Text2Light: Zero-Shot Text-Driven HDR Panorama Generation

ZHAOXI CHEN, S-Lab, Nanyang Technological University, Singapore
GUANGCONG WANG, S-Lab, Nanyang Technological University, Singapore
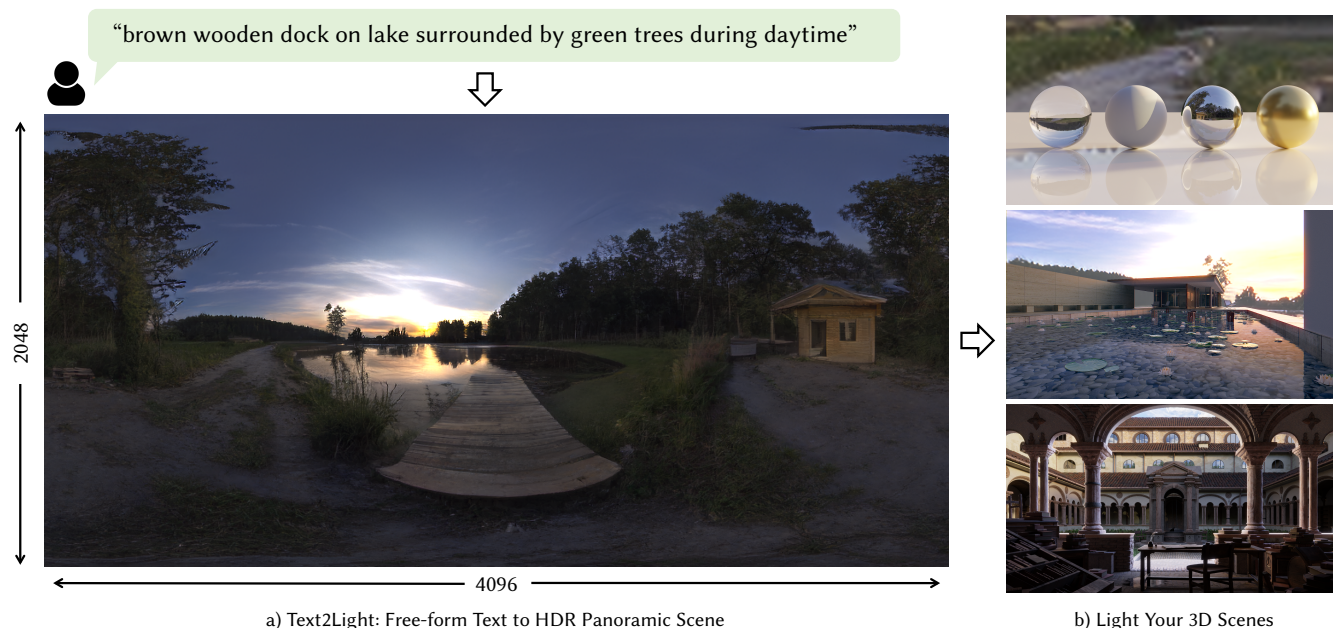ZIWEI LIU*, S-Lab, Nanyang Technological University, Singapore

a) Text2Light: Free-form Text to HDR Panoramic Scene

b) Light Your 3D Scenes

Fig. 1. **Zero-shot text-driven HDR panorama generation**. a) Text2Light can generate the HDR panorama in 4K+ resolution using the free-form text solely. b) Our high-quality results can be directly applied to downstream tasks, *e.g.,* light 3D scenes and immersive virtual reality.

High-quality HDRIs (High Dynamic Range Images), typically HDR panoramas, are one of the most popular ways to create photorealistic lighting and 360-degree reflections of 3D scenes in graphics. Given the difficulty of capturing HDRIs, a versatile and controllable generative model is highly desired, where layman users can intuitively control the generation process. However, existing state-of-the-art methods still struggle to synthesize high-quality panoramas for complex scenes. In this work, we propose a zero-shot text-driven framework, **Text2Light**, to generate 4K+ resolution HDRIs without paired training data. Given a free-form text as the description of the scene, we synthesize the corresponding HDRI with two dedicated steps: **1)** text-driven panorama generation in low dynamic range (LDR) and low resolution (LR), and **2)** super-resolution inverse tone mapping to scale up the LDR

panorama both in resolution and dynamic range. Specifically, to achieve zero-shot text-driven panorama generation, we first build dual codebooks as the discrete representation for diverse environmental textures. Then, driven by the pre-trained Contrastive Language-Image Pre-training (CLIP) model, a text-conditioned global sampler learns to sample holistic semantics from the global codebook according to the input text. Furthermore, a structure-aware local sampler learns to synthesize LDR panoramas patch-by-patch, guided by holistic semantics. To achieve super-resolution inverse tone mapping, we derive a continuous representation of 360-degree imaging from the LDR panorama as a set of structured latent codes anchored to the sphere. This continuous representation enables a versatile module to upscale the resolution and dynamic range simultaneously. Extensive experiments demonstrate the superior capability of Text2Light in generating high-quality HDR panoramas. In addition, we show the feasibility of our work in realistic rendering and immersive VR.

CCS Concepts: • **Computing methodologies** → **Computer vision**.

Additional Key Words and Phrases: Image Generation, Text-Driven Generation, Panorama Generation, High Dynamic Range Imaging

*corresponding author

1 INTRODUCTION

The emergence of the metaverse and virtual reality has fueled the demands for photorealistic rendering of 3D scenes [Chen and Liu 2022; Pandey et al. 2021]. Among all techniques, HDRIs are one of the most efficient and popular ways to produce realistic illuminations of computer graphics (CG) scenes and immersed environment textures. A high-quality HDRI that could be directly used in the graphic pipeline is expected to have sufficient details with high resolution and high dynamic range. However, the acquisition of HDRIs is difficult, since the capture process is both expensive and constrained. It motivates us to design an automatic pipeline.

Unlike common scenery images, HDRIs represent the radiance of the scene in high dynamic range and 360-degree view, indicating that they are more diverse and fine-grained in content. Furthermore, from the perspective of application and interaction, it is also desired to provide intuitive control of the whole generation process for layman users. For instance, due to the massive amount of detailed information contained in a single HDRI, manually constructing and refining it by the user is tough work. A more intuitive way would be controlling the model with explicit textual descriptions, *e.g.,* "brown wooden dock on lake surrounded by green trees during daytime", as shown in **Fig. 1**, which is friendly for users without expert knowledge.

Despite the great potential, text-driven HDR panorama generation has been rarely explored in recent years due to the following challenges. **1) Resolution:** Existing state-of-the-art generative models [Esser et al. 2021; Karras et al. 2021, 2019, 2020b; Lin et al. 2021] struggle to synthesize scene-level contents with ultra high resolution (4K+) and diverse yet sufficient details. **2) Coherence:** Unlike object-level images, scene-level panoramas often contain a large number of objects and structural layouts. It is hard to maintain the structural coherence and holistic semantics during panorama synthesis. **3) Aligning with texts:** It is expensive to collect panorama-text pairs for training. Generating scene-level content from free-form texts in a zero-shot manner remains unexplored. **4) HDR:** HDRIs cover a high dynamic range of lighting values, which often lead to unstable LDR-to-HDR regression. Coping with high dynamic range is still a non-trivial task during generation.

To solve these problems, we propose, **Text2Light**, the first framework for zero-shot text-driven HDR panorama generation, as shown in **Fig. 2**. Due to the complexity of HDR panorama in detailed textures, spherical structure, and high dynamic range, we decompose the generation process into two dedicated stages. **1) Stage I** generates a 360-degree panorama in low dynamic range (LDR) and low resolution (LR) given a description of the scene as the input. **2) Stage II** upscales the result of Stage I both in resolution and dynamic range by a super-resolution inverse tone mapping operator (SR-iTMO). **Stage I** focuses on preserving the structural coherence and holistic semantics without using text-panorama paired training data, and **Stage II** aims at generating high-quality HDRIs with ultra high resolution (4K+) and high dynamic range. With the proposed two-stage framework, users can generate a high-quality HDR panorama in 4K+ resolution using only the text description. The results can be used in the rendering engine to produce realistic illuminations and immersed environment textures, as shown in **Fig. 1**.

In Stage I, given the extremely high diversity of environmental textures, we leverage the codebook, which is commonly used in vector-quantized models [Esser et al. 2021; Oord et al. 2018], as a discrete representation of LDR textures. Specifically, to handle both the global coherence and local details, we introduce a new dual-codebook representation in a coarse-to-fine manner for distinct objectives. The global codebook aims to represent the coarse global appearance and scene-level semantics, while the local codebook focuses on encoding fine-grained details. With this hierarchical discrete representation, the task for Stage I is formulated as sampling appropriate embeddings from dual codebooks given the input text. To this end, we first derive the text embedding from the scene description with the aid of the pre-trained CLIP model [Radford et al. 2021]. Then, a text conditioned global sampler, which learns to align descriptions with scenes without text-panorama paired annotations, is employed to sample the holistic condition from the global codebook. Furthermore, a structure-aware local sampler is leveraged to sample features from the local codebook subject to the holistic condition and spherical positional encoding, which are finally decoded to LDR panoramas.

In Stage II, we treat each panorama as a continuous field distributed on a sphere, which serves as a complementary representation to codebooks in Stage I. We first encode the LDR panorama in a patch-wise manner to structured latent codes that are anchored on the sphere. Then, for any position (θ, ϕ) on the sphere, we derive its latent feature via interpolating among its nearest neighbors, which is further fed into an MLP-based SR-iTMO to produce the corresponding HDR value in higher resolution.

To facilitate scene generation, we construct a high-quality dataset, HDR360-UHD, which contains 4393 HDR panoramas of both outdoor and indoor scenes, ranging from 4K to 8K resolution.

