HY-World 2.0: A Multi-Modal World Model for Reconstructing, Generating, and Simulating 3D Worlds

Tencent Hunyuan*

https://3d.hunyuan.tencent.com/sceneTo3D
https://github.com/Tencent-Hunyuan/HY-World-2.0

Abstract

We introduce **HY-World 2.0**, a multi-modal world model framework that advances our prior project HY-World 1.0. HY-World 2.0 accommodates diverse input modalities, including text prompts, single-view images, multi-view images, and videos, and produces 3D world representations. With text or single-view image inputs, the model performs *world generation*, synthesizing high-fidelity, navigable 3D Gaussian Splatting (3DGS) scenes. This is achieved through a four-stage method: a) **Panorama Generation** with HY-Pano 2.0, b) **Trajectory Planning** with WorldNav, c) **World Expansion** with WorldStereo 2.0, and d) **World Composition** with WorldMirror 2.0. Specifically, we introduce key innovations to enhance panorama fidelity, enable 3D scene understanding and planning, and upgrade WorldStereo, our keyframe-based view generation model with consistent memory. We also upgrade WorldMirror, a feed-forward model for universal 3D prediction, by refining model architecture and learning strategy, enabling *world reconstruction* from multi-view images or videos. Also, we introduce **WorldLens**, a high-performance 3DGS rendering platform featuring a flexible engine-agnostic architecture, automatic IBL lighting, efficient collision detection, and training-rendering co-design, enabling interactive exploration of 3D worlds with character support. Extensive experiments demonstrate that HY-World 2.0 achieves state-of-the-art performance on several benchmarks among open-source approaches, delivering results comparable to the closed-source model Marble. We release all model weights, code, and technical details to facilitate reproducibility and support further research on 3D world models.
Project Page: https://3d-models.hunyuan.tencent.com/world/

* HY-World team contributors are listed at the end of the report.

Contents

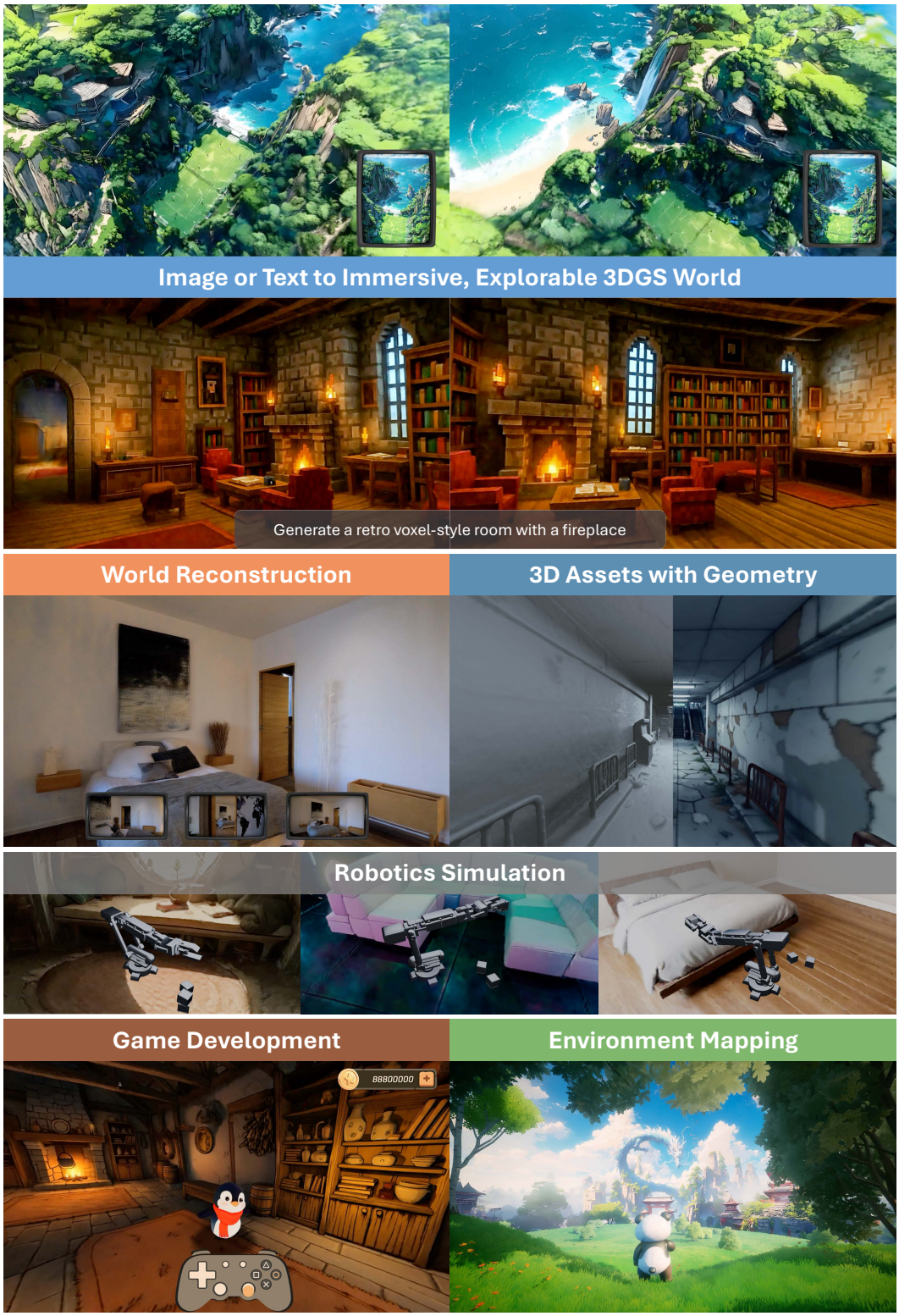


Figure 1: **Versatile applications of HY-World 2.0.** Our framework unifies *world generation* (synthesizing immersive, explorable 3D worlds from text or single-view images) and *world reconstruction* (recovering 3D representation from multi-view inputs). These capabilities empower diverse applications, including robotics simulation, game development, and environment mapping.

1 Introduction

"What Is Now Proved Was Once Only Imagined"
— William Blake

World models have rapidly evolved into a transformative paradigm for AI, enabling agents to simulate, understand, and interact with complex 3D environments [20, 14]. By capturing the physical and spatial dynamics of the real world, these models are unlocking unprecedented possibilities across diverse applications, including virtual reality [23], embodied robotics [27], and video games [19, 24].

Our previous explorations of generative world models involved two primary paradigms: (1) **HY-World 1.0** [23] established a robust foundation for *offline 3D-based world generation* [23, 72, 78, 39, 55, 44], explicitly modeling explorable 3D worlds with inherent 3D consistency, making them seamlessly compatible with standard computer graphics pipelines. (2) **HY-World 1.5** [24, 60, 70] advanced the frontier of *online video-based world generation* [19, 48, 24, 61], enabling real-time, interactive world modeling driven by user actions.

Despite these remarkable advancements, the current landscape of 3D world modeling remains largely bifurcated. Existing solutions typically specialize in either *generation* or *reconstruction*. Generative approaches excel at synthesizing impressive, explorable scenes from sparse inputs like texts or single-view images, but often struggle to maintain strict reconstruction accuracy [80, 61]. Conversely, reconstruction methods focus on recovering precise 3D structures (*e.g.*, depth, normals, and point clouds) from dense multi-view images or videos, yet they lack the generative priors necessary to hallucinate unseen regions [65, 69, 44, 40]. Furthermore, while recent closed-source pioneers [72] have demonstrated impressive capabilities in unifying these tasks, the open-source community still lacks a comprehensive, multi-modal foundational world model that bridges the gap between imaginative generation and accurate physical reconstruction.

To address these fundamental challenges, we introduce **HY-World 2.0**, the first open-source, systematic multi-modal world model that seamlessly unifies both "generation" and "reconstruction" within an *offline 3D world model* paradigm, as illustrated in Fig. 1. Designed to accommodate diverse input modalities—ranging from texts and single-view images to multi-view images and videos—HY-World 2.0 dynamically adapts its behavior based on the available conditions.

For sparse inputs (texts or single-view images), the model performs *world generation* to synthesize high-fidelity, navigable 3D Gaussian Splatting (3DGS) worlds. Formally, this generation capability is driven by a novel four-stage pipeline: panorama generation, trajectory planning, world expansion, and world composition. Crucially, although HY-World 2.0 is fundamentally designed as an *offline 3D world model*, it successfully bridges the gap between the geometric rigor of 3D representations and the rich, dynamic priors of video generation. By leveraging the powerful generative priors of *video diffusion models* during the expansion stage, HY-World 2.0 achieves significantly expanded exploratory spaces and superior visual quality compared to the previous HY-World 1.0.

For richer visual observations (multi-view images or videos), the framework performs *world reconstruction* to recover geometrically consistent and accurate 3D structures. Notably, rather than functioning as an isolated module, this *world reconstruction* capability also serves as a foundational component of *world generation*, powered by our upgraded feed-forward 3D reconstruction.

Beyond paradigm integration, we systematically push every component of HY-World 2.0 to its limits. First, we scale up *Panorama Generation* to **HY-Pano 2.0** in terms of both data and model capacity, enabling adaptive perspective-to-equirectangular (ERP) transformations from input images at arbitrary viewpoints. Next, a scene-parsing enhanced *Trajectory Planning* algorithm, called **WorldNav**, is introduced to produce camera trajectories for subsequent world expansion, considering both information maximization and obstacle avoidance. For *World Expansion*, we upgrade our previous controllable video model [62] to **WorldStereo 2.0**: 1) Rather than video generation, we perform generation within a keyframe space, thereby achieving superior visual fidelity. 2) We introduce a more consistent and robust memory mechanism. In the final stage of *World Composition*, we reconstruct the 3D environment using the upgraded **WorldMirror 2.0**: improved through generalized position encoding and enhanced training strategy. Unlike standard 3DGS learning for reconstruction [33], we incorporate tailored enhancements to strengthen 3DGS training on generated views, effectively bridging the gap between 3D reconstruction and generative world modeling.

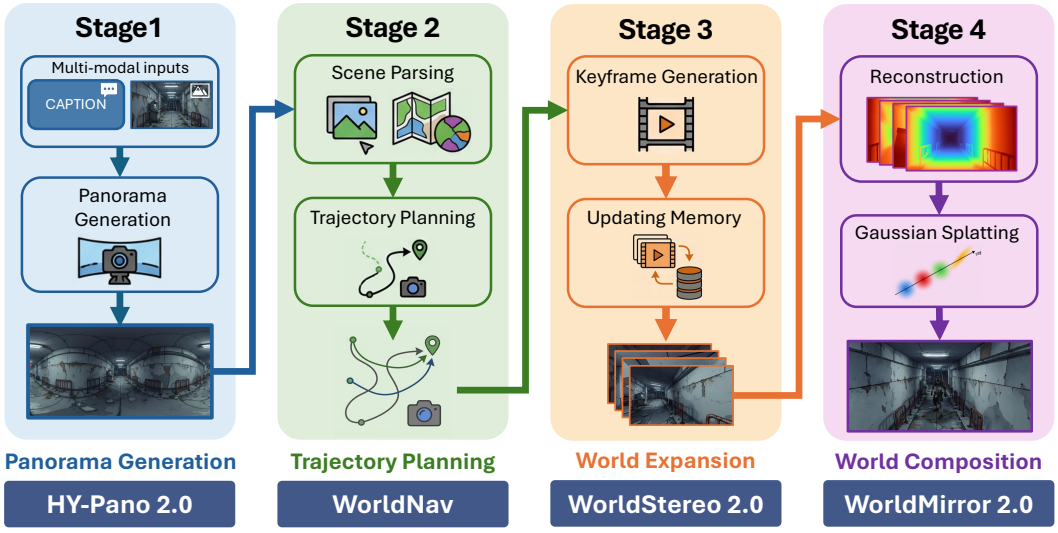


Figure 2: **Architecture of HY-World 2.0.** Our framework presents a four-stage process to transform multi-modal inputs into immersive 3D worlds: (1) initializing the world via *Panorama Generation*, (2) deriving exploration camera paths through *Trajectory Planning*, (3) expanding the world observations via memory-driven *World Expansion*, and (4) constructing the final 3DGS assets using *World Composition*. The core model/algorithm used in each stage is denoted at the bottom.

By unifying all aforementioned capabilities into a cohesive system, HY-World 2.0 achieves state-of-the-art performance in 3D-based world modeling. Extensive experiments demonstrate our model's superiority over existing open-source competitors and competitiveness with closed-source commercial products like Marble [72]. We release all models, codes, and technical details, aiming to democratize spatial intelligence and provide a robust, open-source foundation for the research on world models.

2 Overview

We show the overview of HY-World 2.0 in Fig. 2, which introduces the multi-modal world model as a four-stage pipeline, simulating the process of understanding, synthesizing, and reconstructing worlds. Specifically, the pipeline begins with **Panorama Generation** (Sec. 3), which translates arbitrary text or image inputs into a high-fidelity 360° world initialization. Subsequently, the elaborate **Trajectory Planning** (Sec. 4) is performed to parse and understand the initialized world, deriving optimal and information-rich observation paths. Following these planned routes, the generative **World Expansion** (Sec. 5) utilizes a memory-updating mechanism to ensure precise camera control and multi-view consistency across generated keyframes. Finally, **World Composition** (Sec. 7) is achieved by feeding these generated sequences into **WorldMirror 2.0** (Sec. 6) for robust 3D reconstruction, followed by tailored 3DGS optimization to yield immersive 3D worlds.

3 World Generation Stage I: Panorama Generation

A panorama captures a complete $360° \times 180°$ field-of-view (FoV) from a fixed viewpoint, offering a comprehensive and information-rich representation of entire scenes. Unlike standard perspective images that provide only a limited view of the physical world, 360° panoramas preserve global spatial contexts and intricate semantic relationships. Consequently, this holistic representation is increasingly recognized as a cornerstone for large-scale 3D world generation, providing the essential spatial consistency required for coherent viewpoint synthesis and immersive virtual exploration.

In this stage, we propose **HY-Pano 2.0**, which aims to synthesize high-fidelity panoramas from multi-modal conditions, including texts and single-view images. To achieve this, we optimize our generative pipeline across two orthogonal dimensions: (1) implementing an advanced data curation pipeline to overcome the inherent scarcity of panoramic data by curating high-resolution and diverse

samples; and (2) introducing a dedicated 360° generative model that implicitly learns the spatial mapping between perspective inputs and panoramic targets in a geometry-free manner, facilitating the synthesis of structurally coherent environments without requiring explicit camera metadata.

3.1 Data

To construct a robust foundation for high-fidelity panoramic synthesis, our data curation pipeline builds upon the established framework of HY-World 1.0 [23] while significantly scaling up the richness and diversity of the training data. Specifically, our upgraded dataset integrates two primary data sources: *(1) Real-world captures:* We incorporate a massive collection of high-resolution, real-world panoramas to instill the model with authentic lighting, complex textures, and natural structural priors. *(2) Synthetic assets:* To complement the real-world data, we utilize a large-scale set of synthetic environments rendered via high-end engines such as Unreal Engine (UE). These assets provide precise geometric labels and diverse, imaginative scene configurations that are otherwise difficult to obtain in the wild. To ensure data integrity, we implement a rigorous data filtering stage to eliminate low-quality samples, particularly those exhibiting noticeable stitching artifacts or exposed capturing equipment (*e.g.*, panoramic camera). This hybrid data strategy effectively broadens the semantic distribution of our dataset and mitigates the domain gap between synthetic and real-world distributions, enabling the model to generalize robustly across complex indoor and outdoor environments.

3.2 Model

To achieve high-fidelity panorama synthesis from perspective inputs, we move beyond conventional methods that rely on explicit geometric warping, a paradigm previously employed in HY-World 1.0 [23]. This traditional pipeline typically needs precise camera intrinsic parameters (*e.g.*, focal length and FoV) to perform spatial alignment between the perspective and equirectangular projection (ERP) domains. However, such metadata is frequently unavailable or inaccurate in real-world scenarios. This bottleneck inherently limits the flexibility of the HY-World 1.0 framework and often leads to noticeable projection distortion. To address this, we adopt an *implicit, adaptive mapping strategy* powered by a Multi-Modal Diffusion Transformer (MMDiT), as illustrated in Fig. 3. Instead of relying on explicit camera priors, we process both the conditional input and the panoramic target within a unified latent space. By concatenating the conditional image latent with the panoramic noise latent as a unified sequence of tokens, the MMDiT leverages its self-attention mechanism to autonomously learn the underlying perspective-to-ERP transformation. This purely data-driven approach allows the network to establish spatial correspondences directly within the feature space, enabling it to flexibly hallucinate missing environmental details and maintain global structural coherence, even with uncalibrated and diverse input images.

A common challenge in ERP generation is the discontinuity at the left and right edges. To eliminate these boundary artifacts, we introduce a combined refinement strategy comprising circular padding and pixel blending, as shown in the right of Fig. 3. At the latent level, we apply circular padding to the latent features, enforcing periodic boundary conditions during the denoising process. The padded latent is then decoded into the pixel space, where a linear pixel blending strategy is employed along the equirectangular edges. This combined harmonization effectively smooths the 360° wrap-around transition, ensuring a perfectly seamless and structurally coherent panoramic output.

4 World Generation Stage II: Trajectory Planning

Task Formulation. Following the synthesis of a high-fidelity panorama (Sec. 3), the subsequent objective is to derive exploration trajectories that maximize the coverage of navigable space. To bridge this with the upcoming world expansion stage (Sec. 5), we introduce **WorldNav**, a comprehensive trajectory planning strategy. WorldNav not only generates diverse camera paths to ensure extensive viewpoint coverage but also pairs them with precise textual instructions, thereby providing explicit guidance for the downstream generative process.

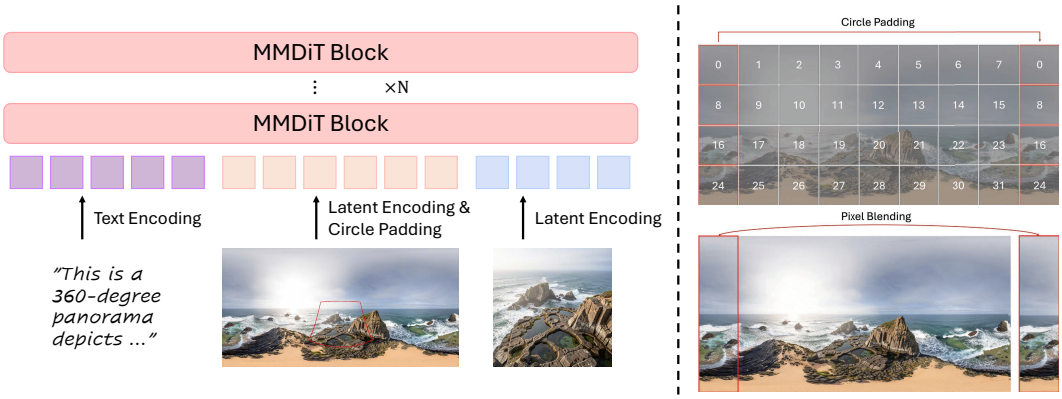


Figure 3: **Overview of the panorama generation architecture of HY-Pano 2.0.** The Left side shows the framework pipeline of panorama generation, while the right side illustrates the circle padding (latent space) and the pixel blending (pixel space) for seamless panorama generation.

