

Seoul World Model: Grounding World Simulation Models in a Real-World Metropolis

Junyoung Seo^{*1}, Hyunwook Choi^{*1}, Minkyung Kwon¹, Jinhyeok Choi¹,
Siyeon Jin¹, Gayoung Lee², Junho Kim², JoungBin Lee¹, Geonmo Gu²,
Dongyoon Han^{2,1}, Sangdoo Yun^{2,3}, Seungryong Kim^{†1}, and Jin-Hwa Kim^{†2,3}

¹ KAIST AI

² NAVER AI Lab

³ SNU AIIS

<https://seoul-world-model.github.io/>

Abstract. What if a world simulation model could render not an imagined environment but a city that actually exists? Prior generative world models synthesize visually plausible yet artificial environments by imagining all content. We present **Seoul World Model (SWM)**, a city-scale world model grounded in the real city of Seoul. SWM anchors autoregressive video generation through retrieval-augmented conditioning on nearby street-view images. However, this design introduces several challenges, including temporal misalignment between retrieved references and the dynamic target scene, limited trajectory diversity and data sparsity from vehicle-mounted captures at sparse intervals. We address these challenges through cross-temporal pairing, a large-scale synthetic dataset enabling diverse camera trajectories, and a view interpolation pipeline that synthesizes coherent training videos from sparse street-view images. We further introduce a Virtual Lookahead Sink to stabilize long-horizon generation by continuously re-grounding each chunk to a retrieved image at a future location. We evaluate SWM against recent video world models across three cities: Seoul, Busan, and Ann Arbor. SWM outperforms existing methods in generating spatially faithful, temporally consistent, long-horizon videos grounded in actual urban environments over trajectories reaching hundreds of meters, while supporting diverse camera movements and text-prompted scenario variations.

Keywords: World Models · Video Generation · Digital Twin

1 Introduction

World models aim to learn internal representations of environments and predict their future states [12]. With recent advances in video generation, such models have rapidly evolved toward video world simulation, where sequences of frames

^{*}Equal contribution. [†]Co-corresponding authors. The first author conducted this work as part of the NAVER Cloud Residency Program.

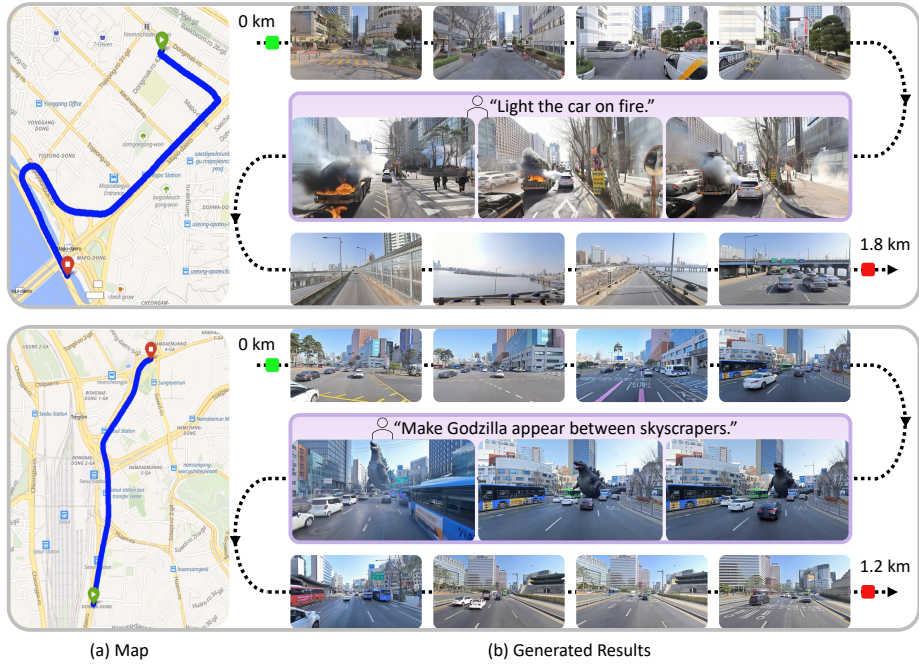


Fig. 1: Seoul World Model (SWM) generates videos over a kilometer grounded in a real city. A camera trajectory placed on a map produces continuous dynamic video depicting actual surroundings along the route. Users can further reshape the scene through text prompts, enabling imaginative scenarios.

are generated conditioned on images, text prompts, and user actions, treating each frame as a predicted state of a simulated world [1, 7, 8, 13, 20, 24, 32, 46, 61, 62]. These models can generate dynamic and interactive environments, including object motion, weather changes, and physical interactions. Yet they operate entirely within imagined worlds: given a starting image, everything beyond it, *e.g.*, the geometry of unseen streets, distant buildings, is imagined by the model.

What if a world model could operate on a world that physically exists? Users could navigate familiar city streets and experience hypothetical scenarios, such as a massive wave engulfing one’s own city, or exploring familiar streets under a golden sunset. In addition, such a real-world grounded simulation would enable urban planning visualization, autonomous driving scenario generation, and location-based exploration [9, 17, 39]. Yet this direction remains unexplored: while large-scale 3D reconstruction systems model real cities [29, 44], they are fundamentally static and lack generative simulation capabilities, and no world simulation model has been grounded in a specific real-world location.

We formalize this goal as **real-world grounded video world simulation** and instantiate it in Seoul, a large and densely structured metropolis, introducing **Seoul World Model (SWM)**. Our key observation is that widely available street-view photographs provide a scalable source of location-specific visual references. SWM fine-tunes a pretrained video world simulation model [1] on 440k

Seoul street-view images, real-world driving videos [43], and synthetic urban data. During generation, SWM performs retrieval-augmented generation: given geographic coordinates, camera actions, and text prompts, it retrieves nearby street-view images and conditions generation on complementary geometric and appearance references. This anchors each generated chunk to the real geometric layout and appearance of the location. Fig. 1 shows an example trajectory generated in Seoul with frames mapped to their corresponding city locations.