We summarize our contributions as follows:

- We build a zero-shot text-driven framework, Text2Light, for HDR panorama generation, which produces photo-realistic illuminations and scene textures in 3D CG scenes controlled by free-form texts.
- We propose a dual-codebook discrete representation to describe both global scene-level semantics and fine-grained details. With dual codebooks, we introduce a text-conditioned global sampler, a structure-aware local sampler, and a spherical positional encoder to synthesize high-quality panoramas aligned with texts.
- We propose an efficient super-resolution inversed tone mapping operator to upscale the resolution and the dynamic range of images jointly.
- Text2Light shows the unprecedented ability to generate HDR panoramas in ultra 4K resolution with high fidelity.

2 RELATED WORK

High-resolution Generative Models. Recent years have witnessed the rapid growth of generative models in high-resolution image synthesis [Bond-Taylor et al. 2021; Chang et al. 2022; Gur et al. 2020; Ho et al. 2020; Huang et al. 2021; Karras et al. 2021, 2019, 2020b; Razavi et al. 2019; Rombach et al. 2021a; Skorokhodov et al. 2021; Wang et al. 2022b]. StyleGAN [Karras et al. 2019] and its variants
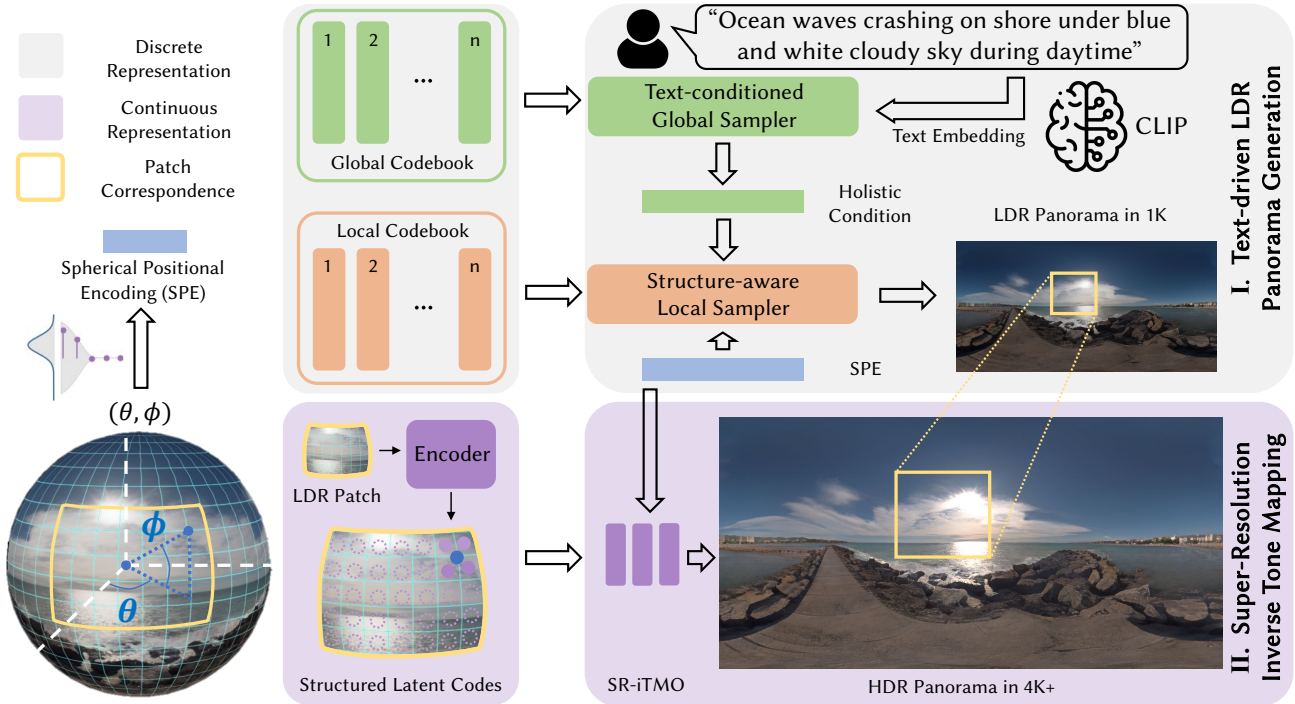
Fig. 2. **Overview of Text2Light**. We decompose the generation process of HDR panorama into two stages. **Stage I** translates the input text to LDR panorama based on a dual-codebook discrete representation. First, the input text is mapped to the text embedding by the pre-trained CLIP model [Radford et al. 2021]. Second, a text-conditioned global sampler learns to sample holistic semantics from the global codebook according to the input text. Then, a structure-aware local sampler synthesizes local patches and composites them accordingly. **Stage II** upscales the LDR result from Stage I based on structured latent codes as continuous representations. We propose a novel Super-Resolution Inverse Tone Mapping Operator (SR-iTMO) to simultaneously increase the spatial resolution and dynamic range of the panorama.

(e.g., StyleGAN2 and StyleGAN3) [Karras et al. 2021, 2020b] have substantial improvements in resolution and level of details, succeeding in synthesizing photorealistic images with a limited field of view. Yet, such methods are not able to generate rich and diverse content at the scene level. Besides, these methods train on the full image frame, which makes them memory intensive when scaling up. A further limitation is that they fail to keep the fidelity of both holistic appearance and fine-grained details due to the limited receptive field. In contrast, patch-based methods [Esser et al. 2021; Lin et al. 2019, 2021; Skorokhodov et al. 2021] are trained on small image patches to get rid of computational resource bottlenecks, succeeding in generating high-resolution scenery images. However, Taming Transformer [Esser et al. 2021] can only synthesize megapixel images with dense conditions, *e.g.,* segmentation maps or depth maps. InfinityGAN [Lin et al. 2021] claims to be able to generate infinite pixels, yet it fails to maintain the holistic semantic of the scene, which leads to low fidelity when zooming in.

Text to Image. Most existing text-to-image models [Ding et al. 2021; Nichol et al. 2022; Ramesh et al. 2022, 2021; Rombach et al. 2021b] are learned with supervision, using text-image pairs. Recently, the pre-trained CLIP model [Radford et al. 2021] enables zero-shot text-to-image generation and manipulation [Blattmann

et al. 2022; Hong et al. 2022; Jiang et al. 2022; Kim et al. 2022; Liu et al. 2021; Patashnik et al. 2021; Schaldenbrand et al. 2021; Wang et al. 2022a; Zhou et al. 2022] without paired data. However, directly guiding the generative model using CLIP may suffer from the semantic gap between the image and text embedding spaces in complex scenes. Given the complex semantics of a single scene, aligning the actual semantics of the input scene description with the generated images is still a non-trivial task.

Inverse Tone Mapping. In the context of photorealistic rendering, it is crucial for the environmental texture to have a high dynamic range that parameterizes the radiance of the real world. Intuitively, one could reconstruct HDR assets from the LDR counterpart via inverse tone mapping [Banterle et al. 2006]. In recent years, many convolutional methods [Chen et al. 2021; Kim et al. 2020; Lee et al. 2018; Liu et al. 2020; Marnerides et al. 2019; Raipurkar et al. 2021; Wang and Yoon 2021; Wei et al. 2021; Yu et al. 2021; Zhang and Lalonde 2017] aim to generate high-quality HDR from LDR images. However, these works leverage fully convolutional network which is not robust to different scales [Xu et al. 2014]. We propose to represent HDR panorama as a continuous field on a sphere. With this novel representation, we could efficiently perform inverse tone

mapping from arbitrary resolution to arbitrary resolution simply by multi-layer perceptrons.

3 TEXT2LIGHT

Our goal is to generate a high-quality HDR panorama in 4K+ resolution given a scene description as the input. The difficulty of this task has three folds, i.e., scene-level coherent textures with high-resolution, zero-shot text-driven generation, and HDR mapping. Existing state-of-the-art generators, such as StyleGAN3 [Karras et al. 2021] and InfinityGAN [Lin et al. 2021], still struggle to generate realistic scene-level panoramas in 1K resolution. We present **Text2Light**, a novel controllable framework to generate a high-quality HDR panorama from a free-form text as the scene description, which is illustrated in **Fig. 2**.

Text2Light mainly consists of two stages. In **Stage I** (Sec. 3.2), we propose a hierarchical framework to generate LDR panorama driven by the text T, which is based on a dual-codebook discrete representation. Driven by the CLIP [Radford et al. 2021] model, a text-conditioned global sampler learns to sample holistic semantics from the global codebook conditioned on the input text. Furthermore, a structure-aware local sampler learns to synthesize LDR panoramas patch-by-patch, guided by the holistic condition. In **Stage II** (Sec. 3.3), we propose a new super-resolution inverse tone mapping operator, termed as SR-iTMO. It upscales the result from Stage I both in spatial resolution and dynamic range to HDR 4K+ version using a continuous representation. With the SR-iTMO, Text2Light generates ultra-high-resolution and high-dynamic panoramas, which can be directly used in rendering engines or VR applications. Besides, different from common scenery images with limited field of views, panoramas can be regarded as 360-degree spherical fields. To integrate this prior into panorama generation, we introduce a spherical positional encoding (Sec. 3.1), or SPE for short, to represent 360-degree scenes in the entire pipeline.

3.1 Spherical Positional Encoding

Considering the intrinsic spherical structural property of panoramic images, we need a careful design to incorporate such a prior into panorama synthesis, which allows the framework to learn position-sensitive signals of a 3D scene. Given a panoramic image $I \in \mathbb{R}^{H \times W \times 3}$, we represent each pixel as a 3D point on an unit sphere S. Let $I(i, j)$ be the pixel value at (i, j) in I where $i \in \{1, ..., H\}$ and $j \in \{1, ..., W\}$, we represent its spherical counterpart as $S(\theta, \phi, r)$:

$$
\begin{aligned}
S(\theta, \phi, r) &= I(i, j), \\
\theta &= (2i/H - 1)\pi, \\
\phi &= (2j/W - 1)\pi/2, \\
r &= 1.
\end{aligned}
\tag{1}
$$

Intuitively, Eqn. 1 inverses the process of the equirectangular projection which converts a 360-degree raw image into a 2D panorama. The center of the sphere S is exactly the position of the camera that takes the shot in the 3D world.

