HunyuanWorld 1.0: Generating Immersive, Explorable, and Interactive 3D Worlds from Words or Pixels

Tencent Hunyuan*

https://3d.hunyuan.tencent.com/sceneTo3D
https://github.com/Tencent-Hunyuan/HunyuanWorld-1.0

Abstract

Creating immersive and playable 3D worlds from texts or images remains a fundamental challenge in computer vision and graphics. Existing world generation approaches typically fall into two categories: video-based methods that offer rich diversity but lack 3D consistency and rendering efficiency, and 3D-based methods that provide geometric consistency but struggle with limited training data and memory-inefficient representations. To address these limitations, we present HunyuanWorld 1.0, a novel framework that combines the best of both worlds for generating immersive, explorable, and interactive 3D scenes from text and image conditions. Our approach features three key advantages: 1) 360° immersive experiences via panoramic world proxies; 2) mesh export capabilities for seamless compatibility with existing computer graphics pipelines; 3) disentangled object representations for augmented interactivity. The core of our framework is a semantically layered 3D mesh representation that leverages panoramic images as 360° world proxies for semantic-aware world decomposition and reconstruction, enabling the generation of diverse 3D worlds. Extensive experiments demonstrate that our method achieves state-of-the-art performance in generating coherent, explorable, and interactive 3D worlds while enabling versatile applications in virtual reality, physical simulation, game development, and interactive content creation.

∗ HunyuanWorld 1.0 team contributors are listed in the end of report.

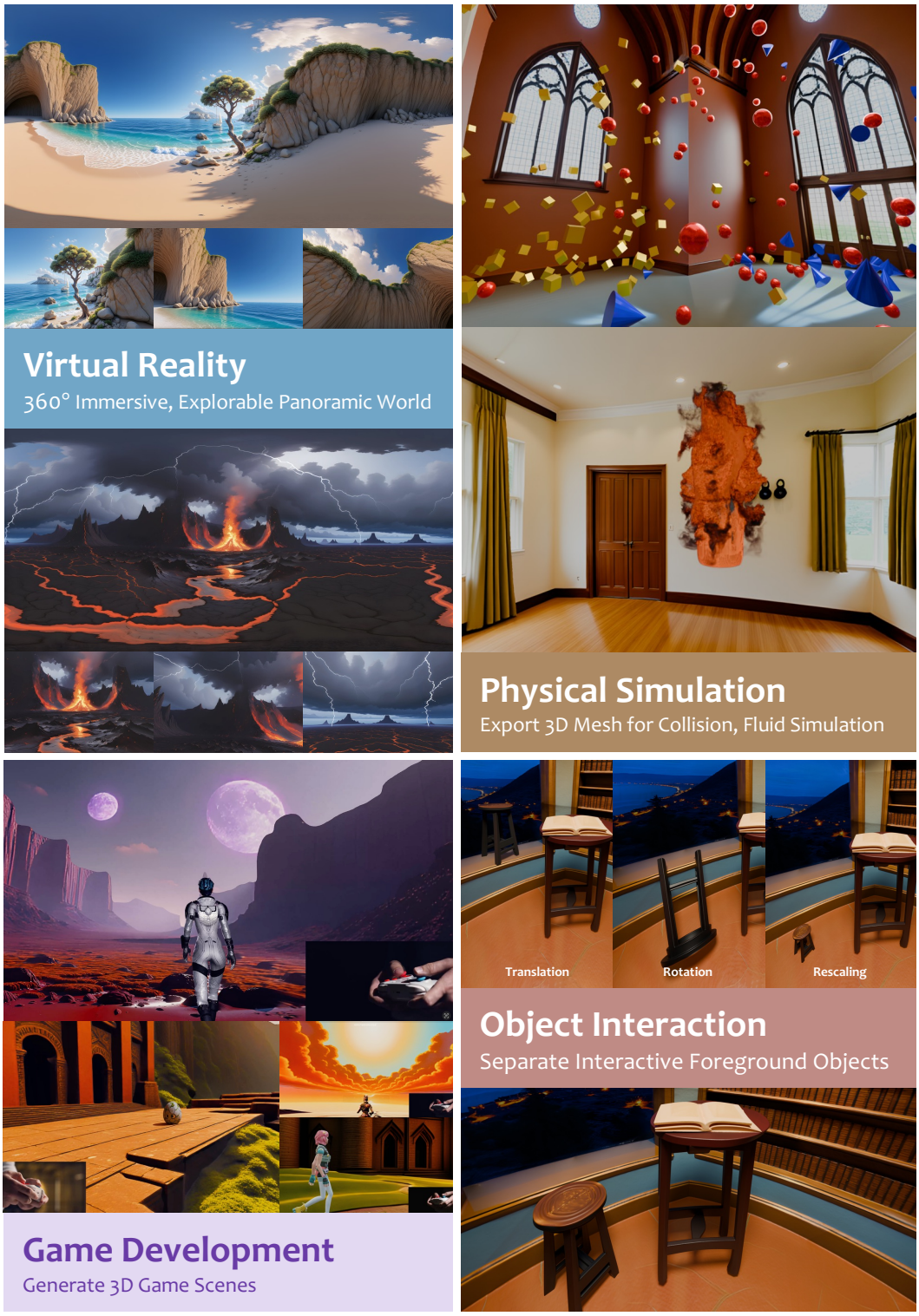


Figure 1: An overview of HunyuanWorld 1.0 applications.

1 Introduction

"To see a World in a Grain of Sand, and a Heaven in a Wild Flower"
— William Blake

World models have emerged as a fundamental paradigm for understanding and generating complex 3D environments, with applications spanning virtual reality, autonomous driving, robotics, and video gaming. The ability to create immersive, explorable, and interactive 3D worlds from natural language descriptions or visual inputs represents a crucial milestone toward democratizing 3D content creation and enabling new forms of human-computer interaction.

Recently, world generation has achieved remarkable progress, and the main solutions can be divided into two categories: video-based world generation methods and 3D-based world generation methods. Video-based methods leverage the inherent world knowledge of video diffusion models [24, 65, 51, 42, 50, 28, 3] to understand the temporal and spatial relationships of the generated world. The rich training data available for video models enables them to capture complex real-world dynamics for generating visually compelling results across diverse scenarios. Some works further incorporate 3D constraints such as camera trajectories [13, 1, 2, 43] or explicit 3D scene point clouds [19, 61] to spatially control the generated video sequences and produce plausibly 3D consistent video worlds.

However, video-based approaches face several fundamental limitations that constrain their practical performance. First, they inherently lack true 3D consistency due to their underlying 2D frame-based representation. This leads to temporal inconsistencies, particularly when generating long-range video scenes where accumulated errors result in severe content drift and incoherence. Second, the rendering costs of video-based methods are prohibitive, as each frame should be generated sequentially. Third, the frame-based video format is fundamentally incompatible with existing computer graphics pipelines, making it hard to be incorporated into game engines, VR applications, and other interactive systems.

In contrast, 3D-based world generation methods directly model geometric structures and offer superior compatibility with current computer graphics pipelines. These approaches provide inherent 3D consistency with efficient real-time rendering. Despite rapid advances in object-level 3D generation [37, 54, 45, 57, 16, 14, 44, 62, 47, 48, 49], world-level 3D generation remains significantly underexplored. Although some recent works [4, 69, 68, 64, 63] have demonstrated the potential of generating multi-view consistent 3D scenes from text descriptions, world-level 3D synthesis remains constrained by several critical challenges. The primary limitation is the scarcity of high-quality 3D scene data compared to the abundant image and video datasets. Additionally, existing 3D representations for generative models are either unstructured or memory-inefficient for large-scale scenes. In addition, previous methods typically generate monolithic 3D scenes where individual objects are not separated, limiting their applicability to interactive manipulations.

To address these fundamental limitations, we propose HunyuanWorld 1.0, a novel world generation framework that combines the best of both worlds, rather than treating 2D and 3D generation as separate paradigms. At the core of our approach is a semantically layered 3D mesh representation that enables structured 3D world generation with instance-level object modeling. Our method delivers three advanced features that distinguish it from existing approaches: (1) **360° immersive experiences** through panoramic world proxies that provide complete 360° scene coverage; (2) **mesh export capability** for seamless compatibility with existing computer graphics pipelines and industry-standard workflows; (3) **disentangled object representations** for object-level interaction within generated scenes.

To generate immersive and interactive 3D worlds with semantic layers, HunyuanWorld 1.0 incorporates several key designs. First, we introduce *panoramic world image generation* that serves as a unified world proxy for both text-to-world and image-to-world generation, leveraging the diversity of 2D generative models while providing immersive 360° environmental coverage. Second, we leverage *agentic world layering* for automating the decomposition of complex scenes into semantically meaningful layers, preparing for the subsequent layer-wise 3D reconstruction and disentangled object modeling. Third, we utilize *layer-wise 3D world reconstruction*, which estimates aligned panoramic depth maps for generating a mesh-based 3D world across all extracted layers. The hierarchical world mesh representation contains explicitly separated objects while maintaining efficient memory usage and rendering performance. Finally, we present *long-range world exploration* with a novel

world-consistent video diffusion model and a world caching mechanism, facilitating user navigation through extensive unseen scene areas far beyond the original viewpoints.

