Scene Diffusion: Text-driven Scene Image Synthesis Conditioning on a Single 3D Model

Xuan Han YiHao Zhao Mingyu You*

同濟大學 TONGJI UNIVERSITY

ACM MULTIMEDIA 2024 (MM 24) Oral Presentation 28: Complex Scene Processing

Melbourne, Australia | 28 Oct. - 1 Nov.

ACM MM' 24 Oral

Single & Multi Object

Condition Image
(Rendered from Single or Multiple 3D Models)

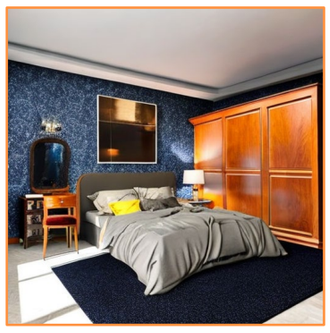
A living room composition, showcasing a loveseat sofa and an armchair, arranged to create a harmonious and inviting space.

A living room highlighted by an armchair, a corner cabinet and a corner table, meticulously placed to enhance the space's comfort and aesthetics.

Scene Description

Synthesized Scene Image

Variable Scene Description



A traditional bedroom with a bed, textured blue walls, a wood wardrobe, an rug and a wooden chair.

Bedroom with a bed, navy walls, marble floor, a round mirrors, green potted plants, brass lighting fixtures.

Room with a bed, soft pink walls, nightstands, a plush carpet, a wood wide closet and a art picture.

A modern bedroom with a bed, grey walls, an art picture, a table lamp stands on metallic nightstand.

A minimalist bedroom with a bed, white walls, a black wardrobe, a mirror and a soft rug lies on the floor.

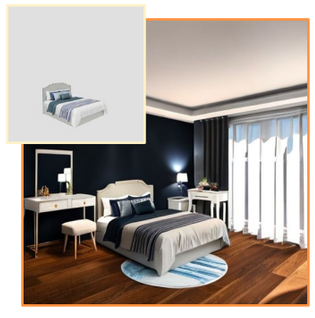
An eclectic bedroom with a bed, yellow walls, a black chair, a collection of plants and brown cushions.

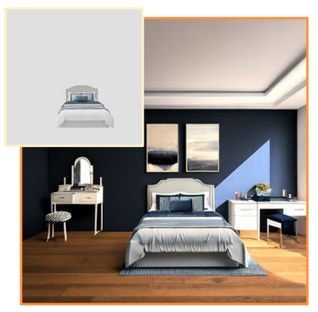
◆ Object Ctrl. ◆ Style Ctrl. ◆ Color Ctrl. ◆ Material Ctrl.

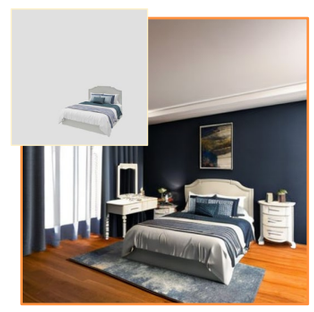
Variable Object Pose

A image of a bedroom with a bed, navy walls, wooden floor ,a dressing Table, a dressing Chair.

(Scene Description)







A. Motivation

• A.1 Creating the Scene Image More Efficiently

*Traditional 3D Rendering (**hours**)* *Our Method (**seconds**)*

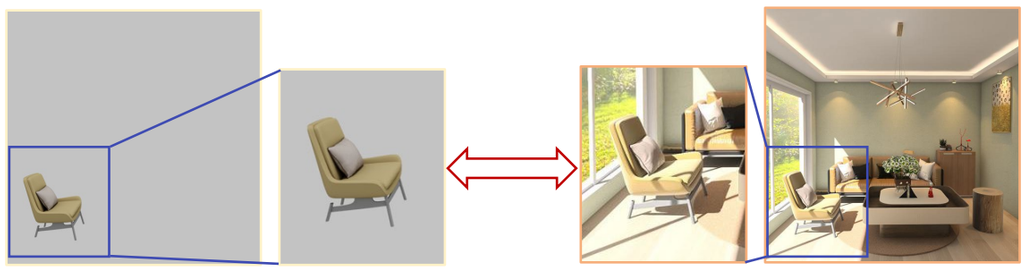
- The traditional 3D rendering pipeline involves many manual steps to render the scene image, which is time-intensive and calls for expertise. We construct a conditional T2I model to help designers express early stage ideas **without spending time on massive 3D rendering work**.