Moreover, as deep networks are biased towards learning lower frequency functions [Rahaman et al. 2019], we leverage Fourier

positional encoding $\gamma(\cdot)$ to map (θ, ϕ) to a higher dimensional space:

$$
\begin{aligned}
\gamma(\theta) &= [\sin(2^0 \pi \theta), \cos(2^0 \pi \theta), ..., \sin(2^{L-1} \pi \theta), \cos(2^{L-1} \pi \theta)], \\
\gamma(\phi) &= [\sin(2^0 \pi \phi), \cos(2^0 \pi \phi), ..., \sin(2^{L-1} \pi \phi), \cos(2^{L-1} \pi \phi)],
\end{aligned}
\tag{2}
$$

where L is set to 4 in our experiments.

Such a representation acts as a strong inductive bias in our model, in the way of preserving high-frequency details and spherical structure in Stage I and introducing geographical priors in Stage II. We will introduce the usage of this representation in related sections.

3.2 Stage I: Text-driven LDR Panorama Generation

In Stage I, we propose a hierarchical framework that focuses on generating a low resolution (1024×512) panorama in low dynamic range using solely natural language descriptions. It can be decomposed into three components. **i) Dual-codebook representation:** Building dual codebooks to represent the holistic appearance and local fine-grained details. **ii) Text-conditioned global sampler:** Sampling holistic semantics from the global codebook conditioned on the input text. **iii) Structure-aware local sampler:** Synthesizing structure-aware LDR panoramas guided by the holistic condition patch-by-patch. An overview of Stage I is shown in **Fig. 3**.

3.2.1 Image Tokenization. Vector-Quantized Variational AutoEncoder (VQVAE) [Oord et al. 2018] is proposed to learn a discrete representation, a codebook $\mathcal{Z}$, by reconstructing images. Given an input image I, VQVAE leverages an encoder E to obtain the continuous feature $\hat{z}$, i.e., $\hat{z} = E(I) \in \mathbb{R}^{h \times w \times c_z}$. Then a quantization is performed onto its closest codebook entry z_k to obtain the discrete representation z_q:

$$
z_q = \underset{z_k \in \mathcal{Z}}{\mathrm{argmin}} ||\hat{z} - z_k||_2.
\tag{3}
$$

Then the image is reconstructed by a decoder G as $\hat{I} = G(z_q)$. To improve the perceptual quality at a high compression rate [Esser et al. 2021], a discriminator D [Isola et al. 2017] is also introduced to discriminate between real and generated image patches. Thus, the overall training objective is:

$$
\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + ||\mathrm{sg}(z_q) - \hat{z}||_2^2 + ||\mathrm{sg}(\hat{z}) - z_q||_2^2 + \mathcal{L}_{\mathrm{GAN}},
\tag{4}
$$

where $\mathcal{L}_{\mathrm{rec}}$ is the perceptual loss [Zhang et al. 2018] between I and $\hat{I}$, $\mathcal{L}_{\mathrm{GAN}} = \log D(I) + \log(1 - D(\hat{I}))$, and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation.

3.2.2 Dual-codebook Discrete Representation. Given that textures of outdoor and indoor scenes are diverse and complicated, a single VQVAE (or StyleGAN) often fails to synthesize coherent appearance and sufficient local details simultaneously (see **Fig. 6**). Therefore, we derive a hierarchical discrete representation of LDR panoramas as dual codebooks. Specifically, to represent holistic semantics and coarse but global features, we first build a *global codebook* $\mathcal{Z}_g$ via training a VQVAE (Eqn. 4) to reconstruct a downsampled panorama $I_g \in \mathbb{R}^{128 \times 256 \times 3}$. Let E_g and G_g be the corresponding encoder and decoder, respectively. The encoder E_g turns I_g into a continuous feature $\hat{z}_g = E_g(I_g) \in \mathbb{R}^{8 \times 16 \times c_z}$, then the coarse feature z_g is derived from $\mathcal{Z}_g$ via quantization (Eqn. 3). Furthermore, to represent sufficient details and diverse textures across different scenes, we
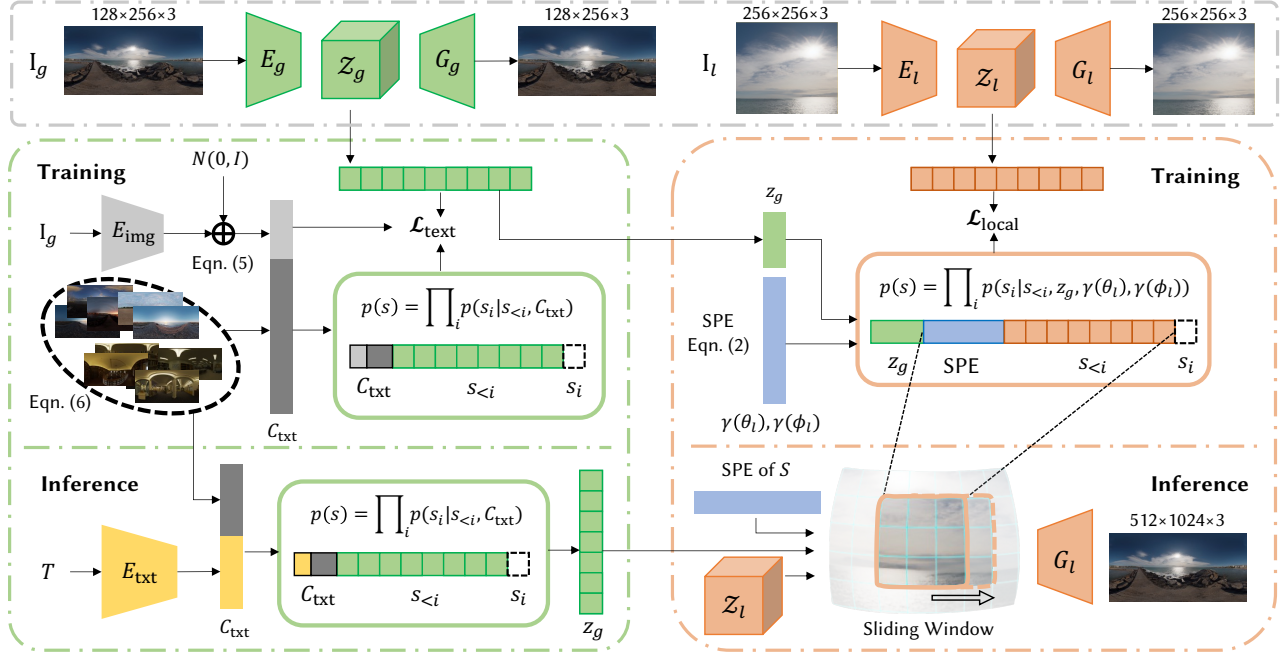
Fig. 3. **Overview of Stage I**. In Stage I, we aim to generate an LDR panorama using solely natural language descriptions. The hierarchical framework can be decomposed into three components. **i) Dual-codebook Discrete Representation:** Building dual codebooks, $\mathcal{Z}_g, \mathcal{Z}_l$, to represent the holistic appearance and local fine-grained details. **ii) Text-conditioned Global Sampler:** Sampling holistic semantics from the global codebook according to the input text. **iii) Structure-aware Local Sampler:** Synthesizing LDR panorama guided by the holistic condition in a structure-aware manner. The process of training and inference is introduced, respectively.

build a *local codebook* $\mathcal{Z}_l$ via training another VQVAE to reconstruct high-resolution patches $I_l \in \mathbb{R}^{256 \times 256 \times 3}$. Let E_l and G_l be the corresponding encoder and decoder, we can obtain a fine-grained feature $z_l \in \mathbb{R}^{16 \times 16 \times c_z}$ by E_l and $\mathcal{Z}_l$. During training, the patch I_l is randomly cropped from the full LDR panorama $I \in \mathbb{R}^{512 \times 1024 \times 3}$, and c_z is set to 256.

3.2.3 Text-conditioned Global Sampler. Intuitively, zero-shot text-driven synthesis can be achieved by directly optimizing the generative model using the guidance from CLIP [Liu et al. 2021; Patashnik et al. 2021]. However, such a strategy is not robust in generating large-scale scenes, *e.g.,* the top two rows of **Fig. 7**, as the model has no sense of the holistic semantics. Therefore, we need to condition the generation process on holistic semantics derived from input texts. Driven by this observation, we propose a text-conditioned global sampler that samples features from the global codebook, to align the text with scenes in a language-free training.

