 기법의 효과를 검증하기 위해 수행되었으며, 비교 대상은 기존의 생성 장면 모델인 DiffuScene [11]이다. 모든 모델은 동일한 데이터셋과 조건에서 학습 및 평가되었다.

Method	FID↓	IoU↓	Accessibility↑
DiffuScene	15.10	0.0033	0.44
Ours	15.67	0.0024	0.59

Table 1. Quantitative comparisons on the task of unconditional scene synthesis. Our method outperforms DiffuScene in terms of IoU and Accessibility Score, indicating enhanced usability and interaction feasibility.

4.1 데이터셋

실험에는 대규모 3D 실내 장면 데이터셋인 3D-FRONT [5]를 사용하였다. 본 데이터셋은 다양한 가구와 공간 배치를 포함하며, 실제 가정 환경과 유사한 구성을 제공한다. 본 연구에서는 이 중에서도 LivingDiningRoom 과 LivingRoom 카테고리에 해당하는 장면을 학습 및 평가에 활용하였다. 이 두 공간은 다양한 상호작용이 발생하는 대표적인 공용 공간으로, 모델이 사람 중심의 배치를 학습하기에 적합한 환경을 제공한다.

4.2 평가 지표

생성된 장면의 품질은 사람 중심 상호작용성, 시각적 자연스러움, 공간적 정합성의 세 가지 관점에서 정량적으로 평가되었다. 이를 위해 본 연구에서는 FID (Fréchet Inception Distance), IoU (Intersection over Union), Accessibility Score 의 세 가지 지표를 사용하였으며, 각 지표는 다음과 같이 정의된다.

먼저, Accessibility Score 는 사람이 실제로 상호작용할 수 있는지를 반영하는 지표로, 접근 가능성 맵을 기반으로 계산된다. 구체적으로는 장면 내 소파 또는 의자에 대해 사전 정의된 앉은 자세를 기준으로 위치를 정렬한 뒤, 그 앞쪽의 상호작용 영역에 대해 접근 가능성 값을 평균하고, 이를 전체 좌석에 대해 다시 평균하여 최종 점수를 계산하였다. 이 지표는 객체의 도달 빈도와 동작 수행 에너지를 함께 고려함으로써,

Figure 4. **Overview. Qualitative comparison of unconditional scene synthesis.** Compared to DiffuScene, our method places objects such as chairs and tables, or sofas and coffee tables, in a way that better supports user interaction.

사람이 실제로 앉거나 선 상태에서 주변 사물과 상호작용하기 쉬운지 정량적으로 평가하는 데 사용된다.

다음으로, FID 는 생성된 장면의 시각적 자연스러움과 다양성을 평가하기 위해 사용된다. 본 연구에서는 FID 계산을 위해, 생성된 장면과 실제 장면 각각에 대해 top-view 에서 렌더링된 이미지 1000 장을 생성하여 비교에 활용하였다. 렌더링된 이미지를 사전 학습된 Inception-V3 모델을 통해 특징 벡터를 추출하고, 실제 장면과 생성된 장면 간의 특징 분포 차이를 계산하여 FID 점수를 얻었다.

마지막으로, IoU 는 생성된 장면의 공간적 정합성을 평가하기 위한 지표로, 객체 간의 충돌 정도를 정량화 하는 데 사용된다. 예측된 객체의 3D 상자간의 중첩 영역을 기반으로 IoU 를 계산하고, 그 평균 값을 통해 장면 내 객체들이 구조적으로 타당하게 배치되었는지를 판단하였다. 충돌이 많을수록 IoU 점수는 낮아지며, 이는 물리적으로 비현실적인 배치를 의미한다.

이러한 세가지 지표는 시각적 유사성을 넘어서, 장면이 실제 인간 사용자의 관점에서 기능적 타당성을 갖추었는지를 종합적으로 평가한다.