• A.2 Faithful and Harmonious Synthesized Image

- The synthesized scene images are expected to **faithfully present the minutiae of given object** and have a **harmonious visual effect**.

B. Challenge

• Shading Difference between the Condition and Target Images



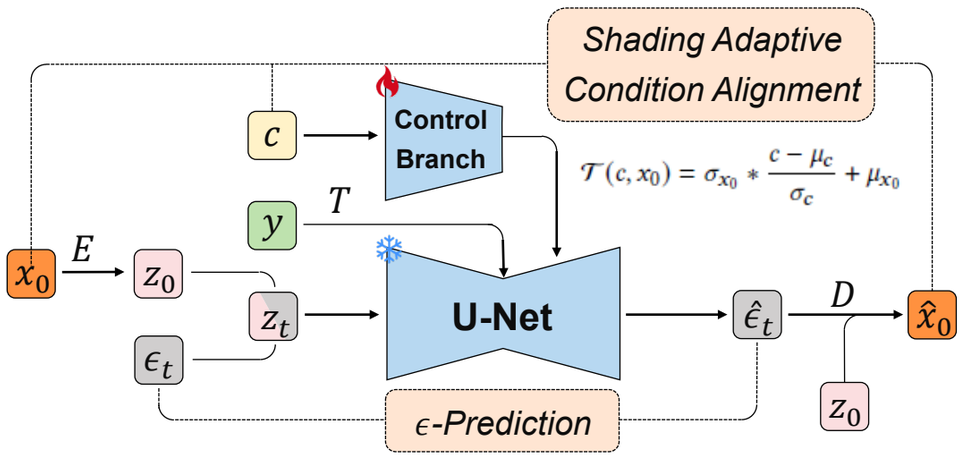
Condition Image c
(Rendered without Context)

Target Image x_0
(Rendered with Context)

- Since the condition image is rendered without scene context, the displayed colors of object, known as its shading, is bound to be different from that in the target image. The objects' shading in the latter is coherent with the surrounding environment. **Model like can hardly to learn the complex correspondence between them.**

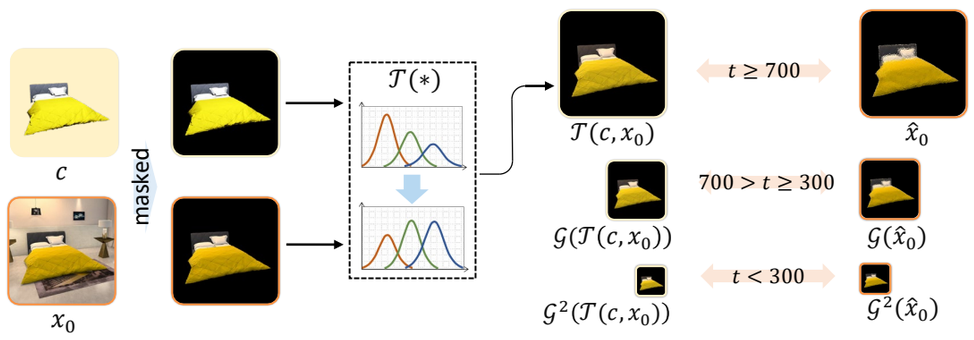
C. Methodology

• C.1 Shading Adaptive Condition Alignment (Fidelity ↑)



- Our approach utilizes the identical network as ControlNet. Shading Adaptive Condition Alignment (SACA) is introduced as an intensive training objective. In SACA, the shading of the condition image will be adapted to resemble that of the targeted output image through a **parametric shading adaptive transformation**. Then the pixel-level error between converted condition and current output is optimized to achieve alignment.

• C.2 Frequency Progression Training Schedule (Harmony ↑)



- The disharmony issue can be attributed to **the abnormal growth of high-frequency signals in the object area**. To tackle the problem, the Frequency Progression Training Schedule (FPTS) is developed. When the timestep is large, the SACA is only executed between the low-frequency bands of the condition and output image.

Xuan Han (hanxuan@tongji.edu.cn), Yihao Zhao (zhaoyihao@tongji.edu.cn) and Mingyu You (myyou@tongji.edu.cn)