Our key insight is to align the sampler with the text by unsupervised learning techniques, *i.e.,* K-nearest neighbors and contrastive learning. The idea is to generate a text condition C_{txt} that aims to approximate the real semantic meaning of the target panorama I without text-panorama paired training data. Denote E_{txt} and E_{img} as the text encoder and image encoder of CLIP [Radford et al. 2021], respectively. Inspired by [Zhou et al. 2022], we first perturb the

image feature $E_{img}(I)$ to generate a pseudo text feature:

$$\hat{C}_{txt} = (1 - \alpha)E_{img}(I) + \alpha\epsilon||E_{img}(I)||_2 / ||\epsilon||_2, \qquad (5)$$

where α is a fixed hyper-parameter and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the Gaussian noise. Furthermore, we fetch the top-K nearest image embeddings $C_{knn} = \{E_{img}(I_k)\}_{k=1}^K$ of $\hat{C}_{txt}$ as the additional condition. Consequently, the text condition C_{txt} is written as:

$$C_{txt} = \{C_{knn}|\hat{C}_{txt}\}. \qquad (6)$$

Specifically, we adapt the transformer [Vaswani et al. 2017] into a text-conditioned sampler, which samples holistic features from the global codebook $\mathcal{Z}_g$ given the input text T. We train the sampler conditioned on C_{txt} in an autoregressive manner. The discrete representation z_g is equivalent to a sequence $s \in \{0, ..., |\mathcal{Z}_g| - 1\}^{8 \times 16}$ of indices from the holistic codebook $\mathcal{Z}_g$. Sampling global features is formulated as autoregressive next-index prediction conditioned on C_{txt}. Given indices $s_{<i}$, the transformer-based sampler learns to predict the distribution of possible next indices, *i.e.,* $p(s_i|s_{<i}, C_{txt})$, to compute the likelihood of the full representation as $p(s) = \prod_i p(s_i|s_{<i}, C_{txt})$. We maximize the likelihood to obtain a text-aligned sampler:

$$\mathcal{L}_{txt} = \mathbb{E}_{I \sim p(I)}[-\log p(s)] + \mathcal{L}_{con}, \qquad (7)$$
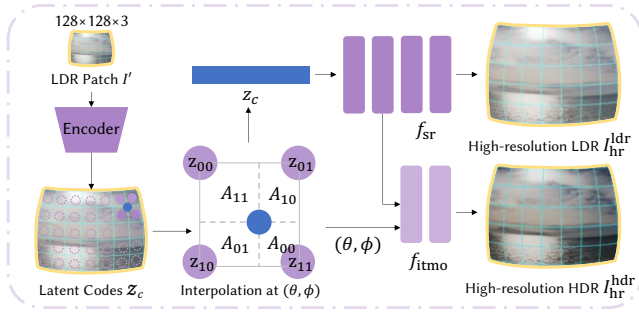
Fig. 4. **Overview of Stage II.** In Stage II, we upscale the result from Stage I both in spatial resolution and dynamic range using a continuous representation. For any position on the sphere (θ, ϕ) in a continuous domain, we derive the latent codes $\mathcal{Z}_c$ and get the feature vector z_c via area-based interpolation. Then, simple MLPs are adopted to query HDR values at (θ, ϕ).

where $\mathcal{L}_{\text{con}}$ is the contrastive regularization:

$$\mathcal{L}_{\text{con}} = -\tau \sum_{i=1}^{n} \log \frac{\exp((E_{\text{img}}(I_i) \cdot \hat{C}_{\text{txt}})/\tau)}{\sum_{j=1}^{n} \exp((E_{\text{img}}(I_j) \cdot \hat{C}_{\text{txt}})/\tau)}, \quad (8)$$

where we maximize the cosine similarity between the real image feature and the pseudo text feature to reduce the gap of the joint text-image space of CLIP. $\tau \in (0, 1]$ is a hyper-parameter.

Trained by Eqn. 7, the sampler is aligned with real text features, which enables zero-shot transferability by replacing $\hat{C}_{\text{txt}}$ with $E_{\text{txt}}(T)$ during inference (**Fig. 3**). Intuitively, as the global codebook represents the full panorama at low resolution (256×128), our text-conditioned global sampler takes a glance at the entire scene and outputs the holistic condition z_g to guide the following step on local patch synthesis.

3.2.4 Structure-aware Local Sampler. Given the holistic condition z_g, we focus on synthesizing local patches with sufficient fine-grained details. Thanks to the local codebook $\mathcal{Z}_l$, we are able to generate individual patches with high fidelity. Furthermore, the synthesizer needs to be aware of the intrinsic structure of panorama to properly composite patches according to the holistic condition, where the spherical positional encoding (Sec. 3.1) comes in.

In specific, we employ another transformer as a structure-aware sampler that samples features from the local codebook $\mathcal{Z}_l$. Given the holistic condition z_g, we want to synthesize the local patch $I_l \in \mathbb{R}^{256 \times 256 \times 3}$ with spherical representation $S(\theta_l, \phi_l, 1)$, where $\theta_l \in \mathbb{R}^{256 \times 256}, \phi_l \in \mathbb{R}^{256 \times 256}$ are derived from Eqn. 1. The discrete representation of I_l is z_l which is equivalent to a sequence $s \in \{0, ..., |\mathcal{Z}_l| - 1\}^{16 \times 16}$ of indices from the local codebook $\mathcal{Z}_l$. SPE is directly prepended to the top of the sequence s after the input tokens are mapped to embeddings. The corresponding spherical coordinates of s is a vector with a dimension of $2 \times 256 = 512$. And the SPE with Fourier features according to Eqn. 2 is derived and stacked to form a tensor with a size of 18×512, which is then concatenated with the embeddings in the transformer. The structure-aware sampler learns to predict the next possible indices, *i.e.*, $p(s_i|s_{<i}, z_g, \gamma(\theta_l), \gamma(\phi_l))$. We

maximize the log-likelihood over the full representation:

$$\mathcal{L}_{\text{local}} = \mathbb{E}_{I_l \sim p(I_l)} \left[-\log p(s) \right], \quad (9)$$

where $p(s) = \prod_i p(s_i|s_{<i}, z_g, \gamma(\theta_l), \gamma(\phi_l))$.

During inference, we utilize this sampler in a sliding-window manner, as illustrated in **Fig. 3**. For example, to generate a full panorama with the holistic condition z_g, we first generate its SPE as $\theta \in \mathbb{R}^{512 \times 1024}$ and $\phi \in \mathbb{R}^{512 \times 1024}$. Then, we patch-wisely slide the attention window of the sampler with size of 256×256 to obtain the full representation $z_l \in \mathbb{R}^{32 \times 64 \times c_z}$. Finally, the full panorama in 1K resolution can be generated using the decoder of the local codebook, *i.e.*, $I = G_l(z_l)$.

3.3 Stage II: Super-Resolution Inverse Tone Mapping

In Stage II, Text2Light upscales the result from Stage I both in spatial resolution and dynamic range and outputs a high-quality HDR panorama in 4K+. On complementary of the discrete representation in Sec. 3.2.2, we build a continuous representation that can be used to query HDR values at any spherical position on S:

$$S(\theta, \phi, 1) = f_c(z_c, \theta, \phi), \quad (10)$$

where z_c is the feature derived from the LDR patch, and $S(\theta, \phi, 1)$ is introduced in Sec. 3.1. An overview of Stage II is shown in **Fig. 4**.

3.3.1 Structured Latent Codes as Continuous Representations. The feature z_c in Eqn. 10 is derived from a set of latent codes $\mathcal{Z}_c$ which represents the features distributed in the continuous domain of the sphere S. Given an LDR panorama I, we randomly crop an image patch $I' \in \mathbb{R}^{128 \times 128 \times 3}$ whose corresponding field of view on the sphere is $S'(\theta', \phi', 1)$. Then the structured latent codes $\mathcal{Z}_c$ are derived as a pixel-aligned feature map $\mathcal{Z}_c = E(I') \in \mathbb{R}^{128 \times 128 \times c'_z}$, where c'_z is the dimension of each latent code. Then we compute the feature vector z_c at any (θ, ϕ) as an area-based interpolation:

$$z_c = \sum_{i \in \{00, 01, 10, 11\}} \frac{A_i}{A} z_i, \quad z_i \in \mathcal{Z}_c, \quad (11)$$

where z_i is the nearest latent code at (θ, ϕ), and A_i is the area of the rectangle as illustrated in **Fig. 4**.

3.3.2 Super-Resolution Inversed Tone Mapping Operator. With the help of the continuous feature z_c, we can leverage a composited function as $f_c = f_{\text{sr}} \circ f_{\text{itmo}}$, termed as Super-Resolution Inversed Tone Mapping Operator (SR-iTMO), to map value points in LDR patches to corresponding value points in high-resolution HDR patches. Specifically, the SR-iTMO consists of two tiny MLPs, f_{sr} and f_{itmo}, that sequentially upscales the spatial resolution and dynamic range, as illustrated in **Fig. 4**. f_{sr} is an MLP consisting of four layers, which upscales the LDR patch in resolution:

$$I_{\text{hr}}^{\text{ldr}}, c_{\text{hr}} = f_{\text{sr}}(z_c), \quad (12)$$

where c_{hr} is the intermediate feature of the second layer of f_{hr}. Then, f_{itmo}, another MLP consists of two layers, upscales the dynamic range of the high-resolution patch:

$$I_{\text{hr}}^{\text{hdr}} = f_{\text{itmo}}(c_{\text{hr}}, \theta, \phi), \quad (13)$$

where θ and ϕ are the spherical coordinates of the queried point.
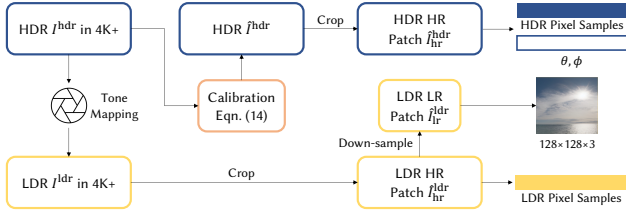
Fig. 5. **Data preparation and calibration of Stage II**. As HDR360-UHD contains high-resolution HDR panoramas ranging from 4K to 8K, we need to carefully generate training samples to train SR-iTMO (Sec. 3.3.4), which can be decomposed into three steps which are tone mapping, calibration, and obtaining pixel samples, respectively.

3.3.3 Data Preparation and Calibration.

To train our SR-iTMO module, whose goal is reconstructing HDR patches with higher resolution from LDR patches, we need to generate paired data, *i.e.,* $\hat{I}^{\text{ldr}}_{\text{lr}} - \hat{I}^{\text{hdr}}_{\text{hr}}$ pairs, as training samples. We illustrate the whole process in **Fig. 5**.