4.3 비조건부 장면 합성

정량적 평가 결과는 표 1 에 제시하였다. 본 방법은 IoU 와 접근성 점수에서 기존 방법인 DiffuScene [11] 을 상회하며, 이는 생성된 장면이 사용자 중심의 상호작용 가능성과 배치의 실용성 측면에서 더 우수함을 의미한다. 특히, 단순한 시각적 유사성뿐 아니라, 실제 사용자가 공간을 어떻게 활용할 수 있는지를 정량적으로 반영할 수 있는 지표에서의 향상은, 본 모델이 실제 사용성을 고려한 배치 구성에 효과적임을 시사한다.

정성적 비교는 그림 4 에 제시되어 있다. DiffuScene [11] 과 비교했을 때, 본 모델은 의자-테이블이나 소파-커피테이블과 같은 주요 객체 쌍들을 보다 상호작용에 적합한 거리와 방향으로 배치하는 경향을 보인다. 이는 앉기, 물건 놓기, 손이 닿는 등의 일상적인 행위를 자연스럽게 고려하며, 생성된 장면의 기능적 타당성과 인간 중심 설계의 실현 가능성을 뒷받침한다.

5. 결론

본 연구는 3 차원 실내 장면 생성을 위한 새로운 접근으로 인간의 상호작용 가능성을 반영한 효율성 기반 guidance 기법을 Diffusion 모델 샘플링 과정에 통합하였다. 이를 통해 단순히 시각적으로 자연스럽거나 기하학적으로 타당한 배치를 넘어, 실제 사용자가 공간 내에서 효과적으로 상호작용할 수 있는 장면 구성을 도출하는 데에 성공하였다.

우리는 VPoser [12] 기반의 자세 샘플링을 통해 다양한 상호작용 자세를 생성하고, 손의 도달 가능성을 바탕으로 공간 내 효율성을 정량화하였다. 이후 이 정보를 확산 모델의 샘플링에 직접 활용하여, 사람이 실제로 접근 가능한 위치에 객체가 배치되도록 유도하였다. 실험 결과는 제안한 방법이 기존 모델에 비해 상호작용 가능성, 공간적 일관성 측면에서 모두 향상된 성능을 보임을 확인하였다.

그러나 본 연구에는 몇 가지 한계도 존재한다. 먼저, 도달 가능성은 정적 자세를 기반으로 평가되므로, 인간의 실제 행동 흐름이나 동적 상호작용 과정을 충분히 반영하지 못한다. 예를 들어, 앉은 후 손을 뻗어 물건을 잡는 등 복합적인 행동 시퀀스를 고려한 평가 체계는 향후 과제로 남아 있다. 또한, 본 연구는 객체 간의 공존 패턴을 기반으로 제한된 상호작용 유형(예: 의자-테이블)을 가정하고 효율성을 계산하였기 때문에, 다양한 객체군에 걸친 상호작용 가능성을 일반화하는 데에는 제약이 존재한다.

종합하면, 본 연구는 확산 기반 장면 생성 모델에 인간 중심의 상호작용 조건을 통합함으로써 실제 사용성을 고려한 새로운 장면 생성 프레임워크를 제시하였다. 향후에는 다양한 행동 시나리오에 대한 확장, 동적 상호작용의 반영 등을 통해 보다 범용적이고 실용적인 장면 생성 기술로 발전시킬 수 있을 것이다.

감사의 글

이 논문은 2025 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2025-00564137).

References

[1] P. Purkait, C. Zach, and I. Reid, “SG-VAE: Scene grammar variational autoencoder to generate new indoor scenes,” Proceedings of the European Conference on Computer Vision (ECCV), pp. 155–171, 2020.

[2] H. Yang, Z. Zhang, S. Yan, H. Huang, C. Ma, Y. Zheng, C. Bajaj, and Q. Huang, “Scene synthesis via uncertainty-driven attribute synchronization,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5630–5640, 2021.