While retrieval-augmented grounding provides a natural way to anchor generation to real-world locations, it introduces three key challenges, each addressed by a corresponding design choice:

(1) Temporal misalignment. Street-view images capture a specific moment, while the simulated world should remain dynamic. Retrieved references may therefore contain transient elements inconsistent with the generated scene. We address this with **cross-temporal pairing**, which pairs references and targets from different timestamps during training, encouraging the model to disentangle persistent structure from transient content.

(2) Limited trajectory coverage and temporal sparsity. Real street-view data is captured by vehicle-mounted cameras at sparse intervals, restricting both trajectory types and temporal continuity. We construct a synthetic urban dataset using an Unreal-Engine-based simulator [10] that provides paired street-view references and target videos with diverse camera trajectories, including pedestrian paths. We additionally develop a view interpolation pipeline, namely an **intermittent freeze-frame** strategy, that synthesizes temporally coherent video between sparse street-view keyframes.

(3) Long-horizon error accumulation. Over long trajectories, autoregressive generation accumulates drift that weakens spatial grounding. Prior methods mitigate this with an attention sink, a fixed global context frame, typically the first frame, that persists throughout generation [28, 40]. However, this static anchor becomes less informative as the camera moves away from the starting locations. We instead propose a **virtual lookahead sink**: at each generation chunk, we retrieve a nearby street-view image and insert it at a future temporal position, acting as a virtual destination that re-anchors generation to upcoming locations, inspired by recent talking-head methods [21, 38].

SWM demonstrates that world simulation can be faithfully grounded in real, physically existing environments at city scale. We evaluate SWM across three cities: Seoul, Busan, and Ann Arbor, where the latter two cities are entirely absent from training, testing cross-city generalization without any fine-tuning. SWM outperforms recent video world models in visual quality, camera adherence, temporal coherence, and structural fidelity to real locations, and maintains stable generation over trajectories reaching hundreds of meters, demonstrating text-prompted scenarios and diverse camera trajectories.

2 Related Work

2.1 Video Generative Models

Video generation has advanced rapidly with diffusion models [16,35,41], enabling high-fidelity video synthesis. Early video diffusion models typically used UNet backbones with temporal modules [3,4,15], while more recent work [22,30,49,55] has shifted toward Diffusion Transformers [11,33] for improved scalability and quality. Recently, long-horizon video generation has emerged as a central target, motivating autoregressive and streaming formulations, where the model rolls out videos chunk-by-chunk while conditioning each new chunk on the generated context [6,18,28,40,56,57]. As generation extends, however, these models increasingly suffer from exposure bias and error accumulation. To address this, several methods [28,40,54,56] preserve long-range information with persistent global anchors such as attention sinks [53], which keep a fixed set of tokens and improve long-range temporal consistency without attending to the full history.

2.2 Video World Models

Building on the progress in video generation, video world simulation models [1,7,8,13,20,23–25,32,36,45,46,48,52,58,61,62] use a generative model as a dynamic model. Conditioned on past observations and actions, they predict future observations to simulate how the environment evolves [1,20]. Recent models generate interactive visual observations conditioned on user actions across diverse settings, including game environments [13,45,48], autonomous driving [23], and open-domain scenarios [20,32,46,62]. Action representations vary from discrete keyboard and mouse inputs [13,45] to continuous camera trajectories [8,23,52,58,61,62] and natural language instructions [32]. To maintain coherent world states over extended interactions, recent methods incorporate persistent memory beyond the local context window [7,25,52,58]. Despite these advances, existing world models operate entirely within imagined or synthetic environments, generating futures without grounding in external real-world observations. This becomes a key limitation when the simulated environment must stay faithful to a specific physical location.

2.3 Geometry-Aware Video Generation

A separate line of work incorporates 3D geometric reasoning into video generation to improve spatial consistency. In novel view synthesis, recent methods render point clouds from predicted depth to achieve geometric consistency for single-scene reconstruction [34,59]. Recent world models integrate geometry into autoregressive generation through joint video-3D prediction [7,8,61], 3D scene representations maintained across generation [24,50], and memory or spatial retrieval mechanisms that reuse previously generated context [7,25,36,52,58]. These approaches build geometric representations from the model’s predictions or generation history, and are typically focused on nearly static settings [25,34,50,59].

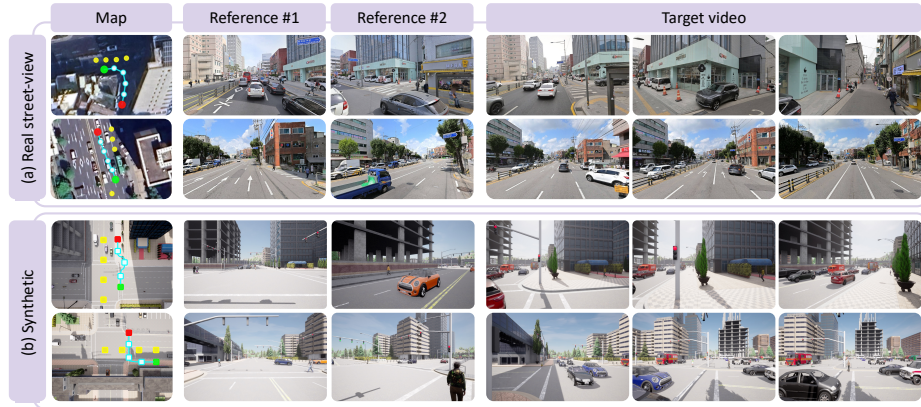


Fig. 2: Data overview: (a) real street-view and (b) synthetic datasets. The satellite-view map (leftmost column) shows the target video trajectory, with green and red indicating the start and end points, and yellow indicating reference view locations selected via cross-temporal pairing.