First, we generate LDR panoramas $I^{\text{ldr}} \in [0, 1]$ from HDR panoramas $I^{\text{hdr}} \in [0, +\infty)$ via Reinhard tone mapping [Reinhard and Devlin 2005]. Then, we calibrate raw HDR panoramas to remove scale ambiguity caused by different luminance scales across different scenes. Inspired by [Yu et al. 2021], we generate HDR ground truth based on the scale invariance of luminance. Given two HDR images I^{hdr}_1 and I^{hdr}_2, we define that they are luminance scale-invariant if and only if $I^{\text{hdr}}_1 = \kappa I^{\text{hdr}}_2, \kappa \in \mathbb{R}^+$. For each HDR panorama I^{hdr}, we downscale it to its luminance scale-invariant version $\hat{I}^{\text{hdr}}$ as the following:

$$\hat{I}^{\text{hdr}} = \frac{\sum(M \odot I^{\text{ldr}})}{\sum(M \odot I^{\text{hdr}})} \cdot I^{\text{hdr}}, \qquad (14)$$

where $\odot$ denotes the element-wise multiplication. And M is a mask to determine the position of being calibrated based on $I^{\text{ldr}} \in \mathbb{R}^{h \times w \times 3}$:

$$M(i, j) = \begin{cases} 1, & \sum_{k=1}^{3} I^{\text{ldr}}(i, j, k) < 3\sigma, \\ 0, & otherwise, \end{cases} \qquad (15)$$

where $i = 0, 1, ..., h - 1$ and $j = 0, 1, ..., w - 1$. We set $\sigma = 0.83$ in our experiments.

Furthermore, we generate training samples at different resolutions. Here we take a single training pair as an example. Given a calibrated HDR panorama $\hat{I}^{\text{hdr}}$, we randomly crop a patch $\hat{I}^{\text{hdr}}_{\text{hr}} \in \mathbb{R}^{128\beta \times 128\beta \times 3}$ as the high-resolution HDR, where $\beta \in [1, 4]$ is a random scaling factor. The input LDR patch is generated by downsampling $\hat{I}^{\text{ldr}}_{\text{hr}}$ to $\hat{I}^{\text{ldr}}_{\text{lr}} \in \mathbb{R}^{128 \times 128 \times 3}$. Note that, $\hat{I}^{\text{ldr}}_{\text{hr}} \in \mathbb{R}^{128\beta \times 128\beta \times 3}$ is cropped from the corresponding tone-mapped LDR panorama I^{ldr} at the same position as $\hat{I}^{\text{hdr}}_{\text{hr}}$. To match the size of pixel samples, we randomly choose 128×128 points from $\hat{I}^{\text{hdr}}_{\text{hr}}$ and their corresponding spherical coordinates as high-resolution HDR ground-truth.

3.3.4 Training.

We jointly train the encoder E and the SR-iTMO f_c in an end-to-end manner. Note that, with the help of our continuous representation, we are able to directly train the networks at any

resolution. Our dataset contains HDR panoramas with resolutions ranging from 4K to 8K, and we do not scale them down to a fixed resolution during training. We first encourage the model to learn to increase the spatial resolution via minimizing:

$$\mathcal{L}_{\text{sr}} = \frac{1}{n} \sum ||I^{\text{ldr}}_{\text{hr}} - \hat{I}^{\text{ldr}}_{\text{hr}}||_1, \qquad (16)$$

where n is the number of pixel samples, and $I^{\text{ldr}}_{\text{hr}}$ is the network prediction from Eqn. 12.

Besides, the model learns to expand the dynamic range by minimizing the following scale-invariant objective:

$$\mathcal{L}_{\text{itmo}} = \frac{1}{n} \sum D^2(I^{\text{hdr}}_{\text{hr}}, \hat{I}^{\text{hdr}}_{\text{hr}}) - [\frac{1}{n} \sum D(I^{\text{hdr}}_{\text{hr}}, \hat{I}^{\text{hdr}}_{\text{hr}})]^2, \qquad (17)$$

where $D(I_1, I_2) = \log(I_1) - \log(I_2)$ is a distance in logarithmic scale.

Thus, the overall learning objective of Stage II is a summation as $\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{sr}} + \mathcal{L}_{\text{itmo}}$. By this objective, the SR-iTMO learns to increase the resolution and dynamic range simultaneously.

4 HDR360-UHD DATASET

Currently, there is no publicly available dataset for **both** indoor and outdoor scenes. [Gardner et al. 2017] contributes an HDR dataset that contains only 2100 indoor scenes, and [Zhang and Lalonde 2017] contributes around 200 outdoor HDR panoramas. To achieve a high-quality generative model, we newly collect thousands of panoramas from online resources and combine them with aforementioned datasets to form a new high-quality HDRI dataset, named HDR360-UHD, which contains 4392 HDR panoramas with resolutions ranging from 4096×2048 (4K) to 8192×4096 (8K).

Specifically, there are 1891 outdoor scenes and 2501 indoor scenes in HDR360-UHD. We further augment the data based on the intrinsic horizontal cyclicity of panoramas by clockwise rotating the panorama that wraps on the sphere ten times. Thus, the total amount of panoramas are 48,312. To facilitate text-driven scene synthesis, we also collect 1901 pieces of scene descriptions. Note that, in our experiments, these texts are only used as input during inference.

5 EXPERIMENTS

5.1 Implementation Details

Our full model is trained stage by stage. We only utilize the tone-mapped LDR samples with 1K resolution in HDR360-UHD for the generation task in Stage I. And we split the dataset into a training set and a testing set for the inverse tone mapping task in Stage II. Specifically, the training set contains 45,056 samples, and the testing set contains 3,256 samples. We adopt Adam [Kingma and Ba 2014] optimizer for training.

5.1.1 Stage I.

We first train the codebooks $\mathcal{Z}_g$ and $\mathcal{Z}_l$ for 60k iterations with a batch size of 12. Note that we replace all convolution layers of E_g and G_g with circular convolution layers in order to incorporate the cyclicity of the panorama. When training the text-conditioned global sampler, we freeze all parameters of CLIP [Radford et al. 2021]. The hyper-parameter α is set to 0.25, and the number of nearest neighbors is set as $K = 5$. The text-conditioned global sampler is trained with a batch size of 16 for 100k iterations. Furthermore, we train the structure-aware local sampler with all other modules fixed. The batch size is set as 24.
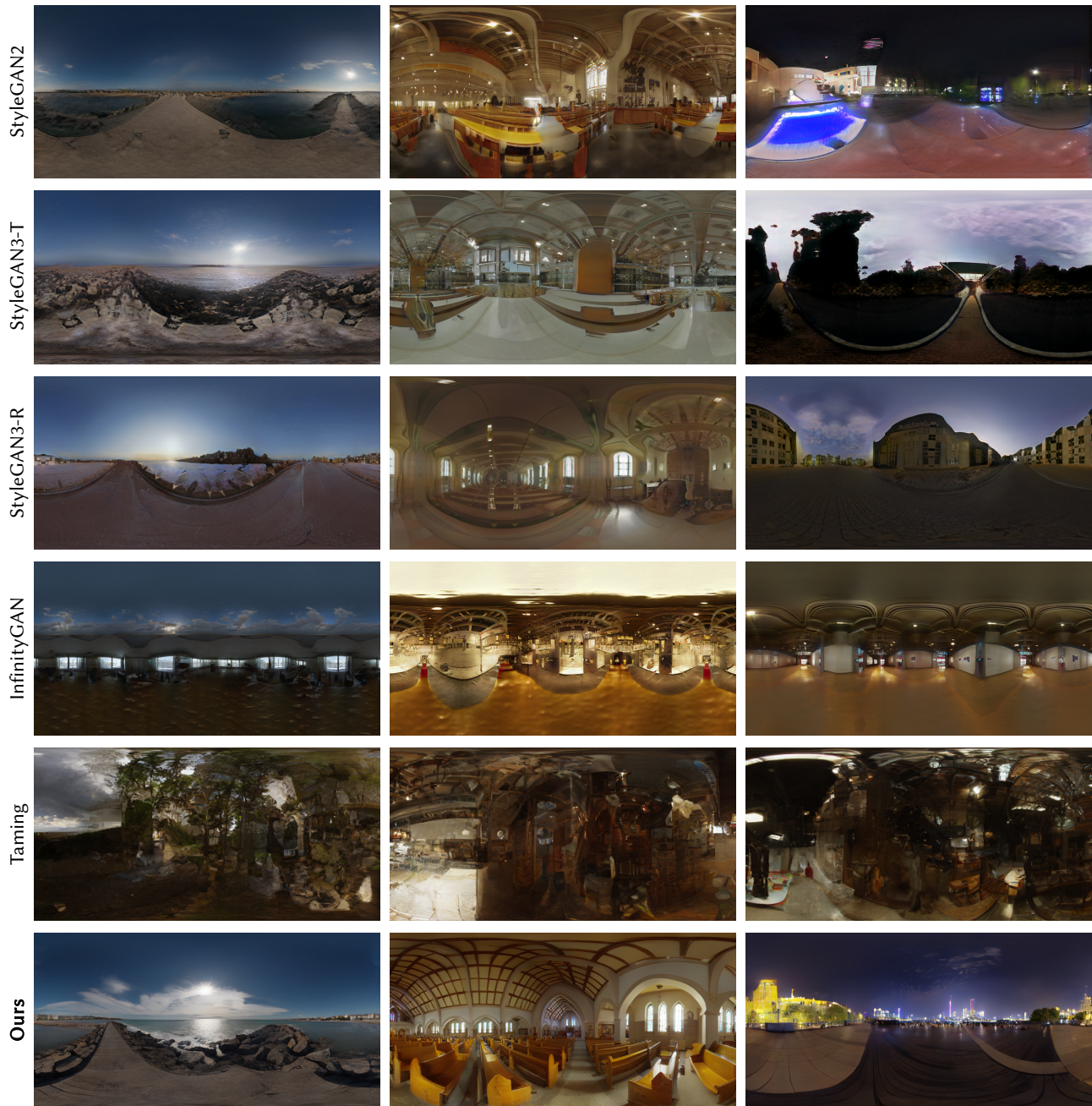
Fig. 6. **Visual Comparisons of LDR panoramas generation. Zoom in for details**. StyleGAN-based methods [Karras et al. 2021, 2020b] are directly trained on the full panorama with 1K resolutions. Their results seem to be realistic, yet contain local artifacts when zooming in. Plus, they fail to maintain the spherical structure of panoramic scenes. InfinityGAN [Lin et al. 2021] generates repeated patterns near the horizon while synthesizing textureless pixels near the pole. The coordinate-conditioned Taming Transformer [Esser et al. 2021] fails to capture the diverse and structural intrinsic of panoramas. Text2Light succeeds in generating high-quality panoramas with sufficient details and diverse scenes.