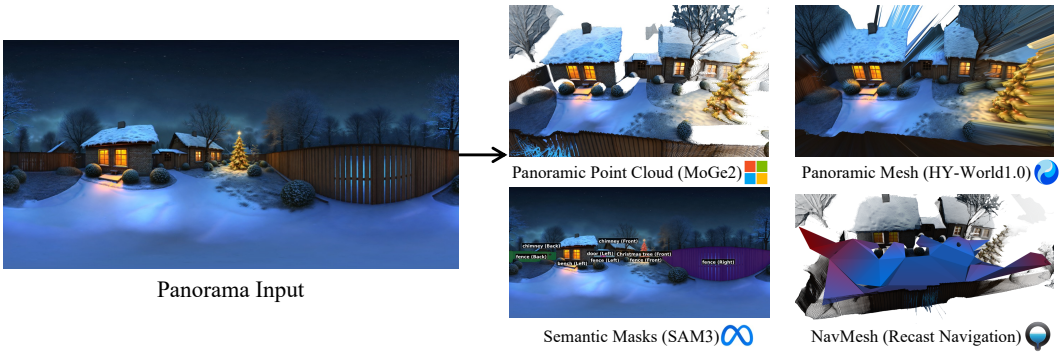


Figure 4: **The initial scene parsing for trajectory planning.** We obtain panoramic point clouds, meshes, semantic masks, and NavMesh via several pioneering works [67, 10, 23, 50].

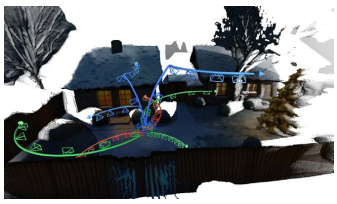
4.1 Geometric and Semantic Scene Parsing

Given the panoramas, we first employ scene parsing to obtain panoramic point clouds, meshes, semantic masks, and navigation meshes for the subsequent trajectory planning, as shown in Fig. 4.

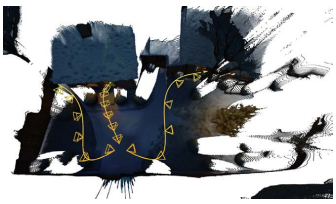
Geometry-Aware Initialization. We initialize the scene geometry by constructing a global panoramic point cloud, $\mathbf{P}^{pan}$. Leveraging the optimization framework from MoGe2 [67], we align monocular depth maps via the Least-Squares Minimal Residual (LSMR) across perspective views subdivided from the ERP space. Crucially, to enhance the geometric quality, we increase the sampling density from the default 12 views to 42, managing the computational overhead via a GPU-accelerated LSMR solver. Furthermore, we employ a hybrid filtering strategy, utilizing a vision-language grounding pipeline [43, 34] to mask unbounded sky regions, and then removing depth discontinuities (*i.e.*, edge floaters). This panoramic point cloud $\mathbf{P}^{pan}$ serves as the fundamental geometric representation across the subsequent trajectory planning, world expansion, and composition stages. Following HY-World 1.0 [23], we build the panoramic mesh at a lower resolution, which works for strict collision detection during trajectory planning.

Semantic Grounding and Navigability Analysis. To facilitate scene-aware camera control, we perform both semantic parsing and topological analysis of the panoramic scene. Specifically, we apply Qwen3-VL [76] to identify key spatial landmarks and obstacles within the panorama. Subsequently, SAM3 [10] is utilized to yield 2D semantic masks for these objects. We then localize their centroids into the 3D space as 3D masks, applying statistical filtering to eliminate background outliers.

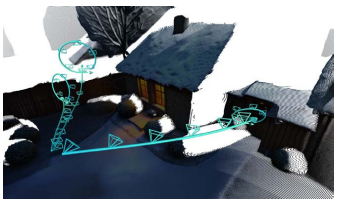
Simultaneously, we construct a Navigation Mesh (NavMesh) using Recast Navigation [50] to define the traversable regions for the camera agent. To ensure physically plausible camera movement, we



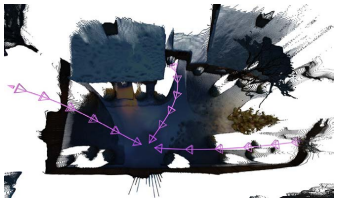
(a) Regular trajectories



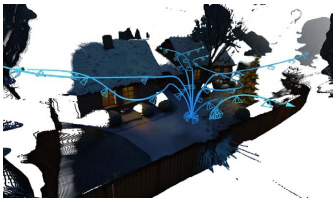
(b) Surrounding trajectories



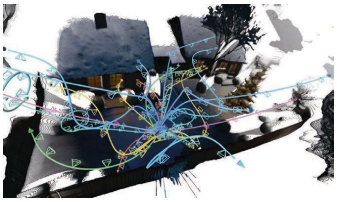
(c) Reconstruct-aware trajectories



(d) Wandering trajectories



(e) Aerial trajectories



(f) Full trajectories

Figure 5: **Illustration of five modes of trajectories planned in WorldNav.** Some trajectories are omitted for a simplified visualization.

Table 1: **Trajectory details of WorldNav.** The aerial category comprises both surrounding and wandering trajectories. Note that the maximum number for surrounding and reconstruct-aware trajectories is determined by the count of object segments detected within the panorama.

	Regular	Surrounding	Recon-Aware	Wandering	Aerial	Total
Max Number	9	5	10	3	8	35
Attached to Object	×	✓	✓	×	–	–
Iterative	×	×	✓	×	×	–

apply several geometric refinements to the raw NavMesh. First, we correct surface irregularities by snapping misaligned vertices to the physical ground via dense ray-casting. Second, we perform boundary erosion using a KD-Tree accelerated search to prevent the camera from moving too close to obstacles. Finally, we connect isolated navigable areas by detecting boundary nodes and synthesizing bridge polygons, thereby ensuring a continuous and fully navigable NavMesh.

4.2 WorldNav

Given the panoramic mesh, the NavMesh, and the 3D semantic landmarks, we design five heuristic trajectory modes for WorldNav. These trajectories start from the panorama's center and are designed to comprehensively cover diverse viewpoints while ensuring collision-free movement, as illustrated in Fig. 5.

Regular Trajectories. We employ regular trajectories to generally expand the visual coverage beyond the fixed origin of the panoramic space, as visualized in Fig. 5(a). First, we uniformly subdivide the panorama into three perspective views with a 120° FoV-x. For each view, we define an orbital target at the center point, positioned at the median depth of this view. The camera then orbits this target with a pitch angle of $+45^\circ$ and azimuth offsets of $\pm 120^\circ$. Specifically, we prioritize generating the pitch rotation before the azimuthal ones; this sequence ensures a global overview and facilitates consistent background generation. To further strengthen coverage with aerial perspectives, we apply an additional $+60^\circ$ azimuth rotation to the pitched orbits. Crucially, we utilize ray-casting to prevent the camera from clipping into the panoramic mesh. Trajectories that result in negligible movement due to collision detection are discarded.

Surrounding Trajectories. To facilitate the visual quality of foregrounds during the scene generation, we design surrounding trajectories that circle around the most significant objects, as shown in Fig. 5(b). The orbit radius is adaptively adjusted based on the object's 3D size: larger landmarks are observed from a greater distance to ensure the entire target fits within the FoV. To ensure collision-free

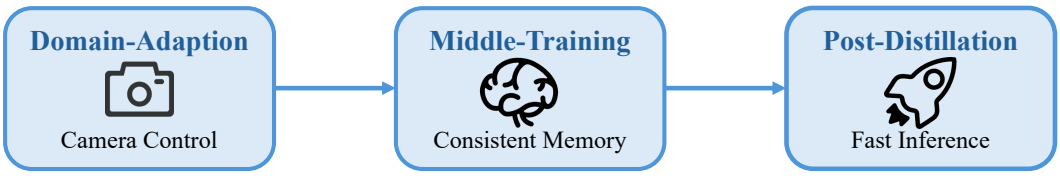


Figure 6: **Three training stages of WorldStereo 2.0**, progressively enabling camera control, memory-based consistency, and fast inference.

navigation, we uniformly sample 72 candidate nodes along the ideal circle and validate them via ray-casting against the NavMesh. Valid nodes are then connected to form a continuous arc using a bidirectional greedy search. To maintain a smooth path, we apply a tail pruning mechanism that removes the ends of the trajectory if they diverge significantly from the intended circular direction. Finally, we connect the start node to the nearest endpoint of the arc using the Dijkstra algorithm [13] on the NavMesh.

Reconstruct-Aware Trajectories. To mitigate the gaps for the subsequent 3D reconstruction, we introduce iterative reconstruction-aware trajectories that specifically target under-observed regions, as illustrated in Fig. 5(c). In the panoramic mesh, these missing areas typically manifest as stretched and sharp faces (refer to Fig. 4). We detect these regions by identifying mesh faces that exceed a heuristic aspect ratio threshold. To prioritize significant reconstruction targets, we employ Non-Maximum Suppression (NMS) to extract representative cluster centers of these degenerate faces and associate them with their nearest semantic landmarks, establishing them as key reconstruction nodes. Similar to surrounding trajectories, we generate candidate viewpoints around these nodes, selecting the endpoint that aligns its vertical viewing angle with the missing region. When multiple candidates exist, we prioritize the one offering the maximum visible range within the NavMesh. Moreover, to increase the ratio of novel views, we append an iterative orbiting trajectory: starting from the selected endpoint, the camera orbits the reconstruction node while maintaining a fixed gaze direction toward the target.

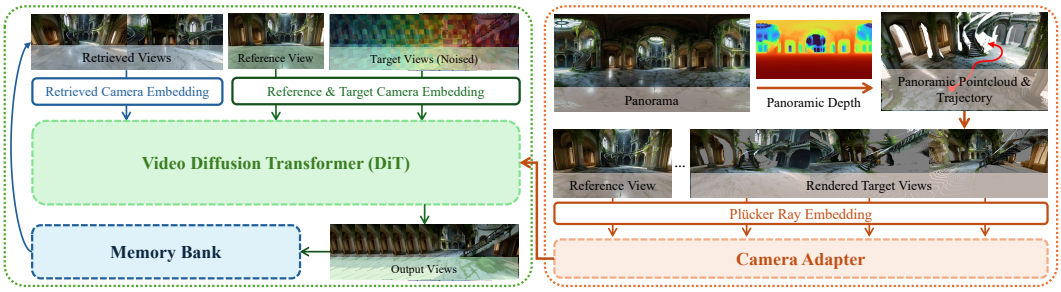
Wandering Trajectories. To maximize scene coverage and reach the environmental boundaries of the scene, we present wandering trajectories as shown in Fig. 5(d). These paths simulate the exploration of an autonomous agent, specifically targeting the farthest reachable points within the panoramic scene. This trajectory is particularly effective for extending visibility in narrow environments, such as streets and corridors. Formally, we partition the NavMesh into eight uniform angular sectors relative to the origin. Within each reachable sector, we utilize the Dijkstra distance field to identify and direct the camera toward the node farthest from the starting point.

Aerial Trajectories. Finally, we introduce auxiliary aerial trajectories to eliminate remaining blind viewpoints, as visualized in Fig. 5(e). Specifically, we augment the existing surrounding and wandering trajectories by applying a $+45^{\circ}$ upward pitch. To ensure geometric validity, this pitch angle is dynamically reduced when the camera view intersects the panoramic mesh, thereby preventing collisions.

5 World Generation Stage III: World Expansion

Task Formulation. Building upon the high-quality panoramas (Sec. 3) and broad-coverage camera trajectories (Sec. 4), we propose **WorldStereo 2.0**. As an upgrade to WorldStereo 1.0 [62], it leverages camera-guided video generation to synthesize extensive novel views for world expansion. As shown in Fig. 6, the training process consists of three stages, designed to enable camera control, memory-based consistency, and efficient inference, respectively.

Overview of WorldStereo 2.0. WorldStereo 2.0 bridges camera-conditioned Video Diffusion Models (VDMs) and 3D scene reconstruction by enabling *consistent multi-trajectory video generations with geometry-aware memories in the keyframe latent space*, as summarized in Tab. 2 and visualized in Fig. 7. Specifically, we first rethink the limitations of the standard Video-VAE in Sec. 5.1, whose spatio-temporal compression often leads to artifacts that degrade downstream reconstruction—and instead formulate WorldStereo 2.0 in a keyframe latent space with precise camera control to preserve



(a) Main DiT branch (b) Camera control branch

Figure 7: **Overall pipeline of WorldStereo 2.0.** (a) The main Video Diffusion Transformer (DiT) branch is enhanced by the retrieval-based improved Spatial-Stereo Memory (SSM++) for fine-grained consistency. (b) The camera control branch is guided by the panoramic point cloud, serving as Global-Geometric Memory (GGM) to confirm precise camera trajectory following and geometry-aware consistency. Here, we omit the VAE encoding/decoding for simplicity.

Table 2: **Different video generation schemes for 3D reconstruction.** Native video diffusion models (VDMs) need to produce long trajectories in a single pass to cover diverse viewpoints as much as possible. Autoregressive (AR) models sequentially generate long videos. WorldStereo 2.0 achieves multiple consistent generations based on a high-fidelity keyframe latent space with complementary viewpoints and memory mechanisms for subsequent reconstruction.

Paradigms	**Native VDM**	**AR**	**WorldStereo 2.0**
Receptive Field	Bidirectional	Autoregressive	Bidirectional
Trajectory Length/Num	Long/Single	Long/Single	Medium/Multiple
Latent Space	Video Clip	Video Clip	Keyframe Image
Frame Quality	😐	☹	🙂
Frame Redundancy	☹	☹	🙂
Precise Camera Control	🙂	😐	🙂
Consistency	😐	😐	🙂
Efficiency	☹	🙂	🙂

high-frequency appearance and geometric cues better. To further ensure coherent expansion across trajectories, it incorporates two complementary memory modules in Sec. 5.2: *Global-Geometric Memory* (GGM) that maintains globally consistent coarse scene structure, and *Spatial-Stereo Memory* (SSM) that reinforces local correspondence and fine-grained details. Together, these designs enable visually faithful and geometrically consistent world expansion suitable for subsequent 3D reconstruction. Finally, we introduce the acceleration of our model (Sec. 5.3).

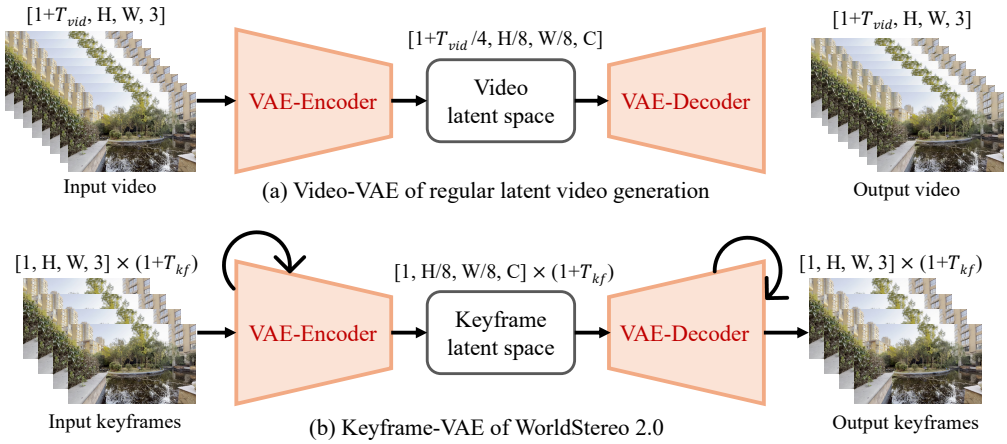
5.1 Domain-Adaption: Camera-Guided Keyframe Generation

During the phase of domain-adaption training, we tame the VDM into a camera-controlled keyframe generator to follow pre-defined camera trajectories. We first introduce the keyframe latent space to confirm high-fidelity generation, followed by the explicit camera control with a unified point cloud and camera ray guidance.

Keyframe-based Spatial Variational Autoencoder. Existing camera-guided VDMs often generate redundant frames when camera motion is slow or smooth, thereby failing to satisfy the requirements of *broad* and *diverse* viewpoints for reliable 3D reconstruction. We attribute this issue largely to a common design choice in latent-based VDMs [81, 64, 63]: videos are compressed by a spatio-temporal Video-VAE. In such spatio-temporally compressed latent spaces, fast camera motion tends to cause severe quality degradation in both generation and reconstruction, as shown in Fig. 8. Inspired by FlashWorld [39], we rethink the importance of preserving the latent fidelity and propose to perform scene generation in a *keyframe latent space* using Keyframe-VAE (see Fig. 9(b)). Formally, given keyframes $\{\mathbf{V}_i\}_{i=1}^{1+T_{kf}} \in \mathbb{R}^{1\times H\times W\times 3}$, we apply the causal-padding image encoder independently to each keyframe to obtain latent features $\{\mathbf{F}_i\}_{i=1}^{1+T_{kf}} \in \mathbb{R}^{1\times \frac{H}{8}\times \frac{W}{8}\times C}$ for training WorldStereo 2.0,

(a) Reconstructed frames from different VAE variants (b) Generated novel views from models with different VAE variants

Figure 8: **Reconstruction and novel-view generation with different VAE variants.** Keyframe-VAE preserves appearance consistency in reconstructions and substantially improves the fidelity of generated novel views, particularly under large viewpoint changes. Please zoom in for details.



(a) Video-VAE of regular latent video generation

(b) Keyframe-VAE of WorldStereo 2.0

Figure 9: **Keyframe-VAE in WorldStereo 2.0 versus a standard Video-VAE [64].** Unlike (a) Video-VAE, which performs spatio-temporal compression, (b) Keyframe-VAE applies *spatial-only* compression to better preserve high-frequency details and reduce artifacts essentially caused by Video-VAE encoding (*e.g.*, motion blur and geometric distortion). Specifically, Keyframe-VAE loops the causal padding-based *image encoding* over $(1 + T_{kf})$ times with a sparse frame set ($T_{kf} \ll T_{vid}$) that spans the same viewpoint changes by sampling at larger temporal intervals.

where H, W, C indicate frame height, width, and latent feature channel, respectively. Thanks to the high-fidelity image preservation of most open-released Video-VAEs [81, 37, 64, 63][1], we can directly inherit their model architectures by treating each keyframe as an image, *i.e.*, applying iterative spatial compression without temporal compression. A potential concern is that, for the same token length, keyframe latents contain fewer frames than standard video latents ($T_{kf} \ll T_{vid}$), which may reduce the viewpoint coverage available for camera control. Empirically, we increase the keyframe sampling interval to maintain the same viewpoint coverages, while Keyframe-VAE achieves superior fidelity with comparable camera controllability as verified in Fig. 8(b) and Tab. 7. Furthermore, we claimed that most discarded video frames are visually repetitive and thus largely *redundant* for the subsequent reconstruction stage. Additionally, the independent property of Keyframe-VAE enables good parallelizability, thereby largely strengthening both VAE encoding and decoding.