3 Data Construction

For SWM training, we build aligned pairs between street-view references and target video sequences. Each reference is associated with its camera pose and depth map, providing geometric conditions that ground the generated video to real-world geometric structure. We construct these pairs from two primary sources: real street-view images captured in Seoul (Sec. 3.1) and synthetic urban data from an Unreal-Engine-based simulator (Sec. 3.2). We additionally incorporate a publicly available driving video dataset [43] to increase scenario diversity. Fig. 2 shows examples from the real and synthetic datasets.

3.1 Street-View Dataset

Collection. We collect 1.2M panoramic images covering major urban areas of Seoul. Each image is associated with GPS coordinates and capture timestamps as metadata, obtained from a local map service provider with permission for research purposes. License plates and pedestrians are blurred for de-identification. After the processing steps below, 440K images are used for training.

Cross-temporal pairing. We define a training sequence as N consecutive street-view images along a route, which serves as the target sequence for supervision, and assign K spatially nearby panoramas as references that condition generation (Sec. 4.1). Each panorama is rendered into a pinhole view: training sequences face forward with a random horizontal rotation up to 90° , and references are rendered to match the viewing direction of the paired training frame.

A key design choice is **cross-temporal pairing**: references must be captured at a different timestamp from the target sequence. This mirrors inference, where retrieved street-view images share the same location as the target but often differ

in transient content such as vehicles or pedestrians. Without this constraint, co-captured references share identical transient content with the target, making persistent structure and transient objects indistinguishable; the model has no incentive to separate them and learns to reproduce both. Cross-temporal pairing removes this ambiguity during training: because transient content differs between reference and target, the model must learn to rely on persistent spatial structure that remains consistent across timestamps. Fig. 2(a) shows representative cross-temporal pairs; Fig. 6 visualizes the resulting attention pattern.

View interpolation. City-scale street-view databases provide panoramic images at sparse spatial intervals (typically 5–20 m between views) rather than continuous video. Training a video generation model directly on such sparse sequences is challenging, as pretrained video diffusion models learn to produce temporally smooth, continuous motion; abrupt jumps between distant view-points break this temporal continuity. We therefore develop an interpolation pipeline that synthesizes T -frame videos from N sparse keyframes ($T \gg N$), leveraging a pretrained latent video generative model [1].

A straightforward way to enable keyframe interpolation is to concatenate the keyframe latents along the channel dimension of the latents at their corresponding timestamps, while zero-padding the conditioning channels at non-keyframe timestamps, as shown in Fig. 3(a) (*e.g.* Wan2.1-FLF2V [49]). However, we observe that this approach yields weak adherence to the keyframes, with generated frames deviating from the inputs. We attribute this to a mismatch with the video 3D VAE’s temporal compression: the encoder compresses every 4 consecutive frames into a single latent, whereas an isolated keyframe does not form a valid 4-frame group.

To address this, we propose an **Intermittent Freeze-Frame** strategy that ensures each keyframe forms a complete 4-frame group matching the 3D VAE’s temporal stride. During training, the pixel frame at each keyframe position is repeated 4 consecutive times, so the 3D VAE encodes it into exactly one latent; the resulting training videos alternate between smooth motion and brief freeze segments. At inference, each given keyframe is similarly repeated 4 times and encoded into a single latent, which then replaces the corresponding position in the noisy input latent of the diffusion model. After generation and decoding, the three repeated frames per keyframe are discarded to recover the intended video, as illustrated in Fig. 3(b). Quantitative results are in Appendix C.1.

Annotation. We generate text captions for all videos using Qwen2.5-VL-72B [2] and augment them with predefined camera actions (straight, stop, left turn, right turn). While GPS metadata provides approximate positions, it lacks sufficient accuracy and does not include camera pose information. We use Depth Anything V3 [27] to estimate per-keyframe depth maps and camera poses, and align them to real-world scale using GPS metadata. Details are provided in Appendix A.2.

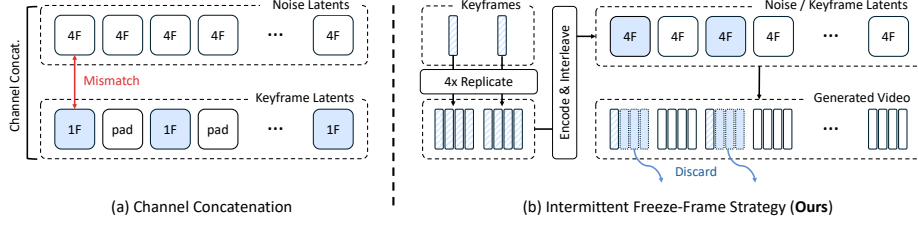


Fig. 3: View interpolation pipeline: (a) Keyframe conditioning via channel concatenation, and (b) Keyframe conditioning with Intermittent Freeze-Frame strategy (**Ours**). 4F and 1F denote latents derived from four frames and one frame, respectively, prior to the $4\times$ temporal compression of the 3D VAE.

3.2 Synthetic Dataset

To complement the driving-like trajectories of real street-view data with diverse camera paths, we construct a synthetic dataset from CARLA [10], an Unreal Engine-based urban simulator. We render 10K videos from 5 urban maps spanning approximately $431,500 \text{ m}^2$ of city area across three trajectory types:

1. **Pedestrian trajectories:** first-person videos rendered from autonomous pedestrian agents, covering sidewalk movement, street crossing, and similar on-foot paths.
2. **Vehicle trajectories:** driving-perspective videos captured across diverse road types, including highways, urban streets, and elevated roads. The trajectories cover lane changes, turns, and straight driving.
3. **Free-camera trajectories:** random paths that freely navigate the scene while avoiding collisions with buildings, terrain, and other scene geometry.

Street-view reference. For each map, we render street-view reference images at regular intervals of 10m along all roads, with eight directional views (uniformly covering 360° horizontal view) for each location. Following the same cross-temporal pairing principle as the real data, reference images and target video sequences are rendered at different simulated timestamps. Examples are shown in Fig. 2(b); additional details are in Appendix A.3.