Taken together, these innovations enable our framework to achieve state-of-the-art performance in generating immersive, explorable, and interactive 3D worlds across diverse artistic styles and vairous scene types. Extensive experiments demonstrate that our method achieves superior performance compared to existing approaches. Besides, HunyuanWorld 1.0 supports versatile applications ranging from virtual reality and game development to physical simulation and object interaction. HunyuanWorld 1.0 represents a significant step toward living out everyone's imagination on creating, exploring, and manipulating 3D worlds, bridging the gap between 2D content creation and immersive 3D experiences.

2 Technical Details

Our goal is to enable 3D world generation that supports both image and text inputs, adapting to diverse user needs. Compared to 3D objects, 3D worlds exhibit far greater diversity — encompassing indoor and outdoor environments, varying styles, and a wide range of scales, from a single room to an entire city. However, 3D scene data is scarce and hard to scale, making 3D world generation challenging. To address this, we propose combining 2D generative models with 3D generation by leveraging panoramas as a proxy representation for worlds. As illustrated in Fig. 2, HunyuanWorld 1.0 is a staged generative framework, where we first utilize a diffusion model to generate a panorama as the world initialization, followed by world layering and reconstruction. We detail the whole 3D world generation pipeline in the following subsections.

2.1 Generating Panoramas As World Proxy

Panoramas capture 360° visual information of a scene and can be formatted as equirectangular projection (ERP) images, making them an ideal proxy for 3D world generation. Thus, we generate a panorama as proxy for 3D world generation from text conditions or image conditions.

Text Conditions. For text-to-panorama generation, the inputs are user-provided sentences. A critical challenge arises here: natural language inputs from users often differ significantly from the caption styles on which the models are trained. To bridge this gap, we employ a LLM to translate the given text prompts if they are in Chinese, and then enhancing their details. This transformation ensures the prompts are well-aligned with the training data distribution of the generative model, thereby facilitating the generation of high-quality panoramas.

Image Conditions. Given a user-provided pinhole image, we aim generate coherent contents in the missing parts to obtain a complete 360° panorama while preserving the input image's content. To achieve this, we unproject the input image into panoramic space via equirectangular projection (ERP) with camera intrinsics estimated by a pretrianed 3D reconstruciton model, such as MOGE [53] or UniK3D [36].

Panorama Generation. The architecture of our panorama generation model (Panorama-DiT) is based on the Diffusion Transformer (DiT) framework [35]. For text-to-panorama generation, only the enhanced text prompts are fed into the diffusion model as condition. For image-to-panorama generation, we first project the input image into panoramic space and then encode it into the latent space using a variational autoencoder (VAE). Next, we concatenate the condition image with the noisy latent for diffusion model. To enhance the generation quality and provide additional control, the image-to-panorama generation process is also conditioned on an auxiliary textual description produced by the scene-aware prompt generation method introduced in Sec. 2.2.

Compared to general image generation, generating panoramic images faces unique challenges: 1) geometric distortion from spherical projection and 2) incontinuous boundary due to panoramic stitching. To address these issues, we introduce two key boundary artifact mitigation strategies: 1) elevation-aware augmentation. During training, we randomly shift ground-truth panoramas vertically (with probability p and displacement ratio r) to enhance the robustness to viewpoint variations. 2) circular denoising [7]. During inference, we apply circular padding with progressive blending in denoising process to preserve structural and semantic continuity across panorama boundaries.

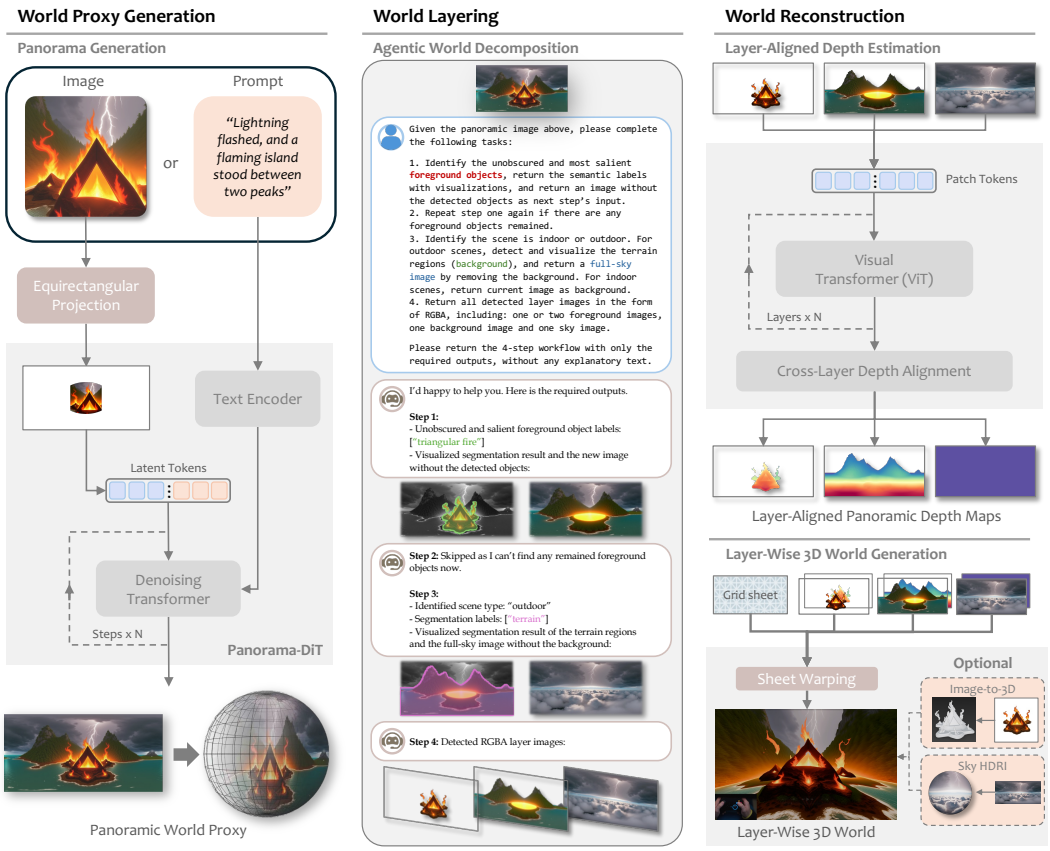


Figure 2: An overview of HunyuanWorld 1.0 architecture for 3D world generation. Given a conditioned scene image or textual description, HunyuanWorld 1.0 generates layer-wise 3D worlds in mesh through a staged generative framework. We first leverage a diffusion model (Panorama-DiT) to generate a panoramic image, which serves as an initial world proxy for providing full 360° scene information. We then obtain semantically layered scene representations via world layering and reconstruction. To ensure layer-wise alignment of the reconstructed 3D world, we enhance the panoramic depth estimation model with a cross-layer depth alignment strategy. Also, users can obtain full 3D objects via image-to-3D generation or represent the sky as HDRI maps for downstream applications.

2.2 Panoramic Data Curation Pipeline

Data Curation. The pipeline for our training data curation is illustrated in Fig. 3. Panoramic images are sourced from commercial acquisitions, open data downloads, and custom renders via Unreal Engine (UE). Each panorama undergoes an automatic quality assessment framework, including watermark, aesthetic score, clarity, resolution, and distortion, *etc*. Panoramas failing to meet predefined quality baselines will be discarded. We also invite expert annotators to manually inspect remaining samples, filtering examples with artifacts such as: 1) geometric artifacts (*e.g.*, obvious distortion, visible boundary seam), 2) scene irregularities (*e.g.*, narrow/unrepresentative spaces), and 3) content inconsistencies (*e.g.*, abnormal object repetition, anomaly human bodies and objects).

Training Caption. Existing VLMs face challenges when generating captions for panoramic images, which contain richer visual details than general perspective images. It either generates overly simplified descriptions that fail to capture sufficient scene details or produces repetitive text with hallucinated elements. To mitigate these issues, we propose a three-stage captioning pipeline. We first leverage the re-captioning technique [27] to produce regularized descriptions with rich details for panoramas. We then utilize LLM to distill these descriptions into a collection of captions with varying lengths, spanning from high-level scene summaries to fine-grained object annotations. Finally, we

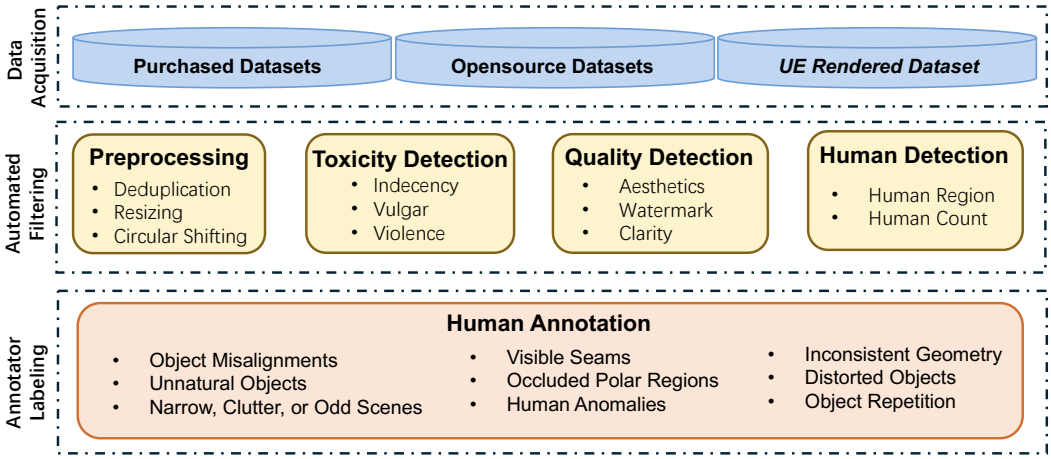


Figure 3: An overview of our panoramic data curation pipeline.

invite professional annotators to verify the generated captions to eliminate image-text misalignment, ensuring semantic fidelity and minimizing hallucinations.

