A Survey on Text-Driven 360-Degree Panorama Generation

Hai Wang, Xiaoyu Xiang, Weihao Xia, and Jing-Hao Xue, *Senior Member, IEEE*

***Abstract*—The advent of text-driven 360-degree panorama generation, enabling the synthesis of 360-degree panoramic images directly from textual descriptions, marks a transformative advancement in immersive visual content creation. This innovation significantly simplifies the traditionally complex process of producing such content. Recent progress in text-to-image diffusion models has accelerated the rapid development in this emerging field. This survey presents a comprehensive review of text-driven 360-degree panorama generation, offering an in-depth analysis of state-of-the-art algorithms. We extend our analysis to two closely related domains: text-driven 360-degree 3D scene generation and text-driven 360-degree panoramic video generation. Furthermore, we critically examine current limitations and propose promising directions for future research. A curated project page with relevant resources and research papers is available at https://littlewhitesea.github.io/Text-Driven-Pano-Gen/.**

***Index Terms*—360-degree panorama generation, text-driven generation, 360-degree 3D scene generation, 360-degree panoramic video generation.**

I. Introduction

Rapid growth of immersive technologies, such as virtual reality (VR) and augmented reality (AR), has dramatically increased the demand for high-quality panoramic visual content. Among such content, 360-degree panoramas are pivotal in delivering realistic and immersive experiences by capturing a complete spherical view of an environment. Traditionally, producing these panoramas requires specialized camera equipment and considerable technical expertise. However, recent advances in text-driven 360-degree panorama generation [14], [40], [56]–[59] have introduced groundbreaking capabilities, enabling the synthesis of 360-degree panoramic images directly from textual descriptions. This innovation not only revolutionizes content creation over diverse domains [85]–[87], [89] including VR/AR applications, gaming, and virtual tours, but also serves as a foundational technology for new creative frontiers.

Unlike conventional 2D images, 360-degree panoramic images, often represented through equirectangular projection [7], encompass the entire $360° \times 180°$ field of view, as shown in Fig. 1. This distinctive format poses unique challenges for text-driven generation, requiring not only accurate image synthesis but also excellent preservation of geometric consistency and seamless visual coherence across the full $360°$ horizontal and $180°$ vertical extents.

360-degree panoramic image | conventional 2D image

Fig. 1. Visual comparison between a 360-degree panoramic image and a conventional 2D image.

The availability of large-scale paired image-text datasets has facilitated the development of text-to-image latent diffusion models (LDMs) [63], which excel at synthesizing high-quality, visually compelling images aligned with given text descriptions [64]–[67], [72]. Leveraging the powerful generative capabilities of pre-trained LDMs, researchers have developed methods specifically tailored to address the unique challenges of text-driven 360-degree panoramic image generation [42], [57]–[60], [73]. Although broader surveys on panoramic vision and 3D scene-generation [88], [104] briefly discuss some text-driven 360-degree panorama generation methods, they treat them only as peripheral topics.

To the best of our knowledge, a focused and systematic analysis devoted specifically to text-driven 360-degree panoramic image generation has not yet been presented. To address this gap, this paper presents a holistic survey and analysis of text-driven 360-degree panorama generation, its direct applications, and related emerging fields.

This survey is structured as follows: First, we establish a foundational understanding of this field by introducing the principal representations of 360-degree panoramas, presenting prominent datasets commonly used in this area, and outlining key evaluation metrics employed to assess the quality and fidelity of generated panoramic content. Next, we review state-of-the-art (SOTA) methods for text-driven 360-degree panorama generation, categorizing them into two primary paradigms: (a) *Text-Only Generation* and (b) *Text-Driven Narrow Field-of-View (NFoV) Outpainting*. Fig. 2 and Fig. 3 provide a systematic taxonomy and a chronological overview of these SOTA methods, respectively. Following this, we explore two key emerging directions that are closely related to this field: (a) text-driven 360-degree 3D scene generation, which uses 360-degree panoramic images as an intermediate step; and (b) text-driven 360-degree panoramic video generation, a parallel and more complex task. Finally, we discuss the

H. Wang, W. Xia and J.-H. Xue are with the Department of Statistical Science, University College London, London WC1E 6BT, U.K. (e-mail: hai.wang.22@ucl.ac.uk).

X. Xiang is with the Core AI team at Meta Reality Labs, Menlo Park, CA 94025, USA.

Corresponding author: Hai Wang

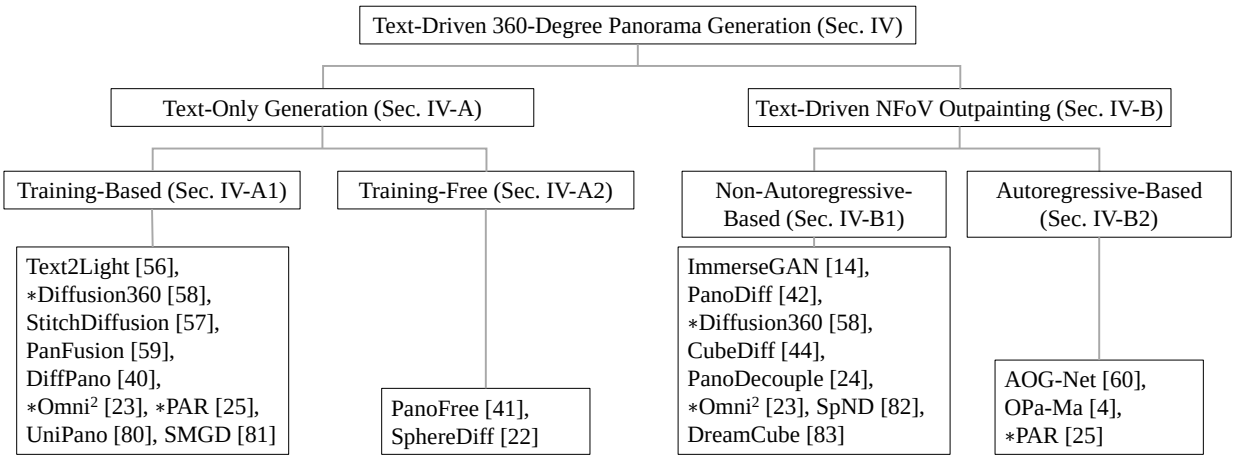

Fig. 2. A systematic taxonomy proposed in this survey of text-driven 360-degree panorama generation methods. Methods marked with ∗ support multiple input modalities and therefore appear in more than one branch.

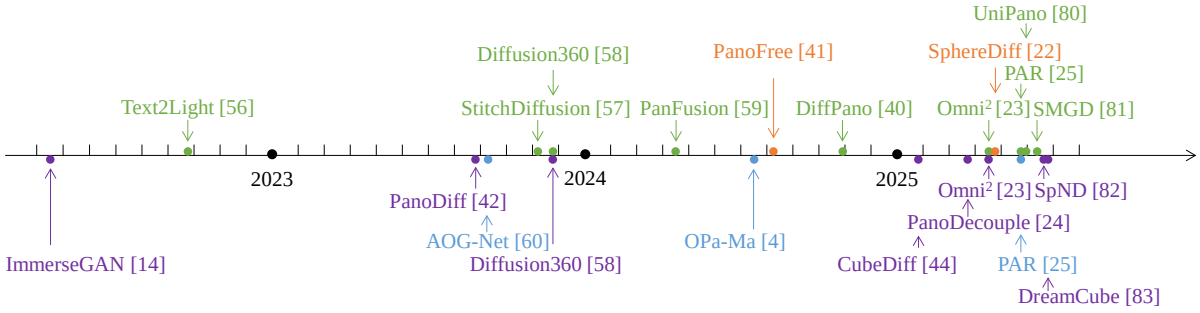

Fig. 3. Chronological overview of text-driven 360-degree panorama generation approaches. Methods in lime, orange, violet, and cyan are from Sec. IV-A1, Sec. IV-A2, Sec. IV-B1, and Sec. IV-B2, respectively.

prevailing challenges in this developing field and propose potential directions for future research.

In short, this paper offers the first dedicated and comprehensive survey on text-driven 360-degree panoramic image generation, systematically reviewing its state-of-the-art techniques, key datasets and evaluation metrics. Furthermore, we explore its two closely related emerging directions: text-driven 360-degree 3D scene generation and text-driven 360-degree panoramic video synthesis. We also identify critical challenges and outline future research directions, aiming to offer a valuable resource to researchers and practitioners in this area.

II. Related Work

A. Text-to-Image Diffusion Models

Text-to-image (T2I) diffusion models [11]–[13], [21], [63] have achieved remarkable progress in generating high-fidelity and photorealistic images from textual descriptions. These models have garnered widespread attention because of their intuitive text-based conditioning as a user-friendly interface for diverse image generation tasks.

T2I diffusion models can be broadly categorized into pixel-space and latent-space models. Pixel-space models, such as GLIDE [11] and Imagen [12], operate directly in the pixel space, producing visually impressive results at the expense of substantial computational resources, limiting their scalability. In contrast, latent diffusion models (LDMs) [63] address these limitations by leveraging pre-trained autoencoders like VQGAN [43] to map images into a compact latent space, where the diffusion process is conducted. This reduces computational overhead while maintaining high-quality outputs, making LDMs a preferred framework for text-driven 360-degree panorama generation, as surveyed in this work.

B. 3D Scene Representation

Efficient and accurate 3D scene representation is a critical challenge in computer graphics and vision. Traditional explicit representations, including point clouds, meshes, and voxel grids, often suffer from high memory requirements and struggle with complex topologies and unbounded scenes.

Neural implicit functions [9], [10], [29], which represent 3D scenes as continuous functions encoded within neural network parameters, offer a compact and flexible paradigm for scene representation. Notably, Neural Radiance Fields (NeRFs) [8] stand out for their ability to achieve high-quality novel view synthesis. However, NeRF's reliance on dense volumetric sampling along camera rays results in slow training, hindering its practicability.