4 Model

SWM generates videos grounded in a real city through retrieval-augmented conditioning from a user-specified starting location, camera motion, and a text prompt. We build on a pretrained Diffusion Transformer (DiT) [1, 33] that operates in a latent space compressed from pixel-space T frames $\mathbf{X} = \{x_t\}_{t=0}^{T-1}$ via a 3D VAE. Generation proceeds autoregressively in chunks with frame length T . For the i -th chunk, the model receives a camera trajectory $\mathbf{C}^{(i)} = \{c_t\}_{t=0}^{T-1}$, a text prompt $P^{(i)}$, and noisy latents $\mathbf{Z}^{(i)} = \{z_l^{(i)}\}_{l=0}^{L-1}$ to produce target latents $\hat{\mathbf{Z}}^{(i)} = \{\hat{z}_l^{(i)}\}_{l=0}^{L-1}$, where L is the number of compressed latents per chunk. Each subsequent chunk additionally conditions on H history latents $\mathbf{Z}_{\text{hist}}^{(i)} = \{\hat{z}_l^{(i-1)}\}_{l=L-H}^{L-1}$ from the tail of the preceding chunk’s output, providing temporal continuity.

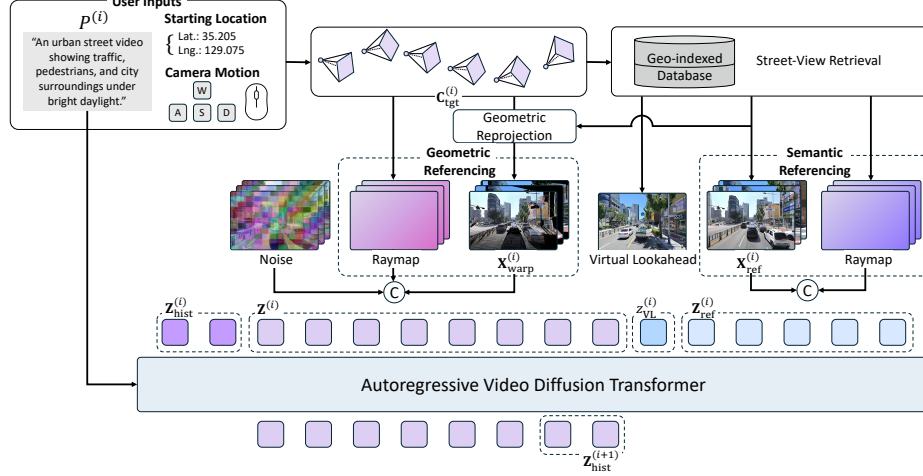


Fig. 4: Model overview. Given a start location, SWM autoregressively generates video grounded in a real city based on text prompt $P^{(i)}$, and target camera trajectory $\mathbf{C}^{(i)}$, retrieving the relevant street-view images from a geo-indexed database. These retrieved images provide a Virtual Lookahead Sink for long-horizon stability and serve as conditioning for geometric referencing and semantic referencing to ground the generation in real-world geometry and appearance. The model generates each chunk autoregressively conditioned on self-generated history.

For each chunk, nearby street-view images are retrieved from a geo-indexed database (Sec. 4.1). These retrieved images serve two roles: as a virtual lookahead sink (Sec. 4.2) that prevents error accumulation in city-scale long-horizon generation, and as conditioning for geometric and semantic referencing (Sec. 4.3) that grounds the generated video to the geometry and appearance of real locations. Since our retrieval-augmented framework is orthogonal to the training strategy for autoregressive generation, we evaluate it under Teacher Forcing [51] and Self-Forcing [18] as two separate configurations. Fig. 4 provides an architectural overview.

4.1 Street-View Retrieval

The retrieval database consists of 1.2M panoramic images covering Seoul. Each panorama is rendered into 8 equi-angular pinhole views, with metric-scale depth maps and 6-DoF camera poses estimated via Depth Anything V3 [27], and aligned to real-world scale using GPS metadata, following the same preprocessing as the training data (Sec. 3.1). For each i -th target chunk, given target camera trajectory $\mathbf{C}^{(i)}$, we retrieve reference images in two stages: (1) nearest-neighbor search identifies candidate street-view locations along the target trajectory, and (2) depth-based reprojection filtering retains only those whose projected pixels exceed a coverage threshold in the nearest target view. This yields up to K pinhole references $\mathbf{X}_{\text{ref}}^{(i)} = \{x_{\text{ref},k}^{(i)}\}_{k=0}^{K-1}$ with their camera poses $\mathbf{C}_{\text{ref}}^{(i)} = \{c_{\text{ref},k}^{(i)}\}_{k=0}^{K-1}$

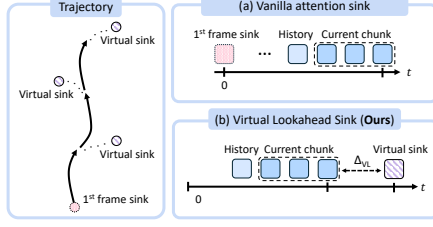


Fig. 5: Virtual Lookahead Sink: (a) Vanilla attention sink and (b) virtual lookahead sink (**Ours**). (a) anchors to the initial frame, whose guidance weakens as the camera moves farther away. (b) dynamically retrieves the nearest street-view as a virtual future destination.

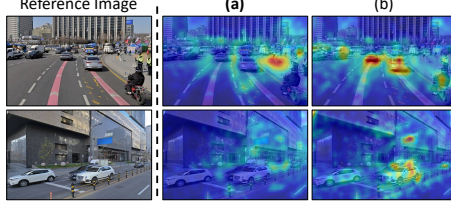


Fig. 6: Attention scores on references. (a) Ours with cross-temporal pairing, and (b) Ours without cross-temporal pairing. Cross-temporal pairing makes the model less attentive to dynamic objects (*e.g.*, cars) at the reference.

and depth estimates $\mathbf{D}_{\text{ref}}^{(i)} = \{d_{\text{ref},k}^{(i)}\}_{k=0}^{K-1}$, each aligned to the viewing direction of the matched target viewpoint.