Scene-Aware Prompt for Image Conditions. As stataed in Sec. 2.1, for image-to-panorama generation, we utilize both image and text conditions. A straightforward approach to obtain the text prompt would be to use a vision-language model (VLM) to generate a caption for the input image. However, user-provided images often contain prominent objects (*e.g.*, a statue), and conditioning the generation process on the input image's caption may lead to unwanted replication of these elements in the synthesized panorama (*e.g.*, duplicate statues).

To address this, we introduce a scene-aware prompt generation strategy. We first instruct the VLM to identify salient objects in the input image and incorporate these as negative prompts to prevent the model from redundantly reproducing existing objects. We then instruct the VLM to envision a complete 360° scene that extends beyond the input image. Finally, we instruct the VLM to produce a refined and complete prompt that describe the scene hierarchically, from foreground to background and artistic styles to environmental atmosphere.

2.3 Agentic World Layering

While panoramas effectively serve as world proxies, they inherently lack information in occluded regions, supporting only viewpoint rotation rather than free exploration (*e.g.* viewpoint translation). Inspired by human modeling practices, where artists typically model a 3D world scene as sky (sky box or dome), terrain mesh, and multiple object assets, we introduce a semantically layered 3D world representation that decomposes a scene into a sky layer (for outdoor scenes), a background layer, and multiple object layers. To automate the layering process, we develop an agentic world decomposition method consisting of instance recognition, layer decomposition, and layer completion.

Instance Recognition. In the context of a playable 3D world, interaction with specific scene objects is essential, while background elements typically remain static. To enable an interactive 3D scene, we must model each interactive object individually in 3D. Our approach begins with identifying which scene objects require such modeling, which demands semantic understanding and spatial relationship reasoning. Given the diversity of our generated scenes—spanning indoor/outdoor environments, natural landscapes, and game-style settings, we leverage a VLM to harness its rich world knowledge for semantic object recognition. Following instance recognition, we categorize objects into distinct sub-layers based on their semantic and spatial relationships. For instance, in an urban scene, we target segregating nearby vehicles from distant buildings into separate layers.

Layer Decomposition. Once obtaining semantic labels, the next step is to determine the precise positions of these recognized objects. However, conventional visual grounding models cannot be directly applied to panoramas due to their inherent spatial discontinuities, where objects may be fragmented across the left and right boundaries of an equirectangular projection (ERP) panorama.

To address this, we preprocess the panoramic image using circular padding before inputting it to an object detector (*e.g.*, Grounding DINO [30]). This transformation ensures that objects spanning the panorama's boundaries are treated as contiguous entities. Following detection, we remap the bounding box coordinates from the padded space back to the original panorama. Subsequently, we pass the detected bounding boxes to a segmentation model (*e.g.*, ZIM [22]) to generate pixel-wise masks. To handle overlapping or fragmented detections, we apply Non-Maximum Suppression (NMS) based on object area size. This ensures that part-level objects separated by the panorama's boundaries are merged, enabling more accurate 3D modeling of interactive scene elements.

Layer Completion. Following object segmentation, we decompose the panorama into background (*e.g.*, terrain) and sky layers through an autoregressive "onion-peeling" process. This involves iteratively removing recognized objects and completing each layer by inpainting occluded regions.

To train a layer completion model, we curate a panoramic object removal dataset consisting of triples of a object mask, the original panorama containing the object, and a target panorama with the object removed. Using this dataset, we fine-tune our Panorama-DiT model to learn the conditional generation of occluded regions. Similarly, we also finetune a completion model for sky layer on a dataset of sky HDRIs.

2.4 Layer-Wise World Reconstruction

Given the hierarchical world layers, we reconstruct the 3D world in a layer-wise manner. As shown in Fig. 2 (right), the reconstruction process includes two stages: (1) layer-aligned depth estimation and (2) layer-wise 3D world generation.

Layer-Aligned Depth Estimation. Given the panoramic world proxies, we predict the depth of each layer and then conduct cross-layer depth alignment. Specifically, we first obtain a base depth map by applying the depth estimation model [53, 36] to the original panorama. The depth of objects in the first foreground layer can be extracted from the base panoramic depth map.

For subsequent layers (*e.g.*, the subsequent foreground layers and the background layer after foreground removal), we predict their depths separately and align them with the base panoramic depth map using depth matching techniques that minimizing the distance of overlapped regions across different layers. This ensures consistent depth relationships across different layers to maintain the geometric coherence of the reconstructed 3D scene. For the sky layer, we set its depth to a constant value that slightly lager than the maximum depth value observed across all existing layers, ensuring the sky appears at the farthest distance.

Layer-Wise 3D World Generation. Given the layered images with aligned depth maps, we reconstruct the world via sheet warping with a grid mesh representation in WorldSheet [18]. The reconstruction process follows a hierarchical approach.

Foreground Object Reconstruction. For each foreground layer, we offer two reconstruction strategies: (1) *Direct projection*, where we convert the foreground objects directly to 3D meshes via sheet warping based on their depths and semantic masks. To ensure the quality of meshes warped from a panoramic image with masks, we also introduce special handling for polar region smoothing and mesh boundary anti-aliasing; (2) *3D generation*, where we generate complete 3D objects in the foreground layers and then place them into the 3D world. To obtain the foreground 3D objects, we extract individual object instances from the foreground layers based on their instance masks, and leverage image-to-3D generation models (*e.g.*, Hunyuan3D [47, 48, 49]) to create high-quality 3D object assets. We also propose an automatic object placement algorithm to place generated objects into 3D scenes considering spatial layout.

Background Layer Reconstruction. for background layer, we first apply adaptive depth compression to handle depth outliers and ensure proper depth distribution. We then convert the background panoramic image into a 3D mesh via sheet warping using the processed background depth map.

Sky Layer Reconstruction. The sky layer is reconstructed using the sky image with uniform depth values set to be slightly larger than the maximum scene depth. In addition to traditional mesh representation obtained via sheet warping, we also support HDRI environment map representation for more realistic sky rendering in VR applications.

We also support 3D gaussian splatting as an alternative to the mesh representation by optimizing a layered 3DGS representation based on the depth. To handle cross-boundary consistency in equirectangular projections, we apply circular padding during the reconstruction process, ensuring seamless transitions at the panorama boundaries. The final layered 3D world maintains proper occlusion relationships and depth ordering, enabling realistic VR experiences with proper parallax effects.

2.5 Long-Range World Extension

While layer-wise world reconstruction enables world exploration, challenges remain with occluded views and limited exploration range. To address these limitations, we introduce Voyager [19], a video-based view completion model that enables consistent world extrapolation. Voyager combines world-consistent video diffusion with long-range exploration mechanisms to synthesize spatially coherent RGB-D videos from an initial world view and user-specified camera trajectories.

World-Consistent Video Diffusion. Voyager employs an expandable world caching mechanism to maintain spatial consistency and prevent visual hallucination. The system constructs an initial 3D point cloud cache with the generated 3D scene, then projects this cache into target camera views to provide partial guidance for the diffusion model. The generated frames continuously update and expand the world cache, creating a closed-loop system that supports arbitrary camera trajectories while preserving geometric coherence.

Long-Range World Exploration. To overcome the limitations of generating long videos in a single pass, we propose a world caching scheme combined with smooth video sampling for auto-regressive scene extension. The world cache accumulates point clouds from all generated frames, with a point culling method that removes redundant points to optimize memory usage. Using cached point clouds as spatial proxies, we develop a smooth sampling strategy that auto-regressively extends video sequences while ensuring seamless transitions between clips.

2.6 System Efficiency Optimization

To ensure practical deployment and real-time performance, HunyuanWorld 1.0 incorporates comprehensive system optimizations across both mesh storage and model inference components.

Mesh Storage Optimization. Meshes for a 3D scene are large for loading and storing. We thus employ dual compression strategies for both offline usage and online deployment scenarios to achieve efficient storage and fast loading while maintaining visual quality.

Mesh Decimation with Advanced Parameterization. For offline mesh usage, we employ a multi-stage pipeline consisting of mesh decimation, texture baking, and UV parameterization. We evaluate an XAtlas-based solution [67] for UV parameterization, which keeps good UV quality while eliminating rendering seams compared with naive parameterization methods. The compression pipeline achieves 80% size reduction, making it suitable for high-quality offline content preparation despite extended processing times.

Web-Optimized Draco Compression. For online web deployment scenarios, we adopt the Draco [12], which delivers exceptional compression efficiency while preserving visual fidelity. This approach demonstrates superior size reduction (90%) capabilities and maintains rendering quality comparable to uncompressed meshes. The format provides native WebAssembly support, ensuring seamless integration with web-based graphics pipelines and broad browser compatibility.