Recently, 3D Gaussian Splatting (3DGS) [26] has emerged as an efficient alternative to 3D scene representation. By combining an explicit representation of 3D Gaussians with a highly

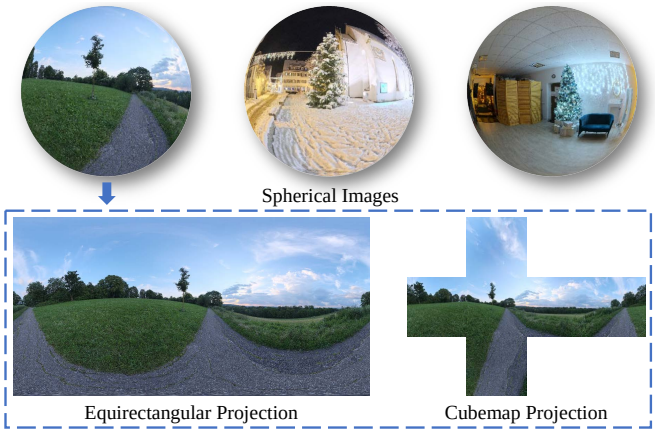

Fig. 4. Visual comparison between the equirectangular and cubemap projections of spherical images (360-degree panoramic images).

efficient differentiable rasterization pipeline, 3DGS facilitates rapid scene reconstruction and rendering. This advancement has opened up new possibilities, including recent explorations in text-to-3D 360-degree scene synthesis [32], [33], which leverage text-driven 360-degree panorama generation techniques.

III. Preliminaries

A. Representations of 360-Degree Panoramas

The representation of 360-degree panoramic content poses a fundamental challenge: How to accurately map spherical visual information onto a two-dimensional plane? To address this, a variety of projection methodologies [7], [88] have been developed, each with distinct advantages and trade-offs. Below, we first outline two widely used formats for 360-degree panorama representation: Equirectangular Projection (ERP) and Cubemap Projection (CMP), as illustrated in Fig. 4, and then introduce Multi-Perspective Projection (MPP), a category we define for the classification purposes of this survey.

1) Equirectangular Projection (ERP): As the most prevalent representation format for 360-degree panoramas, ERP establishes a direct mapping between spherical and planar coordinates: longitude corresponds to the horizontal axis, spanning the full 360° range, while latitude maps to the vertical axis, covering 180° from -90° (south pole) to +90° (north pole). ERP's simplicity and compatibility with web viewers and VR headsets make it the preferred choice for numerous applications. Additionally, its representation as a single, continuous image allows the direct application of image manipulation techniques, such as text-driven 360-degree panorama-to-panorama translation [28]. Despite these advantages, ERP introduces pronounced geometric distortions, particularly at the polar regions, where the visual content appears stretched. Furthermore, the texel density of a spherical image in ERP is non-uniformly distributed: it is comparatively lower in the equatorial regions and markedly higher towards the poles. This inhomogeneity can be particularly problematic in scenarios where critical visual information is predominantly located away from the poles, leading to inefficient utilization of image resolution for regions of interest. In this survey, unless explicitly stated otherwise, 360-degree panoramas are represented using ERP.

2) Cubemap Projection (CMP): CMP offers an alternative representation that mitigates the distortions inherent in the ERP format, particularly at the poles. In CMP, the spherical image is projected onto the six faces of a cube, with each face representing a 90°×90° field of view. This division significantly reduces geometric distortions, making CMP more compatible with diffusion priors from text-to-image diffusion models trained on standard perspective images [44]. However, CMP introduces several challenges: (1) it increases the complexity of image manipulation compared to the single-image format of ERP; (2) it may necessitate additional conversion for compatibility with platforms or viewers that primarily support ERP. Despite these practical challenges, CMP is well-suited for applications that demand reduced distortion and higher fidelity. The width and height of an ERP image are four and two times the side length of the corresponding CMP, respectively, reflecting the geometric relationship between the two formats.

3) Multi-Perspective Projection (MPP): MPP is defined as the use of multiple individual perspective images to collectively represent a 360-degree panoramic view, where these images may or may not overlap. This category is characterized by configurations that deviate from the standard six-face, 90°×90° field of view CMP.

The advantages and limitations of these representation formats are further analyzed in Sec. IV-C1.

TABLE I
Summary of Popular Datasets Used for Text-Driven 360-Degree Panorama Generation: Includes categories, publication year, sample size, resolution (Res.), and license. Categories are indoor (I), outdoor (O), or hybrid (I, O). Datasets marked with ⋆ are sourced from public websites.

Dataset (Category)	Year	#Samples	Res.	License
SUN360 (I, O)	2012	67,583	9K	Custom
Matterport3D (I)	2017	10,800	2K	Custom
Laval Indoor (I)	2017	2,233	7K	Custom
Laval Outdoor (O)	2019	205	7K	Custom
Structured3D (I)	2020	196,515	1K	Custom
Pano360 (I, O)	2021	35,000	8K	Custom
⋆Polyhaven (I, O)	2025	786	8K	CC0
⋆Humus (I, O)	2025	139	8K	CC BY 3.0

B. Datasets

360-degree panoramic image generation from text prompts presents unique challenges due to the complete 360°×180° field of view that these images encompass. Text-to-image diffusion models [12], [21], [35], [63], predominantly trained on perspective images with a narrower field of view, often struggle to synthesize high-quality 360-degree panoramas. To address this, several specialized datasets have been developed to facilitate research in this domain. Tab. I summarizes these datasets, with further details provided below.

SUN360 [48] is a comprehensive database comprising 67,583 high-resolution (9104×4552) panoramic images sourced from the Internet. Each image covers a 360°×180° field of view

in ERP format and is manually categorized into 80 distinct classes. Originally created for scene viewpoint recognition, SUN360 now serves as a valuable resource for a wide range of computer vision, computer graphics, and related research areas.

Matterport3D [49] offers 10,800 indoor 360-degree panoramic images with corresponding depth maps, all at a resolution of 2048×1024 pixels. These panoramas are derived from 194,400 RGB-D images of 90 buildings, making it a rich dataset for studying indoor environments.

Laval Indoor [45] consists of 2,233 high-dynamic-range, high-resolution (7768×3884) 360-degree panoramic images, specifically curated for the study of extensive indoor scenes, such as factories, apartments, and houses.

Laval Outdoor [46] complements its indoor counterpart, offering 205 high-dynamic-range, high-resolution (7768×3884) 360-degree panoramic images that capture diverse outdoor environments, including urban and natural scenes.

Structured3D [47] contains 196,515 360-degree panoramas with varying configurations and lighting conditions, representing 21,835 distinct rooms. Rendered at a resolution of 1024×512 from 3D scenes of original house design files, Structured3D is ideal for research on structured 3D modeling and understanding.

Pano360 [19] contains 35,000 360-degree panoramic images with a resolution of 8192×4096. Of these, 34,000 are sourced from Flickr, with the remainder rendered from photorealistic 3D scenes. Pano360 was originally proposed for training camera calibration networks.

Polyhaven [5] contributes 786 real-world high-resolution (8192×4096) 360-degree panoramas encompassing a variety of indoor and outdoor scenes.

Humus [6] includes 139 real-world 360-degree panoramas represented using cubemap projection, with each face having a resolution of 2048×2048 pixels. This dataset includes indoor and outdoor environments.

C. Evaluation Metrics

A rigorous evaluation of text-driven 360-degree panorama generation methods typically requires to combine (a) *universal* and (b) *panorama-specific* metrics. The universal metrics, comprising Fréchet Inception Distance (FID), Kernel Inception Distance (KID), Inception Score (IS), and CLIP Score (CS), are widely applicable to both perspective and panoramic images.

FID [53] measures the distance between feature distributions of generated and real images using a pre-trained Inception-v3 network [36]. Lower FID scores indicate better perceptual quality and closer alignment with the real image distribution.

KID [20] measures the difference between real and generated image distributions by computing the maximum mean discrepancy of their features extracted from Inception-v3 [36]. Similar to FID, lower KID values indicate better image quality.

IS [54] measures both the quality and diversity of generated images by leveraging Inception-v3 [36]. It calculates the KL divergence between the conditional class distribution of generated images and the marginal distribution over all generated samples. Higher IS suggests better visual quality and diversity.

CS [52] evaluates consistency between text prompts and generated images using the CLIP model [52]. It calculates the cosine similarity between the text embedding of the prompt and the visual embedding of the generated image. A higher CS reflects stronger text-image alignment and semantic coherence.

However, these universal metrics have significant limitations when applied to 360-degree panoramic content. The core issue lies in their underlying encoders (e.g., Inception-v3, CLIP's ViT), which were trained primarily on large datasets of standard perspective images. These encoders are not trained to account for the geometric distortions inherent in equirectangular projections. Consequently, FID, KID, and IS may penalize geometrically correct panoramic features as artifacts, leading to an inaccurate assessment of perceptual quality. Similarly, the CS may fail to accurately measure text-image alignment, as its encoder might misinterpret distorted objects or spatial relationships within the spherical scene. This gap highlights the inadequacy of universal metrics for capturing properties unique to 360-degree panoramas, such as seamlessness and global geometric fidelity. To address these shortcomings, several panorama-specific metrics have been proposed, including Fréchet Auto-Encoder Distance (FAED), Omnidirectional FID (OmniFID), Discontinuity Score (DS), and Distortion-perception FID (Distort-FID).

FAED [50] computes Fréchet distances between features extracted from generated and real panoramas. Unlike FID, it employs an autoencoder [34] specifically trained on 360-degree panoramic images. Lower FAED scores reflect better perceptual and geometric quality tailored to the unique panoramic properties.

OmniFID [51] adapts FID to specifically evaluate 360-degree panoramas. It converts equirectangular panoramas into cubemap representations and calculates FID across three disjoint subsets of cubemap faces, averaging the results. Lower OmniFID scores indicate higher geometric fidelity in 360-degree panorama generation.

DS [51] measures the seam alignment across the borders of generated panoramas by applying a kernel-based edge detection algorithm. A lower DS corresponds to fewer visible seam artifacts, indicating better perceived consistencies across the seam.

Distort-FID [24] measures the distance between feature distributions of generated and real images based on a distortion-perception CLIP [24] that is fine-tuned on 360-degree panoramic images. Lower Distort-FID scores reflect better distortion accuracy of the generated 360-degree panoramic images.

Despite these advances, a critical gap remains. While metrics like OmniFID improve the measurement of geometric fidelity by mitigating ERP distortions and DS checks for visual seamlessness, they do not holistically evaluate the global semantic and structural coherence of a full 360-degree panoramic scene. Developing new metrics that capture this

Fig. 5. Paradigms for Text-Driven 360-Degree Panorama Generation. (a) Text-Only Generation synthesizes 360-degree panoramas from textual descriptions only. (b) Text-Driven NFoV Outpainting uses prompts and initial NFoV images as input to generate 360-degree panoramic images.

TABLE II
SUMMARY OF TEXT-ONLY GENERATION METHODS. 'LDM-B' INDICATES WHETHER THE METHOD IS BASED ON LATENT DIFFUSION MODELS. 'TF' SPECIFIES IF IT IS TRAINING-FREE. 'CODE' DENOTES WHETHER BOTH THE SOURCE CODE AND THE PRE-TRAINED MODEL CHECKPOINT ARE PUBLICLY ACCESSIBLE. 'N/A' MEANS NOT APPLICABLE.