4.2 Virtual Lookahead Sink

Autoregressive generation accumulates errors across chunks, as each step feeds the last few output latents as history latents to generate the next chunk. At the city scale, where the camera may travel hundreds of meters, per-chunk drift compounds into misalignment between retrieved references and the generated scene. We observe that world models trained with forcing-based distillation [6, 18] still degrade under these conditions. Prior work mitigates long-horizon degradation by maintaining an attention sink (Fig. 5(a)), typically the initial frame, as a fixed global context throughout generation [28, 40]. However, this static anchor becomes increasingly irrelevant as the camera moves farther from the starting point in our scenario.

To address this, we propose a **Virtual Lookahead (VL) Sink**, tailored for retrieval-augmented long-horizon generation, dynamically updating the sink with a retrieved street-view image. Specifically, given the target trajectory end point $c_{T-1}^{(i)}$ of each chunk, we retrieve the nearest street-view image to this end-point and treat it as a virtual future destination. By placing a clean, error-free frame ahead of the chunk being generated, the model has a stable anchor to converge toward; retrieving this anchor from a spatially nearby location further ensures that the grounding remains relevant to the region being generated. Because the anchor is not a reconstruction target, it need not coincide with the exact future trajectory; each chunk refreshes it during generation. Fig. 5(b) illustrates this mechanism.

We encode the retrieved image into a single latent $z_{\text{VL}}^{(i)}$ and assign it a RoPE [42] temporal position beyond the current generation chunk. The latent sequence fed to the model $\mathbf{Z}_{\text{seq}}^{(i)}$ and its RoPE temporal positions $\mathbf{p}_{\text{seq}}^{(i)}$ are given

by:

$$\mathbf{Z}_{\text{seq}}^{(i)} = [\mathbf{Z}_{\text{hist}}^{(i)}; \mathbf{Z}^{(i)}; z_{\text{VL}}^{(i)}], \quad \text{and} \quad (1)$$

$$\mathbf{p}_{\text{seq}}^{(i)} = \left[\underbrace{1, \dots, H}_{\text{history}}; \underbrace{H+1, \dots, H+L}_{\text{target}}; \underbrace{H+L+\Delta_{\text{VL}}}_{\text{sink}} \right],$$

where Δ_{VL} is a temporal offset hyperparameter. During training, a ground-truth future frame is sampled at a random temporal offset, exposing the model to varying lookahead distances so that it learns how the anchor’s proximity affects generation [38]; at inference, Δ_{VL} is fixed and the ground-truth frame is replaced by a retrieved street-view image.

4.3 Geometric and Semantic Referencing

Since each retrieved reference and the target observe the same scene from known camera poses, their geometric relationship enables two forms of conditioning. Geometric warping reprojects a reference into the target viewpoint, providing dense spatial layout cues, but loses fine appearance detail due to depth errors and occlusion [24, 34, 37, 52]. Conversely, injecting the original reference preserves appearance but lacks explicit spatial alignment. We therefore condition generation through two complementary pathways: geometric referencing for spatial layout and semantic referencing for appearance detail. Given K retrieved street-view images $\mathbf{X}_{\text{ref}}^{(i)}$ with poses $\mathbf{C}_{\text{ref}}^{(i)}$ and depth estimates $\mathbf{D}_{\text{ref}}^{(i)}$, the two pathways operate as follows.

Geometric referencing. For each target frame $x_t^{(i)}$, we reproject the spatially nearest reference $x_{\text{ref},j}^{(i)}$ into the target viewpoint via depth-based forward splatting [34], to get warped target image $x_{\text{warp},t}^{(i)}$:

$$x_{\text{warp},t}^{(i)} = \text{Render}(\text{Unproj}(x_{\text{ref},j}^{(i)}, d_{\text{ref},j}^{(i)}), c_{\text{ref},j \rightarrow t}^{(i)}), \quad (2)$$

where $c_{\text{ref},j \rightarrow t}^{(i)}$ is the relative camera transformation, $\text{Unproj}(\cdot)$ lifts the reference image to 3D using depth, and $\text{Render}(\cdot)$ projects the 3D points into the target view. Each target frame uses only its single nearest reference to avoid depth conflicts across multiple warped views. The warped video $\mathbf{X}_{\text{warp}}^{(i)} = \{x_{\text{warp},t}^{(i)}\}_{t=0}^{T-1}$ is encoded by the 3D VAE and channel-wise concatenated with the noisy target latent at the DiT input.

Semantic referencing. To preserve appearance detail, the original references are injected into the transformer’s latent sequence. Each $x_{\text{ref},k}^{(i)}$ is encoded into a single latent $z_{\text{ref},k}^{(i)}$, patch-embedded, and concatenated with the target latents along the temporal axis at RoPE position $p_{\text{ref},k}^{(i)} = H+L+G+k\Delta_{\text{ref}}$, where G is a large temporal gap separating references from the generation window and Δ_{ref} is the inter-reference spacing. Unlike geometric referencing, which uses only the nearest reference per frame, semantic referencing allows each target

latent to attend to all K references, enabling the model to gather complementary appearance cues. Camera poses for all latents, including target, reference, and sink, are encoded via Plücker ray embeddings, projected into the latent space through a convolutional encoder, and concatenated with the latent channels.

Since references and targets are captured at different times, dynamic objects in the references need not match those in the generated frame. Instead of introducing explicit mechanisms for this, we leverage the cross-temporal pairing strategy (Sec. 3.1), which encourages the model to focus on persistent scene structure while ignoring transient objects, as shown in Fig. 6.