[3] X. Wang, C. Yeshwanth, and M. Nießner, “SceneFormer: Indoor scene generation with transformers,” Proceedings of the International Conference on 3D Vision (3DV), pp. 106–115, 2021.

[4] D. Paschalidou, A. Kar, M. Shugrina, K. Kreis, A. Geiger, and S. Fidler, “ATISS: Autoregressive transformers for indoor scene synthesis,” Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 12013–12026, 2021

[5] H. Fu, B. Cai, L. Gao, L. X. Zhang, J. W. Cao, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao, H. Zhang, et al., “3D-FRONT: 3D furnished rooms with layouts and semantics,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10933–10942, 2021.

[6] Y. Yang, B. Jia, P. Zhi, and S. Huang, “PhyScene: Physically Interactable 3D Scene Synthesis for Embodied AI,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[7] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[8] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” arXiv preprint arXiv:1801.01401, 2018.

[9] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840–6851, 2020.

[10] M.-J. Yang, Y.-X. Guo, B. Zhou, and X. Tong, “Indoor scene generation from a collection of semantic-segmented depth images,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15203–15212, 2021.

[11] J. Tang, Y. Nie, L. Markhasin, A. Dai, J. Thies, and M. Nießner, “DiffuScene: Denoising diffusion models for generative indoor scene synthesis,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

[12] N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas, G. Pavlakos, V. Choutas, M. Black, C. Bolkart, and M. Tzionas, “VPoser: Variational Human Pose Prior for Body Inverse Kinematics,” arXiv preprint arXiv:1904.05866, 2019.

[13] G. Pavlakos*, V. Choutas*, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, “Expressive Body Capture: 3D Hands, Face, and Body from a Single Image,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10975–10985, 2019.

〈저자소개〉

김미송

- 2024년 경희대학교 소프트웨어융합학과 학사
- 2024년 ~ 현재 경희대학교 일반대학원 인공지능학과 석사과정
- 관심 분야: 3D Indoor Scene Generation, 3D Generative Model
- https://orcid.org/0009-0009-3714-4685

정승재

- 2021년 경희대학교 소프트웨어융합학과 학사
- 2021년 ~ 현재 경희대학교 일반대학원 소프트웨어융합학과 석박사 과정
- 관심 분야: 3D Generation, 3D Reconstruction, 3D Gaussian Splatting
- https://orcid.org/0009-0000-5306-6896

황효석

- 2004년 연세대학교 기계공학과 학사
- 2009년 한국과학기술원 로봇공학학제전공 석사
- 2017년 한국과학기술원 전기및전자공학과 박사
- 2009년 ~ 2017년 삼성전자 종합기술원 연구원
- 2018년 ~ 2021년 가천대학교 소프트웨어학과 조교수
- 2021년 ~ 2024년 경희대학교 소프트웨어융합학과 조교수
- 2024년 ~ 현재 경희대학교 소프트웨어융합학과 부교수
- 관심 분야: Domain Generalization, Sim2Real, Reinforcement learning, Robotics
- https://orcid.org/0000-0003-3241-8455

강형엽

- 2012년 고려대학교 컴퓨터 · 통신공학부 학사
- 2014년 고려대학교 컴퓨터학과 석사
- 2017년 고려대학교 컴퓨터학과 박사
- 2017년 ~ 2018년 고려대학교 컴퓨터학과 연구교수
- 2018년 ~ 2019년 고려대학교 차세대가상증강현실연구소 연구 교수
- 2019년 ~ 2020년 강원대학교 소프트웨어미디어 · 산업공학부 조교수
- 2020년 ~ 2024년 경희대학교 소프트웨어융합학과 조교수
- 2024년 ~ 현재 고려대학교 컴퓨터학과 부교수
- 관심 분야: Computer Graphics, Extended Reality, Agent, World Model, Human-computer Interaction, Holography
- https://orcid.org/0000-0001-5292-4342