Method	Publication	LDM-B	TF	Code	Training Datasets	Representation
Text2Light [56]	TOG 2022	×	×	✓	Polyhaven [5], Laval Indoor [45], Laval Outdoor [46] iHDRI [70], HDRI Skies [68], HDRMaps [69]	ERP
Diffusion360 [58]	arxiv 2023	✓	×	✓	SUN360 [48]	ERP
StitchDiffusion [57]	WACV 2024	✓	×	✓	Polyhaven [5]	ERP
PanFusion [59]	CVPR 2024	✓	×	✓	Matterport3D [49]	ERP
PanoFree [41]	ECCV 2024	✓	✓	×	N/A	MPP
DiffPano [40]	NeurIPS 2024	✓	×	×	Habitat Matterport 3D [15]	ERP
Omni^2 [23]	ACM MM 2025	×	×	×	SUN360 [48], Structured3D [47]	MPP
SphereDiff [22]	arxiv 2025	✓	✓	×	N/A	Spherical
PAR [25]	arxiv 2025	×	×	✓	Matterport3D [49]	ERP
UniPano [80]	ICCV 2025	✓	×	×	Matterport3D [49]	ERP
SMGD [81]	CVPR 2025	✓	×	✓	Matterport3D [49]	Spherical

global consistency is therefore an important direction for future research.

We provide a comprehensive comparison of state-of-the-art methods, introduced in the following section, using the outlined metrics in Sec. IV-D2.

IV. STATE-OF-THE-ART METHODS FOR IMAGES

Existing text-driven 360-degree panorama generation methods can be broadly categorized into two paradigms according to input modalities: (a) *Text-Only Generation* aims to synthesize 360-degree panoramas from textual prompts only, while (b) *Text-Driven Narrow Field-of-View (NFoV) Outpainting* leverages both textual descriptions and initial NFoV images to guide the generation process, offering enhanced user control. Fig. 5 provides an intuitive illustration for both paradigms. We detail the literature for both as follows.

A. Text-Only Generation

This paradigm focuses on synthesizing 360-degree panoramas from textual descriptions only. Tab. II provides a comparative overview of representative text-only methods. These methods can be broadly divided into training-based and training-free approaches.

1) Training-Based: Text2Light [56], an early notable effort, explores a hierarchical text-driven framework, using VQ-GAN [43] and CLIP [52] to address this challenge based on training data aggregated from multiple sources, such as Polyhaven [5], Laval Indoor [45] and Laval Outdoor [46]. Recently, the advent of latent diffusion models (LDMs) [63] for text-to-image synthesis marks a significant advancement, enabling more sophisticated 360-degree panorama generation techniques. LDMs are typically trained on vast datasets consisting of standard perspective images with a limited field of view and corresponding textual descriptions. Despite demonstrating robust capabilities in generating perspective images from text prompts, these models face significant difficulties when creating 360-degree panoramas with a complete 360°×180° field of view, which differ substantially from traditional perspective images.

To adapt pre-trained LDMs for 360-degree panorama synthesis, a common strategy is to fine-tune these models with specialized 360-degree panorama datasets. Diffusion360 [58] exemplifies this approach by leveraging the DreamBooth technique [38] to fine-tune a pre-trained LDM [63] on SUN360 [48]. To ensure geometric consistency of boundaries, Diffusion360 uses a circular blending strategy during both the denoising process and the VAE [74] decoding stage, effectively reducing seam artifacts. In addition, it introduces a super-resolution module to enable the generation of high-resolution (6114×3072) 360-degree panoramas. While this full fine-tuning approach adopted by Diffusion360 effectively embeds panoramic geometry into the model, its primary trade-off is the high computational cost and risk of quality drop by altering the model's original generative priors.

In contrast, LoRA [39] has recently gained attention as a parameter-efficient fine-tuning method. LoRA works by injecting trainable low-rank matrices into the pre-trained model's weights, allowing for rapid adaptation to new tasks with minimal additional parameters. For example, StitchDiffusion [57] employs LoRA to fine-tune a pre-trained LDM on a dataset of 120 paired 360-degree images (sourced from Polyhaven [5]) and corresponding textual descriptions gen-

erated using BLIP [17]. Its key contribution is formulating panorama generation as a latent-space stitching problem, using a MultiDiffusion-based [55] method to enforce boundary continuity. However, the geometry fidelity of 360-degree panoramas generated by StitchDiffusion is relatively low due to the small fine-tuning dataset.

Other works [40], [59], [80], [84] have similarly adopted the LoRA fine-tuning technique. PanFusion [59] contributes a novel dual-branch architecture, trained on the Matterport3D [49] dataset, with separate LoRA layers to integrate both global panoramic and local perspective views. It introduces an equirectangular-perspective projection attention module to facilitate information exchange between the two branches, aiming to alleviate visual inconsistencies in the generated 360-degree panoramas. However, PanFusion's output often exhibits blurriness at the top and bottom regions, due to its training dataset. To avoid this issue, DiffPano [40] uses the Habitat Matterport 3D dataset [15] to produce multi-view consistent 360-degree panoramas with clearer top and bottom details. For generating more precise textual descriptions, DiffPano adopts BLIP2 [18] and Llama2 [16] sequentially, resulting in a panoramic video-text dataset. Based on this dataset, it fine-tunes a pre-trained LDM [63] using LoRA for single-view text-driven 360-degree panorama generation. Furthermore, to enable multi-view 360-degree panorama generation based on text prompts and camera viewpoints, DiffPano introduces a spherical epipolar-aware attention module, a key innovation for enforcing multi-view geometric consistency. Based on the observation that value and output weight matrices are crucial during the LoRA-based fine-tuning of cross-attention blocks within pre-trained LDMs for 360-degree panoramic image generation, UniPano [80] proposes a uni-branch framework for panorama generation. This framework exclusively fine-tunes these specific matrices using LoRA, while keeping the original query and key weight matrices in the cross-attention blocks frozen.

Diverging from fine-tuning the commonly used pre-trained LDMs [63] (as in PanFusion [59] and DiffPano [40]), several studies have explored alternative pre-trained models for synthesizing 360-degree panoramic images from textual descriptions. Inspired by OmniGen [76], Omni^2 [23] adopts a diffusion model consisting of a VAE [74] and a pre-trained Transformer [75]. To adapt the pre-trained Transformer for synthesizing 360-degree panoramic images, the LoRA fine-tuning technique is employed. Instead of processing the entire 360-degree panoramic image at once, Omni^2 generates six overlapping viewports. In the inference phase, these synthesized perspective images are integrated to reconstruct 360-degree panoramic images. Addressing the challenge that spatial distortions in ERP 360-degree panoramic images violate the identically and independently distributed (i.i.d.) Gaussian noise assumption inherent in many diffusion models, PAR [25], inspired by masked autoregressive modeling (MAR) [77], proposes an autoregressive modeling approach for text-based 360-degree panoramic image generation. This method is not constrained by the i.i.d. assumption. Specifically, PAR fine-tunes a pre-trained autoregressive model [78] on the Matterport3D dataset [49] and develops a dual-space circular padding technique to mitigate boundary discontinuities.

Most aforementioned approaches rely on ERP representations, which struggle to adequately deal with the inherent spherical distortions. To mitigate these distortions and maintain global geometric coherence, SMGD [81] proposes the use of spherical manifold convolution within a spherical manifold U-Net combined with VQGAN [43], enabling more accurate synthesis of 360-degree panoramic images, but the primary trade-off is reduced transferability of pre-trained diffusion priors, since the specialized spherical convolutions are not directly compatible with standard text-to-image diffusion architectures [21], [63].

2) Training-Free: In contrast to training-based approaches, training-free methods avoid any model fine-tuning and instead repurpose powerful pre-trained text-to-image diffusion backbones. PanoFree [41] pioneered a tuning-free multi-view image generation framework based on a pre-trained LDM [63]. Guided by textual descriptions, PanoFree leverages iterative warping and inpainting steps to produce multi-view perspective images, which are subsequently stitched into 360-degree panoramas, thus avoiding the need for specialized 360-degree panorama datasets. While this avoids the need for specialized training data, its multi-step process can be slow and risks accumulating errors that harm global consistency. Unlike PanoFree, which operates on perspective latent representations, the recent SphereDiff [22] contributes a more theoretically grounded approach by extending the MultiDiffusion [55] framework to a spherical latent space. To mitigate minor distortions arising from the spherical-to-perspective projection during its process, SphereDiff further incorporates a distortion-aware weighted averaging method. Although these approaches inherit the rich prior knowledge of large text-to-image models and require no additional training or 360-degree panoramic data, their patch-based synthesis mechanisms can lead to global inconsistencies and comparatively longer inference times than training-based models. Future work may explore stronger global guidance and more efficient inference designs to overcome these limitations.

B. Text-Driven NFoV Outpainting

This paradigm enhances user control by conditioning the 360-degree panorama generation process on both textual prompts and initial NFoV images. The NFoV image, representing a limited portion of the scene, serves as a visual starting point, which the generative model subsequently expands into a complete 360-degree panoramic image guided by the textual description. Tab. III offers a summary of representative text-driven NFoV outpainting approaches. These approaches can be broadly classified into non-autoregressive-based (NAR-based) and autoregressive-based (AR-based) methods according to their underlying frameworks.

1) NAR-Based: An early attempt in this paradigm, ImmerseGAN [14], uses a GAN-based inpainting architecture [79] for this task. To achieve text-guided outpainting, ImmerseGAN adopts a pre-trained discriminative network to produce a latent vector representing the given textual description. This latent vector subsequently guides the generator to produce a 360-degree panorama semantically consistent with the text

TABLE III
SUMMARY OF TEXT-DRIVEN NFOV OUTPAINTING METHODS. FOR EXPLANATIONS OF THE 'LDM-B', 'TF' AND 'CODE' COLUMNS, SEE TAB. II.