5 Experiments

5.1 Implementation Details

Model and training setup. SWM fine-tunes Cosmos-Predict2.5-2B [1]. We train with AdamW [31] with a learning rate of $4.8\text{e-}5$ for 10K iterations at a total batch size of 48 across 24 NVIDIA H100 GPUs. For the Self-Forcing (SF) variant [18], we first perform ODE initialization with 1K pairs for 6K steps from the Teacher-Forcing (TF) model checkpoint, followed by 10K iterations of fine-tuning. Under TF, each chunk ($T=77$ frames) conditions on $H=5$ ground-truth history latents during training with $K=5$ references and $G=50$ gap; at inference, the history is replaced by self-generated output. Under SF, the model generates from self-produced history ($H=3$) in KV cache, with shorter chunks (12 frames) and $K=1$, achieving 15.2 fps with a single H100 GPU. Both configurations apply the Virtual Lookahead Sink with $\Delta_{\text{VL}}=5$. Further details are in Appendix A.1.

Evaluation benchmarks. For evaluation, we construct two benchmark datasets, Busan-City-Bench and Ann-Arbor-City-Bench (from the MARS dataset [26]), since our model is trained on Seoul data. Each benchmark contains 30 test sequences, each consisting of 365 frames (approximately 100 meters each). References are retrieved from nearby locations but exclude any street-view image belonging to the test sequence itself, ensuring that the model cannot access the ground-truth viewpoint during generation.

Baselines. We introduce a new task, real-world grounded world simulation, a setting that requires inputs not fully supported by existing world models. We therefore evaluate recent video world models capable of generating dynamic environments by providing each baseline with its supported inputs; per-baseline adaptation details are in Appendix B.2.

Metrics. We evaluate generation quality from three aspects. **Visual and temporal quality** is measured with FID [14] and FVD [47], and we report Image Quality from VBench [19]. **Camera-following accuracy** is measured with Rotation Error (RotErr) and Translation Error (TransErr), which quantify how

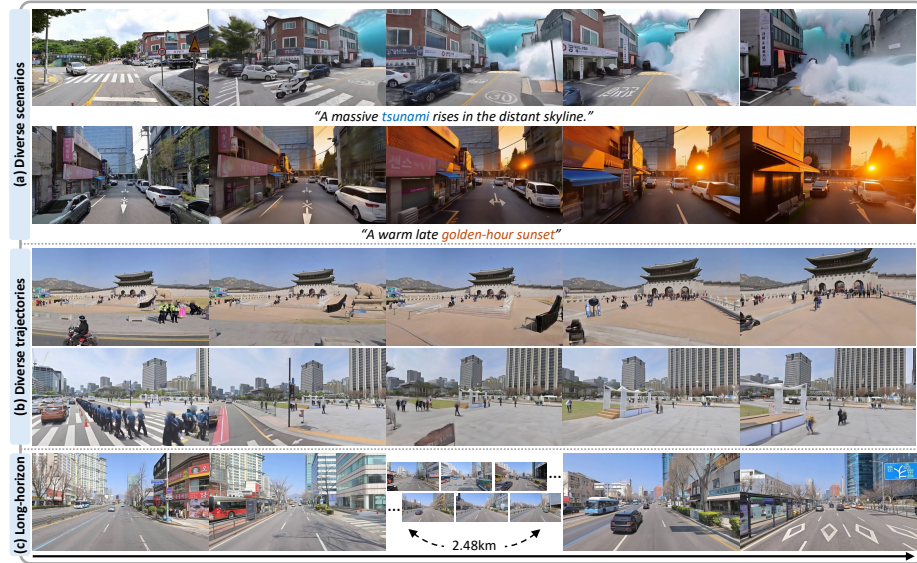


Fig. 7: Qualitative results. Grounded in a real-world city, SWM can (a) generate high-fidelity videos across diverse scenarios from user-guided text prompts, (b) remain robust to diverse camera trajectories, and (c) reduce error accumulation, enabling long-horizon generation over several kilometers.

accurately the generated camera motion follows the target trajectory. **3D adherence** is evaluated with masked PSNR and LPIPS [60] computed only on static regions, since the generated dynamics do not need to match the ground-truth, by applying SAM3 [5] to segment moving objects.

5.2 Results

Generation results. Fig. 7 demonstrates the capabilities of SWM across diverse scenarios, trajectories, and long-horizon generation in Seoul city. Fig. 7(a) shows that, despite grounding generation in real-world references, SWM remains controllable via text prompts and can produce diverse scene conditions while preserving the underlying city layout. Fig. 7(b) demonstrates that training with synthetic urban data enables SWM to follow trajectories beyond standard driving paths, including pedestrian-style camera motions. Fig. 7(c) shows stable long-horizon generation, maintaining spatial consistency without noticeable error accumulation.

Comparison with other models. As shown in Tab. 1, SWM achieves the best performance on both Busan-City-Bench and Ann-Arbor-City-Bench [26] across visual and temporal fidelity, camera-following accuracy, and 3D adherence to real locations. In contrast, existing world models often drift over long trajectories, leading to misalignment in both camera motion and scene structure and resulting in blurred videos, reduced motion, or complete collapse. By leveraging



Fig. 8: Qualitative comparisons with other methods. Across Busan- and Ann-Arbor-City-Bench, SWM produces physically grounded videos consistent with real urban structure and camera motion. Other world models, not originally designed for real-world grounding, tend to struggle with scene coherence, trajectory alignment, or long-horizon stability in this setting. (Best viewed when zoomed in.)

Table 1: Quantitative comparison with other methods. We evaluate visual and temporal fidelity, camera-following accuracy, and 3D adherence. Values are reported as Busan-City-Bench / Ann-Arbor-City-Bench.