Table 1. **Quantitative results on LDR panorama generation.** The top three techniques are highlighted in red, orange, and yellow, respectively.

Methods	FID ↓	IS ↑	PQ ↑	SI ↑
StyleGAN2	14.62	5.07±0.17	2.43	2.20
StyleGAN3-T	16.61	5.06±0.18	2.57	2.42
StyleGAN3-R	15.13	5.01±0.19	2.19	1.87
InfinityGAN	104.62	4.96±0.16	2.03	1.80
Taming Transformer	112.38	3.04±0.08	2.73	2.55
w/o global	308.38	1.82±0.02	1.41	1.36
w/o SP (Eqn. 1)	56.37	4.22±0.12	3.45	3.18
w/o SPE (Eqn. 2)	31.19	6.32±0.27	3.23	3.09
Ours	10.72	6.65±0.22	4.26	4.25

5.1.2 Stage II. We leverage the backbone network in EDSR [Lim et al. 2017] as the encoder E to generate structured latent codes from LDR patches. f_{sr} is implemented as a four-layer MLP with 256 hidden units at each layer. f_{itmo} is another MLP that consists of two layers. During inference, we use the exponential function to map the output of f_{itmo} from the logarithmic scale to the original scale.

5.2 Comparison Methods and Evaluation Metrics

For a fair comparison, we show our significant improvements over prior arts in panorama generation, zero-shot text-driven synthesis, and inverse tone mapping, respectively.

5.2.1 Panorama Generation. We use Fréchet Inception Distance (**FID**) [Heusel et al. 2017] and Inception Score (**IS**) [Salimans et al. 2016] as metrics to evaluate the quality of generated panoramas.

Moreover, a user study is also conducted to evaluate the quality of panoramas, where each panorama is presented individually to the user for scoring. A total of 25 users are asked to: 1) score the perceptual quality (**PQ**), and 2) score the structural integrity (**SI**) of the panoramas. All scores are in the range of 5, and 5 indicates the best. We compare four state-of-the-art generators as follows:

- **StyleGAN2** [Karras et al. 2020b] is a style-based generative model which synthesizes images from a latent distribution. We directly train it on the LDR panoramas with 1K resolution. Note that, we employ the adaptive augmentation technique [Karras et al. 2020a] to stabilize training.
- **StyleGAN3** [Karras et al. 2021] is an alias-free version of StyleGAN2 whose internal representation is equivariant to translation and rotation. We consider two configurations (StyleGAN3-T and StyleGAN3-R) in our experiments.
- **InfinityGAN** [Lin et al. 2021] is a patch-based method that aims at arbitrary-sized image generation. We use the default setting to train their model on our LDR patches.
- **Taming Transformer** [Esser et al. 2021] is a VQVAE-based generative model, succeeding in synthesizing high-resolution images with dense conditions, *e.g.,* depth maps and segmentation maps. As there is no direct condition for HDR panoramas, we train a simple model that conditions on pixel coordinates.

5.2.2 Zero-shot Text-driven Synthesis. We further evaluate the ability of our framework on zero-shot text-driven panorama synthesis,

validating that simple extensions of previous methods cannot handle this challenging task. Same as Sec. 5.2.1, we use **FID** and **IS** to evaluate the quality of generation. Besides, another user study is performed for evaluation. The users are asked to: 1) score the perceptual quality (**PQ**), and 2) score the text consistency (**TC**) between the input text and the output panorama. All scores are in the range of 5, and 5 indicates the best. We compare two text-driven synthesis methods as follows:

- **StyleCLIP** [Patashnik et al. 2021] is a method that directly optimizes the latent space of StyleGAN using the guidance of CLIP. It iterates multiple times to maximize the cosine similarity of the generated image and the input text.
- **FuseDream** [Liu et al. 2021] is an updated version of Style-CLIP, which applies augmentations on images to improve the robustness of aligning free-form texts to the generated image.

Note that, we use the StyleGAN2 model in Sec. 5.2.1 as the underlying generator for these two methods.

5.2.3 Inverse Tone Mapping. We use Mean Absolute Error (**MAE**) and Root Mean Square Error (**RMSE**) as metrics to evaluate the performance of inverse tone mapping operators. Besides, to evaluate the quality of HDRIs in lighting 3D assets, we use them to render four shader balls with different materials, comparing the visual quality. We compare with three methods as follows:

- **LANet** [Yu et al. 2021] is a luminance attentive network for HDR reconstruction from a single LDR image, which uses convolutional networks with a branch to output segmentation masks of the area with high luminance.
- **ExpandNet** [Marnerides et al. 2019] is a fully convolutional model with three branches. It decomposes the image into local, global, and dilate features, which are fused at the bottom of the network.
- **HDR-CNN** [Eilertsen et al. 2017] is a convolutional autoencoder for HDR reconstruction. Its encoder converts an LDR input to a latent feature representation, and the decoder reconstructs this into an HDR image in the log domain.

5.3 Evaluations on LDR Panorama Generation

We first report the quantitative results in terms of LDR panoramas generation compared with representative methods, as shown in Table 1. Note that, three variants of StyleGAN are directly trained on the full panorama with 1K resolution. The rest two baselines are patch-based during training, and only synthesize the full panorama during evaluation. In a nutshell, Text2Light achieves the lowest FID and the highest IS among all methods, which demonstrates the high quality of our generation. Plus, the result of user study indicates that our method achieves the best perceptual quality to human users while also maintaining the structural coherence of panoramic scenes.

Furthermore, we qualitatively compare our method with baselines in **Fig. 6**. Although three variants of StyleGAN, which are trained on the full-framed panoramas, achieve relatively low FID metric, they cannot accurately catch the intrinsic structural property of panorama, leading to low fidelity. For example, the church interior synthesized by StyleGAN2 seems to be realistic at first glance,

Fig. 7. Visual Comparisons of zero-shot text-driven panorama synthesis. Zoom in for details. (Above) **Comparison with baselines.** Compared with two optimization-based methods, the proposed method generates diverse textures in high-fidelity that perfectly match the input text, even for indoor scenes with high complexity. (Below) **Ablation studies.** The use of KNN (K-nearest neighbors) significantly improves the robustness of zero-shot text-driven synthesis. Plus, contrastive regularization helps the model better match the fine-grained details with specific texts, *e.g.,* "sunset" with sunset clouds in the first column, and "body of water" in the third column.

but the ceiling and bench are distorted and broken. InfinityGAN generates repeated patterns near the horizontal areas while synthesizing textureless pixels near the pole. The coordinate-conditioned Taming Transformer fails to synthesize the full panorama when trained on patches without any dense condition. The results are attributed to the lack of inductive bias of a plain transformer. In contrast, Text2Light learns to synthesize high-quality LDR panoramas thanks to careful designs on global semantic alignment and structure-aware local patch synthesis.

In addition, we show nearest neighbors of our generated results based on the cosine similarity in the image space of CLIP [Radford

Table 2. **Quantitative results on zero-shot text-driven synthesis.** The top three techniques are highlighted in red, orange, and yellow, respectively.

Methods	FID ↓	IS ↑	PQ ↑	TC ↑
StyleCLIP	57.86	4.28±0.18	1.98	1.33
FuseDream	38.16	5.89±0.28	2.41	1.57
w/o KNN	52.85	4.77±0.23	2.14	1.23
w/o $\mathcal{L}_{con}$	34.70	6.45±0.26	3.81	3.50
Ours	32.01	6.46±0.37	4.38	4.56

Table 3. **Quantitative results on inverse tone mapping.** The top two techniques are highlighted in red, and orange, respectively.

Methods	MAE ↓	RMSE ↓
LANet	0.3218	51.73
ExpandNet	0.1894	26.68
HDR-CNN	0.1713	26.40
w/ single MLP	0.1798	31.36
Ours	0.1442	26.38

et al. 2021], as shown in **Fig. 8**. The generated panoramas are not exactly identical to the training samples.

5.4 Evaluations on Zero-shot Text-driven Synthesis

We further evaluate the capability of Text2Light in generating text-aligned content in high fidelity and zero-shot manner. The quantitative results compared with two CLIP+GAN methods are presented in Table 2. Our careful design of the text-conditioned global sampler significantly improves the generation quality.

Besides, the visual comparisons are shown in **Fig. 7**. It is obvious that the simple combination of generative models and CLIP cannot catch the accurate semantics of the entire scene given the free-form text, which is validated by the failure of StyleCLIP [Patashnik et al. 2021]. Indeed, the robust optimization process in FuseDream [Liu et al. 2021] can somewhat improve the consistency between the content and the text. However, it still cannot convert all information from the text into a panoramic scene. For example, driven by the text "sunset by the ocean", FuseDream only synthesizes "ocean" but ignores the "sunset". In contrast, our model generates diverse textures that perfectly match the input text, even for indoor scenes with high complexity. Our hierarchical scheme enables the synthesizer both to align with the scene description globally and generate coherence complex details locally, *e.g.,* cathedral interior with coherent structures of benches and ceilings.

5.5 Evaluations on Inverse Tone Mapping

The quantitative comparisons on the testing set are presented in Table 3. Note that, we set the upscaling factor $\beta = 1$ of our method for fair comparisons. We validate the potential of realizing iTMO as MLPs to get rid of leveraging heavy fully convolutional architectures. Our iTMO achieves the lowest error on the testing set.