Method	Publication	LDM-B	TF	Code	Training Datasets	Representation
ImmerseGAN [14]	3DV 2022	×	×	×	360Cities [71]	ERP
PanoDiff [42]	MM 2023	✓	×	✓	SUN360 [48]	ERP
Diffusion360 [58]	arxiv 2023	✓	×	✓	SUN360 [48]	ERP
AOG-Net [60]	AAAI 2024	✓	×	×	Laval Indoor [45], Laval Outdoor [46]	ERP
OPa-Ma [4]	arxiv 2024	✓	×	×	Laval Indoor [45], Laval Outdoor [46]	ERP
CubeDiff [44]	ICLR 2025	✓	×	×	Polyhaven [5], Humus [6], Structured3D [47] Pano360 [19]	CMP
PanoDecouple [24]	CVPR 2025	✓	×	×	SUN360 [48]	ERP
Omni^2 [23]	ACM MM 2025	×	×	×	SUN360 [48], Structured3D [47]	MPP
PAR [25]	arxiv 2025	×	×	×	Matterport3D [49]	ERP
SpND [82]	ICML 2025	✓	×	✓	Matterport3D [49], Structured3D [47]	ERP
DreamCube [83]	ICCV 2025	✓	×	✓	Structured3D [47]	CMP

prompt. More recent approaches have primarily focused on leveraging the power of pre-trained latent diffusion models (LDMs) for their strong image generation priors acquired from training on large-scale datasets.

PanoDiff [42], the first LDM-based method for text-driven NFoV outpainting, is trained on the SUN360 [48] dataset. It initially converts the input NFoV images into partial panoramas with visibility masks, and then employs a ControlNet-based LDM [37] for text-guided panorama completion. To ensure geometric continuity at the borders of the generated panorama, PanoDiff further implements a circular padding scheme during inference. Similarly, Diffusion360 [58] adopts a ControlNet-based LDM [37] to generate 360-degree panoramas from perspective images and textual descriptions. However, instead of circular padding, Diffusion360 leverages a circular blending strategy during the denoising and VAE decoding stages for improved boundary continuity of the generated 360-degree panorama. Recognizing that a single network (as employed in PanoDiff and Diffusion360) often struggles to simultaneously learn the inherent 360-degree panoramic distortion and perform content completion, PanoDecouple [24] introduces a decoupled diffusion model as its core contribution. This framework separates the NFoV outpainting process into distortion guidance and content completion. While this modular design is effective, it increases model complexity by requiring a separately trained Distort-CLIP model. Contrastingly, SpND [82] incorporates structural prior information from 360-degree panoramic images processed through a spherical network into its diffusion model to guide the 360-degree panoramic image outpainting process.

Depart from the aforementioned methods [24], [42], [58], which predominantly process and generate 360-degree panoramas using an equirectangular representation throughout their networks, several recent studies have explored leveraging alternative representations for 360-degree panoramic synthesis. CubeDiff [44], inspired by multi-view diffusion models [61], [62], generates 360-degree panoramas in cubemap format. This cubemap representation enables CubeDiff to more effectively leverage the diffusion priors learned by the LDM from extensive perspective images during the generation process. CubeDiff fine-tunes a pre-trained LDM on a mixed dataset of Structured3D [47], Pano360 [19], Polyhaven [5], and Humus [6], using a single conditional view (NFoV image) and textual embeddings as input. Its central innovation is the inflation of 2D attention layers in the LDM into 3D attention layers, enabling the model to explicitly learn inter-face dependencies. This effectively trades the ERP distortion problem for a complex inter-view consistency challenge. Similarly, Omni^2 [23] produces six overlapping perspective images, each with a $110°\times110°$ field of view, using a Transformer [75] architecture fine-tuned with LoRA on the SUN360 [48] and Structured3D [47] datasets. These overlapping images are subsequently stitched together to synthesize the final 360-degree panoramic image. To address the computational redundancy and resolution constrains introduced by these overlapping images, DreamCube [83] designs a multi-plane synchronization approach, enabling seamless and consistent RGB-D cubemap generation without overlaps as its main contribution.

2) AR-Based: AOG-Net [60] introduces an autoregressive framework, building upon a pre-trained LDM [63], to progressively outpaint NFoV images into complete panoramas under textual guidance. This stepwise approach enhances the generation of fine-grained visual content and improves alignment with the input textual descriptions. Its key innovation is a global-local conditioning mechanism that integrates multiple guidance signals at each step to ensure consistency across the generated panorama. AOG-Net is trained on Laval Indoor [45] and Laval Outdoor [46] for indoor and outdoor scenarios, respectively. Following the training dataset settings in AOG-Net, OPa-Ma [4] uses a pre-trained LDM with Mamba [3], a state-space model known for its efficiency in handling long sequences, to iteratively outpaint local regions in each step. It introduces two modules: the Visual-textual Consistency Refiner, which enhances input utilization during generation, and the Global-local Mamba Adapter, which ensures global coherence across the generated panorama. In contrast to methods leveraging pre-trained LDMs (such as AOG-Net and OPa-Ma), the recent PAR [25] adopts a text-conditioned masked autoregressive model [78] for 360-degree panoramic image outpainting. Guided by text descriptions, PAR incorporates its designed dual-space circular padding and a translation consistency loss to improve output quality. For all AR-based

methods, this step-by-step generation enforces strong logical coherence but at the cost of slower inference and a susceptibility to error propagation.

C. Analysis of Fundamental Design Choices

The methods surveyed above make fundamentally different design choices that entail significant trade-offs in performance, efficiency, and quality. A critical examination of these choices is essential for understanding the current research landscape and identifying future directions. Below, we systematically analyze the trade-offs and implications of three core architectural decisions: (1) the choice of panoramic representation, (2) the generation framework, and (3) the strategy for adapting pre-trained models.

1) Panoramic Representation – Simplicity vs. Prior Compatibility: As detailed in Sec. III-A, methods typically adopt ERP, CMP, or MPP. This selection reflects a trade-off between representational simplicity and compatibility with the generative priors of large pre-trained text-to-image models.

- **Implications of Choosing ERP:** ERP offers a single continuous image, making it directly compatible with standard 2D architectures such as U-Nets and Transformers. However, this choice presents two significant challenges. (1) Standard 2D convolutions are not inherently periodic, so to enforce seamlessness at the 360-degree wrap-around boundary, methods must incorporate special treatments such as circular padding (PanoDiff [42]), circular blending (Diffusion360 [58]), or dedicated stitching steps (StitchDiffusion [57]). (2) ERP introduces non-uniform texel density (geometric distortions), stretching content near the poles and complicating learning of accurate 360-degree panoramas. Some methods address this with spherical manifold convolution (SMGD [81]) or a distortion guidance branch (PanoDecouple [24]), albeit at additional computational cost.
- **Implications of Choosing CMP/MPP:** The decision to use a collection of perspective views (CMP or MPP) aims to leverage the priors of pre-trained models more effectively, as these models were trained predominantly on low-distortion perspective images. Representing the 360-degree panoramic scene in this format, as done in CubeDiff [44] and Omni^2 [23], maximizes compatibility and often yields higher-fidelity details. The trade-off is a shift in complexity: the problem of distortion is replaced by the problem of inter-view consistency. This necessitates specialized mechanisms to ensure seamlessness between views, such as inflating 2D attention layers to model 3D relations (CubeDiff [44]) or employing complex multi-step stitching and inpainting pipelines (PanoFree [41]), which can increase computational overhead and introduce new types of boundary artifacts.

2) Generation Framework – NAR-based vs. AR-based: The underlying mechanism for generating 360-degree panoramic content represents a key architectural choice, centered on a trade-off between generation speed and global coherence.

- **Non-Autoregressive (NAR-based):** Most recent methods are built upon non-autoregressive LDMs [63], in which the entire latent representation of the 360-degree panorama is denoised in parallel over a series of steps. The primary advantage of this approach is computational efficiency, enabling relatively fast generation of full panoramas, as demonstrated by Diffusion360 [58]. The main drawback is the difficulty of ensuring global coherence across the vast 360-degree scene. Because the whole image is generated in a single holistic process, the model may struggle to maintain consistent long-range context, sometimes producing mismatched regions or repetitive textures.
- **Autoregressive (AR-based):** In contrast, an autoregressive framework generates the panorama sequentially, explicitly conditioning each new region on the content generated in previous steps. This step-by-step process naturally enforces both local and global consistency, and methods like AOG-Net [60] and PAR [25] often exhibit strong logical coherence and fine-grained detail. The trade-off for this improved consistency is a substantial increase in inference time, as the sequential process is inherently slower than parallel generation. Moreover, these methods can suffer from error propagation, where low-quality generation in an early step can negatively impact the quality of all subsequent parts.

3) Adaptation of Pre-trained Models – Fine-tuning vs. Training-Free: Since no large-scale generative model is natively trained on 360-degree panoramic image-text pairs, researchers have to adapt models originally trained on standard perspective image-text datasets. Three distinct strategies have emerged, spanning a spectrum of trade-offs among computational cost, data requirements, and model specialization.

- **Full/Substantial Fine-Tuning:** One strategy is to fine-tune a large portion of the model parameters on a specialized 360-degree panoramic image-text dataset. For instance, Diffusion360 [58] leverages the DreamBooth [38] technique to adapt a pre-trained LDM [63]. This approach enables the model to learn the unique geometric properties and data distributions of 360-degree panoramas, often leading to strong performance. However, the trade-offs are significant: it is usually computationally expensive and time-consuming, requires a substantial panoramic dataset, and may induce catastrophic forgetting, in which the model loses some of the rich generative knowledge from its original training.
- **Parameter-Efficient Fine-Tuning (PEFT):** A more economical alternative trains only a small set of additional parameters while keeping the backbone of the pre-trained model frozen. LoRA [39], employed by methods such as StitchDiffusion [57] and DiffPano [59], typifies this class. The chief advantage is a drastic reduction in computational and memory requirements, making adaptation more accessible and largely preserving the original model's powerful priors. The main limitation is that the frozen backbone can restrict the model's ability to learn fundamentally new concepts, such as complex spherical geometry, making it potentially less expressive than a fully fine-tuned model.