Method	FID↓	FVD↓	Img.Q.↑	RotErr↓	TransErr↓	mPSNR↑	mLPIPS↓
Aether [61]	141.24/132.77	1096.50/1214.84	0.55/0.51	0.030/ 0.078	0.083/ 0.192	11.10/13.03	0.671/0.635
DeepVerse [7]	130.32/182.95	892.63/1524.97	0.53/0.46	0.062/0.251	0.103/0.469	12.20/13.43	0.679/0.727
Yume1.5 [32]	54.82/85.62	425.24/993.62	0.73/ 0.61	0.153/0.326	0.104/0.271	12.09/14.15	0.667/0.623
HY-World1.5 [20]	49.63/67.02	544.04/864.76	0.78 /0.54	0.044/0.193	0.079/0.221	11.87/ 14.26	0.588/0.575
FantasyWorld [8]	83.51/67.72	783.11/917.57	0.63/0.49	0.056/0.215	0.141/0.302	10.01/11.97	0.654/0.592
Lingbot [46]	62.14/57.99	717.44/1039.50	0.75/0.60	0.081/0.269	0.073/0.239	10.48/12.51	0.645/0.641
SWM (TF)	28.43 / <u>56.61</u>	301.76 / 640.17	0.78 / 0.66	0.020 / 0.055	0.015 / 0.154	14.56 / 15.18	0.392 / 0.481
SWM (SF)	<u>32.50</u> / 43.97	<u>325.87</u> / <u>779.94</u>	<u>0.77</u> /0.57	<u>0.028</u> /0.217	<u>0.033</u> /0.208	<u>13.52</u> /14.20	<u>0.478</u> /0.573

retrieved images, SWM remains anchored to real-world scene layout and preserves alignment with the target trajectory, demonstrating stronger real-world grounding, as illustrated in Fig. 8.

5.3 Ablation Study

Tab. 2 summarizes ablations on three components of SWM, including dataset construction, the referencing strategy and the attention sink design. Representative qualitative ablations of the conditioning strategies are shown in Fig. 9, with additional examples provided in Appendix C.2.

Effect of dataset construction. Among all variants, removing cross-temporal pairing leads to the largest degradation across metrics, indicating that the model fails to disregard dynamic objects that are mismatched between retrieved references and generated frames. Removing synthetic data slightly improves FID, but substantially harms camera-following accuracy and 3D adherence since the model no longer learns diverse trajectories during training.

Table 2: Ablation on data strategy, conditioning strategy, and attention sink design. We evaluate visual and temporal fidelity, camera-following accuracy, and 3D adherence. Values are reported as Busan-City-Bench.

Variant	FID↓	FVD↓	Img.Q.↑	RotErr↓	TransErr↓	mPSNR↑	mLPIPS↓
Full model	<u>28.43</u>	301.76	<u>0.78</u>	<u>0.020</u>	0.015	14.56	0.392
w/o cross-temporal pairing	44.74	487.87	0.77	0.057	0.123	12.54	0.519
w/o synthetic data	27.74	365.24	<u>0.78</u>	0.021	0.020	13.52	0.427
w/o real street-view data	29.82	467.58	0.77	0.059	0.050	13.99	0.411
w/o geometric referencing	33.01	398.74	0.79	0.036	0.051	12.33	0.525
w/o semantic referencing	30.27	<u>326.18</u>	<u>0.78</u>	0.032	0.022	14.08	0.442
w/o any attention sink	33.06	342.81	<u>0.78</u>	0.021	<u>0.016</u>	14.16	0.406
w/ first frame attention sink	32.71	378.92	<u>0.78</u>	0.018	0.018	14.25	<u>0.388</u>
w/ first position attention sink	32.41	354.61	<u>0.78</u>	0.026	0.027	<u>14.35</u>	0.379

Effect of referencing. Ablations on the conditioning pathways confirm that geometric and semantic referencing play complementary roles. Geometric referencing supports camera alignment and static structural consistency, while semantic referencing improves appearance fidelity by injecting visual details. Removing either pathway degrades overall quality, yielding suboptimal results.

Effect of attention sink. We compare four configurations: (1) our full model with the Virtual Lookahead (VL) Sink, (2) without any attention sink, (3) with a conventional First-Frame (FF) attention sink, and (4) with a retrieved street-view image used as a conventional First-Position (FP) sink. As shown in Tab. 2 and Fig. 10, removing the sink causes drift in camera motion and scene structure, while FF and FP sinks reduce this drift but remain limited as the camera moves far from the anchor. The VL Sink achieves the lowest sliding-window FID and the slowest degradation over time.

6 Conclusion

We presented Seoul World Model, a video world model that grounds generation in a real city through retrieval-augmented conditioning on street-view images. Cross-temporal pairing, synthetic urban data, and a Virtual Lookahead Sink collectively address the temporal, spatial, and long-horizon challenges of city-scale grounding. We hope this work encourages further exploration of world simulation that operates in the physical world beyond imagined environments.

Acknowledgements

We would like to sincerely thank everyone at NAVER and NAVER Cloud who contributed their time, expertise, and feedback throughout this work. We are



Fig. 9: Qualitative ablation results. (a) Our full model, (b) ours without VL sink, and (c) ours without semantic referencing. VL Sink prevents error accumulation over long trajectories, while semantic referencing preserves appearance details.

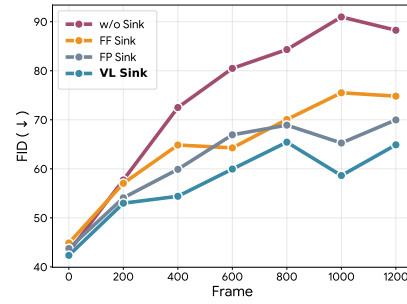


Fig. 10: Performance over time. We present sliding window FID with a 200-frame window size for different attention sink strategies. Our VL Sink achieves the lowest FID.

grateful to Jinbae Im, Moonbin Yim, and Bado Lee for their help with data pre-processing. We also thank Jongchae Na, Hyunjoon Cho, Hochul Hwang, Myoung-Suk Chae, and Kwangkwan Kim for their support with data processing and guidance on the use of map data and metadata. Finally, we thank Jonghak Kim, Jieun Shin, and Hyeun Shin for valuable discussions and thoughtful feedback. Their support and collaboration are greatly appreciated. This research was supported by the NAVER Cloud Residency Program and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279) and National Research Foundation of Korea (RS-2024-00346597).