We make further visual comparisons to evaluate the quality of generated HDR panorama by different iTMOs, as shown in **Fig. 9**. Compared with the ground-truth, only our method catches the high value of the radiance in the real world scene, validated as the bright region when the exposure value is $EV = -4.0$. The rendering results of the predicted HDR panoramas are also presented. We use four shader balls with different materials to evaluate the quality of panoramas for image-based lighting. Among all methods, ours is the closest one compared with ground truth. Different from iTMO for HDR displays, the one for HDR lighting needs to restore the scale of real-world luminance in order to produce photorealistic rendering, *e.g.,* casting shadows and inducing specular highlights.

Moreover, we demonstrate the versatility of our SR-iTMO in **Fig. 10**. With the benefits of the continuous representation z_c, the SR-iTMO can perform on arbitrary resolutions. Even the scaling factor β (Sec. 3.3.4) is only sampled within the interval of $[1, 4]$, we show the capability to scale up the LDR image eight times (x8), which can be simply achieved by increasing the resolution of the spherical coordinates (θ, ϕ).

5.6 Ablation Studies

5.6.1 Effectiveness of the global codebook. We show the significant improvement brought by our hierarchical framework, especially the usage of the global codebook $\mathcal{Z}_g$. The term "w/o global", denotes a variant of our model where the global codebook is removed. As shown in Table 1 and **Fig. 11**, the quality generation drastically decreased. Since the local codebook and the structure-aware sampler are trained on image patches, there is no way that the model can directly synthesize full panoramas without knowing the global appearance. Plus, without the global codebook, our method loses the ability to align the text with the content.

5.6.2 SPE is important for preserving structural coherence. To validate that our proposed SPE for panoramic images is crucial for preserving structural coherence, we implement two baselines. The first one, "w/o SP", directly removes representation in Eqn. 1 and Eqn. 2, and outpaints horizontally to synthesize full panoramas. The second one, "w/o SPE", only removes the Fourier feature of SPE in Eqn. 2. As shown in Table 1, our spherical representation of panoramas is important for generation in terms of overall quality and structural integrity. A visual explanation is presented in **Fig. 11**. We could notice that, without SPE, our model tends to generate repeated textures. Plus, the encoding in Eqn. 2 brings significant improvements in preserving the structural coherence, *i.e.,* our full model generates continuous ceilings and lanes, which is the key to the fidelity of panoramic scenes.

5.6.3 Text-conditioned sampler improves zero-shot transferability. To better align the free-form text with the generated scene via language-free training, we explore the unsupervised learning techniques, *i.e.,* K-nearest neighbors (KNN in Eqn. 6) and contrastive learning (Eqn. 8). We additionally train our model without these two techniques to ablate their impacts on zero-shot transferability. As shown in Table 2, KNN significantly improves the generation quality and the consistency with the text, while the contrastive regularization brings a slight benefit. The qualitative results are

Fig. 8. **Nearest neighbors of two generated results, based on the cosine similarity in the image space of CLIP [Radford et al. 2021].** (Left) Panoramas synthesized by Text2Light. (Right) Nearest neighbors in HDR360-UHD measured by the cosine similarity between two images in the CLIP image space.
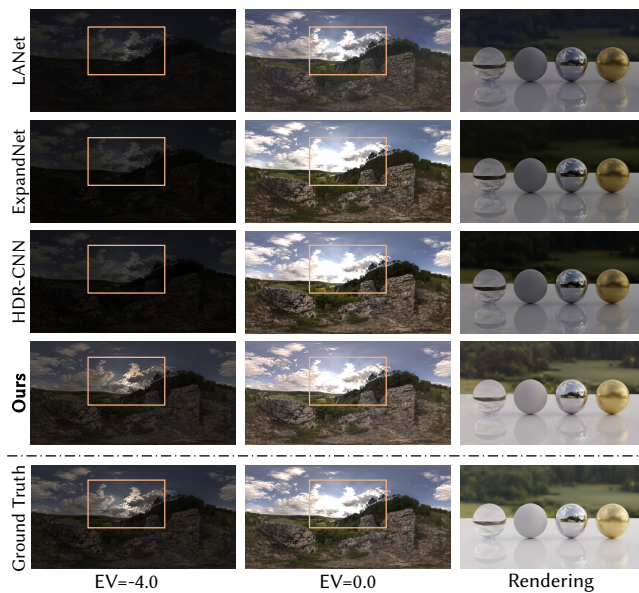

Fig. 9. **Visual Comparisons of iTMOs.** We qualitatively compare the predicted HDR panoramas at two exposure values (EV= −4.0, EV= 0.0). In addition, we use the HDR panorama to render four shader balls to evaluate their capability for image-based lighting. The material of the four balls is glass, diffuse, glossy, and a mixture of diffuse and glossy, respectively.
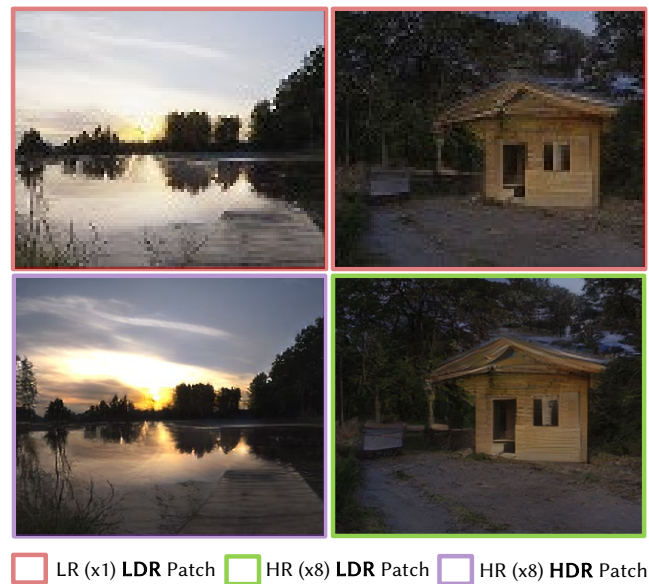

Fig. 10. **Illustration of our versatile SR-iTMO.** With the aid of our continuous representation, the SR-iTMO is capable of scaling up the LDR patch eight times (x8) to the HDR domain.

presented in **Fig. 7**. It is clear that KNN helps the model to eliminate generating artifacts when aligning with the text. Moreover, the contrastive regularization further improves Text2Light to locate and match the fine-grained details in text and images.

5.6.4 Benefits of the continuous representation z_c. Besides the arbitrary scaling factor (**Fig. 10**), the continuous representation brings other benefits to our SR-iTMO. There is a naive way to realize our SR-iTMO, which directly learns the coordinate-based MLPs f_{sr}, f_{itmo} without using continuous latent representation z_c. We show the quantitative comparison in **Fig. 12**. Given the same input LDR patch,

one can observe that the reconstructed HDR contains blurry and checkerboard artifacts when the model is trained without z_c. We attribute this phenomenon to the lack of capability to learn a continuous representation that fits different scales.

5.6.5 Why Stage II uses two separated MLPs. Intuitively, treating the task of upscaling LDR images to super-resolved HDR images as a whole is always possible, which might motivate the use of a single MLP in Stage II. However, we turn to use separated MLPs, f_{sr} and f_{itmo}, to model two different process independently. One is super-resolution while the other is inverse tone mapping. Separating two processes allows us to apply different losses to different branches to facilitate optimization. For example, the gradient from

Fig. 11. **Qualitative ablation studies of LDR panoramas generation. Zoom in for details**. The global codebook and the corresponding sampling step is important for a global coherent generation. The spherical positional encoding (Eqn. 1 and Eqn. 2) is crucial for preserving structural coherence and unrepeated textures. Without such a representation, our model tends to generate repeated textures that expand horizontally.

Fig. 12. **Ablation study of the continuous representation in Stage II**. Given an input LDR patch, the reconstructed HDR patch becomes blurry, and even contains checkerboard artifacts when our SR-iTMO is trained without z_c. We attribute this phenomenon to the lack of capability to learn a continuous representation that fits different scales.

super-resolution loss $\mathcal{L}_{sr}$ will not affect parameters of f_{itmo}. We ablate this design choice by retraining a single MLP model with a comparable number of parameters as the original model, as reported in **Table 3**. The use of two separated MLPs significantly improves the performance as the single MLP model lacks a disentangled modeling of super-resolution and inverse tone mapping. Besides, the

physical process of inverse tone mapping has spatial bias. The radiance with high dynamic range, *e.g.,* sun and lamp, often comes from the upper hemisphere of the spherical panorama. Thus, the spherical coordinates fed into f_{itmo} also serve as an inductive bias for inverse tone mapping but not super-resolution, which further motivates the use of two MLPs.

5.7 Applications

5.7.1 Photo Realistic Rendering. The generated panoramas can be directly used in modern graphic pipelines to produce photorealistic rendering on 3D assets. We present additional results in **Fig. 13** for indoor and outdoor scenes, shown in the top two rows. Thanks to our careful designs in generating high-quality textures and inverse tone mapping, driven by only simple texts, Text2Light can render high-fidelity photos and provide realistic 360-degree reflections and illuminations.