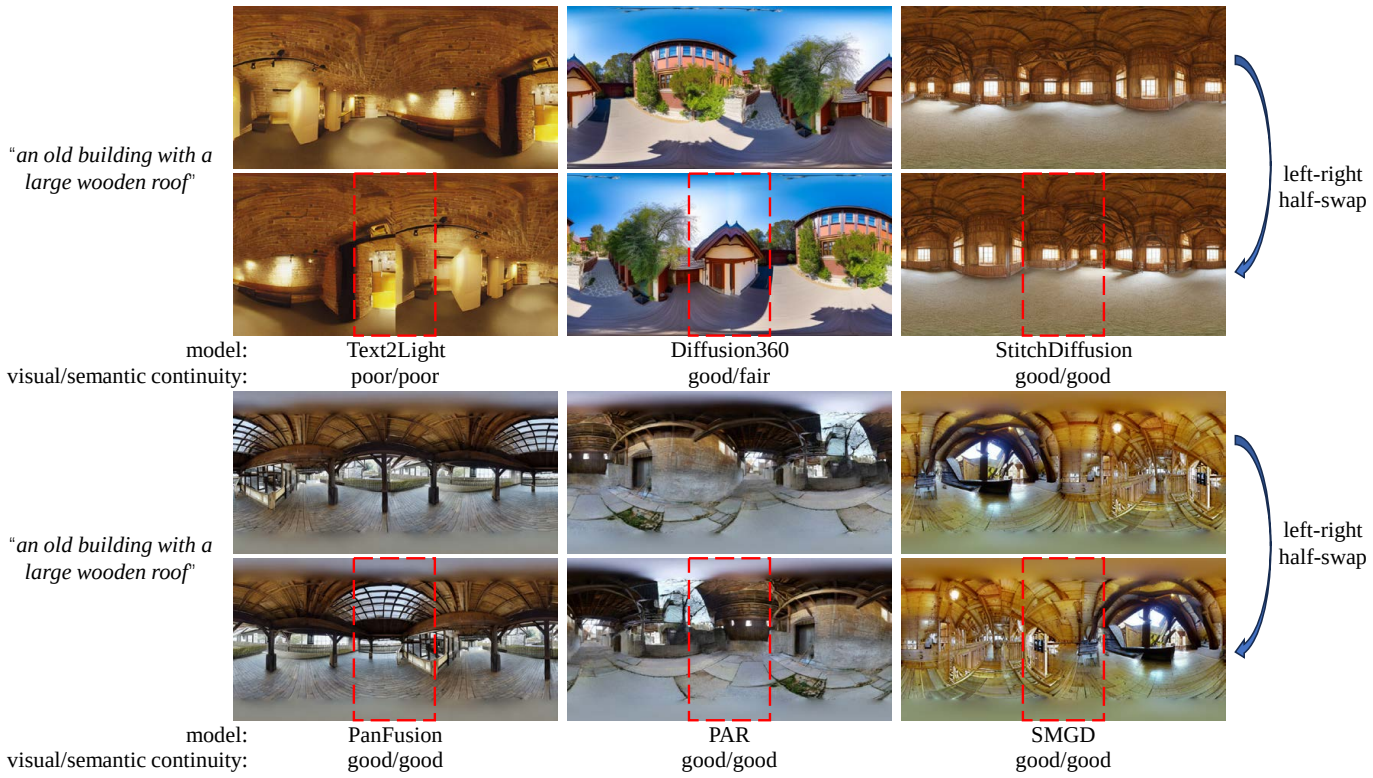

Fig. 6. Visual Comparison of Text-Only Generation Methods. To facilitate the assessment of visual and semantic continuities at the image borders, we generate "swapped" versions by exchanging the left and right halves of each generated image, and then use red dashed boxes in these swapped images to highlight the boundary regions. The continuity results are categorized as good, fair, or poor.

- **Training-Free:** The third strategy completely bypasses fine-tuning. Training-free methods like PanoFree [41] and SphereDiff [22] use pre-trained text-to-image models directly. They achieve panorama generation through careful orchestration of existing capabilities, such as iterative inpainting, view stitching, and specialized sampling algorithms. This approach requires no 360-degree panoramic training data and avoids costly training, fully leveraging the inherent power of the base models. The trade-offs include typically slower inference due to their multi-step, patch-based processes and a higher risk of global inconsistency, as there is no end-to-end training to enforce panoramic coherence.

D. Comparisons

To systematically evaluate strengths and weaknesses of representative methods, we conduct a benchmark on methods with publicly available inference code and model checkpoints. Our primary objective is to provide a fair and consistent comparison of the official releases, which requires running all models on the same hardware and evaluating them using a consistent set of metrics. We acknowledge that this approach introduces a selection bias, which means our quantitative comparison does not include several recent and important contributions that lack public code or model checkpoints. While this constraint limits the breadth of our comparison, it critically ensures the validity and fairness of the presented results. We encourage readers to consult the original papers for the reported performance of methods not included here. For Text-Only Generation, we compare Text2Light [56], Diffusion360 [58], StitchDiffusion [57], PanFusion [59], PAR [25], and SMGD [81]. For Text-Driven NFoV Outpainting, we compare PanoDiff [42], Diffusion360 [58], SpND [82] and DreamCube [83].

1) Qualitative Comparison: Fig. 6 presents a visual comparison of the six text-only generation methods. To facilitate the assessment of visual and semantic continuities at the image borders, we generate "swapped" versions by exchanging the left and right halves of each generated image. Red dashed boxes are employed in these swapped images to highlight the boundary regions. Among the evaluated methods, Text2Light exhibits obvious seams in the highlighted areas of its swapped images, indicating a failure to maintain visual and semantic continuities at the borders. While Diffusion360 effectively mitigates such observable seams and preserves visual continuity, it occasionally demonstrates local semantic discontinuity at the borders, where content across image boundaries lacks consistent logical or meaningful coherence. In contrast, the other four methods achieve strong visual and semantic continuities at the borders of their synthesized images.

Fig. 7 displays the visual comparison for the four text-driven NFoV outpainting approaches. Similar to the text-only comparison, swapped image versions were created to scrutinize the visual and semantic continuities at the borders. PanoDiff, Diffusion360, and DreamCube demonstrate commendable visual and semantic continuities at the borders of

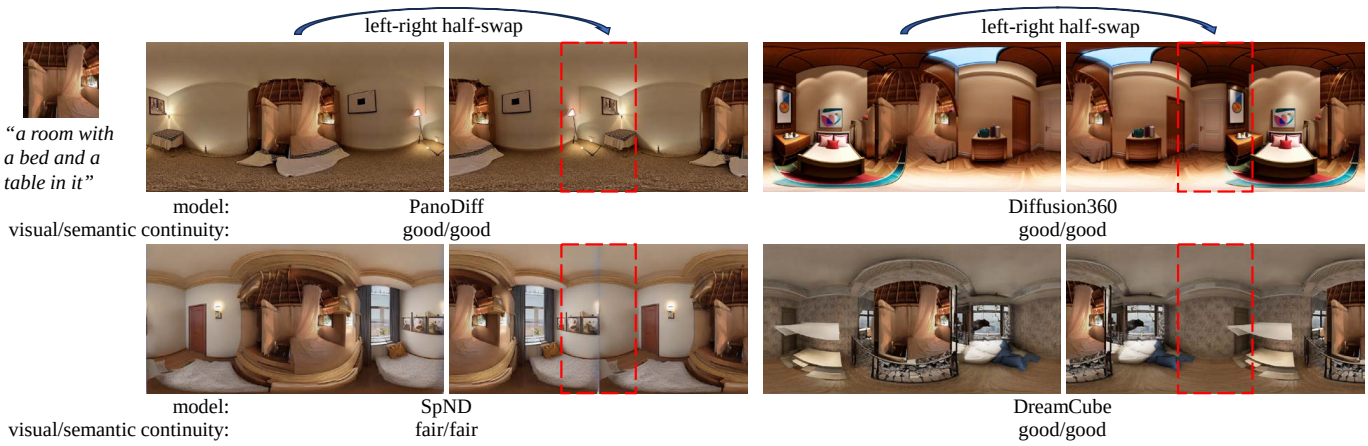

Fig. 7. Visual Comparison of Text-Driven NFoV Outpainting Approaches. The rest caption is the same as that for Fig. 6. Note that Diffusion360 provides two separate models. The model for text-driven NFoV outpainting generates images with good semantic continuity at the borders.

TABLE IV
QUANTITATIVE COMPARISON OF REPRESENTATIVE TEXT-DRIVEN 360-DEGREE PANORAMA GENERATION METHODS. THE FIRST BLOCK OF ROWS ARE FOR METHODS IN THE PARADIGM OF TEXT-ONLY GENERATION, WHILE THE SECOND BLOCK OF ROWS ARE FOR TEXT-DRIVEN NFOV OUTPAINTING. WE USE METRICS OUTLINED IN SEC. III-C FOR COMPREHENSIVE EVALUATION. THE INFERENCE TIME AND GPU MEMORY, REQUIRED BY EACH METHOD TO GENERATE A 1024×512 360-DEGREE PANORAMA, ARE ALSO REPORTED. THE **BEST** AND SECOND-BEST RESULTS ARE HIGHLIGHTED FOR THE TWO PARADIGMS, RESPECTIVELY.

Method	FID ↓	KID ($\times 10^{-2}$) ↓	IS ↑	CS ↑	FAED ↓	OmniFID ↓	DS ↓	Inference (s) ↓	GPU Memory (GB) ↓
Text2Light	72.63	1.54	5.35	19.20	18.10	99.81	5.38	33	12.5
Diffusion360	70.32	2.00	5.29	18.74	**12.43**	92.23	0.94	3	**3.5**
StitchDiffusion	76.69	2.04	**7.36**	19.20	15.58	108.63	1.07	28	3.6
PanFusion	**61.23**	**1.07**	6.16	18.96	13.16	**92.22**	0.85	30	26.3
PAR	64.96	1.49	6.68	18.91	13.99	104.02	0.76	17	18.6
SMGD	74.91	2.00	4.23	**19.22**	16.78	106.68	**0.75**	**2**	8.0
PanoDiff	65.94	2.44	4.72	19.02	10.24	122.30	1.10	48	36.0
Diffusion360	**64.19**	**2.05**	4.53	17.92	**5.50**	**101.39**	**0.72**	**4**	**3.7**
SpND	69.54	3.00	3.77	19.17	8.67	119.05	1.40	22	16.7
DreamCube	66.15	**2.05**	**4.88**	**19.26**	15.87	115.52	1.10	12	16.7

the produced images, while SpND exhibits fair visual and semantic continuities.

2) Quantitative Comparison: While the qualitative examples in Figs. 6 and 7 offer valuable visual insights, a quantitative evaluation is essential for a rigorous and objective comparison. To this end, we conduct a comprehensive comparison using metrics outlined in Sec. III-C.

To ensure an unbiased evaluation of the generalizability of the methods, we employ an out-of-domain dataset, ODI-SR [1]. This dataset was specifically chosen for two primary reasons: (1) None of the evaluated models were trained on it, which guarantees a fair test of generalization to unseen data; and (2) its diverse composition of indoor and outdoor 360-degree panoramas provides a robust benchmark for evaluating performance across varied real-world scenarios. For generating text descriptions, we use BLIP2 [18] to create textual captions for the 360-degree panoramas included in the ODI-SR dataset. These generated text prompts serve as inputs for both Text-Only Generation and Text-Driven NFoV Outpainting tasks. To simulate NFoV images, we first project the equirectangular 360-degree panoramas from the ODI-SR dataset into a cubemap format and then extract the front face of each cubemap. The original 360-degree panoramas from ODI-SR are designated as real images and used as ground truth for the computation of evaluation metrics.