Model Inference Acceleration. Our inference optimization employs a comprehensive TensorRT-based acceleration framework with intelligent caching and multi-GPU parallelization. The system converts diffusion transformer models into optimized TensorRT engines, supporting both cached and uncached inference modes with shared memory allocation to minimize GPU overhead. We implement a selective caching strategy that applies cached inference for non-critical denoising steps while using full computation for key steps that significantly impact generation quality. For classifier-free guidance scenarios, the system leverages multi-GPU parallel processing through threaded execution, simultaneously computing positive and negative prompt conditions on separate devices with synchronized result aggregation. This integrated optimization approach enables fast 3D world generation while maintaining high visual quality across diverse deployment environments.

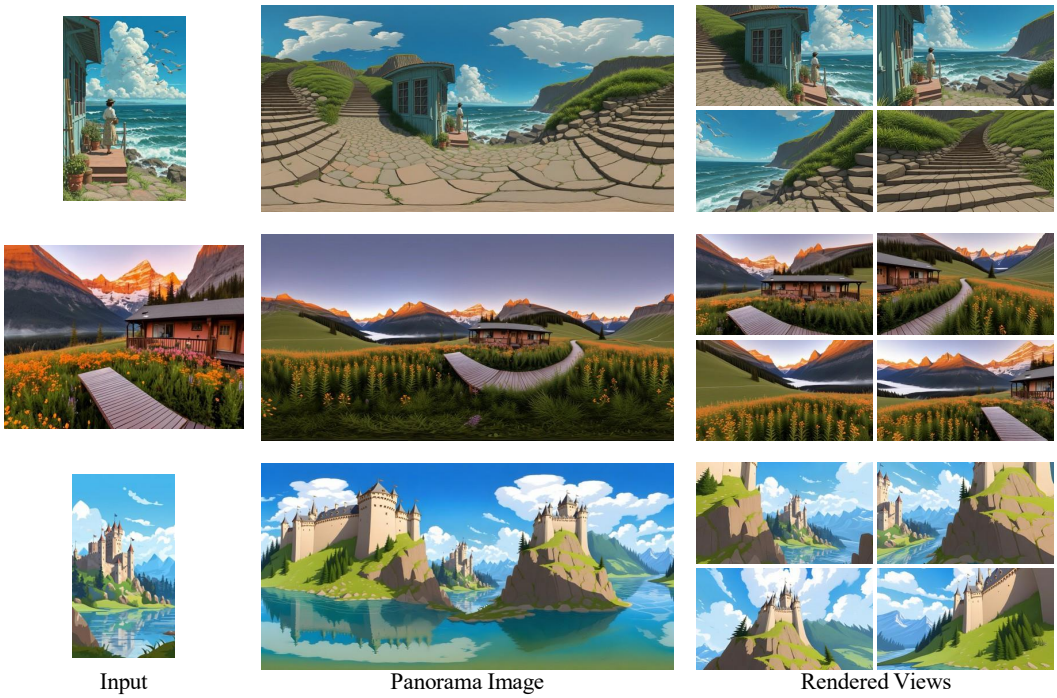


Figure 4: Visual results of image-to-panorama generation by HunyuanWorld 1.0.

	BRISQUE (↓)	NIQE (↓)	Q-Align (↑)	CLIP-I (↑)
Diffusion360 [7]	71.4	7.8	1.9	73.9
MVDiffusion [46]	47.7	7.0	2.7	80.8
HunyuanWorld 1.0 (Ours)	**45.2**	**5.8**	**4.3**	**85.1**

Table 1: Quantitative comparisons for image-to-panorama generation.

3 Model Evaluation

We show the generated panoramic images and explorable 3D worlds from image and text input in Fig. 4, 5, 10, 11. We can see that HunyuanWorld 1.0 generates high-quality panoramic images that precisely follow the conditions, while the generated 3D worlds maintain spatial coherence and enable immersive exploration across diverse camera trajectories, scene types, and artistic styles. For the rest of this section, we first conduct experiments to compare our results with those generated by the state-of-the-art methods on both panorama generation and 3D world generation. We then introduce a series of practical applications to demonstrate the versatility of HunyuanWorld 1.0 on virtual reality, physical simulation, game development, and interactive object manipulation scenarios.

3.1 Evaluation Protocol

Benchmark. For image-conditioned generation, we evaluate on both real and AI-generated images, curating from World Labs [58], Tanks and Temples [23], and images collected from real users. For text-conditioned generation, we curate a prompt benchmark by crowd-sourcing, covering different scene types, styles, and lengths.

Metrics. Due to the absence of ground-truth data for comparisons, we follow the setting of [41] to quantitatively assess the performance of our method from two key aspects: the alignment between input and output, alongside the visual quality. To evaluate the input-output alignment, we utilize the CLIP score [15] to measure the similarity of the generated worlds and the given prompts (CLIP-T) or images (CLIP-I). To evaluate the visual quality, we employ a few non-reference image quality assessment metrics, including BRISQUE [31], NIQE [32], and Q-Align [59].

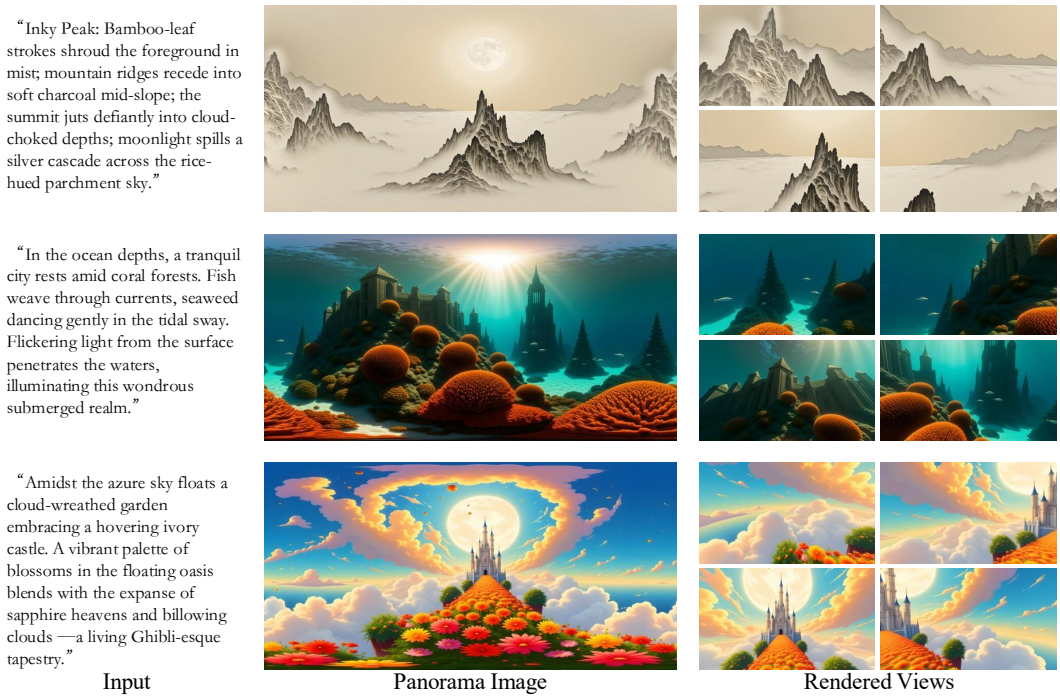


Figure 5: Visual results of text-to-panorama generation by HunyuanWorld 1.0.

	BRISQUE (↓)	NIQE (↓)	Q-Align (↑)	CLIP-T (↑)
Diffusion360 [7]	69.5	7.5	1.8	20.9
MVDiffusion [46]	47.9	7.1	2.4	21.5
PanFusion [70]	56.6	7.6	2.2	21.0
LayerPano3D [64]	49.6	6.5	3.7	21.5
HunyuanWorld 1.0 (Ours)	**40.8**	**5.8**	**4.4**	**24.3**

Table 2: Quantitative comparisons for text-to-panorama generation.

3.2 Panorama Generation

We compare and evaluate both image-to-panorama generation and text-to-panorama generation.

Image-to-Panorama Comparisons. *Settings.* We compare HunyuanWorld 1.0 with two state-of-the-art image-based panorama generation methods, Diffusion360 [7] and MVDiffusion [46]. We measure the visual quality of the generated panoramic images and the similarity between the CLIP image embeddings of novel images rendered from the generated panorama and the input image. Specifically, we render six views with 90° field of view (FOV), strategically positioned to provide complete 360° coverage. Each view is rendered at a resolution of 960×960.

Results. The quantitative results presented in Tab. 1 demonstrate that HunyuanWorld 1.0 consistently outperforms both baseline methods across all evaluation metrics. These findings highlight the effectiveness of our method in producing high-fidelity panoramic images while preserving strong semantic correspondence with the input. Qualitative comparisons illustrated in Fig. 6 and Fig. 7 corroborate these quantitative findings. In contrast to baseline approaches, which frequently exhibit discontinuous artifacts and geometric distortions, our method generates panoramic scenes with enhanced visual coherence and aesthetic quality.