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv e-prints (2025)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
6. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. NeurIPS (2024)

7. Chen, J., Zhu, H., He, X., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Fu, Z., Pang, J., et al.: Deepverse: 4d autoregressive video generation as a world model. arXiv preprint arXiv:2506.01103 (2025)
8. Dai, Y., Jiang, F., Wang, C., Xu, M., Qi, Y.: Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. arXiv preprint arXiv:2509.21657 (2025)
9. Deng, B., Tucker, R., Li, Z., Guibas, L., Snavely, N., Wetzstein, G.: Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In: SIGGRAPH (2024)
10. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: CoRL (2017)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
12. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)
13. He, X., Peng, C., Liu, Z., Wang, B., Zhang, Y., Cui, Q., Kang, F., Jiang, B., An, M., Ren, Y., et al.: Matrix-game 2.0: An open-source real-time and streaming interactive world model. arXiv preprint arXiv:2508.13009 (2025)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS (2017)
15. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
17. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)
18. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
19. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR (2024)
20. HunyuanWorld, T.: Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. arXiv preprint (2025)
21. Jiang, J., Zeng, W., Zheng, Z., Yang, J., Liang, C., Liao, W., Liang, H., Zhang, Y., Gao, M.: Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. arXiv preprint arXiv:2508.19209 (2025)
22. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
23. Li, B., Ma, Z., Du, D., Peng, B., Liang, Z., Liu, Z., Ma, C., Jin, Y., Zhao, H., Zeng, W., et al.: Omninwm: Omniscient driving navigation world models. arXiv preprint arXiv:2510.18313 (2025)
24. Li, G., Zheng, S., Xu, S., Chen, J., Li, B., Hu, X., Zhao, L., Jiang, P.T.: Magicworld: Interactive geometry-driven video world exploration. arXiv preprint arXiv:2511.18886 (2025)
25. Li, R., Torr, P., Vedaldi, A., Jakab, T.: Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In: ICCV (2025)

26. Li, Y., Li, Z., Chen, N., Gong, M., Lyu, Z., Wang, Z., Jiang, P., Feng, C.: Multiagent multitraversal multimodal self-driving: Open mars dataset. In: CVPR (2024)
27. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
28. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025)
29. Liu, Y., Luo, C., Fan, L., Wang, N., Peng, J., Zhang, Z.: Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In: ECCV (2024)
30. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177 (2024)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
32. Mao, X., Li, Z., Li, C., Xu, X., Ying, K., He, T., Pang, J., Qiao, Y., Zhang, K.: Yume-1.5: A text-controlled interactive world generation model. arXiv preprint arXiv:2512.22096 (2025)
33. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
34. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: CVPR (2025)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
36. Schneider, M.A., Höllein, L., Nießner, M.: Worldexplorer: Towards generating fully navigable 3d scenes. In: SIGGRAPH Asia (2025)
37. Seo, J., Fukuda, K., Shibuya, T., Narihira, T., Murata, N., Hu, S., Lai, C.H., Kim, S., Mitsufuji, Y.: Genwarp: Single image to novel views with semantic-preserving generative warping. NeurIPS (2024)
38. Seo, J., Mira, R., Haliassos, A., Bounareli, S., Chen, H., Tran, L., Kim, S., Landgraf, Z., Shen, J.: Lookahead anchoring: Preserving character identity in audio-driven human animation. arXiv preprint arXiv:2510.23581 (2025)
39. Shang, Y., Lin, Y., Zheng, Y., Fan, H., Ding, J., Feng, J., Chen, J., Tian, L., Li, Y.: Urbanworld: An urban world model for 3d city generation. arXiv preprint arXiv:2407.11965 (2024)
40. Shin, J., Li, Z., Zhang, R., Zhu, J.Y., Park, J., Shechtman, E., Huang, X.: Motion-stream: Real-time video generation with interactive motion controls. arXiv preprint arXiv:2511.01266 (2025)
41. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
42. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing (2024)
43. Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020)
44. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: CVPR (2022)
45. Tang, J., Liu, J., Li, J., Wu, L., Yang, H., Zhao, P., Gong, S., Yuan, X., Shao, S., Lu, Q.: Hunyuan-gamecraft-2: Instruction-following interactive game world model. arXiv preprint arXiv:2511.23429 (2025)

46. Team, R., Gao, Z., Wang, Q., Zeng, Y., Zhu, J., Cheng, K.L., Li, Y., Wang, H., Xu, Y., Ma, S., et al.: Advancing open-source world models. arXiv preprint arXiv:2601.20540 (2026)
47. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
48. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. In: ICLR (2025)
49. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
50. Wang, Z., Lin, H., Yoon, J., Cho, J., Zhang, Y., Bansal, M.: Anchorweave: World-consistent video generation with retrieved local spatial memories. arXiv preprint arXiv:2602.14941 (2026)
51. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. *Neural Computation* (1989)
52. Wu, T., Yang, S., Po, R., Xu, Y., Liu, Z., Lin, D., Wetzstein, G.: Video world models with long-term spatial memory. arXiv preprint arXiv:2506.05284 (2025)
53. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: ICLR (2024)
54. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al.: Longlive: Real-time interactive long video generation. arXiv preprint arXiv:2509.22622 (2025)
55. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
56. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. arXiv preprint arXiv:2512.05081 (2025)
57. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: CVPR (2025)
58. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: Context as memory: Scene-consistent interactive long video generation with memory retrieval. In: SIGGRAPH Asia (2025)
59. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. In: TPAMI (2025)
60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)
61. Zhu, H., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Chen, J., Shen, C., Pang, J., He, T.: Aether: Geometric-aware unified world modeling. In: ICCV (2025)
62. Zhu, Y., Feng, J., Zheng, W., Gao, Y., Tao, X., Wan, P., Zhou, J., Lu, J.: Astra: General interactive world model with autoregressive denoising. arXiv preprint arXiv:2512.08931 (2025)