The quantitative results obtained from this comparative evaluation, across the seven evaluation metrics, are presented in Tab. IV. To provide insights into the computational efficiency of each method, we also report the inference time and GPU memory required to generate a 1024×512 360-degree panorama on a consistent machine equipped with an RTX A6000 GPU. Note that our results only represent the performance of the publicly released version of each method on a specific GPU (RTX A6000, 48 GB). The actual performance may be influenced by hardware differences and implementation-specific optimizations. These quantitative findings effectively delineate the strengths and weaknesses of the evaluated approaches, offering valuable guidance for future research in this domain.

V. EMERGING DIRECTIONS

This section details two emerging directions that are closely related to text-driven 360-degree panoramic image generation: text-driven 360-degree 3D scene generation and text-driven 360-degree panoramic video generation, respectively.

Fig. 8. The Framework for Text-Driven 360-Degree 3D Scene Generation using Text-Driven 360-Degree Panorama Generation. This framework accommodates both Text-Only Generation and Text-Driven NFoV Outpainting methods. The input NFoV image is omitted when employing text-driven NFoV outpainting methods for simplicity. "3D lifting" denotes the transformation from a generated 360-degree panoramic image to a 3D scene representation by inferring the underlying geometry of the scene.

TABLE V
SUMMARY OF TEXT-DRIVEN 360-DEGREE 3D SCENE GENERATION METHODS. '360-DEGREE PG' INDICATES WHICH TEXT-DRIVEN 360-DEGREE PANORAMA GENERATION TECHNIQUES ARE ADOPTED TO SYNTHESIZE THE INTERMEDIATE PANORAMIC REPRESENTATION OF THE SCENE. '3DGS' DENOTES 3D GAUSSIAN SPLATTING.

Method	Publication	360-Degree PG	3D Lifting
FastScene	IJCAI 2024	Diffusion360	3DGS
DreamScene360	ECCV 2024	StitchDiffusion	3DGS
HoloDreamer	arxiv 2024	Diffusion360	3DGS
SceneDreamer360	arxiv 2024	PanFusion	3DGS
LayerPano3D	SIGGRAPH 2025	Diffusion360 & PanFusion	3DGS

A. 360-Degree 3D Scene Generation

Recent advances in text-driven 360-degree panorama generation [57]–[59] have catalyzed innovative methods for reconstructing 360-degree 3D scenes from textual descriptions. 360-degree panoramic images inherently capture both global contexts and geometric constraints of a scene, making them an essential intermediate representation for 3D scene generation. Consequently, recent text-driven 360-degree 3D scene generation methods [27], [30]–[33], [90]–[92] use 360-degree panorama generation to bridge the gap between text prompts and 360-degree 3D scene reconstruction.

As depicted in Fig. 8, these methods typically use a two-stage process: (1) 360-Degree Panorama Generation: generating a 360-degree panorama from the input text prompt using a fine-tuned LDM [63], and (2) 3D Scene Reconstruction: inferring a 3D representation, typically with 3D Gaussian Splatting (3DGS) [26], from the generated panorama and corresponding multi-view perspective images. Tab. V provides a comparative summary of the methods using text-driven 360-degree panorama generation techniques introduced in Sec. IV.

Within this framework, emerging methods are primarily differentiated by (a) their choice of 360-degree panorama generators and (b) their strategies for extracting and utilizing 3D information. For instance, FastScene [31] and HoloDreamer [30] both employ Diffusion360 [58] to generate the initial 360-degree panorama depicting a scene from a given text prompt. FastScene [31] then synthesizes multi-view panoramas of this scene for specific camera poses using Coarse View Synthesis and Progressive Novel View Inpainting. With these synthesized multi-view panoramas, FastScene introduces Multi-View Projection to get their perspective views. The point clouds derived from these views are then used as input for 3DGS to reconstruct the 3D scene. HoloDreamer [30] enhances the Diffusion360-generated panorama with two distinct ControlNet-based LDMs [37] and a super-resolution network to create a high-resolution, stylized output. Subsequently, HoloDreamer initializes 3D Gaussians using point clouds derived from a reverse equirectangular projection of the high-resolution panorama combined with its corresponding depth information. Finally, a two-stage 3DGS optimization process is developed to refine the scene rendering, resulting in the desired 3D scene reconstruction.

Furthermore, certain methods deviate from the reliance on Diffusion360 [58]. DreamScene360 [33] uses StitchDiffusion [57] to generate multiple 360-degree panorama candidates and then employs a self-refinement process to select the optimal candidate for initializing panoramic 3D Gaussians with a 3D geometric field. To facilitate visual feature correspondences between different views and maintain geometric consistencies during the 3DGS optimization process, semantic and geometric regularizations are applied. In contrast, SceneDreamer360 [32] uses a fine-tuned PanFusion [59] generator, coupled with a super-resolution module from [58], to produce a high-resolution (6K) panorama aligned with the input text prompt. It then uses optimization-based viewpoint selection to extract multi-view images, which are subsequently used for improved point cloud initialization, ultimately leading to 3DGS-based scene reconstruction.

Other methods explore alternative panorama generation techniques. LayerPano3D [27] begins by generating four orthogonal perspective views with a fine-tuned text-to-image model [21]. These initial views are then combined with text-guided inpainting [63], and further processed by using a fine-tuned Diffusion360 [58] model to outpaint the polar regions, resulting in a reference 360-degree panorama. To handle occlusions in complex scenes, LayerPano3D [27] decomposes the reference panorama into multiple depth-based layers and uses a fine-tuned inpainter [59] to complete unseen content at each layer. These inpainted, layered panoramas then provide supervision for panoramic 3D Gaussian scene optimization.

B. 360-Degree Panoramic Video Generation

Analogous to the natural evolution from text-to-image generation [12], [13], [21], [63] to text-to-video (T2V) generation [98]–[101], recent progress in text-driven 360-degree panoramic image synthesis has spurred research into the more challenging task: text-driven 360-degree panoramic video gen-

TABLE VI
SUMMARY OF TEXT-DRIVEN 360-DEGREE PANORAMIC VIDEO GENERATION METHODS. 'TF' SPECIFIES IF IT IS TRAINING-FREE.

Method	Publication	Training Dataset	TF	Representation
360DVD [93]	CVPR 2024	WEB360 [93]	×	ERP
DynamicScaler [94]	CVPR 2025	N/A	✓	ERP
PanoDiT [95]	AAAI 2025	PHQ360 [95]	×	ERP
SphereDiff [22]	arxiv 2025	N/A	✓	Spherical
VideoPanda [96]	arxiv 2025	WEB360 [93]	×	MPP
PanoWan [97]	arxiv 2025	PanoVid [97]	×	ERP

eration. Its representative methods [93]–[97] are summarized in Tab. VI.

Most works in this area rely on specialized training. 360DVD [93] pioneers this task by first constructing a tailored dataset, WEB360, consisting of 2,114 360-degree panoramic video-text pairs. Using this dataset, it trains an adapter to enable a pre-trained T2V models [99] to synthesize 360-degree panoramic videos from provided text prompts. Inspired by PanoDiff [42], 360DVD adopts a latent rotation mechanism in the inference process to maintain the boundary continuity of the synthesized results. Limitations in 360DVD, particularly the lack of detailed motion descriptions in its dataset, have prompted further refinements. PanoDiT [95] addresses this by curating a higher-quality subset named PHQ360 and replacing the U-Net architecture with a Diffusion Transformer (DiT) [102] and a motion LoRA for improved generation. Similarly, the rotation mechanism from PanoDiff is used in the post-processing phase of PanoDiT to ensure continuity. Recognizing the critical role of large-scale and high-quality data, PanoWan [97] establishes PanoVid, a dataset with over 13,000 video-text pairs. It introduces a latitude-aware sampling technique to mitigate ERP distortions and fined-tune a DiT-based T2V model [103] with LoRA for generation.

In a different vein, VideoPanda [96] seeks to better leverage priors from pre-trained T2V models. Trained on WEB360, it introduces multi-view attention layers to synthesize multiple perspective video outputs, which are then stitched together to form the final 360-degree panoramic video. This approach avoids direction in the equirectangular projection space.

Contrasting with these trained-based methods, several approaches have explored trained-free generation. DynamicScaler [94] designs an offset-shifting denoiser and a panoramic projection technique to synthesize a low-resolution 360-degree panoramic video, which then provides global motion guidance for refining a high-resolution version. However, its reliance on the ERP latent representation can lead to discontinuities near the poles. To address this, SphereDiff [22] introduces a spherical latent representation and extends Multidiffusion [55] to the constructed spherical space, achieving a more uniform distribution and improving quality at the poles. A common challenge to these training-free methods is that their patch-based synthesis mechanism can introduce global inconsistencies. Future work could focus on incorporating global guidance into these frameworks to mitigate this issue.

Datasets for 360-Degree Panoramic Video Generation. While the datasets discussed in Sec. III-B are foundational for static 360-degree panoramic image generation, the task of 360-degree panoramic video generation requires specialized datasets that include temporal information. These datasets are crucial for training and evaluating models capable of producing coherent and immersive 360-degree panoramic video content. Key datasets in this domain include:

- **WEB360** [93] offers 2,114 text-video pairs of 360-degree panoramas. The videos are sourced from existing datasets such as ODV360 [107] and platforms like YouTube. To generate detailed textual descriptions for the videos, a combination of BLIP [17] and ChatGPT was employed.
- **YouTube360** [105] provides 9,557 360-degree panoramic videos sourced from YouTube, featuring diverse scenes such as virtual city tours and wildlife documentaries. The corresponding text prompts were generated using VideoLLaMa-2 [108].
- **360-1M** [106] is a large-scale dataset consisting of 1,076,592 360-degree videos, collected from YouTube and distributed across 15 categories. As it was not originally created for text-driven generation, this dataset does not provide paired textual descriptions.
- **PanoVid** [97] is a high-quality dataset of over 13,000 video clips curated specifically for text-driven 360-degree panoramic video generation. The videos in PanoVid are collected from multiple sources, including WEB360, YouTube360, and 360-1M. Qwen-2.5-VL [109] was adopted to produce rich textual descriptions for the video content.