[1]Most Video-VAEs separately encode the first frame as image encoding via causal zero-padding to preserve high-fidelity information for image-to-video generation.

(a) Global point clouds of GGM training

(b) Retrieval of SSM++ training

Figure 10: **Data construction of WorldStereo 2.0 memory training.** (a) The global point clouds built for the GGM training with one reference view and $T_g = 2$ target views. (b) Trajectory retrieval strategies for SSM++ training tailored to dataset characteristics: temporally misaligned retrieval for existing multi-view data (top), and multi-trajectory retrieval for synthetic data (bottom).

Explicit Camera Control. Following [8, 62], WorldStereo 2.0 is built upon the pre-trained video DiT and integrated with a lightweight transformer-based camera adapter trained from scratch, as shown in Fig. 7(b). Formally, WorldStereo 2.0 incorporates both camera Plücker rays [58] and point clouds as complementary camera guidance to enable explicit and precise camera control for subsequent 3D reconstruction. In the domain-adaption, we only use the point cloud $\mathbf{P}^{ref} \in \mathbb{R}^{N\times 3}$ extracted from the reference view $\mathbf{I}^{ref} \in \mathbb{R}^{H\times W\times 3}$ ($N \leq HW$, after filtering floaters), instead of the panoramic point cloud. We warp it into each target view to obtain $\{\mathbf{P}_i^{tar}\}_{i=1}^{T_{kf}}$, indicated as:

$$\mathbf{P}_i^{tar}(x) \simeq \mathbf{R}_i^{c\to w}\mathrm{D}(x)\mathbf{K}_i^{-1}\hat{x}, \tag{1}$$

where $\mathbf{R}_i^{c\to w}$ and $\mathbf{K}_i$ denote the camera-to-world and intrinsic matrices of target view i; $\mathrm{D}(\cdot)$ is the monocular depth [67] estimated on the reference view at pixel x, and $\hat{x}$ is the homogeneous pixel coordinate. We then render the warped point clouds into view-wise keyframes [53] and encode them into latent features using the Keyframe-VAE. Compared with Uni3C, which trains only the control branch, we also fine-tune a subset of the Diffusion Transformer (DiT) backbone [51] to match the keyframe latent space better. Specifically, we freeze the cross-attention and feed-forward layers during the domain-adaption stage, which gives the best trade-off between performance and generalization in our ablations (see Tab. 7).

5.2 Middle-Training: Memory Mechanism

In the middle-training stage, we adapt the global-geometric and spatial-stereo memory mechanisms proposed in [62], tailoring them for panoramic scenarios and the keyframe-based VDM to ensure frame consistency across diverse trajectories.

5.2.1 Global-Geometric Memory

Global-Geometric Memory (**GGM**) renders extended point clouds into videos as global 3D priors to generate multiple consistent videos, as illustrated in Fig. 7(b). Particularly in panoramic scenes, GGM allows WorldStereo 2.0 to internalize 360° environmental structures, significantly improving geometric consistency. Although point clouds have been used for camera control in WorldStereo 2.0, they previously served merely as *soft camera guidance* rather than forcing the VDM to strictly adhere to these 3D representations [8]. While this behavior is beneficial for preserving the generalization of camera-guided VDMs against degradation caused by inferior monocular depth, it leads the model to ignore most geometric structures in the point clouds, even when the point clouds are perfectly reconstructed. To overcome this, we fine-tune the WorldStereo 2.0 using videos rendered by extended global point clouds $\mathbf{P}^{glo}$ beyond the reference points $\mathbf{P}^{ref} \in \mathbb{R}^{N\times 3}$ as:

$$\mathbf{P}^{glo} = [\mathbf{P}^{ref}, \hat{\mathbf{P}}] \in \mathbb{R}^{(N+\hat{N})\times 3}, \tag{2}$$

where $\hat{\mathbf{P}} \in \mathbb{R}^{\hat{N}\times 3}$ denotes the additional point clouds randomly sampled from T_g novel views, as shown in Fig. 10(a). Furthermore, to prevent overfitting to the point clouds from novel views during training, we employ robust augmentation strategies, as detailed in Sec. 5.2.3. For inference, we

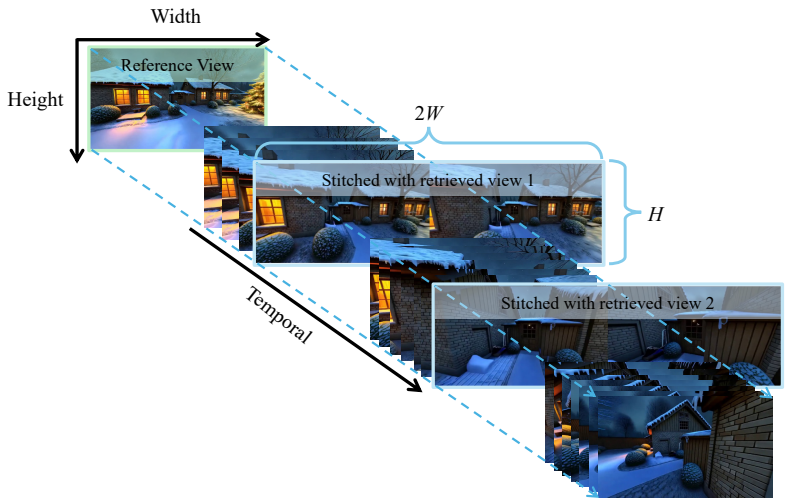


Figure 11: **Illustration of the RoPE [59] modification in SSM++.** Target frames are spatially concatenated with their corresponding retrieved reference views along the horizontal axis (resulting in width $2W$). Crucially, each retrieved view inherits the temporal index of its paired target frame before being fed into the main DiT branch.

define the panoramic point cloud $\mathbf{P}^{pan}$ from Sec. 4.1 as the global point cloud, which covers 360° viewpoints' information as effective geometric guidance.

5.2.2 Improved Spatial-Stereo Memory

While GGM maintains global structural coherence using point clouds, it often struggles to preserve fine-grained details and is prone to accumulating errors. Many previous studies [84, 87, 55, 38] retrieve historical reference frames and jointly model all frames via full-attention. However, we cannot guarantee the continuity of retrieved frames (*e.g.*, panoramic scenarios). These disparate, unordered reference views further hinder the VDM learning process. To overcome these issues, WorldStereo [62] draws inspiration from the traditional stereo matching [49] and the reference-based inpainting [6] and proposes the Spatial-Stereo Memory (**SSM**), which discretely retrieves reference views and spatially stitches each with its corresponding target view. By constraining the attention receptive field to each retrieval-target pair and utilizing pointmap guidance, SSM effectively recovers details by establishing correspondence within the stitched pairs.

In the WorldStereo 2.0, we advance this design with **SSM++**, retaining the core concept of horizontal retrieval stitching while introducing significant improvements. First, we discard the separate memory branch used in WorldStereo and instead directly incorporate retrieved keyframes into the main DiT branch (Fig. 7a). Second, as illustrated in Fig. 11, we modify the Rotary Positional Embedding (RoPE) [59] to accommodate this integration. Each target view is horizontally stitched with its retrieved counterpart, sharing the same temporal index. Unlike WorldStereo, which enforces a retrieval for every view, SSM++ selectively retrieves only the most relevant keyframes from the memory bank. This selective strategy significantly reduces redundant computation and memory overhead. Third, we transition from restricted attention to a full fine-tuning strategy. During the mid-training stage, we remove the constraints on attention receptive fields (except for cross-attention layers), enabling the model to learn global context across all target and retrieved features via full self-attention. Finally, to enhance flexibility, we replace the explicit pointmap guidance of WorldStereo with implicit camera embeddings. Formally, we normalize all input camera poses to a unified world coordinate and represent them as 7-dimensional vectors (quaternion and translation). These vectors are then encoded by a 3-layer MLP into camera tokens, which are added to the target and retrieved keyframe features via zero-initialization to provide geometry-aware perception.

Memory Bank and Retrieval Strategies. We adopt distinct retrieval strategies during the mid-training stage to accommodate varying data properties, as illustrated in Fig. 10(b). Practically, training SSM++ requires multi-view videos to construct target-retrieval pairs, which is difficult to obtain

in practice. To address this, we employ the *temporally misaligned retrieval* to existing multi-view data [41, 74, 68, 1]. Specifically, we randomly select frames from the retrieval trajectories with a specified temporal overlap (30% to 90%) relative to the target frames. Consequently, unlike simple interpolation, this strategy introduces several retrieved frames that lie outside the target trajectory, thereby increasing training difficulty and enhancing model robustness. Additionally, we construct a synthetic dataset using UE, featuring multiple trajectories for each asset. For this synthetic data, we employ *multi-trajectory retrieval*, which selects the most relevant frames from alternative trajectories based on 3D FoV similarity [62]. Furthermore, we apply data augmentation to the retrieved frames to further strengthen SSM++ generalization. Furthermore, we apply data augmentation to the retrieved frames to further strengthen SSM++ generalization, as detailed in Sec. 5.2.3.

During inference, perspective views subdivided from the input panorama serve as the initial memory bank. Subsequently, the memory bank is incrementally updated with generated keyframes, storing both RGB images and camera parameters for 3D-FoV retrieval. To reduce the computational overhead for each video clip's generation, we limit the retrieval to a maximum of T_r keyframes ($T_r < T_{kf}$).

5.2.3 Memory Augmentation

To mitigate the potential error accumulation stemming from imperfect point clouds and the retrieved generation, we employ comprehensive data augmentations during the middle-training stage to improve the robustness of memory components.

For GGM, we employ specific strategies to degrade the training depth, thereby simulating the inference imperfections: 1) We apply bilinear interpolation to downsample 50% of the depth maps, simulating the "depth bleeding" artifacts; 2) For 10% of the samples, we apply a small Gaussian filter to the depth maps to create artificial floaters [79]; 3) We retain the raw monocular aligned depth for 50% of the real-world dataset samples [41, 1, 74] without any post-filtering, preserving native floaters and noises. Notably, we empirically find that aggressive degradation strategies, like point cloud distortion used in [79], are not suitable for GGM. Such strong augmentations excessively weaken the geometric guidance provided by the point clouds, resulting in inconsistent geometry across multiple generated videos. For SSM++, we randomly perform the motion blur and color jitter on the retrieved frames. Moreover, we randomly crop the target and retrieved images to simulate varying visibility ranges and FoV overlaps in the inference scenarios.

5.3 Post-Train: Model Distillation

During the post-distillation, we apply the modified Distribution Matching Distillation (DMD) [83] to accelerate the inference of WorldStereo 2.0. DMD extends the idea of Variational Score Distillation (VSD) [71], distilling a few-step diffusion student G_θ through the approximate Kullback-Liebler (KL) divergence built from the difference between the frozen real score function s_{real} and the trainable fake score function s_{fake}. The updating gradient of DMD can be written as:

$$\nabla \mathcal{L}_{\text{DMD}} = -\mathbb{E}_t \left(\int \left(s_{\text{real}}(x_t, t) - s_{\text{fake}}(x_t, t) \right) \frac{dx_t}{d\theta} dz \right), \tag{3}$$

where $x = G_\theta(z)$ denotes the student generation given random Gaussian noise $z \sim \mathcal{N}(0; \mathbf{I})$ and $t \sim \mathcal{U}(0, 1)$, while $x_t \sim q_t(x_t|x, t)$ indicates the forward diffusion process.

The generator G_θ of WorldStereo 2.0 is distilled into a 4-step DiT. G_θ, s_{real}, s_{fake} are all initialized from the same VDM after the middle-training phase: s_{real} is frozen, while G_θ and s_{fake} are fully trainable. Following [83], we train s_{fake} 5 times per generator update. The stochastic gradient truncation [22] is employed to stabilize the training phase. We omit the GAN loss, as we found its impact to be insignificant while substantially slowing down training. Different from WorldStereo [62], which only distilled on the camera control task with a frozen memory branch due to a shortage of annotated memory data (specifically, the requirement for well-aligned depth for memory guidance). In contrast, benefiting from the flexible, explicit-guidance-free SSM++ and the abundance of high-quality UE rendering data, WorldStereo 2.0 enables full fine-tuning of the post-distillation within the memory-based training. Although this choice is slightly more costly, we find that it simultaneously enhances both camera control precision and memory capability.

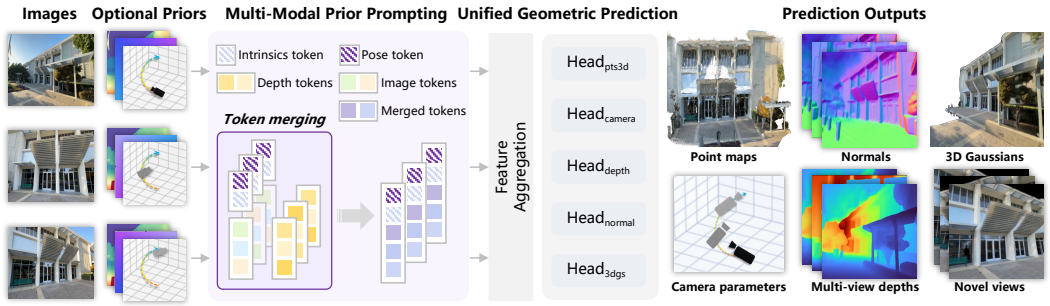


Figure 12: **Model architecture of WorldMirror 2.0,** which is a unified feed-forward model that takes multi-view images with optional geometric priors (camera poses, intrinsics, depth maps) as input, and simultaneously predicts dense point clouds, depth maps, surface normals, camera parameters, and 3DGS through a shared Transformer backbone with task-specific DPT decoder heads.

6 World Reconstruction: WorldMirror 2.0

Before detailing the final world composition stage (Sec. 7), we first introduce our upgraded feed-forward 3D reconstruction model, **WorldMirror 2.0**, which serves as the crucial bridge between 2D keyframe generation and 3D world composition. While *world generation* aims to synthesize explorable 3D worlds from sparse inputs (*e.g.*, single-view images or texts), *world reconstruction* focuses on recovering geometrically accurate 3D spatial relationships from dense 2D visual observations (*i.e.*, multi-view images or videos). In HY-World 2.0, we build this reconstruction capability upon WorldMirror [44], a unified feed-forward model for comprehensive 3D geometric prediction. We address three key limitations of WorldMirror 1.0: (1) degraded performance at non-training resolutions, (2) limited depth geometric consistency due to the lack of explicit depth–normal coupling, and (3) prohibitive memory and latency when scaling to large numbers of views. These are tackled through improvements in model architecture (Sec. 6.2), training data and supervision, and training strategy (Sec. 6.5), respectively. Consequently, WorldMirror 2.0 not only functions as a powerful standalone reconstruction foundation but also acts as the core geometry extractor for the generated views in our pipeline. Fig. 12 illustrates the overall model architecture, and Tab. 3 summarizes the key differences between WorldMirror 1.0 and WorldMirror 2.0.

6.1 Revisiting WorldMirror 1.0

WorldMirror [44] is a unified feed-forward model for comprehensive 3D geometric prediction (see Fig. 12). A core design is *Any-Modal Tokenization*, which encodes all input modalities, including images, camera poses, intrinsics, and depth maps, as tokens within a unified sequence. During training, each prior modality is independently dropped with probability 0.5, enabling flexible control over input modalities at inference time. These tokens are jointly processed by a Transformer backbone with global-local attention mechanisms and decoded by multiple DPT heads [65] to produce 3D point maps, multi-view depth maps, surface normals, camera parameters, and pixel-wise 3D Gaussian Splatting attributes in a single forward pass. To decouple geometry learning from appearance modeling, WorldMirror employs a two-phase curriculum: geometry heads (point map, depth, camera, normal) are jointly trained in the first phase, then all geometry parameters are frozen while only the 3D Gaussian head is trained in the second phase.

6.2 Model Improvements

As summarized in Tab. 3, we introduce three key model-level improvements in WorldMirror 2.0: normalized position encoding for flexible resolution inference, explicit normal-based supervision for depth via a depth-to-normal loss, and a dedicated depth mask prediction head for robust handling of invalid pixels. We further describe data improvements (Sec. 6.3), inference efficiency optimizations (Sec. 6.4), and training strategy improvements (Sec. 6.5) in subsequent subsections.

Table 3: **Comparison between WorldMirror 1.0 and WorldMirror 2.0.** Improvements are organized by model architecture (Sec. 6.2), training data, and training strategy (Sec. 6.5). The bottom section summarizes the resulting capability gains.

Component	WorldMirror 1.0	WorldMirror 2.0
Model Improvements (Sec. 6.2)		
Position Encoding	Absolute RoPE	Normalized RoPE
Depth Supervision	GT depth only	GT depth + GT/Pseudo normal
Invalid Pixel Modeling	Confidence only	Confidence + Depth mask head
Acceleration	None	Token/Frame SP + BF16 + FSDP
Data Improvements (Sec. 6.3)		
Data	Open-sourced	+ Internal UE renderings
Pseudo-Label Enhancement	✗	Normal pseudo-labels
Training Strategy (Sec. 6.5)		
Image Res./Num. Sampling	Independent	Token-budget dynamic
Curriculum Stages	2 stages	3 stages
Resolution Sampling	100K–250K pixels	50K–500K pixels
Resulting Capabilities		
Flexible Resolution Inference	☹	☺
Depth Geometric Consistency	☹	☺
Robust Invalid Pixel Handling	😐	☺
Training Efficiency	😐	☺
Inference Efficiency	☹	☺
Overall Reconstruction Quality	😐	☺

6.2.1 Normalized Position Encoding

Motivation. WorldMirror 1.0 adopts the standard RoPE [59] to inject 2D spatial awareness into multi-head self-attention. Each patch is assigned its absolute integer grid index $(i, j) \in \{0, \dots, H_p-1\} \times \{0, \dots, W_p-1\}$, which is used to compute position-dependent rotation angles. While effective at a fixed resolution, this scheme introduces a fundamental limitation for multi-resolution inference: when the test resolution exceeds the training resolution, a significant portion of patch indices fall outside the range observed during training (*i.e.*, position extrapolation), leading to degraded predictions. Conversely, when the test resolution is lower, the index space is under-utilized, causing a distribution shift in the attention pattern.

Method. Inspired by DINOv3 [57], we replace the absolute integer coordinates with normalized coordinates that map all patch positions into a fixed $[-1, 1]$ range regardless of the input resolution. Specifically, given an input image with a patch grid of size $H_p \times W_p$ (where $H_p = H/p$ and $W_p = W/p$ for patch size p), we compute the normalized coordinates for each patch (i, j) as:

$$\hat{x}_i = \frac{2i+1}{H_p} - 1, \quad \hat{y}_j = \frac{2j+1}{W_p} - 1, \tag{4}$$

where $\hat{x}_i, \hat{y}_j \in [-1, 1]$. The $+1$ offset in the numerator ensures pixel-center alignment, preventing boundary patches from collapsing onto ± 1. We normalize the height and width dimensions independently, which preserves aspect ratio information and generalizes better to non-square inputs. These normalized coordinates are then fed into the standard 2D RoPE computation to produce position-dependent rotations for each query and key token in attention.