Text-to-Panorama Comparisons. *Settings.* We compare HunyuanWorld 1.0 against four state-of-the-art text-conditioned panorama generation methods: Diffusion360 [7], MVDiffusion [46], PanFusion [70], and LayerPano3D [64]. We maintain a consistent rendering strategy with the image-to-panorama experimental settings with six rendered perspective views at 90° FOV and a resolution

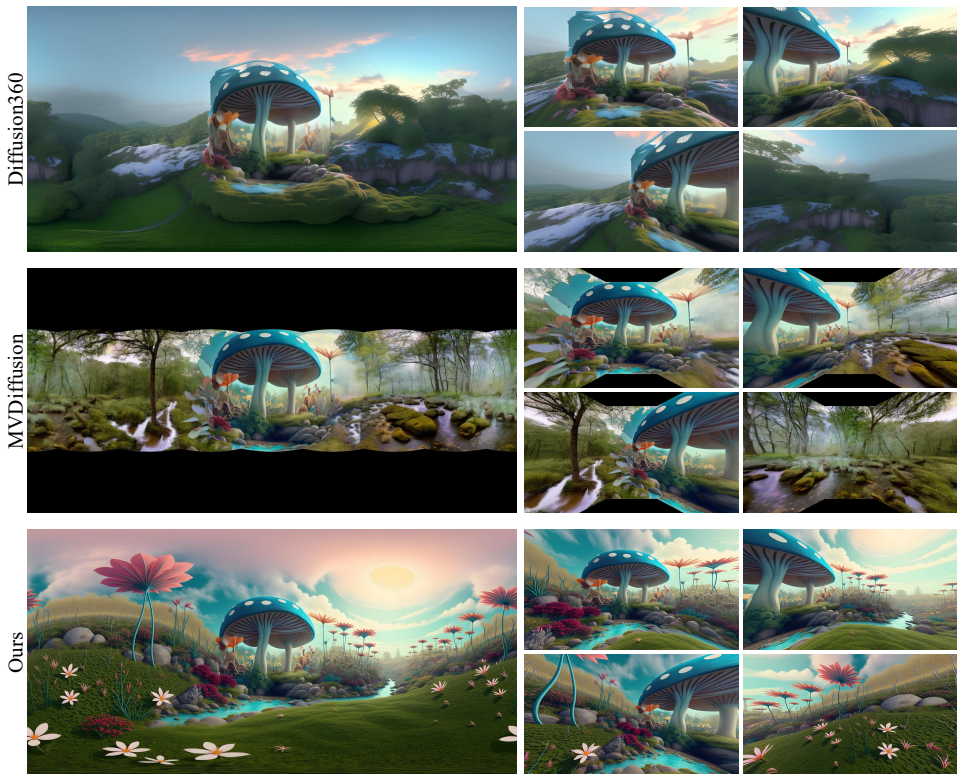


Figure 6: Qualitative comparisons for image-to-panorama generation (World Labs). Left: panoramic images generated from the same input image. Right: Four perspectively rendered views.

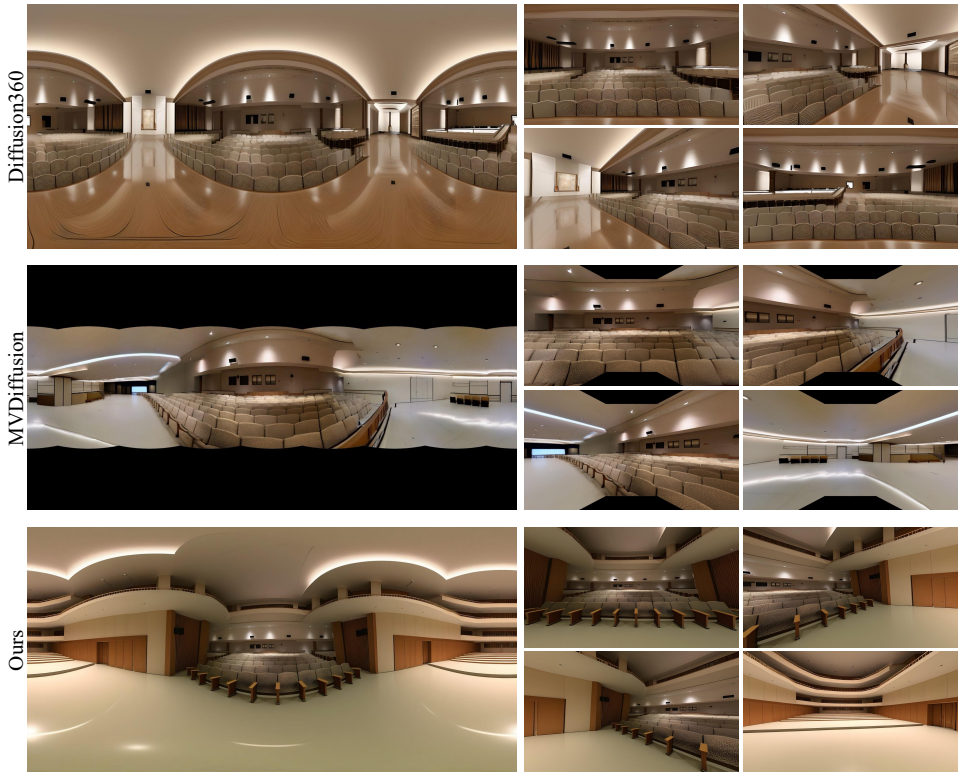
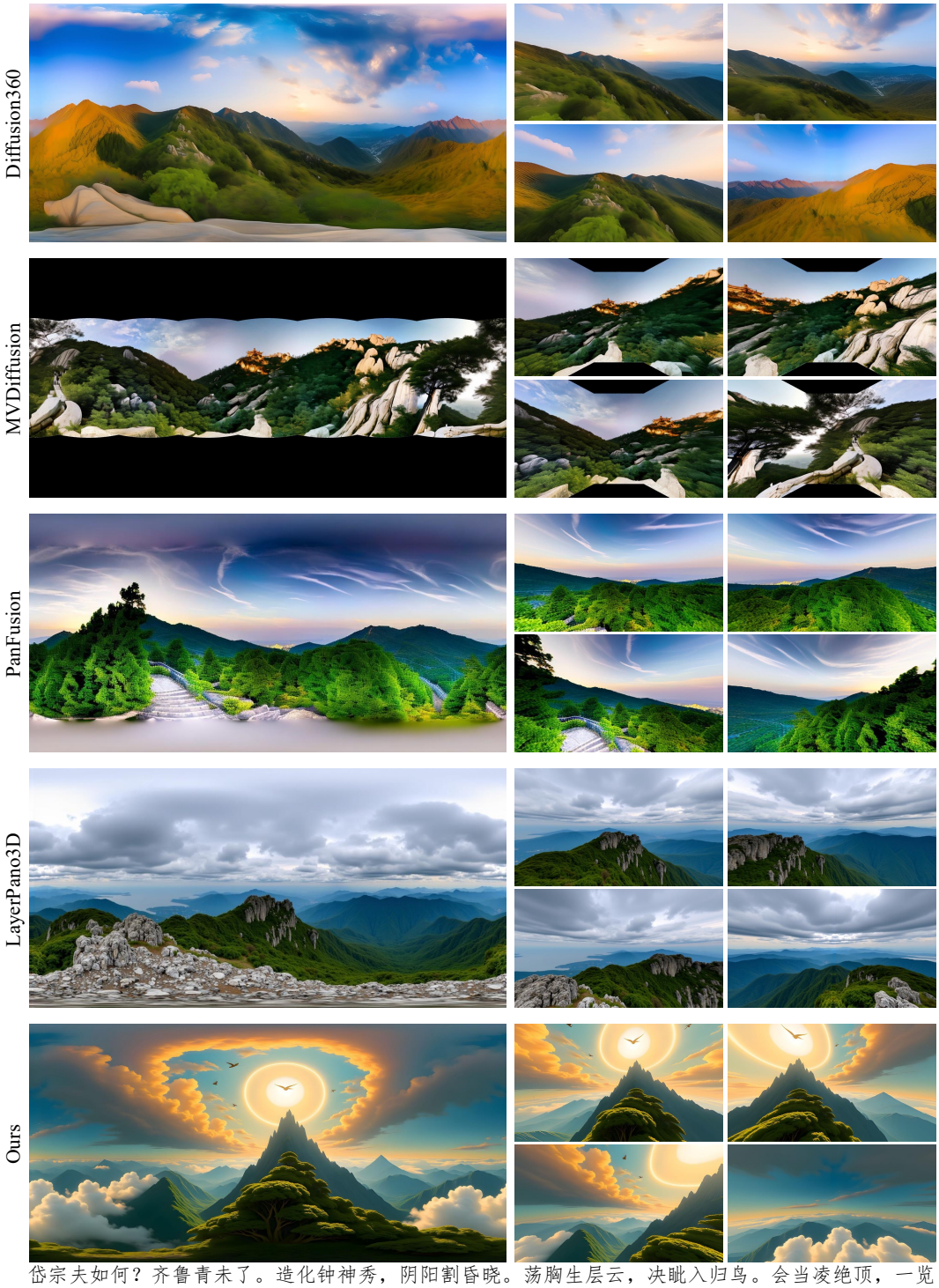