VI. CHALLENGES AND FUTURE DIRECTIONS

Despite the impressive results achieved in text-driven 360-degree panorama generation, challenges remain in evaluation metrics, resolution, controllability, model design, societal impact and industrial adoption. This section identifies these challenges and outlines potential directions for future research.

a) Evaluation Metrics: As established in our analysis (see Sec. III-C), the development of metrics for global scene consistency remains a key challenge. This includes creating more robust, panorama-aware methods for evaluating both text-to-image semantic alignment and the overall structural plausibility of the generated 360-degree panoramic space. Future work could explore using advanced Vision-Language Models (VLMs) for question-based evaluations of complex spatial relationships. Another promising direction is the development of metrics that assess the implicit 3D geometry of the scene to detect logical inconsistencies in scale and layout that 2D metrics currently miss.

b) Higher Resolution: While Diffusion360 [58], which uses a super-resolution module, is among the few methods that can currently achieve a maximum resolution of 6144×3072 (6K), this remains inadequate for demanding applications like VR gaming and high-fidelity 3D scene reconstruction, which often necessitate resolutions of 8K or higher to capture intricate details of landscapes and architecture. However, generating such high-resolution panoramas incurs both high memory consumption and long inference times, which severely limit practical deployment. Addressing these limitations will require the development of more efficient model architectures and

optimization techniques. Promising approaches include the use of window-based operations, model pruning, quantization, knowledge distillation, and advanced neural network designs tailored to resource-intensive tasks. Moreover, the availability of high-resolution, large-scale datasets will be critical for driving progress in this direction.

c) Multi-modal Generation: Existing text-driven methods, despite their ability to produce photorealistic 360-degree panoramas, often lack precise control over global semantic layout and spatial structure of the generated scene. This motivates exploring multi-modal approaches to enhance controllability. Although 360PanT [28] demonstrates panorama-to-panorama translation using auxiliary modalities like edge and segmentation maps alongside text, its outputs deviate from the standard $360° \times 180°$ field of view when these additional modalities are incorporated. Future research should focus on developing multi-modal techniques that effectively integrate diverse inputs (*e.g.* depth maps, segmentation maps, or edge maps) with text prompts to achieve fine-grained spatial control in the generated 360-degree panoramas, while ensuring strict adherence to the standard equirectangular projection format.

d) Model Design: Most text-driven 360-degree panorama generation methods are built upon latent diffusion models (LDMs) [63]. While LDMs have achieved remarkable success in text-to-image synthesis, recent advancements in autoregressive models indicate promising alternative architectures. Specifically, Visual Autoregressive (VAR) models, exemplified by Infinity [2], have exhibited superior performance compared to the leading LDMs in standard text-to-image synthesis. This highlights an exciting avenue for future research: exploring VAR-based models for text-driven 360-degree panorama generation.

e) Ethical and Societal Considerations: The persuasive nature of 360-degree panoramas introduces risks of misuse, including fabricated environments, disinformation, and privacy violations. To mitigate these concerns, future research should pair technical advances with safeguards such as transparent data documentation, responsible licensing, and watermarking mechanisms. Proactive engagement with ethical guidelines and interdisciplinary oversight is essential to ensure that text-driven 360-degree panorama generation benefits society while limiting potential harm.

f) Industry Translation and Adoption: Text-driven 360-degree panorama generation is beginning to move from research to practice, supporting VR/AR content creation, virtual tourism, and game development. Major industry players, including Meta (Quest), Google (Street View), and Apple (Vision Pro), are actively developing related capabilities. Broader deployment is currently limited by requirements for stable 8K plus resolution, low latency on-device inference, and industry-standard benchmarks. Addressing these challenges through more efficient model design and closer collaboration between academia and industry will be critical for production-ready adoption.

VII. Conclusion

This survey has provided a comprehensive overview of the rapidly evolving field of text-driven 360-degree panoramic image generation. We began by introducing primary representation methods of 360-degree panoramic images, along with widely used datasets, and key evaluation metrics in this domain. Subsequently, we presented an in-depth discussion of prevalent methods for text-driven 360-degree panorama generation, and explored its two closely related directions: text-driven 360-degree 3D scene generation and text-driven 360-degree panoramic video synthesis. Despite the significant progress achieved in this field, several challenges remain. To address these challenges, we have articulated promising directions for future research.

References

[1] X. Deng, H. Wang, M. Xu, Y. Guo, Y. Song, and L. Yang, "Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9189–9198.

[2] J. Han, J. Liu, Y. Jiang, B. Yan, Y. Zhang, Z. Yuan, B. Peng, and X. Liu, "Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 15733–15744.

[3] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First Conference on Language Modeling*, 2024.

[4] P. Gao, K. Yao, T. Ye, S. Wang, Y. Yao, and X. Wang, "Opa-ma: Text guided mamba for 360-degree image out-painting," *arXiv preprint arXiv:2407.10923*, 2024.

[5] Poly Haven, "HDRIs," https://polyhaven.com/hdris, 2025, accessed: February, 2025.

[6] E. Persson, "Texture from humus," https://www.humus.name/index.php?page=Textures, 2025, accessed: February, 2025.

[7] T. L. da Silveira, P. G. Pinto, J. Murrugarra-Llerena, and C. R. Jung, "3d scene geometry estimation from 360 imagery: A survey," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–39, 2022.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision*. Springer, 2020, pp. 405–421.

[9] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[10] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4460–4470.

[11] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 784–16 804.

[12] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[13] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[14] M. R. K. Dastjerdi, Y. Hold-Geoffroy, J. Eisenmann, S. Khodadadeh, and J.-F. Lalonde, "Guided co-modulated gan for 360 field of view extrapolation," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 475–485.

[15] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[16] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[17] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[18] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[19] M. Kocabas, C.-H. P. Huang, J. Tesch, L. Müller, O. Hilliges, and M. J. Black, "Spec: Seeing people in the wild with an estimated camera," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 035–11 045.

[20] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," in *International Conference on Learning Representations*, 2018.

[21] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," in *Proceedings of the International Conference on Learning Representations*, 2024.

[22] M. Park, T. Kang, J. Yun, S. Hwang, and J. Choo, "SphereDiff: Tuning-free Omnidirectional Panoramic Image and Video Generation via Spherical Latent Representation," *arXiv preprint arXiv:2504.14396*, 2025.

[23] Y. Yang, H. Duan, Y. Zhu, X. Liu, L. Liu, Z. Xu, G. Ma, X. Min, G. Zhai, and P. L. Callet, "Omni^2: Unifying Omnidirectional Image Generation and Editing in an Omni Model," *arXiv preprint arXiv:2504.11379*, 2025.

[24] D. Zheng, C. Zhang, X.-M. Wu, C. Li, C. Lv, J.-F. Hu, and W.-S. Zheng, "Panorama generation from NFoV image done right," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 21610–21619.

[25] C. Wang, X. Li, L. Qi, X. Lin, J. Bai, Q. Zhou, and Y. Tong, "Conditional Panoramic Image Generation via Masked Autoregressive Modeling," *arXiv preprint arXiv:2505.16862*, 2025.

[26] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[27] S. Yang, J. Tan, M. Zhang, T. Wu, G. Wetzstein, Z. Liu, and D. Lin, "LayerPano3D: Layered 3D panorama for hyper-immersive scene generation," in *Proceedings of the ACM SIGGRAPH Conference Papers*, 2025, pp. 1–10.

[28] H. Wang and J.-H. Xue, "360pant: Training-free text-driven 360-degree panorama-to-panorama translation," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 212–221.

[29] W. Xia and J.-H. Xue, "A survey on deep generative 3d-aware image synthesis," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–34, 2023.

[30] H. Zhou, X. Cheng, W. Yu, Y. Tian, and L. Yuan, "Holodreamer: Holistic 3d panoramic world generation from text descriptions," *arXiv preprint arXiv:2407.15187*, 2024.

[31] Y. Ma, D. Zhan, and Z. Jin, "FastScene: Text-driven fast 3D indoor scene generation via panoramic Gaussian splatting," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2024, pp. 1173–1181.

[32] W. Li, F. Cai, Y. Mi, Z. Yang, W. Zuo, X. Wang, and X. Fan, "Scenedreamer360: Text-driven 3d-consistent scene generation with panoramic gaussian splatting," *arXiv preprint arXiv:2408.13711*, 2024.

[33] S. Zhou, Z. Fan, D. Xu, H. Chang, P. Chari, T. Bharadwaj, S. You, Z. Wang, and A. Kadambi, "Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting," in *European Conference on Computer Vision*. Springer, 2024, pp. 324–342.

[34] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[35] H. Chen, Y. Zhang, X. Wang, X. Duan, Y. Zhou, and W. Zhu, "DisenDreamer: Subject-driven text-to-image generation with sample-aware disentangled tuning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

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

[38] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510.

[39] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[40] W. Ye, C. Ji, Z. Chen, J. Gao, X. Huang, S.-H. Zhang, W. Ouyang, T. He, C. Zhao, and G. Zhang, "DiffPano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion," *Advances in Neural Information Processing Systems*, vol. 37, pp. 1304–1332, 2024.

[41] A. Liu, Z. Li, Z. Chen, N. Li, Y. Xu, and B. A. Plummer, "Panofree: Tuning-free holistic multi-view image generation with cross-view self-guidance," in *European Conference on Computer Vision*. Springer, 2024, pp. 146–164.

[42] J. Wang, Z. Chen, J. Ling, R. Xie, and L. Song, "360-degree panorama generation from few unregistered nfov images," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6811–6821.

[43] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[44] N. Kalischek, M. Oechsle, F. Manhardt, P. Henzler, K. Schindler, and F. Tombari, "Cubediff: Repurposing diffusion-based image models for panorama generation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[45] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde, "Learning to predict indoor illumination from a single image," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, pp. 1–14, 2017.

[46] Y. Hold-Geoffroy, A. Athawale, and J.-F. Lalonde, "Deep sky modeling for single image outdoor lighting estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6927–6935.

[47] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou, "Structured3D: A large photo-realistic dataset for structured 3D modeling," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 519–535.

[48] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, "Recognizing scene viewpoint using panoramic place representation," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2695–2702.

[49] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," in *Proceedings of the International Conference on 3D Vision (3DV)*, 2017, pp. 667–676.

[50] C. Oh, W. Cho, Y. Chae, D. Park, L. Wang, and K.-J. Yoon, "Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 352–371.

[51] A. Christensen, N. Mojab, K. Patel, K. Ahuja, Z. Akata, O. Winther, M. Gonzalez-Franco, and A. Colaco, "Geometry fidelity for spherical images," in *European Conference on Computer Vision*. Springer, 2024, pp. 276–292.

[52] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[53] 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*, vol. 30, 2017.

[54] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[55] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "Multidiffusion: Fusing diffusion paths for controlled image generation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 1737–1752.