5.7.2 Immersive Virtual Reality. Besides, our generation pipeline enables an amazing application that allows users to create their 3D

Fig. 13. **Applications of Text2Light**. **Top:** The generated HDR panoramas can be directly used in lighting 3D scenes to produce photorealistic reflections and illuminations. **Middle:** Text2Light can use text to light 3D assets with high fidelity. **Bottom:** The high quality of the generated panoramas guarantees the immersive experience for users to create a virtual world using solely natural language description.
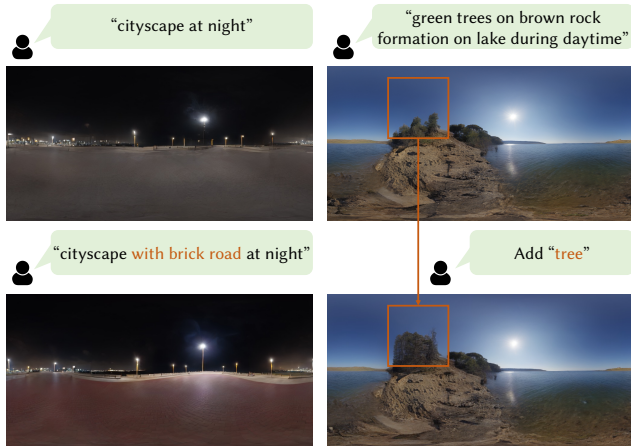
Fig. 14. **Scene editing by Text2Light**. With the hierarchical framework, we can further edit the generated scene by adding phrases or manipulating certain regions by extra descriptions.
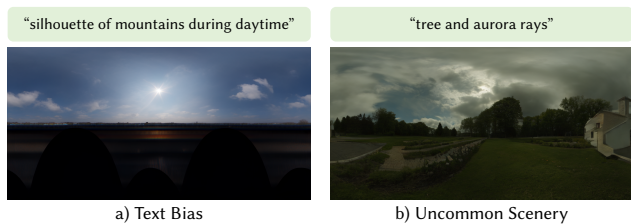

Fig. 15. **Failure cases**. a) The CLIP model is biased to the word "silhouette" which leads to pure dark regions that degrade the quality. b) Uncommon scenery, like "aurora", seems to be ignored during synthesis.

scenes using only free-form texts, where users can view in any direction for an immersive VR experience. Two examples are presented in the bottom row of **Fig. 13**, and more results will be shown in the supplementary videos.

5.7.3 Scene Editing by Texts. Thanks to our hierarchical generation framework, we can further manipulate and edit the generated scenes, as shown in **Fig. 14**. For example, by adding a phrase "with brick road" to "cityscape at night", we can change the appearance of the road in the panoramic scene. Plus, by changing the holistic condition of certain area, we add more "trees" to the generated rock.

6 LIMITATIONS

In this section, we discuss two limitations of Text2Light. **1) Text Bias.** Since we do not use any paired data for text-driven synthesis, the performance relies on CLIP model [Radford et al. 2021]. Interestingly, we observe that CLIP tends to be biased on some word-image pairs in our framework. For example, as shown in **Fig. 15(a)**, our generation pipeline is impacted by the bias of CLIP on the word "silhouette". If the word "silhouette" appears in the input description, our model will synthesize pure dark regions regardless of other

words, which leads to a low fidelity. **2) Uncommon Scenery**. As presented in **Fig. 15(b)**, Text2Light cannot synthesize scene-level contents that are uncommon in the dataset. Therefore, when the sentence "tree and aurora rays" is given, our model only focuses on the term "tree". We attribute it to the use of KNN during training the text-align sampler. One possible solution would be further exploiting the joint image-text space of CLIP, which is still an open and challenging problem in zero-shot text-driven synthesis tasks.

7 CONCLUSIONS

This work proposes a zero-shot text-driven framework, Text2Light, to generate HDR panoramas with 4K+ resolution. Given a free-form text as the description of environment textures, we synthesize the corresponding HDR panorama with two steps: 1) text-driven LDR panorama generation and 2) super-resolution inverse tone mapping. Driven by the powerful CLIP model, we leverage a coarse-to-fine framework to generate LDR panoramas by free-form texts. Furthermore, by representing the panorama as a spherical field, we propose a versatile module to upscale the resolution and dynamic range of images jointly. Besides, we contribute a high-quality HDRI dataset, with rich and diverse scenes for the scene generation task. Text2Light shows the unprecedented ability of generating high-fidelity HDR panoramas in higher than 4K resolution and good alignment with the input text.

REFERENCES

Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. 2006. Inverse tone mapping. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia - GRAPHITE '06*. ACM Press, Kuala Lumpur, Malaysia, 349.

Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. 2022. Retrieval-Augmented Diffusion Models. https://doi.org/10.48550/ARXIV.2204.11824

Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks. 2021. Unleashing Transformers: Parallel Token Prediction with Discrete Absorbing Diffusion for Fast High-Resolution Image Generation from Vector-Quantized Codes. *arXiv:2111.12701 [cs]* (Nov. 2021). arXiv: 2111.12701.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. 2022. MaskGIT: Masked Generative Image Transformer. *arXiv:2202.04200 [cs]* (Feb. 2022). arXiv: 2202.04200.

Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K. Wong, and Lei Zhang. 2021. HDR Video Reconstruction: A Coarse-to-fine Network and A Real-world Benchmark Dataset. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Montreal, QC, Canada, 2482–2491.

Zhaoxi Chen and Ziwei Liu. 2022. Relighting4D: Neural Relightable Human from Videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems* 34 (2021), 19822–19835.

Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger. 2017. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics* 36, 6 (Nov. 2017), 1–15.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12873–12883.

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090* abs/1704.00090 (2017).

Shir Gur, Sagie Benaim, and Lior Wolf. 2020. Hierarchical patch vae-gan: Generating diverse videos from a single sample. *Advances in Neural Information Processing Systems* 33 (2020), 16761–16772.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *arXiv:2006.11239 [cs, stat]* (Dec. 2020). arXiv: 2006.11239.

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–19.

Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. 2021. Multimodal Conditional Image Synthesis with Product-of-Experts GANs. *arXiv:2112.05130 [cs]* (Dec. 2021). arXiv: 2112.05130.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022. Text2Human: Text-Driven Controllable Human Image Generation. *ACM Transactions on Graphics (TOG)* 41, 4, Article 162 (2022), 11 pages. https://doi.org/10.1145/3528223.3530104

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020a. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems* 33, 12104–12114.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-Free Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 852–863.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020b. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. *arXiv:2110.02711 [cs]* (April 2022). arXiv: 2110.02711.

Soo Ye Kim, Jihyong Oh, and Munchurl Kim. 2020. Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 11287–11295.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. 2018. Deep Chain HDRI: Reconstructing a High Dynamic Range Image from a Single Low Dynamic Range Image. *IEEE Access* 6 (2018), 49913–49924. arXiv: 1801.06277.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 136–144.

Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. 2019. COCO-GAN: Generation by Parts via Conditional Coordinating. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. 2021. InfinityGAN: Towards Infinite-Pixel Image Synthesis. *arXiv:2104.03963 [cs]* (Oct. 2021). arXiv: 2104.03963.

Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. 2021. FuseDream: Training-Free Text-to-Image Generation with Improved CLIP+GAN Space Optimization. *arXiv:2112.01573 [cs]* (Dec. 2021). arXiv: 2112.01573.

Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. 2020. Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline. *arXiv:2004.01179 [cs, eess]* (April 2020). arXiv: 2004.01179.

Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. 2019. ExpandNet: A Deep Convolutional Neural Network for High Dynamic Range Expansion from Low Dynamic Range Content. *arXiv:1803.02266 [cs]* (Sept. 2019). arXiv: 1803.02266.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *arXiv:2112.10741 [cs]* (March 2022). arXiv: 2112.10741.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2018. Neural Discrete Representation Learning. *arXiv:1711.00937 [cs]* (May 2018). arXiv: 1711.00937.

Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Häne, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total Relighting: Learning to Relight Portraits for Background Replacement. *ACM Trans. Graph.* 40, 4, Article 43 (jul 2021), 21 pages. https://doi.org/10.1145/3450626.3459872

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. *arXiv:2103.17249 [cs]* (March 2021). arXiv: 2103.17249.

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. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]* (Feb. 2021). arXiv: 2103.00020.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 2019. On the spectral bias of neural networks. In *International Conference on Machine Learning*. PMLR, 5301–5310.

Prarabdh Raipurkar, Rohil Pal, and Shanmuganathan Raman. 2021. HDR-cGAN: Single LDR to HDR Image Translation using Conditional GAN. *arXiv:2110.01660 [cs, eess]* (Oct. 2021). arXiv: 2110.01660.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. (2022), 24.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems* 32 (2019).

E. Reinhard and K. Devlin. 2005. Dynamic range reduction inspired by photoreceptor physiology. *IEEE Transactions on Visualization and Computer Graphics* 11, 1 (2005), 13–24. https://doi.org/10.1109/TVCG.2005.9

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021a. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752 [cs]* (Dec. 2021). arXiv: 2112.10752.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021b. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).

Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. 2021. StyleCLIPDraw: Coupling Content and Style in Text-to-Drawing Synthesis. *arXiv:2111.03133 [cs]* (Nov. 2021). arXiv: 2111.03133.

Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. 2021. Aligning Latent and Image Spaces to Connect the Unconnectable. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Montreal, QC, Canada, 14124–14133.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. 2022b. StyleLight: HDR Panorama Generation for Lighting Estimation and Editing. In *European Conference on Computer Vision (ECCV)*.

Lin Wang and Kuk-Jin Yoon. 2021. Deep Learning for HDR Imaging: State-of-the-Art and Future Trends. *arXiv:2110.10394 [cs, eess]* (Nov. 2021). arXiv: 2110.10394.

Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. 2022a. CLIP-GEN: Language-Free Training of a Text-to-Image Generator with CLIP. *arXiv:2203.00386 [cs]* (March 2022). arXiv: 2203.00386.

Wei Wei, Li Guan, Yue Liu, Hao Kang, Haoxiang Li, Ying Wu, and Gang Hua. 2021. Beyond Visual Attractiveness: Physically Plausible Single Image HDR Reconstruction for Spherical Panoramas. *arXiv:2103.12926 [cs, eess]* (March 2021). arXiv: 2103.12926.

Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. 2014. Scale-invariant convolutional neural networks. *arXiv preprint arXiv:1411.6369* (2014).

Hanning Yu, Wentao Liu, Chengjiang Long, Bo Dong, Qin Zou, and Chunxia Xiao. 2021. Luminance Attentive Networks for HDR Image and Panorama Reconstruction. *arXiv:2109.06688 [cs, eess]* (Sept. 2021). arXiv: 2109.06688.

Jinsong Zhang and Jean-Francois Lalonde. 2017. Learning High Dynamic Range from Outdoor Panoramas. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, 4529–4538.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. 2022. LAFITE: Towards Language-Free Training for Text-to-Image Generation. *arXiv:2111.13792 [cs]* (March 2022). arXiv: 2111.13792.