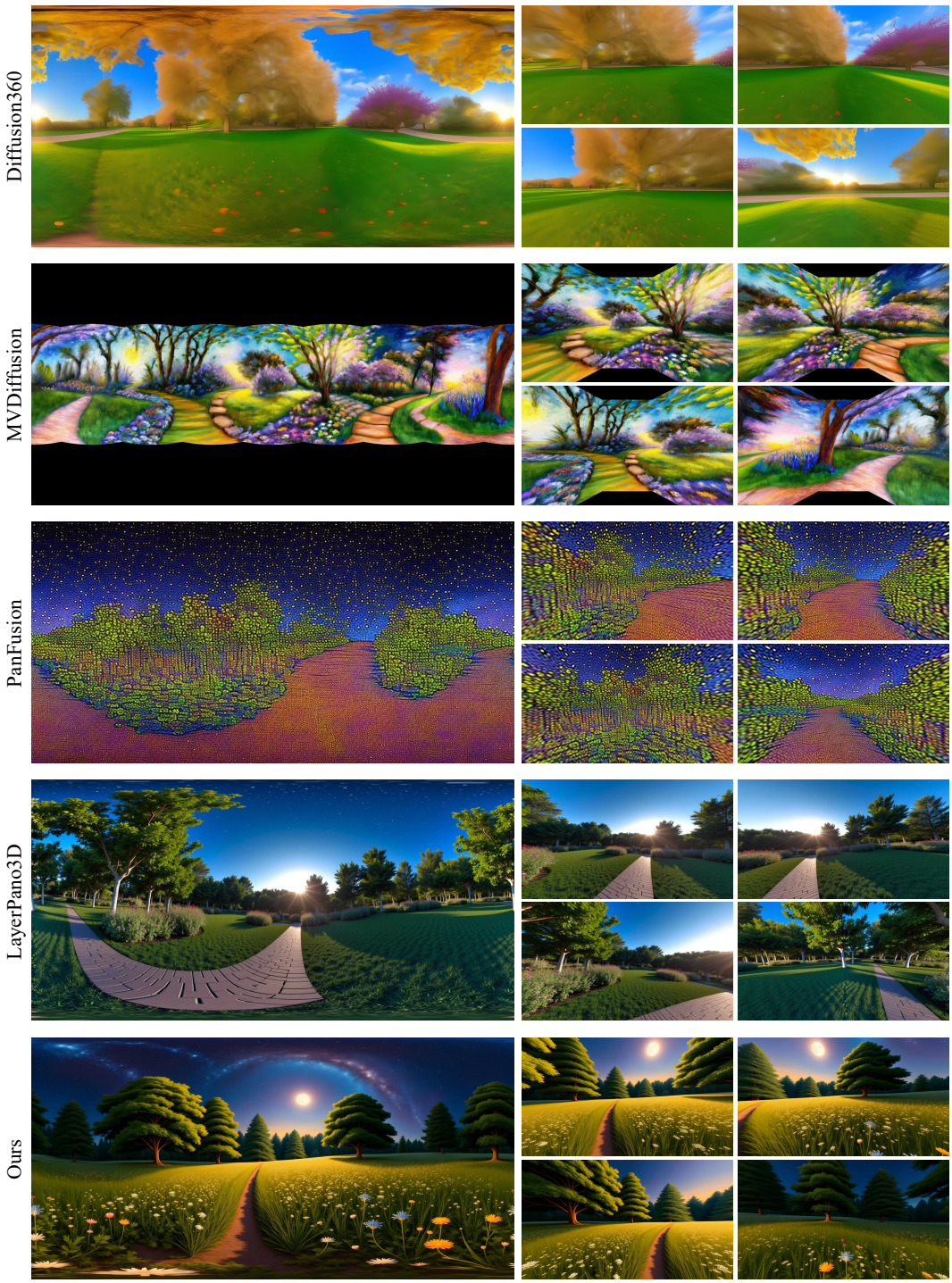
Figure 7: Qualitative comparisons for image-to-panorama generation (Tanks and Temples). Left: panoramic images generated from the same input image. Right: Four perspectively rendered views.



岱宗夫如何？齐鲁青未了。造化钟神秀，阴阳割昏晓。荡胸生层云，决眦入归鸟。会当凌绝顶，一览众山小。非写实风格。

Majestic Mount Tai—eternal green across lands. Creation's might carves dawn from dark. Clouds surge through soul; birds vanish in twilight's glow. Scale its summit: all peaks shrink to stones beneath. Non-representational style.

Figure 8: Qualitative comparisons for text-to-panorama generation (case 1). Left: panoramic images generated from the text at the bottom. Right: Four perspectively rendered views.



印象派风格的花园小径，在星空中清晰可见。

An Impressionist garden path, clear against the starry sky.

Figure 9: Qualitative comparisons for text-to-panorama generation (case 2). Left: panoramic images generated from the text at the bottom. Right: Four perspectively rendered views.

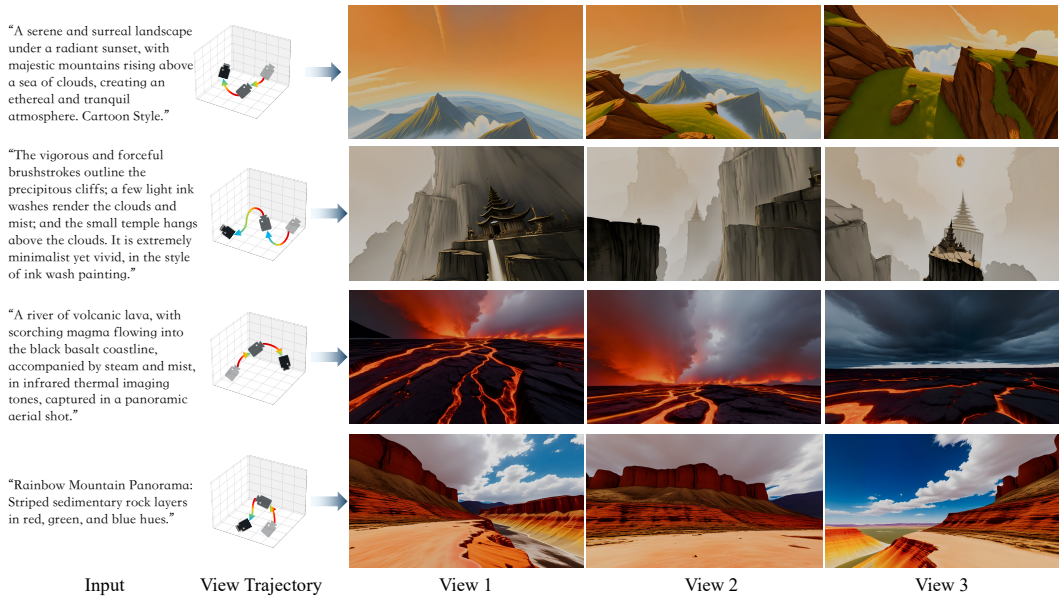


Figure 10: Visual results of text-to-world generation by HunyuanWorld 1.0.

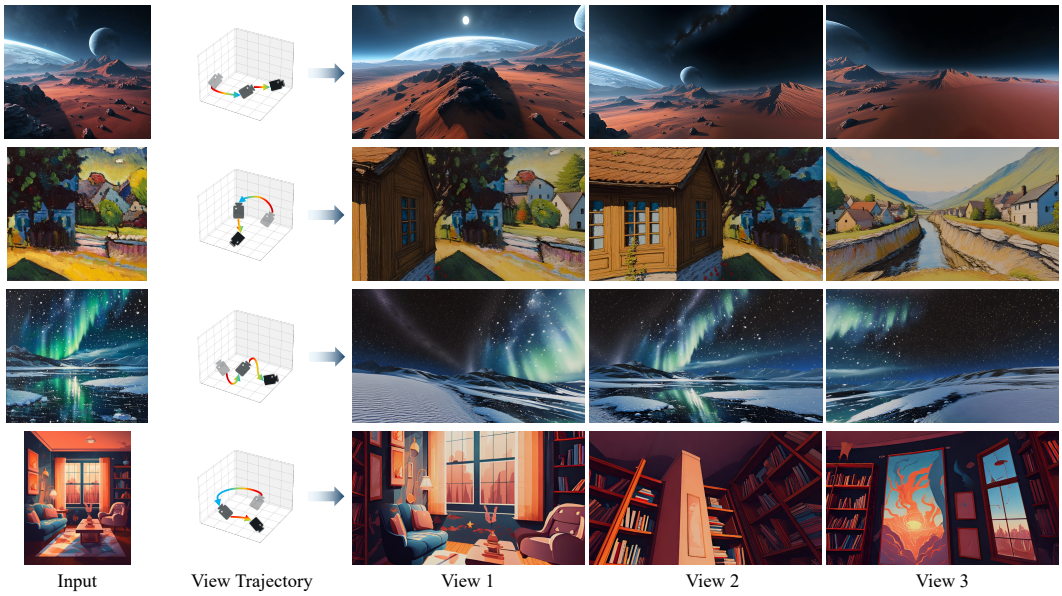


Figure 11: Visual results of image-to-world generation by HunyuanWorld 1.0.

of 960×960 for each generated panoramic image. We utilize CLIP-T scores to quantify the semantic alignment between the generated panoramic image and the corresponding textual input.

Results. The quantitative results in Tab. 2 demonstrate that HunyuanWorld 1.0 achieves superior performance across all evaluation metrics compared to the baseline methods. Qualitative results in Fig. 8 and Fig. 9 reveal that our approach exhibits exceptional fidelity to textual descriptions while maintaining high visual quality standards. Furthermore, HunyuanWorld 1.0 excels at generating panoramic scenes across diverse artistic styles and thematic contexts.