Analysis. The key advantage of normalized position encoding lies in converting resolution extrapolation into interpolation. With standard RoPE, an 8-patch training grid occupies integer indices $[0, 7]$; at inference on a 16-patch grid, indices $[8, 15]$ are entirely out-of-distribution. Normalized RoPE maps both grids into $[-1, 1]$, so inference-time coordinates are simply a denser sampling of the same range. We verify this in Fig. 13: (a) normalized RoPE maintains consistently high cross-resolution cosine similarity (> 0.95), whereas standard RoPE degrades significantly; (b, c) the mean and standard deviation of encoding values remain stable under normalized RoPE, while standard RoPE exhibits systematic mean drift.

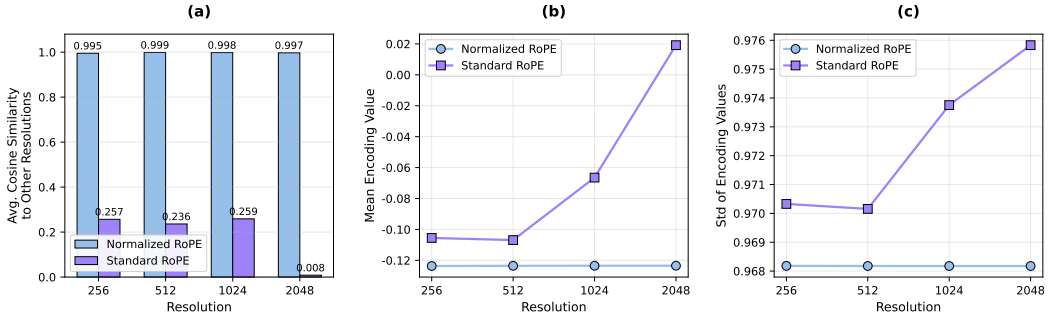


Figure 13: **Analysis of normalized position encoding.** **(a)** Average cosine similarity of center-point RoPE encodings to other resolutions. Normalized RoPE maintains high cross-resolution consistency (> 0.95), while standard RoPE degrades significantly. **(b)** and **(c)** show the mean and standard deviation of encoding values across resolutions, respectively. Normalized RoPE exhibits near-constant statistics, whereas standard RoPE shows systematic mean drift, confirming that normalization converts position extrapolation into interpolation.

6.2.2 Explicit Normal Supervision for Depth Estimation

Motivation. In WorldMirror 1.0, the depth and normal prediction heads are independently supervised without explicit geometric coupling between the two quantities. Moreover, real-world multi-view datasets often contain noisy or incomplete depth annotations, while monocular depth pseudo labels suffer from multi-view inconsistencies. These challenges motivate us to introduce an alternative supervision pathway that explicitly couples depth with normals.

Method. We introduce a depth-to-normal loss ($\mathcal{L}_{\text{d2n}}$) that converts predicted depth into surface normals via back-projection and cross products, and supervises the derived normals against normal targets. Specifically, given a predicted depth map $\hat{\mathbf{D}}_i$ and camera intrinsics $\mathbf{K}$, we compute the derived normal $\tilde{\mathbf{N}}_i$ as:

$$\tilde{\mathbf{N}}_i(x) = \text{normalize}\left(\frac{\partial \mathbf{P}_i}{\partial u} \times \frac{\partial \mathbf{P}_i}{\partial v}\right), \quad \mathbf{P}_i = \mathbf{K}^{-1}\hat{\mathbf{D}}_i \cdot [u, v, 1]^\top, \tag{5}$$

where partial derivatives are approximated by finite differences from four quadrant directions and robustly aggregated. The loss is defined as the angular error between the derived and target normals:

$$\mathcal{L}_{\text{d2n}} = \frac{1}{|\mathcal{V}|}\sum_{x\in\mathcal{V}} \arccos\left(\frac{\tilde{\mathbf{N}}_i(x)\cdot\hat{\mathbf{N}}_i(x)}{\|\tilde{\mathbf{N}}_i(x)\|\|\hat{\mathbf{N}}_i(x)\|}\right), \tag{6}$$

where $\hat{\mathbf{N}}_i$ denotes the normal supervision target and $\mathcal{V}$ is the set of valid pixels. The choice of normal target depends on the data source:

- **Synthetic datasets:** $\hat{\mathbf{N}}_i$ is obtained by applying the same depth-to-normal transform to the ground-truth depth, which provides clean and multi-view consistent supervision.
- **Real-world datasets:** $\hat{\mathbf{N}}_i$ is the pseudo normal predicted by a monocular normal estimation teacher model (see Sec. 6.3), which offers dense and reliable surface orientation supervision without multi-view inconsistency.

Through this mechanism, the depth head receives effective geometric supervision from normals on *all* datasets, even those lacking reliable depth ground truth.

6.2.3 Depth Mask Prediction

Real-world depth data frequently contains invalid pixels due to sensor noise, occlusion boundaries, transparent surfaces, and sky regions. WorldMirror 1.0 handles pixel reliability through learned confidence weights that modulate the training loss, but does not produce an explicit per-pixel validity prediction at inference time, forcing downstream applications to rely on heuristic thresholds. To

address this, we augment WorldMirror 2.0 with a dedicated depth mask prediction head that outputs a per-pixel validity logit $\hat{m}(x)$, trained with a binary cross-entropy loss:

$$\mathcal{L}_{\text{mask}} = -\frac{1}{|\mathcal{M}|} \sum_{x \in \mathcal{M}} \left[m^*(x) \log \sigma(\hat{m}(x)) + (1 - m^*(x)) \log(1 - \sigma(\hat{m}(x))) \right], \tag{7}$$

where $m^*(x) \in \{0, 1\}$ denotes the ground-truth validity label and $\mathcal{M}$ is the set of pixels with known validity. For synthetic datasets, ground-truth masks are derived from rendering pipelines where invalid regions are precisely known. For real-world datasets, we generate pseudo labels by identifying pixels with extreme depth values, large depth discontinuities, or sky regions. At inference time, the predicted mask enables downstream applications to selectively filter invalid pixels, improving the robustness of point cloud fusion and 3D reconstruction.

6.3 Data Improvements

We expand the training data of WorldMirror 2.0 with two key additions. First, we incorporate high-quality synthetic renderings from Unreal Engine scenes, which provide pixel-accurate ground-truth geometry in diverse indoor and outdoor environments. Second, we adopt a *normal-only* pseudo-label enhancement strategy for real-world datasets. A natural approach is to use monocular depth estimators to produce pseudo depth labels; however, we observe that independently predicted per-view depths introduce multi-view geometric inconsistencies (visible as point cloud layering artifacts). Surface normals, by contrast, describe local orientation without requiring global metric consistency, making them inherently more robust as pseudo labels in multi-view settings. We therefore employ a monocular normal estimation teacher model to predict dense normals per view and use them as pseudo supervision targets: directly for the normal head via an angular loss, and indirectly for the depth head through the depth-to-normal loss $\mathcal{L}_{\text{d2n}}$ (Sec. 6.2.2).

6.4 Inference Efficiency Improvements

WorldMirror 1.0 runs on a single GPU with FP32 weights, which limits the maximum number of views and resolution at inference time. WorldMirror 2.0 introduces three complementary acceleration strategies to enable scalable multi-GPU deployment. First, we adopt *sequence parallelism* at two granularities: token-level parallelism for the Transformer backbone, where the input token sequence is partitioned across GPUs and redistributed via All-to-All collectives at each attention layer, and *frame-level* parallelism for the DPT decoder heads, whose convolutional layers operate independently on per-view feature maps and do not benefit from token-level partitioning—per-view features are instead redistributed so that each GPU decodes a disjoint subset of complete frames. Second, following VGGT-X [65], we apply *selective mixed-precision inference* by casting most parameters to BF16 while keeping a small set of precision-critical modules in FP32, halving the memory footprint with negligible accuracy loss. Third, we employ *fully sharded data parallelism* (FSDP) to shard model parameters across GPUs, with each Transformer block and DPT head wrapped as an independent FSDP unit. These three strategies are complementary: sequence parallelism distributes computation and activation memory, mixed-precision reduces per-element cost, and FSDP shards weight memory. Together, they enable WorldMirror 2.0 to process substantially larger inputs while reducing per-GPU memory consumption and wall-clock time (Sec. 8.2).

6.5 Training Strategy Improvements

Token-based Dynamic Batch Sizing. WorldMirror 1.0 independently samples the per-image resolution and the number of views at each training iteration. Since GPU memory must accommodate the worst-case joint maximum (*i.e.*, highest resolution $\times$ maximum view count), this independent sampling strategy leads to substantial GPU memory under-utilization in practice, as most sampled configurations fall well below this ceiling.

We address this with a *token-budget-first* strategy. Concretely, we fix a maximum token budget T_{max} per GPU (*e.g.*, 25,000 tokens). At each iteration, we first sample the per-image resolution (pixel count from a configurable range, *e.g.*, 50K–500K) and aspect ratio, then compute the per-image token count $t = \frac{H}{p} \times \frac{W}{p}$. The maximum number of views is then derived as:

$$N_{\text{max}} = \min\left(N_{\text{cap}}, \left\lfloor \frac{T_{\text{max}}}{t} \right\rfloor \right), \tag{8}$$

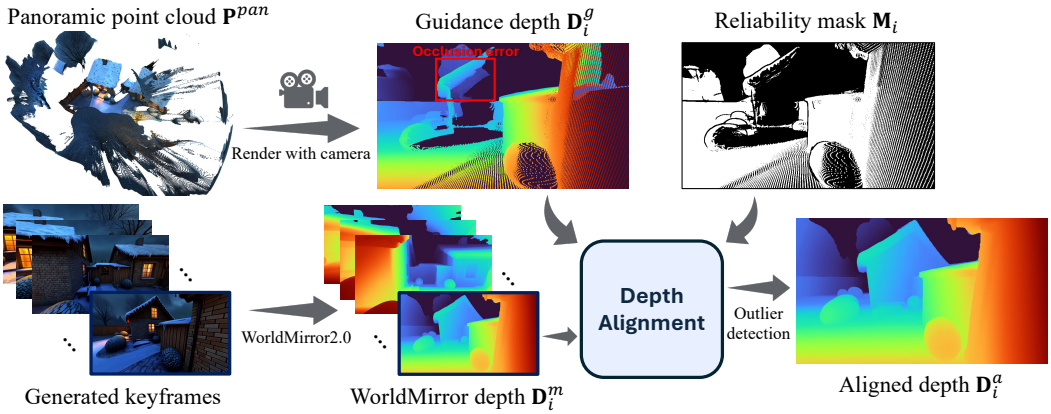


Figure 14: **The pipeline of depth alignment.** We align the WorldMirror depth estimated from generated keyframes with the panoramic point cloud. The point cloud is rendered as geometric guidance, identifying reliable regions to supervise the linear alignment. The outlier detection process effectively eliminated and corrected the alignment coefficients based on global statistics.

where N_{cap} is the architectural view-count cap (*e.g.*, 48). The actual view count is uniformly sampled from $[N_{\text{min}}, N_{\text{max}}]$. When the sampled view count is smaller than N_{max}, multiple samples are packed into the same GPU to fill the token budget, ensuring the tightly bounded token count for each GPU:

$$T_{\text{total}} = N \times \frac{H}{p} \times \frac{W}{p} \leq T_{\text{max}}, \tag{9}$$

where N is the total number of images on one GPU, including multiple samples. This design consistently achieves near-full GPU memory utilization regardless of the sampled resolution, exposes the model to more diverse resolution–view-count combinations during training, and eliminates out-of-memory errors without conservative memory provisioning.

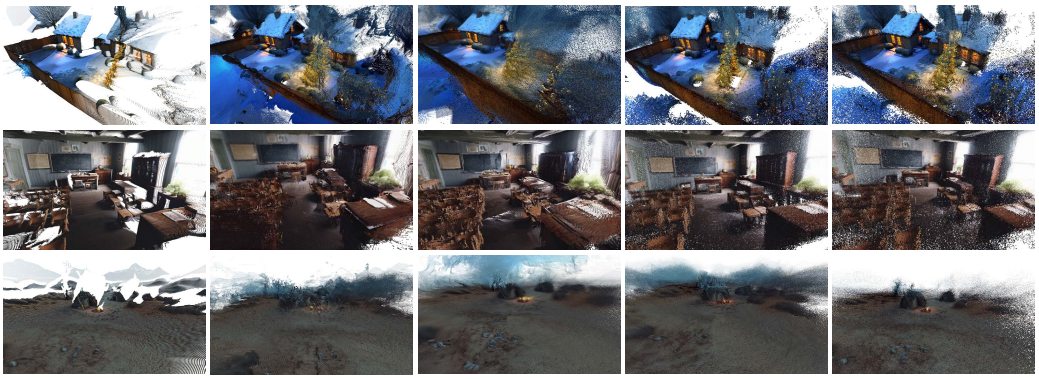
Multi-Stage Curriculum Learning. WorldMirror 1.0 employs a two-phase curriculum: geometry heads are jointly trained first, then all geometry parameters are frozen while only the Gaussian head is trained. In WorldMirror 2.0, we further decompose the geometry training into two sub-stages, yielding a three-stage pipeline: **Stage 1** trains all geometry heads using native annotations without pseudo-label enhancement or the depth-to-normal loss; **Stage 2** introduces the depth-to-normal loss (Sec. 6.2.2), while significantly increasing the proportion of synthetic data to improve geometric precision; **Stage 3** freezes the backbone and all geometry heads, training only the 3DGS head initialized from the depth head weights.

7 World Generation Stage IV: World Composition

Task Formulation. We define the input for this stage as a tuple containing the initial panorama $\mathbf{I}^{pan}$ (Sec. 3), its corresponding panoramic point cloud $\mathbf{P}^{pan}$ (Sec. 4.1), and the whole set of T_{ex} novel keyframes $\{\mathbf{V}_i\}_{i=1}^{T_{ex}}$ generated from WordExpand (Sec. 5) based on pre-defined trajectories $\{\mathbf{C}_i\}_{i=1}^{T_{ex}}$ (Sec. 4). The goal of *World Composition* is to integrate these inputs into a unified, navigable 3D representation. This process consists of two sequential steps:

1) Point cloud expansion (Sec. 7.1): constructing a globally aligned point cloud $\tilde{\mathbf{P}}$ by expanding $\mathbf{P}^{pan}$ with generated keyframes.

2) 3D scene optimization (Sec. 7.2): training a 3DGS, initialized with the expanded point cloud $\tilde{\mathbf{P}}$, to synthesize the complete high-fidelity 3D world.



(a) Panorama (MoGe2) (b) Mapanything (c) DA3 (d) WorldMirror2.0 (e) WorldMirror2.0 (aligned)

Figure 15: **Comparison of reconstructed point clouds.** (a) Reference panoramic point cloud estimated by MoGe2 [67]. (b-d) Results from state-of-the-art feedforward reconstruction methods: Mapanything [32], DepthAnything3 (DA3) [40], and WorldMirror 2.0. (e) WorldMirror 2.0 with the proposed depth alignment. Note that for all feedforward methods (b-d), the 10% of points with the lowest confidence are filtered out, and sky points are manually cropped for better visualization.

7.1 Point Cloud Expansion

7.1.1 Reconstruction via WorldMirror 2.0

We employ the state-of-the-art feedforward reconstruction model, WorldMirror 2.0 (a core component of HY-World 2.0, as detailed in Sec. 6), to reconstruct globally aligned point clouds and depth maps for point cloud expansion, as illustrated in Fig. 14. Specifically, we first downsample a subset of T'_{ex} frames from the fully generated sequence of T_{ex} frames. Subsequently, WorldMirror 2.0 is applied to estimate the per-frame depth and normal maps for this subset, conditioned on their respective camera poses as geometric priors:

$$\{\mathbf{D}_i^m, \mathbf{N}_i^m\}_{i=1}^{T'_{ex}} = \Phi\left(\{\mathbf{V}_j, \mathbf{C}_j\}_{j=1}^{T_{pan}}, \{\mathbf{V}_i, \mathbf{C}_i\}_{i=1}^{T'_{ex}}\right), \tag{10}$$

where $\Phi(\cdot)$ denotes the WorldMirror 2.0 network; $\{\mathbf{V}_j, \mathbf{C}_j\}_{j=1}^{T_{pan}}$ represents the perspective views and their corresponding camera parameters subdivided from the initial panorama $\mathbf{I}^{pan}$. Although WorldMirror 2.0 is not explicitly tailored for panoramic reconstruction, it performs well when combined with our generated keyframe sequences. Furthermore, we empirically demonstrate that WorldMirror 2.0 benefits significantly from camera conditions, outperforming other state-of-the-art feedforward reconstruction methods [32, 40] under identical conditional settings as verified in Fig. 15.

7.1.2 Depth Alignment

Although WorldMirror 2.0 generates high-quality depth maps $\{\mathbf{D}_i^m\}_{i=1}^{T'_{ex}}$, they suffer from scale ambiguity and fail to align with the world coordinate of the panoramic point cloud $\mathbf{P}^{pan}$. Moreover, while WorldMirror outperforms other feed-forward reconstruction methods under camera conditions, it still struggles in highly challenging outdoor scenes, as illustrated in the third row of Fig. 15. Therefore, we propose a robust alignment strategy to rectify WorldMiror depth $\mathbf{D}_i^m$ into an aligned depth map $\mathbf{D}_i^a$ using the panoramic point cloud $\mathbf{P}^{pan}$ as the geometric guidance.

Formally, we render $\mathbf{P}^{pan}$ from the viewpoint of $\mathbf{C}_i$ to obtain the sparse guidance depth $\mathbf{D}_i^g$, as illustrated in Fig. 14. The alignment process is formulated as:

$$\mathbf{D}_i^a = \varphi_{align}\left(\mathbf{D}_i^m, \mathbf{D}_i^g, \mathbf{M}_i\right), \tag{11}$$

where $\mathbf{M}_i$ denotes the reliability mask for view i, indicating valid overlapping regions where the alignment should be enforced. We define $\mathbf{M}_i$ as the intersection across several empirical masks:

$$\mathbf{M}_i = \mathbf{M}_i^m \cap \mathbf{M}_i^g \cap \mathbf{M}_i^n \cap \mathbf{M}_i^p \cap \overline{\mathbf{M}}_i^{sky}, \tag{12}$$

where $\mathbf{M}_i^m$ and $\mathbf{M}_i^g$ represent the valid projection regions of the WorldMirror confidence and the panoramic guidance, respectively, with edge floaters removed. $\mathbf{M}_i^n$ enforces normal consistency,

excluding regions where the angular deviation between the WorldMirror normal $\mathbf{N}_i^m$ and the derived panoramic normal $\mathbf{N}_i^g$ exceeds 90 degrees. To mitigate occlusion artifacts shown in the guidance depth of Fig. 14, we employ a percentile-based statistical filter $\mathbf{M}_i^p$ to discard outliers with significant relative depth discrepancies. Finally, we omit the sky regions using the non-sky mask $\overline{\mathbf{M}}_i^{sky}$ identified by SAM3 [9] in video mode.