[56] Z. Chen, G. Wang, and Z. Liu, "Text2light: Zero-shot text-driven hdr panorama generation," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–16, 2022.

[57] H. Wang, X. Xiang, Y. Fan, and J.-H. Xue, "Customizing 360-degree panoramas through text-to-image diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4933–4943.

[58] M. Feng, J. Liu, M. Cui, and X. Xie, "Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models," *arXiv preprint arXiv:2311.13141*, 2023.

[59] C. Zhang, Q. Wu, C. C. Gambardella, X. Huang, D. Phung, W. Ouyang, and J. Cai, "Taming stable diffusion for text to 360 panorama image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6347–6357.

[60] Z. Lu, K. Hu, C. Wang, L. Bai, and Z. Wang, "Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 13, 2024, pp. 14 211–14 219.

[61] S. Tang, F. Zhang, J. Chen, P. Wang, and Y. Furukawa, "Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[62] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," in *The Twelfth International Conference on Learning Representations*, 2024.

[63] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[64] Z. Wang, O. Li, T. Wang, L. Wei, Y. Hao, X. Wang, and Q. Tian, "Prior Preserved Text-to-Image Personalization without Image Regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[65] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion models in generative AI: A survey," *arXiv preprint arXiv:2303.07909*, 2023.

[66] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023.

[67] F. Bie, Y. Yang, Z. Zhou, A. Ghanem, M. Zhang, Z. Yao, X. Wu, C. Holmes, P. Golnari, D. A. Clifton *et al.*, "RenAIssance: A survey into AI text-to-image generation in the era of large model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 3, pp. 2212–2231, 2025.

[68] HDRI Skies, "Free HDRI Skies," https://hdri-skies.com/, 2025, accessed: February, 2025.

[69] HDRMAPS, "HDRI Maps and Textures," https://hdrmaps.com/, 2025, accessed: February, 2025.

[70] iHDRI, "HDRI Skies – Outdoor," https://www.ihdri.com/hdri-skies-outdoor/, accessed: February, 2025.

[71] 360Cities, "360Cities – World Panoramic Photography," https://www.360cities.net/, accessed: February, 2025.

[72] F. Nazarieh, Z. Feng, M. Awais, W. Wang, and J. Kittler, "A survey of cross-modal visual content generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6814–6832, 2024.

[73] S. Frolov, B. B. Moser, and A. Dengel, "Spotdiffusion: A fast approach for seamless panorama generation over time," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025, pp. 2073–2081.

[74] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proceedings of the International Conference on Learning Representations*, 2014.

[75] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[76] S. Xiao, Y. Wang, J. Zhou, H. Yuan, X. Xing, R. Yan, C. Li, S. Wang, T. Huang, and Z. Liu, "Omnigen: Unified image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 13294–13304.

[77] T. Li, Y. Tian, H. Li, M. Deng, and K. He, "Autoregressive image generation without vector quantization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 56424–56445, 2024.

[78] H. Deng, T. Pan, H. Diao, Z. Luo, Y. Cui, H. Lu, S. Shan, Y. Qi, and X. Wang, "Autoregressive video generation without vector quantization," in *Proceedings of the International Conference on Learning Representations*, 2025.

[79] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I.-C. Chang, and Y. Xu, "Large scale image completion via co-modulated generative adversarial networks," in *Proceedings of the International Conference on Learning Representations*, 2021.

[80] J. Ni, C.-B. Zhang, Q. Zhang, and J. Zhang, "What Makes for Text to 360-degree Panorama Generation with Stable Diffusion?" *arXiv preprint arXiv:2505.22129*, 2025.

[81] X. Sun, M. Xu, S. Li, S. Ma, X. Deng, L. Jiang, and G. Shen, "Spherical manifold guided diffusion model for panoramic image generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5824–5834.

[82] X. Sun, S. Ma, S. Li, M. Xu, J. Xia, L. Jiang, X. Deng, and J. Wang, "Spherical-nested diffusion model for panoramic image outpainting," in *Proceedings of the Forty-second International Conference on Machine Learning*, 2025.

[83] Y. Huang, Y. Zhou, J. Wang, K. Huang, and X. Liu, "DreamCube: 3D panorama generation via multi-plane synchronization," *arXiv preprint arXiv:2506.17206*, 2025.

[84] LatentLabs360, "Latentlabs360," 2023. [Online]. Available: https://civitai.com/models/10753/latentlabs360

[85] Z. Cai, Z. Huang, X. Zheng, Y. Liu, C. Liu, Z. Wang, and L. Wang, "Interact360: Interactive identity-driven text to 360° panorama generation," in *Proceedings of the IEEE Conference on Artificial Intelligence (CAI)*, 2024, pp. 728–736.

[86] E. Brivio, S. Serino, E. Negro Cousa, A. Zini, G. Riva, and G. De Leo, "Virtual reality and 360 panorama technology: A media comparison to study changes in sense of presence, anxiety, and positive emotions," *Virtual Reality*, vol. 25, pp. 303–311, 2021.

[87] C.-L. Fan, W.-C. Lo, Y.-T. Pai, and C.-H. Hsu, "A survey on 360 video streaming: Acquisition, transmission, and display," *ACM Computing Surveys*, vol. 52, no. 4, 2019, pp. 1–36.

[88] H. Ai, Z. Cao, and L. Wang, "A Survey of Representation Learning, Optimization Strategies, and Applications for Omnidirectional Vision," *International Journal of Computer Vision*, vol. 133, no. 8, pp. 4973–5012, 2025.

[89] B. Yang, W. Dong, L. Ma, W. Hu, X. Liu, Z. Cui, and Y. Ma, "Dreamspace: Dreaming your room space with text-driven panoramic texture propagation," in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2024, pp. 650–660.

[90] Z. Xiong, Z. Chen, Z. Li, Y. Xu, and N. Jacobs, "PanoDreamer: Consistent text to 360-degree scene generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 295–304.

[91] J. Yuan, B. Yang, K. Wang, P. Pan, L. Ma, X. Zhang, X. Liu, Z. Cui, and Y. Ma, "ImmerseGen: Agent-guided immersive world generation with alpha-textured proxies," *arXiv preprint arXiv:2506.14315*, 2025.

[92] G. Wang, P. Wang, Z. Chen, W. Wang, C. C. Loy, and Z. Liu, "Perf: Panoramic neural radiance field from a single panorama," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 10, pp. 6905–6918, 2024.

[93] Q. Wang, W. Li, C. Mou, X. Cheng, and J. Zhang, "360dvd: Controllable panorama video generation with 360-degree video diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6913–6923.

[94] J. Liu, S. Lin, Y. Li, and M.-H. Yang, "Dynamicscaler: Seamless and scalable video generation for panoramic scenes," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6144–6153.

[95] M. Zhang, Y. Chen, R. Xu, C. Wang, J. Yang, W. Meng, J. Guo, H. Zhao, and X. Zhang, "PanoDit: Panoramic videos generation with diffusion transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10040–10048.

[96] K. Xie, A. Sabour, J. Huang, D. Paschalidou, G. Klar, U. Iqbal, S. Fidler, and X. Zeng, "VideoPanda: Video panoramic diffusion with multi-view attention," *arXiv preprint arXiv:2504.11389*, 2025.

[97] Y. Xia, S. Weng, S. Yang, J. Liu, C. Zhu, M. Teng, Z. Jia, H. Jiang, and B. Shi, "PanoWan: Lifting diffusion video generation models to 360-degree with latitude/longitude-aware mechanisms," *arXiv preprint arXiv:2505.22016*, 2025.

[98] J. An, S. Zhang, H. Yang, S. Gupta, J.-B. Huang, J. Luo, and X. Yin, "Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation," *arXiv preprint arXiv:2304.08477*, 2023.

[99] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning," in *Proceedings of the International Conference on Learning Representations*, 2024.

[100] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang, *et al.*, "Make-your-video: Customized video generation using textual and structural guidance," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[101] H. Wang, C.-Y. Ma, Y.-C. Liu, J. Hou, T. Xu, J. Wang, F. Juefei-Xu, Y. Luo, P. Zhang, T. Hou, *et al.*, "Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2578–2588.

[102] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[103] Team Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[104] B. Wen, H. Xie, Z. Chen, F. Hong, and Z. Liu, "3D scene generation: A survey," *arXiv preprint arXiv:2505.05474*, 2025.
[105] J. Tan, S. Yang, T. Wu, J. He, Y. Guo, Z. Liu, and D. Lin, "Imagine360: Immersive 360 video generation from perspective anchor," *arXiv preprint arXiv:2412.03552*, 2024.
[106] M. Wallingford, A. Bhattad, A. Kusupati, V. Ramanujan, M. Deitke, A. Kembhavi, R. Mottaghi, W.-C. Ma, and A. Farhadi, "From an image to a scene: Learning to imagine the world from a million 360 videos," *Advances in Neural Information Processing Systems*, vol. 37, pp. 17743–17760, 2024.
[107] M. Cao, C. Mou, F. Yu, X. Wang, Y. Zheng, J. Zhang, C. Dong, G. Li, Y. Shan, R. Timofte, *et al.*, "NTIRE 2023 Challenge on 360° Omnidirectional Image and Video Super-Resolution: Datasets, Methods and Results," *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1731–1745, 2023.
[108] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, *et al.*, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs," *arXiv preprint arXiv:2406.07476*, 2024.
[109] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, *et al.*, "Qwen2.5-VL Technical Report," *arXiv preprint arXiv:2502.13923*, 2025.

Hai Wang received the B.E. degree in electronic engineering from Xidian University in 2019, and the M.E. degree in electronic engineering from Tsinghua University in 2022. He is currently pursuing the Ph.D. degree in statistical science with University College London. His research interests include image generation and manipulation, generative models, video super-resolution and enhancement.

Xiaoyu Xiang received the B.E. degree in engineering physics from Tsinghua University in 2015, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Purdue University in 2021. She is currently a Research Scientist with the Meta Reality Labs, USA. Her primary area of research has been image and video restoration, novel view synthesis and generative models.

Weihao Xia received the BEng degree in automation from Sun Yat-sen University in 2016, the MEng degree in control engineering from Tsinghua University in 2019, and the PhD degree from University College London in 2025. His research interests include computer vision and statistical machine learning, especially in controllable and interpretable visual content creation.

Jing-Hao Xue (Senior Member, IEEE) received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998, and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is a Professor of Statistical Pattern Recognition in the Department of Statistical Science, University College London. His research interests include statistical pattern recognition, machine learning, and computer vision. He is a Senior Area Editor of the IEEE Transactions on Circuits and Systems for Video Technology.