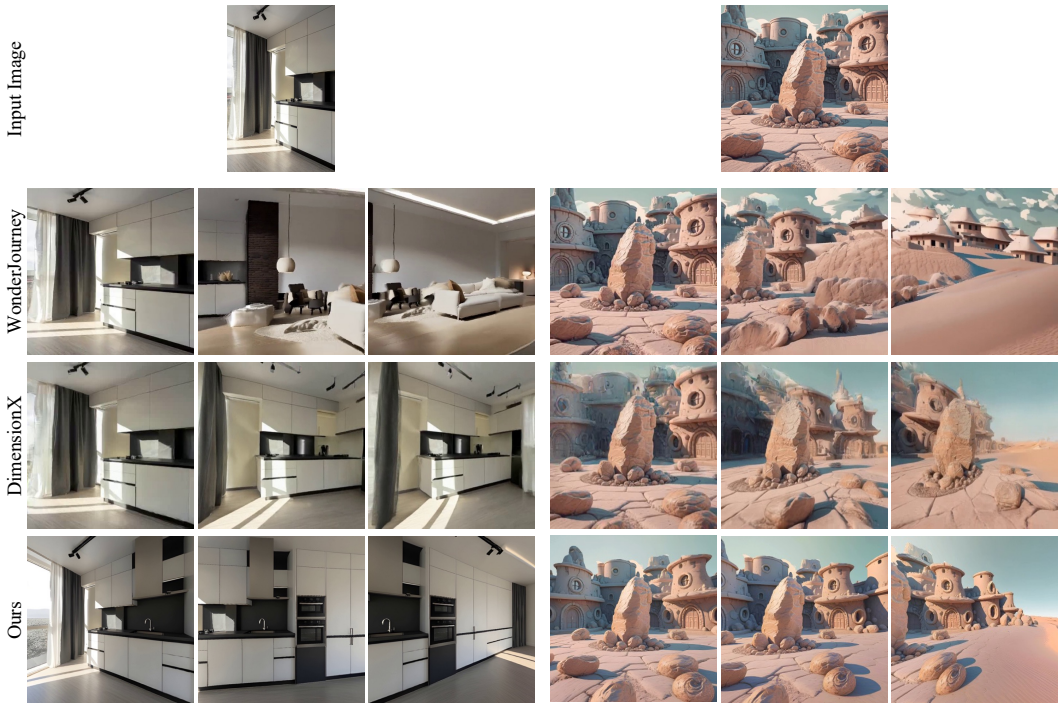


Figure 12: Qualitative comparisons for image-to-world generation. For each case, we render three perspective views from the generated 3D scenes.

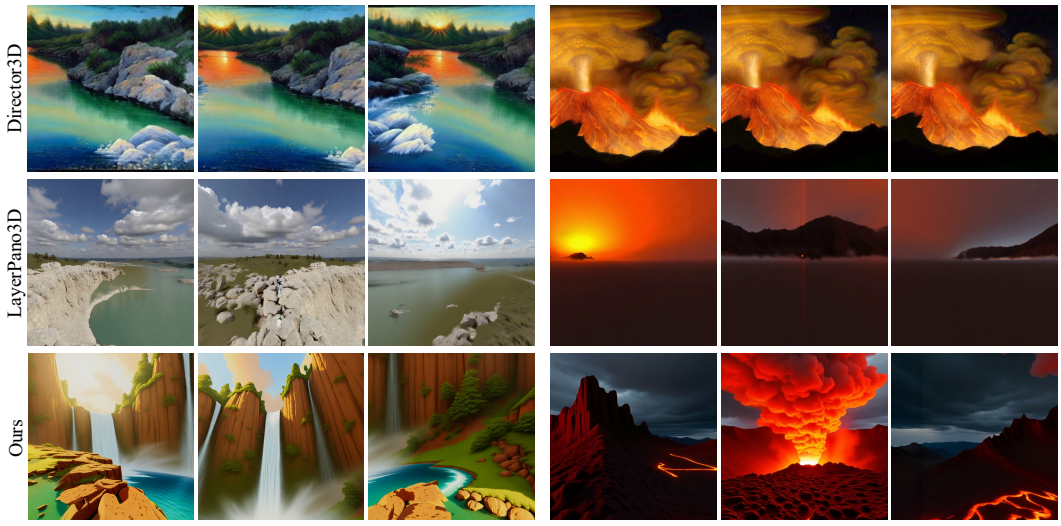


Figure 13: Qualitative comparisons for text-to-world generation. For each case, we render three perspective views from the generated 3D scenes.

	BRISQUE (↓)	NIQE (↓)	Q-Align (↑)	CLIP-I (↑)
WonderJourney [69]	51.8	7.3	3.2	81.5
DimensionX [43]	<u>45.2</u>	<u>6.3</u>	<u>3.5</u>	<u>83.3</u>
HunyuanWorld 1.0 (Ours)	**36.2**	**4.6**	**3.9**	**84.5**

Table 3: Quantitative comparisons for image-to-world generation.

	BRISQUE (↓)	NIQE (↓)	Q-Align (↑)	CLIP-T (↑)
Director3D [26]	49.8	7.5	2.7	<u>23.5</u>
LayerPano3D [64]	<u>35.3</u>	<u>4.8</u>	<u>3.9</u>	22.0
HunyuanWorld 1.0 (Ours)	**34.6**	**4.3**	**4.2**	**24.0**

Table 4: Quantitative comparisons for text-to-world generation.

3.3 3D World Generation

We compare and evaluate both image-to-world generation and text-to-world generation capabilities of HunyuanWorld 1.0.

Image-to-World Comparisons. *Settings.* We compare HunyuanWorld 1.0 with two state-of-the-art image-based 3D world generation methods: WonderJourney [69] and DimensionX [43]. We measure both the visual quality of the generated 3D worlds and the alignment between the rendered novel views and the input images. For HunyuanWorld 1.0, we generate and reconstruct the complete 3D scenes for rendering seven views with 90° FOV at azimuth angles of $\{0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ\}$. For DimensionX, we utilize its open-source *orbit* LoRA for video generation and 3D reconstruction, following its predefined 90° orbital camera trajectory. For WonderJourney, we employ a camera trajectory that also rotates right 90° for evaluation. All novel views are rendered at a resolution of 960×960.

Results. The quantitative results presented in Tab. 3 demonstrate that HunyuanWorld 1.0 consistently outperforms both baseline methods on the visual quality of rendered novel views and semantic alignment with the input image. Qualitative comparisons illustrated in Fig. 12 show our method generates 3D worlds with superior visual quality and geometric consistency compared to the baseline approaches while maintaining enhanced alignment with the input images.

Text-to-World Comparisons. *Settings.* We compare HunyuanWorld 1.0 against two state-of-the-art text-conditioned 3D world generation methods: LayerPano3D [64] and Director3D [26]. For a fair evaluation, we utilize a consistent evaluation protocol for HunyuanWorld 1.0 and LayerPano3D by rendering six views with 90° FOV at azimuth angles of $\{0^\circ, 60^\circ, 120^\circ, 180^\circ, 240^\circ, 300^\circ\}$ from the reconstructed 3D scenes. For Director3D, we utilize its model-predicted camera trajectories for novel view rendering, as its performance heavily depends on its self-predicted camera trajectories. We utilize CLIP-T scores to quantify semantic alignment between generated 3D content and its textual descriptions. All novel views are rendered at a resolution of 960×960.

Results. We present the quantitative results in Tab. 4, with qualitative comparisons shown in Fig. 13. HunyuanWorld 1.0 consistently outperforms both baseline methods across all evaluation metrics. The visual results demonstrate that our approach generates 3D worlds with high visual fidelity and strong semantic alignment with the input text descriptions. Notably, Director3D exhibits limitations in generating long-range camera trajectories for many test cases, which restricts its ability to generalize across diverse input conditions.

3.4 Applications

HunyuanWorld 1.0 enables a wide range of practical applications given its three key advantages: 360° immersive experiences, mesh export capability, and disentangled object modeling, as shown in Fig. 1.

Virtual Reality. Our panoramic world proxy enables the generation of fully immersive 360° environments optimized for virtual reality deployment across contemporary VR platforms, such as Apple Vision Pro and Meta Quest. The comprehensive spatial coverage eliminates visual artifacts and boundary discontinuities, providing seamless omnidirectional browsing capabilities.

Physical Simulation. The generated 3D worlds and separate 3D object representations support direct 3D mesh export, ensuring full compatibility with existing computer graphics pipelines for physical simulation. This enables seamless integration with physics engines for collision detection, rigid body dynamics, and fluid simulation.

Game Development. The generated 3D worlds span diverse scenes with various aesthetic styles, including extraterrestrial landscapes, medieval architectural ruins, historical monuments, and futuristic urban environments. These worlds are exported as standard 3D mesh formats, enabling seamless integration with industry-standard game engines such as Unity and Unreal Engine for game development applications.

Object Interaction. The disentangled object representations enable precise object-level manipulation and interaction within the generated 3D worlds. Users can perform precise 3D transformations, such as translation, rotation, and rescaling, on individual scene components while preserving the integrity of surrounding environmental elements.