Subsequently, we perform a *RANSAC-based linear alignment* over the valid regions defined by $\mathbf{M}_i$ to estimate a scale γ_i and shift β_i, yielding a transformation as $\mathbf{D}_i^a = \gamma_i \mathbf{D}_i^m + \beta_i$[2]. Due to the high-quality initial depth provided by WorldMirror 2.0, we empirically find that per-frame linear alignment is sufficient for our scenarios, thus obviating the need for complex non-linear refinements [54, 21]. However, erroneous alignment coefficients may still occur, particularly when the valid guidance masks are overly sparse or the camera trajectories are highly challenging. To address this, we propose an outlier detection and revision strategy based on the global distribution of the alignment coefficients $\{\gamma_i, \beta_i\}_{i=1}^{T'_{ex}}$. Specifically, we set $Q = 9$ anchor depth values $\{A_q\}_{q=1}^{Q}$ uniformly distributed across the scene's depth range. For each frame i, we compute the transformed anchor values $\mathcal{V}_{i,q} = \gamma_i A_q + \beta_i$. The maximum relative deviation for each coefficient pair (γ_i, β_i) is then formulated as:

$$\mathcal{V}_i^{\max} = \max_q \left(\left| \frac{\mathcal{V}_{i,q} - \hat{\mathcal{V}}_q}{\hat{\mathcal{V}}_q} \right| \right), \quad \hat{\mathcal{V}}_q = \text{median}_{j \in \{1,\dots,T'_{ex}\}}(\mathcal{V}_{j,q}), \tag{13}$$

where $\hat{\mathcal{V}}_q$ represents the median transformed value of anchor q across all frames. Any coefficient pair (γ_i, β_i) whose maximum relative deviation $\mathcal{V}_i^{max}$ exceeds the 90-th percentile is regarded as an outlier. These outliers are then replaced by the nearest inlier coefficient pairs within the same video sequence. If an entire sequence is detected as an outlier, we discard all its depth maps. Finally, upon obtaining the aligned depth maps $\{\mathbf{D}_i^a\}_{i=1}^{T'_{ex}}$, we back-project them into 3D space to construct the extended point cloud $\mathbf{P}^{ex}$. The union of this extension and the original panorama, $\mathbf{P}^{pan} \cup \mathbf{P}^{ex}$, is then further voxel-downsampled to yield the final expanded point cloud $\tilde{\mathbf{P}} \in \mathbb{R}^{N \times 3}$.

7.2 3D Gaussian Splatting

Task Formulation. Given the panorama $\mathbf{I}^{pan}$, expanded point cloud $\tilde{\mathbf{P}}$, and and a set of T_{ex} generated novel keyframes along with their corresponding camera parameters $\{\mathbf{V}_i, \mathbf{C}_i\}_{i=1}^{T_{ex}}$, we optimize a 3DGS model [33] to serve as the final scene representation.

Initialization. We initialize the 3DGS model using the expanded point cloud $\tilde{\mathbf{P}}$. Each 3D Gaussian is parameterized by a set of learnable attributes, including an opacity $\sigma_k \in [0, 1]$, a center position $\boldsymbol{\mu}_k \in \mathbb{R}^{3\times1}$, and a 3D covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{3\times3}$. Following [33], to ensure the covariance matrix remains positive semi-definite during optimization, $\boldsymbol{\Sigma}_k$ is decomposed into a scaling matrix $\mathbf{S}_k$ and a rotation matrix $\mathbf{R}_k$, formulated as $\boldsymbol{\Sigma}_k = \mathbf{R}_k \mathbf{S}_k \mathbf{S}_k^T \mathbf{R}_k^T$. Furthermore, we empirically observe that the generated scenes exhibit negligible view-dependent effects. Therefore, instead of employing Spherical Harmonics (SH) for appearance modeling, we adopt view-independent RGB colors $\mathbf{c}_k \in \mathbb{R}^3$ as the color features for each Gaussian, reducing both redundancy and complexity.

Growth Strategy and MaskGaussian. During the 3DGS training on generated data, we observe a dilemma regarding the adaptive densification mechanism [33]. On the one hand, relying solely on the initial point cloud $\tilde{\mathbf{P}}$ without densification leads to a conflict between rendering efficiency and detail preservation. Specifically, $\tilde{\mathbf{P}}$ exhibits a spatially uneven distribution, over-populating low-frequency regions (*e.g.*, sky and flat surfaces) with redundant Gaussians that degrade real-time rendering efficiency. While applying a uniform voxel downsampling (with a voxel size v) can mitigate this redundancy, it severely undermines the reconstruction quality in high-frequency regions, which inherently require denser Gaussian coverage to faithfully capture fine-grained textures. On the other hand, enabling the standard growth strategy, which periodically densifies Gaussians via cloning and splitting based on view-space positional gradients, successfully recovers these high-frequency details but inevitably introduces severe floating artifacts (floaters). These artifacts predominantly originate from sky regions, where the generated depth supervision is unavailable.

[2] In practice, we apply the alignment in the disparity space instead of the depth space for better foreground alignment. To avoid confusion, we omit the disparity transformation for simplicity.

To resolve this dilemma and simultaneously achieve rendering efficiency and high-fidelity detail reconstruction, we adopt a two-pronged approach. First, we segment the initial point cloud $\tilde{\mathbf{P}}$ into sky and scene subsets, denoted as $\tilde{\mathbf{P}}_{\text{sky}}$ and $\tilde{\mathbf{P}}_{\text{scene}}$, respectively. The standard growth strategy is applied exclusively to $\tilde{\mathbf{P}}_{\text{scene}}$, enabling necessary densification in texture-rich regions while strictly preventing the sky from spawning floaters. To further eliminate redundant Gaussians in over-populated areas and suppress residual floating artifacts, we integrate MaskGaussian [45]. Instead of relying on heuristic hard-pruning, MaskGaussian models the existence of each Gaussian as a probabilistic entity. Concretely, for the k-th Gaussian, a binary mask $M_k \in \{0, 1\}$ is sampled via Gumbel-Softmax [25] from learnable mask logits. This mask is then incorporated into the tile-based rasterizer through a *masked-rasterization* scheme. For a given pixel $\mathbf{x}$, the rendered color $\mathbf{c}(\mathbf{x})$ and transmittance evolution T_{k+1} are reformulated as:

$$\mathbf{c}(\mathbf{x}) = \sum_{k=1}^{N} M_k \, \mathbf{c}_k \, \sigma_k \, T_k, \quad T_{k+1} = T_k(1 - M_k \sigma_k), \tag{14}$$

where σ_k denotes the opacity and T_k is the accumulated transmittance of the k-th Gaussian in depth order ($T_1 = 1$). When $M_k = 0$, the Gaussian's color contribution is negligible, and it consumes no transmittance. Crucially, thanks to the Gumbel-Softmax relaxation, it still receives gradients during the backward pass, allowing for a dynamic reassessment of its importance as the scene optimization evolves. To encourage sparsity, a squared loss regularizes the average mask activation:

$$\mathcal{L}_{\text{mask}} = \lambda_m \left(\frac{1}{N} \sum_{k=1}^{N} M_k \right)^2, \tag{15}$$

which is added to the overall training objective $\mathcal{L}_{\text{GS}}$. During training, Gaussians whose activation probabilities consistently remain near zero are permanently pruned. This adaptive mechanism preferentially eliminates redundant Gaussians in over-populated low-frequency areas while preserving essential primitives in detail-rich regions. Consequently, it simultaneously accelerates rendering speed and suppresses floaters through the implicit regularization induced by probabilistic masking.

Optimization and Losses. For the i-th training view, the 3DGS renderer produces an RGB image $\hat{\mathbf{I}}_i$ and a depth map $\hat{\mathbf{D}}_i$. The corresponding surface normal $\hat{\mathbf{N}}_i$ is derived analytically as the normalized spatial gradient of $\hat{\mathbf{D}}_i$. The photometric objective is defined as:

$$\mathcal{L}_{\text{color}} = (1 - \lambda_{c1}) \, \mathcal{L}_1(\hat{\mathbf{I}}_i, \mathbf{I}_i) + \lambda_{c1} \, \text{SSIM}(\hat{\mathbf{I}}_i, \mathbf{I}_i) + \lambda_{c2} \, \text{LPIPS}(\hat{\mathbf{I}}_i, \mathbf{I}_i), \tag{16}$$

where the ground truth images $\mathbf{I}_i$ are sampled from the union of views divided from the panorama and the generated keyframes. Here, SSIM and LPIPS denote the structural similarity and perceptual loss [29], respectively. To enforce geometric consistency, we introduce a geometric loss:

$$\mathcal{L}_{\text{geo}} = \lambda_d \, \mathcal{L}_1(\hat{\mathbf{D}}_i, \mathbf{D}_i^a) + \lambda_n \left(1 - \cos(\hat{\mathbf{N}}_i, \mathbf{N}_i)\right), \tag{17}$$

where $\cos(\cdot)$ denotes the pixel-wise cosine similarity. To mitigate computational overhead, depth supervision is applied sparsely, restricted to the partially aligned depth maps $\{\mathbf{D}_i^a\}_{i=1}^{T'_{ex}}$ (Sec. 7.1). In contrast, the high-quality normal maps $\{\mathbf{N}_i\}_{i=1}^{T_{ex}}$ estimated by MoGe2 [67] are inherently alignment-free. This property enables them to be applied across all frames, providing a dense and robust geometric constraint. Furthermore, following [75], we incorporate a scale regularization term $\mathcal{L}_{\text{reg}}$ to penalize excessively sharp Gaussians, encouraging more isotropic shapes. The overall 3DGS training objective is thus given by:

$$\mathcal{L}_{\text{GS}} = \mathcal{L}_{\text{color}} + \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{mask}}. \tag{18}$$

Mesh Extraction. To support downstream applications such as collision detection and physics simulation, we further extract a mesh from the optimized 3DGS representation. Specifically, we render RGB images and depth maps from all training views and integrate them into a Truncated Signed Distance Function (TSDF) volume. The final mesh is extracted via the marching cube algorithm [46]. To improve mesh quality, we remove small disconnected components and apply mesh simplification, which effectively suppresses floating artifacts and reduces storage overhead.

Table 4: **Quantitative comparisons on text-to-panorama (T2P) and image-to-panorama (I2P).**

Metric	Text-to-Panorama (T2P)				Image-to-Panorama (I2P)			
	DiT360	Matrix3D	HY-World 1.0	HY-Pano 2.0	CubeDiff	GenEx	HY-World 1.0	HY-Pano 2.0
CLIP-T / CLIP-I ↑	0.248	0.238	0.250	**0.258**	0.828	0.831	0.831	**0.844**
Q-Align Qual (Persp) ↑	3.788	2.983	3.992	**4.103**	2.938	2.917	3.317	**4.026**
Q-Align Qual (Equi) ↑	4.436	4.258	**4.493**	4.403	3.814	3.868	4.076	**4.277**
Q-Align Aes (Persp) ↑	2.882	2.126	**3.404**	3.376	2.319	2.445	2.638	**3.208**
Q-Align Aes (Equi) ↑	4.072	3.880	4.186	**4.247**	3.645	3.646	3.767	**4.056**

8 Results: Multi-Modal World Creation

8.1 World Generation from Text or Single Image

In this section, we comprehensively evaluate the world generation pipeline of HY-World 2.0. We first assess its individual components: panorama generation (Sec. 8.1.1), trajectory planning (Sec. 8.1.2), world expansion (Sec. 8.1.3), and 3DGS (Sec. 8.1.4). Then, the final outputs of the integrated system will be showcased in Sec. 8.1.5.

8.1.1 Results & Analysis of HY-Pano 2.0

For both qualitative and quantitative comparisons, we evaluate the panorama generation of HY-Pano 2.0 against several state-of-the-art approaches across text-to-panorama (T2P) and image-to-panorama (I2P). For T2P, we compare with DiT360 [17], Matrix3D [80], and HY-World 1.0 [23]. For I2P, we compare with CubeDiff [30], GenEx [47], and HY-World 1.0 [23].

Quantitative Results. Tab. 4 presents the quantitative comparison for both T2P and I2P tasks. We evaluate generated panoramas using multiple complementary metrics. CLIP-T [52] (T2P) and CLIP-I [52] (I2P) measure text-image and image-image alignment, respectively. Q-Align [73] provides both perceptual quality (Qual) and aesthetic (Aes) scores based on a large multi-modal model aligned with human ratings. For all applicable metrics, we report results on both the equirectangular (Equi) panorama and averaged perspective (Persp) projections, where each panorama is projected onto 12 perspective faces. As shown in Tab. 4, HY-Pano 2.0 achieves the best scores on the majority of metrics across both tasks. For T2P, it obtains the highest CLIP-T score and leads on most Q-Align quality and aesthetics measures. For I2P, it ranks first on all five metrics, with notable improvements over HY-World 1.0 in both perceptual quality and aesthetics. These results demonstrate that HY-Pano 2.0 exhibits stronger adherence to input signals (text prompts or reference images), improved fine-grained detail quality, and enhanced aesthetic score compared to prior methods.

Qualitative Results. We first show some generated panoramas conditioned on image and text inputs in Fig. 16. Then, we present qualitative comparisons for T2P and I2P in Fig. 17 and Fig. 18, respectively. Compared to existing methods, HY-Pano 2.0 generates panoramas with more structurally coherent layouts, exhibiting plausible spatial arrangements and consistent geometric structures across the full 360° field of view. In terms of visual aesthetics, our results demonstrate superior color harmony, lighting consistency, and overall artistic quality. Furthermore, HY-Pano 2.0 produces notably finer details, including sharper textures, cleaner object boundaries, and richer high-frequency content, leading to more realistic and visually appealing panoramas.

8.1.2 Results & Analysis of WorldNav

We present qualitative comparisons in Fig. 19 to intuitively demonstrate the necessity of each trajectory planning component. Training 3DGS solely on panoramic views (Fig. 19b) inevitably suffers from massive geometric voids and poor rendering quality. By sequentially integrating views from different trajectories, the scene completeness progressively improves. Specifically, regular trajectories (Fig. 19c) break the limitation of a fixed viewpoint, providing expanded observations that eliminate most large-scale artifacts. However, regular paths often fail to cover occluded structures, leaving the sides and backs of specific objects (*e.g.*, the car, cabin, and arcade machine) incomplete. This limitation is effectively resolved by introducing surrounding and reconstruction-aware trajectories (Fig. 19d), which explicitly target and complete these complex structures. Furthermore, wandering trajectories (Fig. 19e) enhance the textural details of distant walls and floors, enabling good roaming

Figure 16: **Visual results of panorama generation by HY-Pano 2.0.** Our model supports both text and images of various resolutions as inputs.

Figure 17: **Qualitative comparison on the text-to-panorama (T2P) task.** Our method outperforms previous approaches in terms of layout coherence, fine-grained details, and overall visual aesthetics.

experiences. Finally, aerial trajectories (Fig. 19f) incorporate additional bird's-eye view (BEV) observations, improving the freedom of the 3D world's viewpoint changing.

8.1.3 Results & Analysis of WorldStereo 2.0

Results of Single-View Scene Reconstruction. Following WorldStereo [62], we evaluate WorldStereo 2.0 on the single-view scene reconstruction benchmark in Tab. 5, utilizing the out-of-distribution Tanks-and-Temples [35] and MipNeRF360 [5] datasets. For quantitative evaluation, we compare our results against pseudo ground-truth point clouds reconstructed via Multi-View-Stereo [7] from real multi-view images. To rigorously test our method, we introduce more challenging camera trajectories than the original benchmark: closed-loop paths for object-centric scenes and manually designed, explorable routes for large-scale forward-facing scenes. This significantly increases the difficulty of maintaining multi-view consistency. As demonstrated in Tab. 5, WorldStereo 2.0 achieves the highest point cloud F1 and AUC scores, surpassing all video-based and 3D-based competitors. Although single-view generative reconstruction inherently suffers from high uncertainty, these superior geometric metrics confirm that our approach successfully synthesizes highly consistent and physically plausible 3D structures.

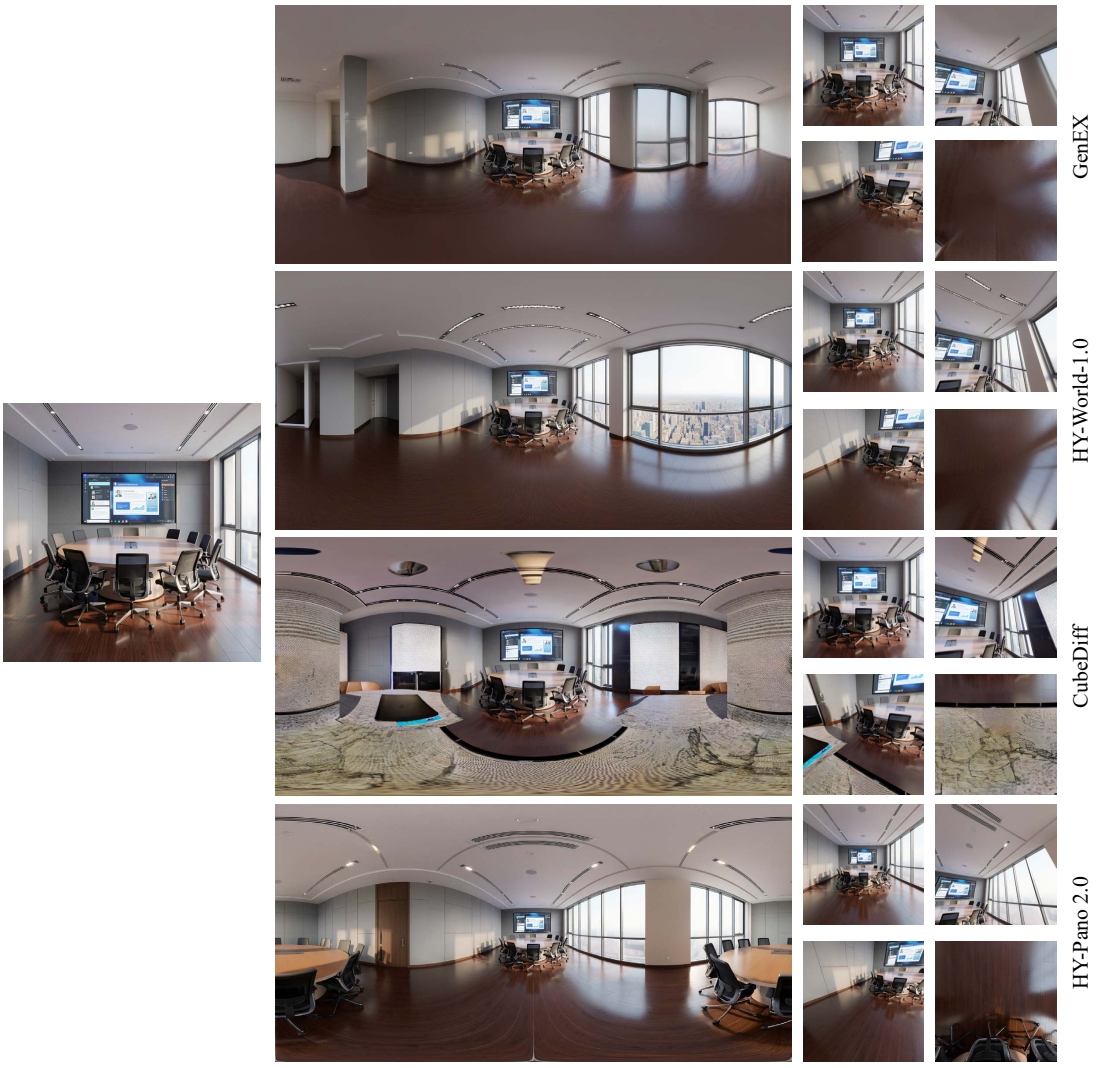
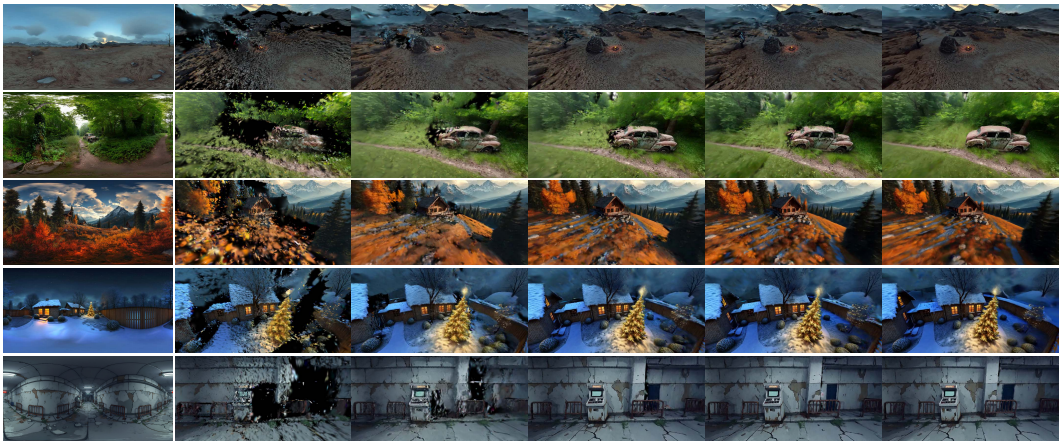


Figure 18: **Qualitative comparison on the image-to-panorama (I2P) task.** Our method outperforms previous approaches in extension plausibility, content richness, and overall quality.

Results of Camera Control Capability. We quantitatively evaluate the camera control capability of WorldStereo 2.0 in Tab. 6, while ablation studies are performed in Tab. 7. Both evaluations are applied with 100 out-of-distribution images selected from [15] with challenging trajectories. Notably, WorldStereo 2.0 outperforms all video-based competitors by achieving the lowest errors across all camera metrics. Furthermore, it also delivers superior visual quality and semantic alignment. For the ablation study in Tab. 7, since Keyframe-VAE introduces significant changes to the latent representations, directly applying it without training the main network is unfair and yields limited improvements. Therefore, we unlock the main DiT for full training (freezing "None"). Compared with the Video-VAE baseline, the fully trained Keyframe-VAE significantly improves visual quality, user-perceived camera control, and most camera metrics. Moreover, we provide qualitative comparisons in Fig. 8, which further support this conclusion. However, we observe an obvious trade-off between performance and generalization for the trainable parts of the main DiT. While full model training maximizes visual metrics, it leads to inferior camera precision and suboptimal user study quality. We find that this is due to overfitting issues, where the global image style slightly drifts during video generation. To address this, we selectively freeze specific layers. As shown in the blue row, freezing the cross-attention and FFN layers achieves the best balance. It effectively mitigates the overfitting, yielding the most precise camera control with the lowest RotErr, TransErr, and ATE, while attaining the highest user preference for visual quality (64.39%).



(a) Panorama (b) Panoramic views only (c) +Regular Traj. (d) +Surround & Recon. Traj. (e) +Wander Traj. (f) +Aerial Traj. (full)

Figure 19: **Qualitative ablation results of trajectory planning.** Relying solely on panoramic views (b) results in severe artifacts and incomplete geometry. By sequentially integrating views generated from (c) regular, (d) surrounding & reconstruction, (e) wandering, and (f) aerial trajectories, our method progressively eliminates blind spots, refines complex object structures, and enhances overall scene completeness. Please zoom in for details.

Table 5: **Point cloud results of one-view-generated 3D reconstruction.** Precision measures the percentage of generated points that fall within a distance threshold of ground-truth points, while Recall measures the percentage of ground-truth points covered by generated points. The F1-score is their harmonic mean, and AUC denotes the area under the curve across varying distance thresholds.

Methods	Tanks-and-Temples [35]				MipNeRF360 [5]			
	Precision↑	Recall↑	F1-Score↑	AUC↑	Precision↑	Recall↑	F1-Score↑	AUC↑
SEVA [87]	33.59	35.34	36.73	51.03	22.38	55.63	28.75	46.81
Gen3C [54]	46.73	25.51	31.24	42.44	23.28	**75.37**	35.26	52.10
Lyra [4]	**50.38**	28.67	32.54	43.05	30.02	58.60	36.05	49.89
FlashWorld [39]	26.58	20.72	22.29	30.45	35.97	53.77	42.60	53.86
WorldStereo 2.0	43.62	41.02	41.43	58.19	**43.19**	65.32	**51.27**	**65.79**
WorldStereo 2.0 (DMD)	40.41	**44.41**	**43.16**	**60.09**	42.34	64.83	50.52	65.64

Table 6: **Quantitative results of camera control capability.** We evaluate performance across camera metrics and visual quality. Methods labeled with * indicate that the model is only trained with camera control (domain-adaption) without memory capabilities.

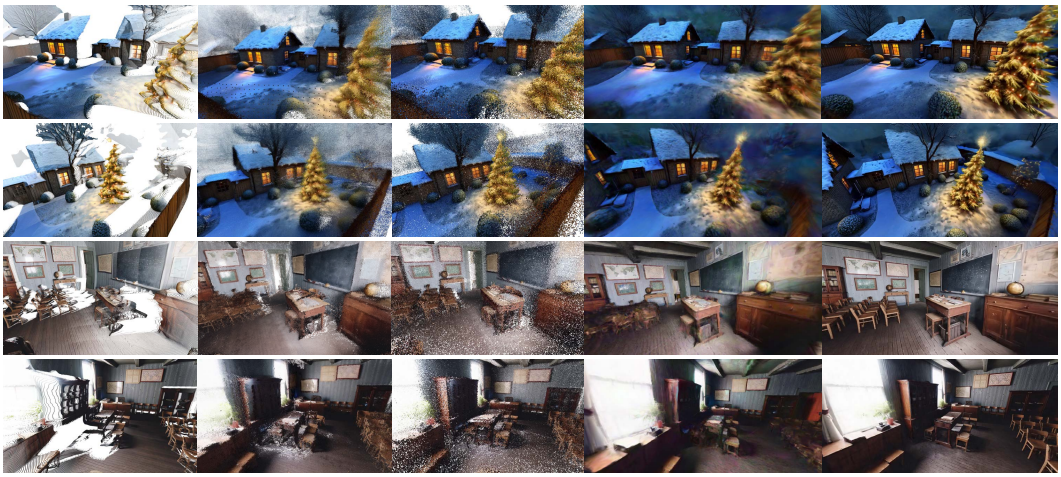
Methods	Camera Metrics			Visual Quality			
	RotErr↓	TransErr↓	ATE↓	Q-Align↑	CLIP-IQA+↑	Laion-Aes↑	CLIP-I↑
SEVA [87]	1.690	1.578	2.879	3.232	0.479	4.623	77.16
Gen3C [54]	0.944	1.580	2.789	3.353	0.489	4.863	82.33
WorldPlay [60]	3.481	1.288	2.722	3.628	**0.552**	5.103	86.79
WorldCompass [70]	3.452	1.068	2.379	3.615	0.548	5.111	85.51
WorldStereo [62]*	0.762	1.245	2.141	4.149	0.547	5.257	89.05
WorldStereo 2.0*	**0.492**	**0.968**	**1.768**	**4.205**	0.544	**5.266**	**89.43**

Table 7: **Camera control ablation studies of WorldStereo 2.0.** We evaluate performance across camera metrics, visual quality, and user study. Given the misalignment between the commonly used metrics and human perception, we prioritize user study results for model selection. The red row indicates the baseline (camera control of WorldStereo [62]), while the blue row denotes our final domain-adaption setting.

Frozen Parts	VAE Types	Camera Metrics			Visual Quality			User Study	
		RotErr↓	TransErr↓	ATE↓	Q-Align↑	CLIP-IQA+↑	Laion-Aes↑	Camera↑	Quality↑
Main DiT	Video-VAE	0.762	1.245	2.141	4.149	0.547	5.257	84.85%	46.46%
Main DiT	Keyframe-VAE	0.768	1.149	2.027	4.060	0.520	5.210	–	–
None	Keyframe-VAE	0.578	1.115	2.245	**4.237**	**0.554**	**5.278**	**93.81%**	60.61%
Cross-Attn	Keyframe-VAE	0.684	1.243	2.111	4.181	0.538	5.235	93.13%	60.95%
Cross-Attn + FFN	Keyframe-VAE	**0.492**	**0.968**	**1.768**	4.205	0.544	5.266	92.44%	**64.39%**

Table 8: **Memory and distillation ablation studies of WorldStereo 2.0.** PSNR_m and SSIM_m are computed within *valid warping mask* regions to evaluate consistency. The configuration A^* denotes a variant where the SSM incorporates reference features via temporal concatenation rather than spatial concatenation. Settings of our final memory and distilled models are colorized in green and yellow.

	Configuration	Photometric Metrics			Consistency		Camera Metrics		
		PSNR↑	SSIM↑	LPIPS↓	PSNR_m ↑	SSIM_m ↑	RotErr↓	TransErr↓	ATE↓
	Camera control baseline	16.13	0.474	0.349	28.81	0.448	0.396	0.053	0.071
A	GGM and SSM++ (A)	20.94	0.640	0.170	30.27	0.623	0.407	0.047	0.046
B	Trainable FFN (A,B)	21.56	0.667	0.162	30.44	0.624	0.351	**0.036**	**0.035**
C	Pointcloud augmentation (A,B,C)	21.36	0.632	0.163	30.72	0.619	0.360	0.050	0.053
D	Reference augmentation (A,B,C,D)	20.86	0.639	0.165	30.66	0.636	0.322	0.049	0.067
E	Camera embedding (A,B,C,D,E)	21.06	0.639	0.164	30.58	0.617	0.329	0.042	0.048
A^*	Temporal-concated SSM (A^*,B,C,D,E)	19.83	0.581	0.219	29.77	0.571	0.545	0.087	0.114
F	Doubled batch-size (64) (A,B,C,D,E,F)	21.63	**0.669**	**0.156**	30.76	0.647	**0.296**	**0.036**	0.041
G	After post-distillation (A,B,C,D,E,F,G)	**21.84**	**0.669**	0.165	**30.93**	**0.656**	0.316	0.052	0.072



(a) Panoramic Pcd. (b) video2world Pcd. (~5h) (c) Our Pcd. (~1.7min) (d) video2world 3DGS (e) Our 3DGS

Figure 20: **Point clouds and 3DGS comparisons with video2world [21].** "Pcd." denotes point cloud. For each scene, both methods are evaluated within 300 images generated by WorldStereo 2.0.

Ablation Studies of Memory Training and Distillation. We comprehensively evaluate the memory training and post-distillation in Tab. 8. Incorporating GGM and SSM++ (Config A) substantially improves the photometric quality and multi-trajectory consistency. Furthermore, unfreezing the FFN (Config B) significantly enhances camera control precision. To mitigate overfitting to point cloud guidance and retrieved reference views, we introduce several augmentation strategies (Configs C and D, detailed in Sec. 5.2.3) to these conditions. While these regularizations cause a little gap in clean-data metrics, they are crucial for overall robustness and maintain highly competitive performance. Moreover, we validate our spatial-stereo stitching design in SSM. Replacing it with temporal concatenation (Config A^*) severely degrades performance across all metrics. Additionally, scaling the training batch size to 64 (Config F) stabilizes training, yielding consistent improvements. Finally, after applying DMD post-distillation (Config G), the model not only retains comparable camera control but even slightly improves photometric and consistency metrics.

8.1.4 Results & Analysis of World Composition

Reconstruction and Alignment. While Sec. 7.1 establishes the effectiveness of WorldMirror 2.0 in point cloud expansion with known camera poses, we further evaluate our overall composition pipeline against the concurrent world reconstruction method, video2world [21] in Fig. 20. To ensure a fair comparison, both methods are evaluated on 300-view images generated by WorldStereo 2.0, which reaches the memory limit of an NVIDIA H20 GPU for video2world. As illustrated in Fig. 20, although

Table 9: **3DGS ablation studies averaged over 10 scenes.** † denotes that the adaptive densification is restricted to non-sky regions.

Voxel Downsample	Adaptive Densification	MaskGaussian	GS Number	PSNR↑	SSIM↑	LPIPS↓
			6.000M	25.176	0.751	0.209
✓			1.000M	24.504	0.720	0.276
✓	✓		5.254M	25.158	0.750	0.210
✓	✓	✓	1.383M	25.017	0.747	0.216
✓	✓†	✓	1.381M	25.023	0.747	0.215

video2world produces impressive point clouds via feature-matched Iterative Closest Point (ICP), this process is inherently difficult to parallelize, resulting in a prohibitive computational overhead of approximately 5 hours per scene. In contrast, our lightweight linear alignment fully leverages camera pose priors to achieve comparable reconstruction quality in less than 2 minutes. Furthermore, our final 3DGS reconstructions exhibit superior geometric and textural details, largely attributed to the tailored optimization strategies proposed in Sec. 7.2. Notably, because our WorldStereo 2.0 generates sequences with significantly higher visual consistency than SEVA [87], the complex non-rigid aware 3DGS required by video2world [21] becomes unnecessary in our pipeline. Additionally, we observe that the SH optimization often leads to undesirable color artifacts rendered in novel views (see Fig. 20(d)). Consequently, our pipeline adopts a direct RGB optimization, which proves to be more robust and effective for generative scenarios.

Gaussian Splattings. We ablate each component of the proposed 3DGS pipeline across 10 scenes, evaluating each on a 20-view validation set (Tab. 9). The baseline initializes from 6M Gaussians randomly sampled from the expanded point cloud $\tilde{\mathbf{P}}$, yielding the highest quality (PSNR 25.176) but incurring substantial rendering overhead. Applying voxel downsampling alone reduces the Gaussian count to 1M, but at a severe cost to quality—a 0.68 dB drop in PSNR and a 32% increase in LPIPS—confirming that uniform decimation disproportionately degrades detail-rich regions. Enabling adaptive densification restores the quality to near-baseline levels (PSNR 25.158), yet inflates the count to 5.254M, largely negating the efficiency gains of downsampling. Integrating MaskGaussian resolves this trade-off: redundant Gaussians in low-frequency areas are pruned, reducing the count by 73.7% (from 5.254M→1.383M) with only -0.14 dB PSNR degradation. Further restricting densification to non-sky regions suppresses floaters where depth supervision is unavailable. The full configuration retains comparable visual quality while reducing 77% Gaussian count compared to the baseline.

8.1.5 Full Results & Comparison with Marble

Explorable and Interactive Worlds. As illustrated in Fig. 21, HY-World 2.0 yields comprehensive multi-modal 3D assets, encompassing panoramas, aligned point clouds for 3DGS initialization, high-fidelity 3DGS renderings, and extracted geometric meshes. Crucially, these rich 3D representations transcend static visualization, serving as foundational environments for explorable and interactive 3D worlds (see Fig. 22). By leveraging the meshes extracted from 3DGS as underlying collision proxies, our system supports real-time physical feedback and spatial interactions. To ensure seamless user experiences and rapid scene loading, we optimize these meshes into lightweight topological structures. This deliberate design strikes a balance between physical plausibility and efficiency, paving the way for downstream applications in gaming, virtual reality, and embodied AI.

Comparisons with the State-of-the-art. We compare our approach against the closed-source commercial world model, Marble [72][3]. The comparison is conducted under two settings: using identical panoramic inputs (Fig. 23) and using the same perspective conditions (Fig. 24). While Marble can produce impressive 3DGS results, it usually deviates from the input guidance, resulting in noticeable discrepancies and lower fidelity in regions explicitly covered by the panoramic or perspective conditions. In contrast, our method achieves high-fidelity results that strictly adhere to the provided conditions. Furthermore, our generation outperforms Marble in terms of detail preservation and geometric consistency of novel views. As illustrated from Fig. 23 and Fig. 24, our results maintain superior structural integrity and smoother textures across fences, cars, furniture, mountains,

[3] In this report, we compare our method with Marble 1.0 (as of March 30, 2026).

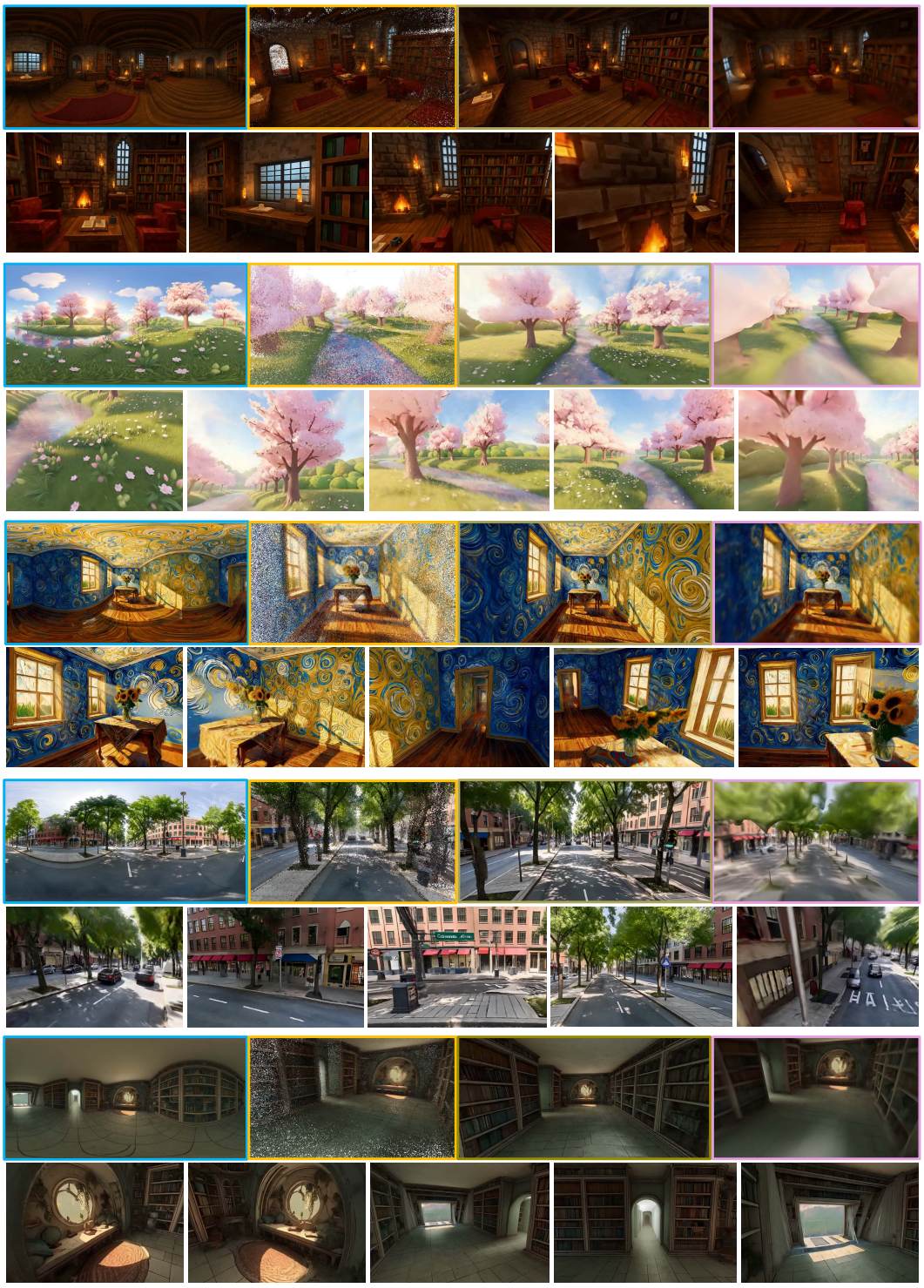


Figure 21: **Results of the overall world generation pipeline.** Each scene is visualized across two rows. The top row displays, from left to right: the generated panorama, the aligned point clouds, a global overview of splattings, and the extracted coarse mesh. The bottom row showcases novel views rendered from various viewpoints.