4 Related Work

Immersive Scene Image Generation. The 360° panoramic image has emerged as a cornerstone for immersive virtual reality (VR) experiences. Recent advancements in latent diffusion models (LDMs) [40, 25] have demonstrated remarkable capabilities in image synthesis. Several studies including MVDiffusion [46], PanoDiff [52], and DiffPano [66] have successfully adapted diffusion models specifically for panorama generation. Subsequent efforts [60, 66, 70, 71] have focused on incorporating spatial priors, such as spherical consistency, depth to fine-tune pre-trained text-to-image (T2I) diffusion models on limited panoramic datasets. CubeDiff [21] begins with a single cubemap face and synthesizes the remaining faces to reconstruct the complete panorama. In our work, we present a scalable panorama generation model for both text-conditioned and image-conditioned generation with a dedicated data curation pipeline.

Video-Based World Generation. Recent advances in video diffusion models (*e.g.* Hunyuan-Video [24], CogVideo-X [65], and Wan-2.1 [51]) have significantly improved high-quality video generation. Building on these models' inherent world knowledge, many methods now incorporate 3D constraints (e.g., camera trajectories, 3D points) to produce 3D-consistent videos for dynamic scene generation [13, 1, 2, 19]. For instance, CameraControl [13] and Cosmos [1, 2] encode camera poses as Plücker coordinates, integrating them with latent embeddings to ensure camera-consistent video sequences. To achieve finer control, Streetscapes [6] employs multi-frame layout conditioning and autoregressive synthesis for long-range scene coherence. Meanwhile, Voyager [19] and Wu et al. [61] leverage explicit 3D scene points to enhance spatial consistency in extended video generation. In game development, methods like Genie [34] and Matrix [8] achieve interactive video generation with keyboard actions. Similarly, many works generate videos in [17, 56, 11, 55] from text prompts, BEV maps, bounding boxes, and driver actions to simulate autonomous driving scenarios. However, these video-based methods suffer from high rendering cost and long-range inconsistency due to the lack of 3D representation.

3D World Generation. Compared to videos, 3D scene assets offer superior compatibility with standard computer graphics pipelines and ensure stronger consistency. Existing 3D world generation methods can be categorized into procedural-based methods [33, 39, 72] and learning-based methods [64, 4, 69, 68, 10, 29, 43]. Procedural generation techniques automate the creation of 3D scenes by leveraging predefined rules or constraints. Among these, rule-based methods [33, 38] employ explicit algorithms to directly generate scene geometry, laying the foundation for visual rendering. In contrast, optimization-based generation [5, 39] frames scene synthesis as an optimization problem, using cost functions based on predefined constraints (*e.g.*, physics or design principles). A more recent paradigm, LLM-based generation (*e.g.*, LayoutGPT [9], SceneX [72]) enables users to define environments through natural language descriptions, offering greater flexibility and user control over scene design. Learning-based approaches focus on reconstructing 3D scenes from visual inputs. These methods typically fine-tune existing diffusion models using 3D-consistent data, enabling the generation of dense multi-view images from sparse known viewpoints. They then employ per-scene 3D/4D Gaussian Splatting optimization to reconstruct the full scene. Dimension-X [43] uses video diffusion models to train separate LoRA models, decoupling spatial and temporal dimensions for multi-view video generation. LucidDreamer [4, 69, 68] generates multi-view images via progressive

inpainting and multi-view diffusion models, respectively. Some methods [41, 64, 20] approach 3D scene reconstruction by first inpainting occluded regions of a panorama, then lifting 2D views to 3D Gaussian Splatting (3DGS) through scene-specific optimization. In contrast, our work targets generating layered mesh assets that can be directly integrated into existing computer graphics pipelines.

5 Conclusion

In this report, we introduced HunyuanWorld 1.0, a novel framework for generating immersive, explorable, and interactive 3D worlds from both text and image inputs. We leverage a semantically layered 3D mesh representation with a panoramic world proxy to create diverse and 3D-consistent worlds with disentangled objects for enhanced interactivity. Extensive experiments demonstrate that HunyuanWorld 1.0 achieves state-of-the-art performance for both text-based and image-based 3D world generation. The key features of our method—360° immersive experiences, mesh export capabilities, and disentangled object representations—enables a wide range of applications in virtual reality, physical simulation, and game development. We believe that HunyuanWorld 1.0 represents a significant step forward in world-level 3D content creation and will serve as a valuable baseline for future research in this exciting and rapidly evolving field.

Contributors

- **Project Sponsors:** Jie Jiang, Linus, Yuhong Liu, Di Wang, Tian Liu, Peng Chen
- **Project Leaders:** Chunchao Guo, Tengfei Wang
- **Core Contributors:** Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li
- **Contributors:**
  - **Engineering:** Sheng Zhang, Yihang Lian, Sicong Liu, Puhua Jiang, Xianghui Yang, Minghui Chen, Zhan Li, Wangchen Qin, Lei Wang, Yifu Sun, Lin Niu, Xiang Yuan, Xiaofeng Yang, Yingping He, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu
  - **Data:** Lifu Wang, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Chao Zhang, Yonghao Tan, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu
  - **Art Designer:** Yulin Tsai, Dongyuan Guo, Yixuan Tang, Xinyue Mao, Jiaao Yu, Junlin Yu

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[4] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023.

[5] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

[6] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[7] Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *arXiv preprint arXiv:2311.13141*, 2023.

[8] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024.

[9] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36:18225–18250, 2023.

[10] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

[11] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023.

[12] Google. Draco: A library for compressing and decompressing 3d geometric meshes and point clouds. `https://github.com/google/draco`, 2017.

[13] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025.

[14] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. `https://github.com/3DTopia/OpenLRM`, 2023.

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[16] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

[17] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[18] Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12528–12537, 2021.

[19] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025.

[20] Zilong Huang, Jun He, Junyan Ye, Lihan Jiang, Weijia Li, Yiping Chen, and Ting Han. Scene4u: Hierarchical layered 3d scene reconstruction from single panoramic image for your immerse exploration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26723–26733, 2025.

[21] Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. Cubediff: Repurposing diffusion-based image models for panorama generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] Beomyoung Kim, Chanyong Shin, Joonhyun Jeong, Hyungsik Jung, Se-Yun Lee, Sewhan Chun, Dong-Hyun Hwang, and Joonsang Yu. Zim: Zero-shot image matting for anything. *arXiv preprint arXiv:2411.00626*, 2024.

[23] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[24] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[25] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[26] Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3d: Real-world camera trajectory and 3d scene generation from text. *Advances in neural information processing systems*, 37:75125–75151, 2024.

[27] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024.

[28] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[29] Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024.

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[31] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[32] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[33] F Kenton Musgrave, Craig E Kolb, and Robert S Mace. The synthesis and rendering of eroded fractal terrains. *ACM Siggraph Computer Graphics*, 23(3):41–50, 1989.

[34] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024.

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[36] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1028–1039, 2025.

[37] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023.

[38] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, et al. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12630–12641, 2023.

[39] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, et al. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21783–21794, 2024.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[41] Katja Schwarz, Denys Rozumnyi, Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. A recipe for generating 3d worlds from a single image. *arXiv preprint arXiv:2503.16611*, 2025.

[42] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[43] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024.

[44] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[45] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22819–22829, October 2023.

[46] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023.

[47] Tencent Hunyuan3D Team. Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation, 2024.

[48] Tencent Hunyuan3D Team. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025.

[49] Tencent Hunyuan3D Team. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details, 2025.

[50] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

[51] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[52] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. *arXiv preprint arXiv:2308.14686*, 2023.

[53] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025.

[54] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023.

[55] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024.

[56] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024.

[57] Zhenwei Wang, Tengfei Wang, Zexin He, Gerhard Hancke, Ziwei Liu, and Rynson WH Lau. Phidias: A generative model for creating 3d content from text, image, and 3d conditions with reference-augmented diffusion. *arXiv preprint arXiv:2409.11406*, 2024.

[58] WorldLabs. Worldlabs blog, 2024. `https://www.worldlabs.ai/blog`, Last accessed on 2025-07-08.

[59] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

[60] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. *arXiv preprint arXiv:2307.03177*, 2023.

[61] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025.

[62] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

[63] Ziyang Xie. Worldgen: Generate any 3d scene in seconds. `https://github.com/ZiYang-xie/WorldGen`, 2025.

[64] Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Yixuan Li, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. *arXiv preprint arXiv:2408.13252*, 2024.

[65] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[66] Weicai Ye, Chenhao Ji, Zheng Chen, Junyao Gao, Xiaoshui Huang, Song-Hai Zhang, Wanli Ouyang, Tong He, Cairong Zhao, and Guofeng Zhang. Diffpano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion. *arXiv preprint arXiv:2410.24203*, 2024.

[67] Jonathan Young. xatlas: Mesh parameterization library. `https://github.com/jpcy/xatlas`, 2018.

[68] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5916–5926, 2025.

[69] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024.

[70] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6347–6357, 2024.

[71] Dian Zheng, Cheng Zhang, Xiao-Ming Wu, Cao Li, Chengfei Lv, Jian-Fang Hu, and Wei-Shi Zheng. Panorama generation from nfov image done right. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21610–21619, 2025.

[72] Mengqi Zhou, Jun Hou, Chuanchen Luo, Yuxi Wang, Zhaoxiang Zhang, and Junran Peng. Scenex: Procedural controllable large-scale scene generation via large-language models. *arXiv e-prints*, pages arXiv–2403, 2024.