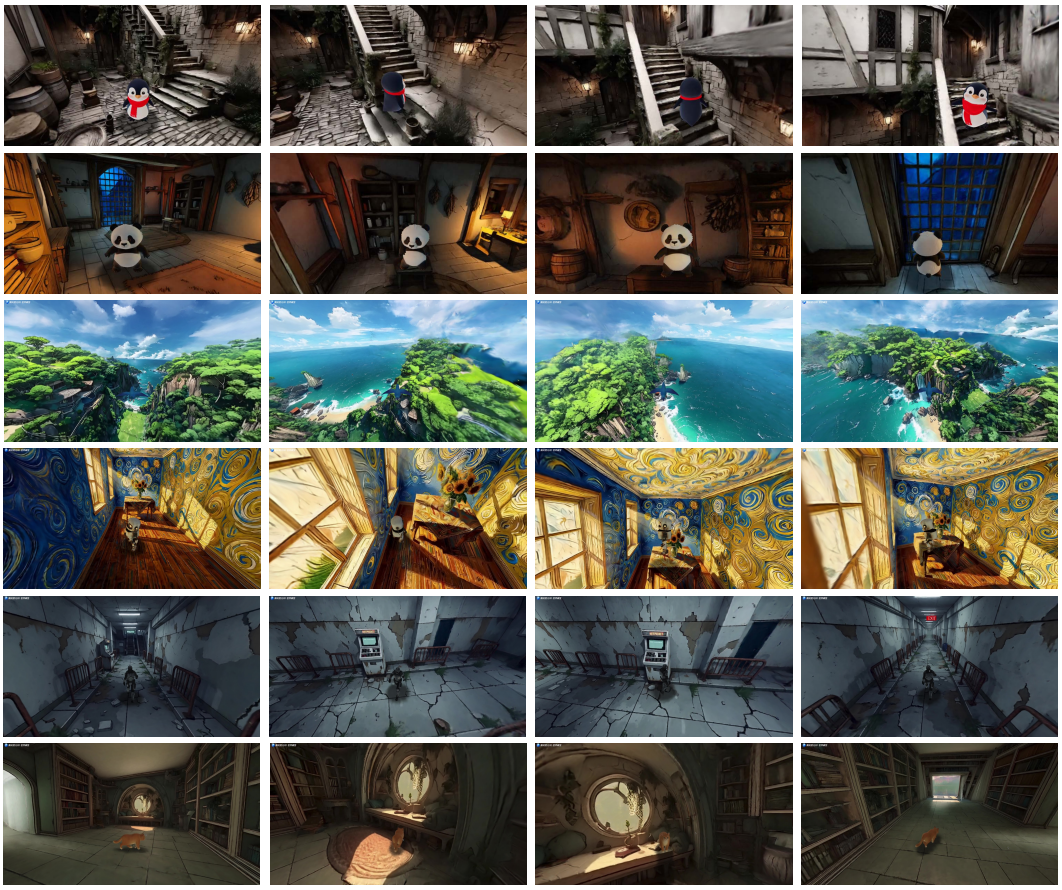


Figure 22: **Interactive exploration within the generated 3D worlds of HY-World 2.0.** By controlling virtual agents, users can navigate complex geometric structures (*e.g.*, stairs and indoor layouts) with real-time collision detection and physically plausible feedback, demonstrating the readiness of our results for interactive applications.

Table 10: **Runtime of each component in HY-World 2.0 for one single world generation.** "Recon" means reconstruction. All efficiency evaluations are performed using NVIDIA H20 GPUs.

Stage	Panorama	Trajectory Plan	World Expansion	Recon and Align	3DGS	Total
Time (sec)	15s	182s	286s	102s	127s	712s

and arcade machines, whereas Marble suffers from severe blurring and geometric missing under large viewpoint changes.

Runtime Analysis. We evaluate the overall runtime of HY-World 2.0 on NVIDIA H20 GPUs, as detailed in Tab. 10. By integrating systematic efficiency optimizations, the end-to-end pipeline for generating a complete 3D world is accelerated, requiring only 10 minutes. Specifically, we employ Sequence Parallelism (SP) to distribute all model inference stages, including panorama generation, Keyframe-VAE, WorldStereo 2.0, and WorldMirror 2.0. Furthermore, the overall efficiency is improved by incorporating other acceleration techniques, such as SageAttention2 [85], FP8 mixed-precision inference, and step caching mechanisms [42].

8.2 World Reconstruction from Multi-View Images or Video

We evaluate WorldMirror 2.0 as a standalone reconstruction foundation model on comprehensive benchmarks covering point map reconstruction (Tab. 11), camera pose estimation, depth estimation, novel view synthesis (Tab. 12), and surface normal estimation (Tab. 13). All tasks are evaluated at

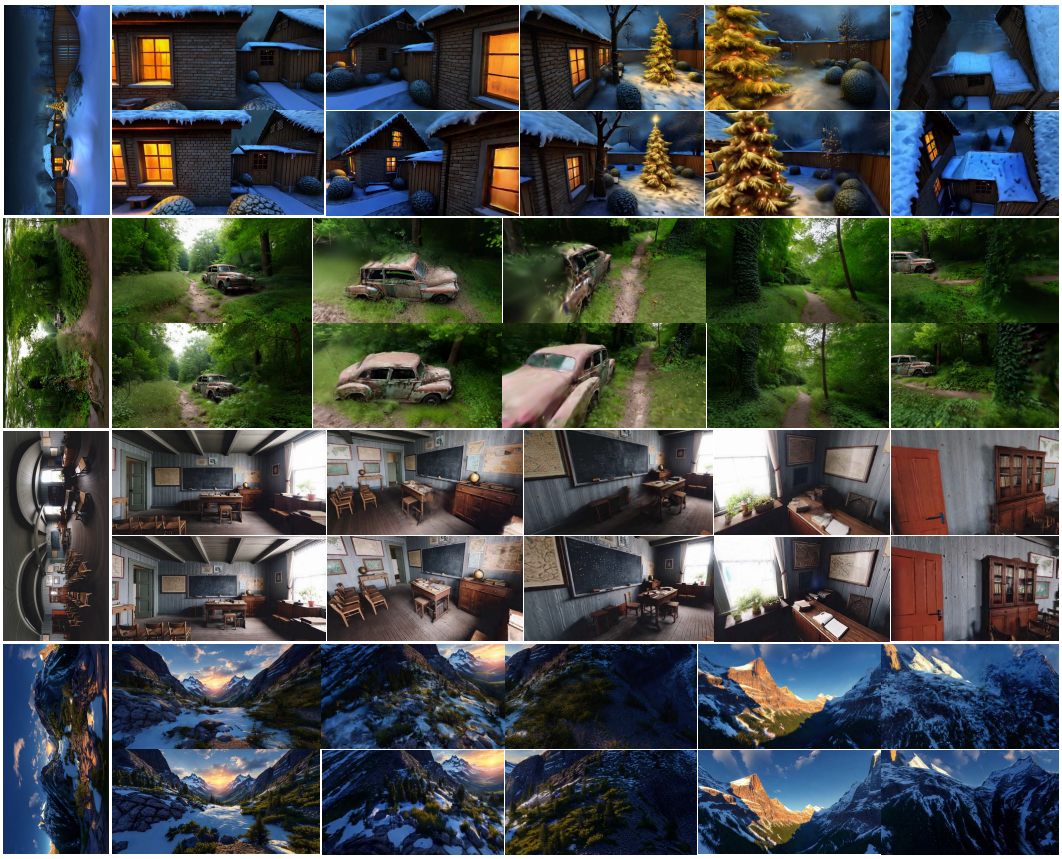


Figure 23: **Qualitative comparison with Marble [72] using the same panoramic inputs.** The input panoramas are displayed on the left. For each scene, we present novel view renderings from the generated 3DGS models of both methods. Compared to Marble, our approach achieves higher fidelity to the input conditions, sharper textures, and superior geometric consistency across diverse viewpoints. Please zoom in for details.

three inference resolutions, *i.e.*, low (189×259), medium (378×518, the default of WorldMirror 1.0), and high (756×1036), to validate the resolution generalization enabled by normalized position encoding (Sec. 6.2.1).

A consistent finding across all tasks is that WorldMirror 1.0 suffers from severe performance degradation at high resolution due to position extrapolation (*e.g.*, camera pose AUC@30 drops from 86.13 to 66.29; 7-Scenes point map accuracy degrades from 0.043 to 0.079), whereas WorldMirror 2.0 maintains, and often improves, performance from medium to high resolution across every benchmark. Beyond the multi-resolution improvements, we further evaluate the effectiveness of flexible geometric prior injection (Sec. 8.2.2) and the inference efficiency optimizations introduced in Sec. 6.4.

8.2.1 Results & Analysis of WorldMirror 2.0

Point Map Reconstruction. We evaluate point map reconstruction on scene-level datasets (7-Scenes, NRGBD) and an object-level dataset (DTU), following the same sequence mappings as [69]. As shown in Tab. 11, WorldMirror 1.0 at medium resolution already surpasses all baselines. WorldMirror 2.0 further improves at every resolution: at medium, it reduces the 7-Scenes accuracy error from 0.043 to 0.033; at high, the gap is even more pronounced (0.079→0.037). Incorporating geometric priors yields additional gains, with WorldMirror 2.0 at high resolution with all priors achieving the best overall results on 7-Scenes and DTU.

Camera Pose, Depth, and Novel View Synthesis. In Tab. 12, we jointly report camera pose estimation and depth estimation on RealEstate10K, and novel view synthesis averaged across RealEstate10K

(a) Perspective input image (b) Generated panorama (c) 3DGS rendered from various viewpoints

Figure 24: **Qualitative comparison with Marble [72] using the same input image.** (a) displays input perspective images. For each scene, we compare both (b) generated panoramas and (c) 3DGS renderings. Compared to Marble, our approach better adheres to the input views while achieving comparable quality and superior completeness in 3DGS. Please zoom in for details.

and DL3DV, following the protocol of [69]. For camera pose, WorldMirror 2.0 improves AUC@30 over WorldMirror 1.0 at every resolution: low (80.55→83.43), medium (86.13→86.48), and high (66.29→86.89). The high-resolution gain of over 20 points being particularly striking. For depth, WorldMirror 2.0 consistently reduces AbsRel (medium: 0.178→0.167; high: 0.195→0.162) and achieves the best δ<1.25 accuracy (0.815 at high resolution). For novel view synthesis, WorldMirror 1.0 at high resolution suffers a dramatic quality collapse (PSNR from 21.34 to 17.78), whereas WorldMirror 2.0 maintains stable performance across resolutions (PSNR of 20.14/20.07/19.98 at low/medium/high) and achieves the best SSIM (0.726 at high resolution), confirming that higher-resolution inference improves structural fidelity.

Surface Normal Estimation. Following [3], we evaluate surface normal estimation on ScanNet [12], NYUv2 [56], and iBims-1 [36]. As shown in Tab. 13, WorldMirror 2.0 achieves the best results across all three benchmarks at medium resolution, surpassing dedicated single-task methods. The improvements of both depth and normal estimation are consistent with the explicit depth-to-normal supervision (Sec. 6.2.2) and pseudo-normal enhancement (Sec. 6.3), which jointly strengthen geometric coupling between depth and normal predictions. The resolution generalization remains consistent: WorldMirror 2.0 at high resolution (ScanNet mean error 12.5) closely matches its medium-resolution optimum (12.3), whereas WorldMirror 1.0 degrades from 13.8 to 17.6.

Table 11: **Point map reconstruction on 7-Scenes, NRGBD, and DTU.** Both Acc. (↓) and Comp. (↓) are lower-is-better. Baseline methods are evaluated at M resolution. WorldMirror 2.0 generalizes across multiple resolutions. "+ all priors" denotes additionally providing camera pose, intrinsics, and depth as input. "L/M/H": low (182×252), medium (378×518), high (756×1036) resolution. Best results are highlighted.

Method	**7-Scenes** (scene)				**NRGBD** (scene)				**DTU** (object)			
	Acc. ↓		Comp. ↓		Acc. ↓		Comp. ↓		Acc. ↓		Comp. ↓	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Fast3R [77]	0.096	0.065	0.145	0.093	0.135	0.091	0.163	0.104	3.340	1.919	2.929	1.125
CUT3R [66]	0.094	0.051	0.101	0.050	0.104	0.041	0.079	0.031	4.742	2.600	3.400	1.316
FLARE [86]	0.085	0.058	0.142	0.104	0.053	0.024	0.051	0.025	2.541	1.468	3.174	1.420
VGGT [65]	0.046	0.026	0.057	0.034	0.051	0.029	0.066	0.038	1.338	0.779	1.896	0.992
π^3 [69]	0.048	0.028	0.072	0.047	0.026	0.015	0.028	0.014	1.198	0.646	1.849	0.607
WorldMirror 1.0												
L	0.043	0.029	0.055	0.029	0.046	0.027	0.049	0.026	1.476	0.889	1.768	0.917
L + all priors	0.021	0.014	0.026	0.016	0.022	0.015	0.020	0.014	1.347	0.854	1.392	0.865
M	0.043	0.026	0.049	0.028	0.041	0.020	0.045	0.019	1.017	0.564	1.780	0.690
M + all priors	0.018	0.011	0.023	0.014	0.016	0.011	0.014	0.010	0.735	0.461	0.935	0.550
H	0.079	0.052	0.087	0.051	0.077	0.047	0.093	0.051	2.271	1.083	2.113	0.825
H + all priors	0.042	0.024	0.041	0.024	0.078	0.053	0.082	0.051	1.773	0.792	1.478	0.782
WorldMirror 2.0												
L	0.041	0.027	0.052	0.027	0.047	0.028	0.058	0.035	1.352	0.824	2.009	0.880
L + all priors	0.019	0.012	0.024	0.014	0.017	0.011	0.015	0.010	1.100	0.748	1.201	0.774
M	0.033	0.020	0.046	0.026	0.039	0.024	0.047	0.027	1.005	0.545	1.892	0.681
M + all priors	0.013	0.008	0.017	0.011	0.013	0.009	0.013	0.009	0.690	0.458	0.876	0.506
H	0.037	0.025	0.040	0.023	0.046	0.026	0.053	0.030	0.845	0.426	1.904	0.632
H + all priors	0.012	0.008	0.016	0.010	0.015	0.010	0.016	0.010	0.554	0.343	0.771	0.398

Table 12: **Results of camera pose estimation and depth estimation (left); Results of novel view synthesis (right).** Camera pose and depth are evaluated on RealEstate10K; novel view synthesis is averaged across RealEstate10K and DL3DV. ↑/↓ indicate higher-/lower-is-better. Baseline methods are evaluated at M resolution. WorldMirror 2.0 generalizes across multiple resolutions. "L/M/H" denote low / medium / high inference resolution. Best results are highlighted.

Method	**Camera Pose** (Avg.)		**Depth** (Avg.)	
	AUC@30↑	RTA@30↑	AbsRel↓	δ<1.25 ↑
Fast3R [77]	61.68	81.86	0.353	0.666
CUT3R [66]	81.47	95.10	0.260	0.704
FLARE [86]	80.01	95.23	0.445	0.551
VGGT [65]	77.62	93.13	0.256	0.789
π^3 [69]	85.90	95.62	0.151	0.805
WorldMirror 1.0 (L)	80.55	93.68	0.225	0.751
WorldMirror 1.0 (M)	86.13	95.47	0.178	0.812
WorldMirror 1.0 (H)	66.29	89.62	0.195	0.797
WorldMirror 2.0 (L)	83.43	94.79	0.199	0.770
WorldMirror 2.0 (M)	86.48	95.55	0.167	0.806
WorldMirror 2.0 (H)	86.89	95.34	0.162	0.815

Method	**NVS** (Avg.)		
	PSNR↑	SSIM↑	LPIPS↓
FLARE [86]	15.84	0.545	0.500
AnySplat [28]	18.57	0.626	0.255
WorldMirror 1.0 (L)	20.38	0.658	0.163
WorldMirror 1.0 (M)	21.34	0.709	0.181
WorldMirror 1.0 (H)	17.78	0.659	0.379
WorldMirror 2.0 (L)	20.14	0.679	0.149
WorldMirror 2.0 (M)	20.07	0.680	0.186
WorldMirror 2.0 (H)	19.98	0.726	0.235

Qualitative Results. We present visual comparisons between WorldMirror 1.0 and 2.0 in Fig. 25 and Fig. 26. As shown in Fig. 25, WorldMirror 2.0 produces sharper and more geometrically coherent surface normals, with finer structural details and fewer artifacts in complex regions. The reconstructed point clouds of WorldMirror 2.0 also exhibit tighter multi-view consistency, reflecting the geometric coupling enforced by the depth-to-normal supervision (Sec. 6.2.2) and more robust invalid-pixel handling via the depth mask prediction head (Sec. 6.2.3).

Fig. 26 further examines multi-resolution robustness under both dense (32 views) and sparse (8 views) input configurations. WorldMirror 1.0 produces reasonable point clouds at medium resolution (518×518), but suffers from severe geometric degradation at high resolution (1036×1036); under the dense 32-view setting, the point cloud structure collapses entirely. In contrast, WorldMirror 2.0 maintains stable and coherent reconstructions across all tested resolutions, directly validating the effectiveness of normalized position encoding (Sec. 6.2.1) for flexible resolution inference.

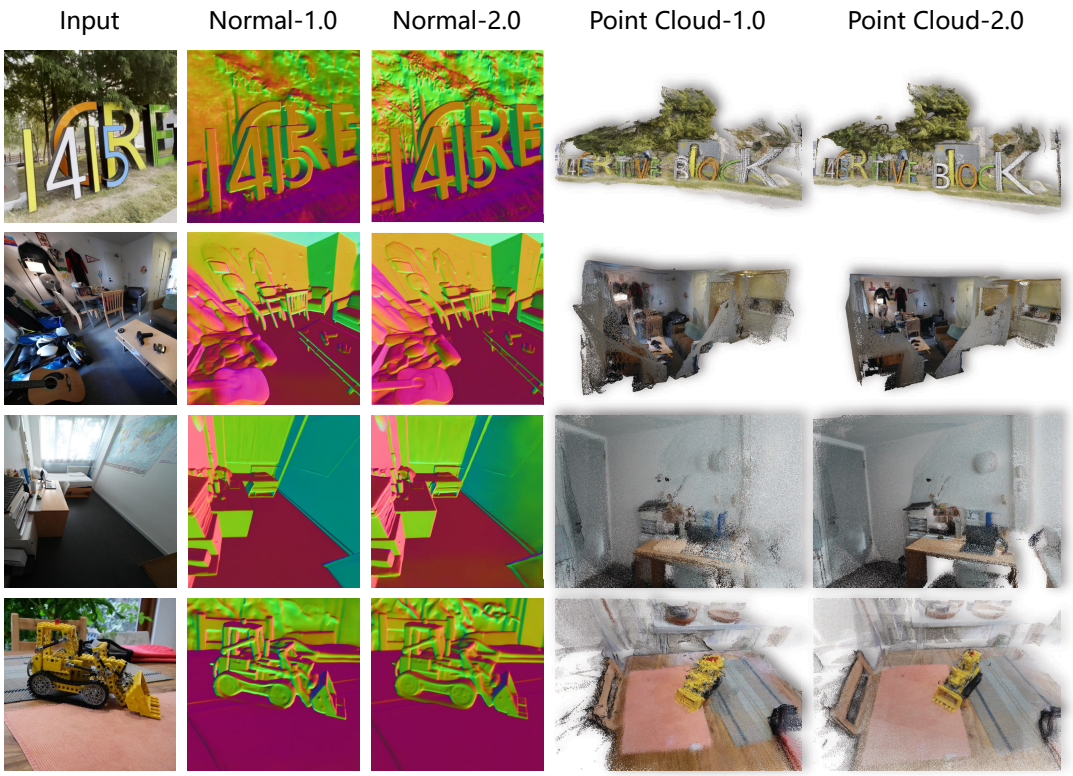


Figure 25: **Visual comparison of WorldMirror 1.0 and 2.0.** We compare predicted surface normals and reconstructed point clouds. WorldMirror 2.0 produces more accurate normals with finer structural details and more consistent multi-view point clouds.

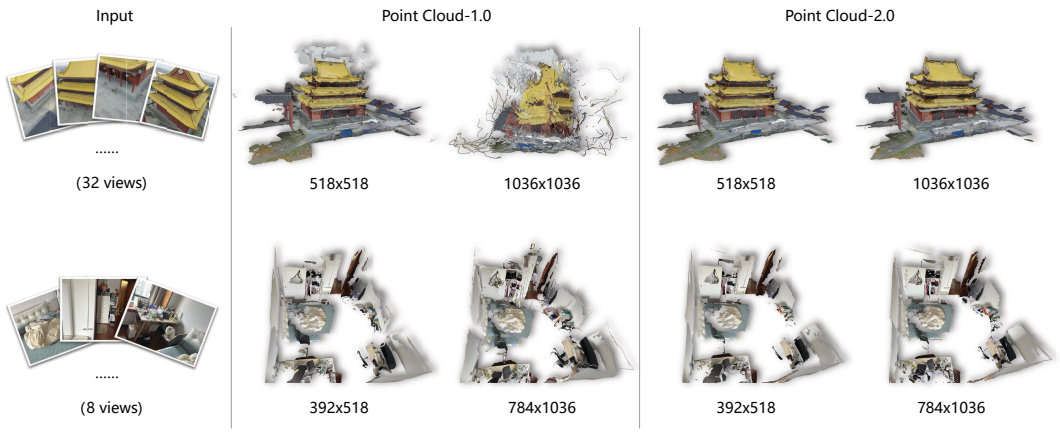


Figure 26: **Multi-resolution point cloud comparison of WorldMirror 1.0 and 2.0.** We evaluate under dense (32 views, top) and sparse (8 views, bottom) settings at different inference resolutions. WorldMirror 1.0 degrades severely at high resolution, while WorldMirror 2.0 maintains consistent reconstruction quality across all resolutions.

Table 13: **Surface normal estimation on ScanNet, NYUv2, and iBims-1.** We compare with both regression-based and diffusion-based approaches. ↑/↓ indicate higher-/lower-is-better. Baseline methods are evaluated at M resolution. WorldMirror 2.0 generalizes across multiple resolutions. "L/M/H" denote low / medium / high inference resolution. Best results are highlighted.

Method	ScanNet			NYUv2			iBims-1		
	mean ↓	med ↓	22.5° ↑	mean ↓	med ↓	22.5° ↑	mean ↓	med ↓	22.5° ↑
OASIS [11]	32.8	28.5	38.5	29.2	23.4	48.4	32.6	24.6	46.6
EESNU [2]	–	–	–	16.2	8.5	77.2	20.0	8.4	73.4
Omnidata v1 [16]	22.9	12.3	66.1	23.1	12.9	66.3	19.0	7.5	76.1
Omnidata v2 [31]	16.2	8.5	79.5	17.2	9.7	76.5	18.2	7.0	77.4
DSine [3]	16.2	8.3	78.7	16.4	8.4	77.7	17.1	6.1	79.0
GeoWizard [18]	16.7	9.5	78.3	19.5	11.7	74.5	20.4	9.4	76.4
StableNormal [82]	16.0	9.9	81.5	18.5	11.2	77.5	17.9	8.5	80.4
WorldMirror 1.0 (L)	14.4	7.4	81.5	16.0	8.2	78.7	19.0	7.2	76.3
WorldMirror 1.0 (M)	13.8	7.3	82.5	15.1	8.0	80.1	16.6	6.4	80.1
WorldMirror 1.0 (H)	17.6	12.5	76.2	19.1	13.3	72.4	19.2	10.9	76.9
WorldMirror 2.0 (L)	12.7	6.8	83.7	14.4	7.8	80.4	15.6	6.2	80.4
WorldMirror 2.0 (M)	12.3	6.5	84.3	13.9	7.6	81.4	14.2	5.6	82.4
WorldMirror 2.0 (H)	12.5	6.8	84.2	14.0	7.8	81.4	14.5	6.1	82.0

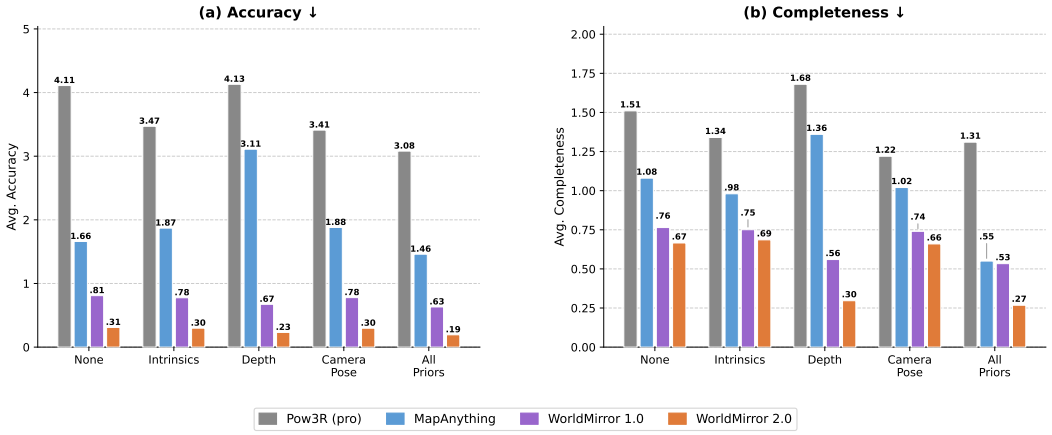


Figure 27: **Comparison with Pow3R and MapAnything under different prior conditions.** All methods are evaluated at high resolution (756×1036). Results are averaged on 7-Scenes, NRGBD, and DTU datasets. Pow3R (pro) refers to the original Pow3R with Procrustes alignment. WorldMirror 2.0 demonstrates stronger geometric reasoning and 3D prior integration capabilities at high resolution.

8.2.2 Inference-Time Evaluation

Geometric Prior Injection. A distinctive feature of WorldMirror is its ability to flexibly incorporate geometric priors, including camera poses, intrinsics, and depth maps, through Any-Modal Tokenization (Sec. 6.1). We compare WorldMirror 1.0 and 2.0 with prior-guided methods Pow3R [26] and MapAnything [32] under different prior conditions at high resolution (Fig. 27). WorldMirror 2.0 consistently outperforms all alternatives, with the largest improvements appearing in the camera-conditioned and all-priors settings. Camera poses capture the global geometric layout, calibrated intrinsics resolve metric scale ambiguity, and depth priors provide pixel-level constraints; combining all priors yields synergistic gains.

The practical benefit of this prior integration is further demonstrated in our world generation pipeline. As illustrated in Fig. 15, when conditioned on the same camera poses, WorldMirror 2.0 produces significantly more coherent and globally consistent point clouds than MapAnything [32] and DepthAnything3 [40], both of which exhibit noticeable structural inconsistencies. This confirms that WorldMirror 2.0's learned multi-modal tokenization enables more effective utilization of geometric cues, making it well-suited as the reconstruction backbone for world generation.

Table 14: **Inference efficiency of WorldMirror 2.0.** We report GPU memory (GB) and wall-clock time (s) for different numbers of input views at 518×378 resolution. "Baseline" refers to single-GPU FP32 inference (WorldMirror 1.0 setting). SP denotes Token/Frame Sequence Parallelism. All measurements on NVIDIA H20 GPUs.

Configuration	#GPUs	32 views		64 views		128 views		256 views	
		Mem. (GB)	Time (s)	Mem. (GB)	Time (s)	Mem. (GB)	Time (s)	Mem. (GB)	Time (s)
Baseline (FP32, 1 GPU)	1	24.95	2.45	38.56	6.27	59.26	18.00	OOM	OOM
+ BF16	1	15.10	2.11	25.06	5.65	41.73	16.96	75.05	56.96
+ SP (×2 GPUs)	2	26.32	1.59	41.31	3.73	64.75	10.53	OOM	OOM
+ SP (×4 GPUs)	4	25.55	1.09	39.71	2.38	61.53	6.27	OOM	OOM
+ SP + BF16 (×4 GPUs)	4	15.81	0.96	26.44	2.21	44.47	5.65	80.54	17.69
+ SP + BF16 + FSDP (×4 GPUs)	4	14.04	0.93	24.67	2.20	42.71	5.60	78.78	17.52

Inference Efficiency. We benchmark the inference efficiency optimizations of WorldMirror 2.0 introduced in Sec. 6.4. Tab. 14 reports per-GPU memory consumption and wall-clock inference time across different view counts at 518×378 resolution, measured on NVIDIA H20 GPUs. The single-GPU FP32 baseline runs out of memory (OOM) at 256 views. BF16 mixed-precision inference reduces per-GPU memory by approximately 40% (*e.g.*, 128 views: 59.26 GB→41.73 GB) and critically enables 256-view inference (75.05 GB) that is infeasible under FP32. Sequence parallelism (SP) provides substantial speedups by distributing computation across GPUs (*e.g.*, 128 views: 18.00s→6.27s with 4-GPU SP). The full combination of SP, BF16, and FSDP on 4 GPUs achieves the best trade-off: 128 views in 5.60s with 42.71 GB per GPU (a 3.2× speedup and 28% memory reduction over the baseline), and 256 views in 17.52s with 78.78 GB per GPU. These complementary strategies enable WorldMirror 2.0 to scale to substantially larger input configurations within practical memory and latency constraints.

9 Conclusion

In this report, we present **HY-World 2.0**, a comprehensive multi-modal world model framework that bridges the longstanding gap between 3D world generation and reconstruction. By dynamically adapting to diverse input modalities—ranging from sparse texts and single images to dense multi-view videos—our framework establishes a unified paradigm for offline 3D world modeling. To achieve this, we introduced a four-stage pipeline. We scaled up panorama generation (**HY-Pano 2.0**) for high-fidelity world initialization and designed a semantic-aware trajectory planning (**WorldNav**) to guide optimal, collision-free routes for scene exploration. Furthermore, we significantly upgraded our generative world expansion (**WorldStereo 2.0**) by operating in a keyframe-latent space with spatially consistent memory. Finally, we employ the world composition via our enhanced 3D reconstruction foundation (**WorldMirror 2.0**) to produce geometrically accurate and navigable 3DGS assets. We also propose a high-performance 3DGS rendering platform (**WorldLens**) to enable interactive exploration of 3D worlds with character support and lighting control. Extensive evaluations demonstrate that HY-World 2.0 achieves state-of-the-art performance among open-source approaches, delivering visual quality, geometric consistency, and exploratory capabilities that are highly competitive with leading closed-source commercial models.

Contribution

- **Project Lead:** Chunchao Guo, Tengfei Wang
- **Core Contributors:** Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Tengfei Wang
- **Contributors (in alphabetical order by first name):**
  - **Engineering & Infra:** Bo Yuan, Coopers Li, Fan Yang, Haiyu Zhang, Jianchen Zhu, Jie Xiao, Lei Wang, Minghui Chen, Penghao Zhao, Qi Chen, Wangchen Qin, Xiang Yuan, Yifu Sun, Yihang Lian, Yuyang Yin, Zhiyuan Min
  - **Data & Art Design:** Chao Zhang, Dongyuan Guo, Hang Cao, Jiaxin Lin, Jihong Zhang, Junlin Yu, Lifu Wang, Lilin Wang, Peng He, Rui Chen, Rui Shao, Sicong Liu, Xiaochuan Niu, Yifei Tang, Yi Sun, Yonghao Tan, Yuhong Liu
- **Project Sponsors:** Linus

References

[1] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pages 690–708. Springer, 2022.

[2] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13137–13146, 2021.

[3] Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9535–9545, 2024.

[4] Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. *arXiv preprint arXiv:2509.19296*, 2025.

[5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022.

[6] Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, and Yanwei Fu. Leftrefill: Filling right canvas based on left reference through generalized text-to-image diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7705–7715, 2024.

[7] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer++: Revealing the devil in transformer's details for multi-view stereo. In *International Conference on Learning Representations*, 2024.

[8] Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. 2025.

[9] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[10] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025.

[11] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 679–688, 2020.

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[13] Edsger W Dijkstra. A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: his life, work, and legacy*, pages 287–290. 2022.

[14] Jiahua Dong, Qi Lyu, Baichen Liu, Xudong Wang, Wenqi Liang, Duzhen Zhang, Jiahang Tu, Hongliu Li, Hanbin Zhao, Henghui Ding, et al. Learning to model the world: A survey of world models in artificial intelligence. 2026.

[15] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

[16] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021.

[17] Haoran Feng, Dizhe Zhang, Xiangtai Li, Bo Du, and Lu Qi. Dit360: High-fidelity panoramic image generation via hybrid training. *arXiv preprint arXiv:2510.11712*, 2025.

[18] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024.

[19] Google DeepMind. Genie 3: A new frontier for world models. https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/, 2025. Blog post, August 5, 2025.

[20] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[21] Lukas Höllein and Matthias Nießner. World reconstruction from inconsistent views. *arXiv preprint arXiv:2603.16736*, 2026.

[22] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

[23] Team HunyuanWorld. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint*, 2025.

[24] Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

[25] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[26] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1071–1081, 2025.

[27] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

[28] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025.

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[30] 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.

[31] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18974, 2022.

[32] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.

[33] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

[34] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23828–23838, 2025.

[35] 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.

[36] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[37] 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.

[38] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. *arXiv preprint arXiv:2506.18903*, 2025.

[39] Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. Flashworld: High-quality 3d scene generation within seconds. *arXiv preprint arXiv:2510.13678*, 2025.

[40] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[41] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024.

[42] Jiacheng Liu, Xinyu Wang, Yuqi Lin, Zhikai Wang, Peiru Wang, Peiliang Cai, Qinming Zhou, Zhengan Yan, Zexuan Yan, Zhengyi Shi, et al. A survey on cache methods in diffusion models: Toward efficient multi-modal generation. *arXiv preprint arXiv:2510.19755*, 2025.

[43] 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.

[44] Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025.

[45] Yifei Liu, Zhihang Zhong, Yifan Zhan, Sheng Xu, and Xiao Sun. Maskgaussian: Adaptive 3d gaussian representation from probabilistic masks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 681–690, 2025.

[46] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998.

[47] Taiming Lu, Tianmin Shu, Junfei Xiao, Luoxin Ye, Jiahao Wang, Cheng Peng, Chen Wei, Daniel Khashabi, Rama Chellappa, Alan Yuille, et al. Genex: Generating an explorable world. *arXiv preprint arXiv:2412.09624*, 2024.

[48] Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025.

[49] David Marr and Tomaso Poggio. Cooperative computation of stereo disparity: A cooperative algorithm is derived for extracting disparity information from stereo image pairs. *Science*, 194(4262):283–287, 1976.

[50] Mikko Mononen and the Recast Navigation Contributors. Recast navigation: State-of-the-art navmesh generation and navigation for games. Open-source software repository, 2009–2026. Accessed: [2026-02].

[51] 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.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[53] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[54] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[55] Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. *arXiv preprint arXiv:2506.01799*, 2025.

[56] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[57] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[58] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021.

[59] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[60] Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

[61] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

[62] Tencent Hunyuan 3D Team. Worldstereo: Bridging camera-guided video generation and scene reconstruction via 3d geometric memories, 2026.

[63] Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report, 2025.

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

[65] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

[66] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025.

[67] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025.

[68] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020.

[69] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. pi3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

[70] Zehan Wang, Tengfei Wang, Haiyu Zhang, Xuhui Zuo, Junta Wu, Haoyuan Wang, Wenqiang Sun, Zhenwei Wang, Chenjie Cao, Hengshuang Zhao, et al. Worldcompass: Reinforcement learning for long-horizon world models. *arXiv preprint arXiv:2602.09022*, 2026.

[71] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023.

[72] World Labs. Marble. https://marble.worldlabs.ai/, 2025. Accessed: 2026-03-30.

[73] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.

[74] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22389, 2024.

[75] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4389–4398, 2024.

[76] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[77] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025.

[78] Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. In *Proceedings of the special interest group on computer graphics and interactive techniques conference conference papers*, pages 1–10, 2025.

[79] Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

[80] Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, et al. Matrix-3d: Omnidirectional explorable 3d world generation. *arXiv preprint arXiv:2508.08086*, 2025.

[81] 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. In *International Conference on Learning Representations*, 2025.

[82] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43(6):1–18, 2024.

[83] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

[84] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025.

[85] Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. In *International Conference on Learning Representations*, 2025.

[86] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21936–21947, 2025.

[87] Jensen Jinghao Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv e-prints*, pages arXiv–2503, 2025.