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

**Junyoung Seo** [*1] **Hyunwook Choi** [*1] **Minkyung Kwon** [1] **Jinhyeok Choi** [1] **Siyoon Jin** [1]
**Gayoung Lee** [2] **Junho Kim** [2] **JoungBin Lee** [1] **Geonmo Gu** [2] **Dongyoon Han** [2 1]
**Sangdoo Yun** [2 3] **Seungryong Kim** [1] **Jin-Hwa Kim** [2 3]
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 challenges including temporal misalignment between retrieved references and the dynamic target scene, limited trajectory diversity and data sparsity from vehicle-mounted captures, and long-horizon error accumulation. We address these challenges through cross-temporal pairing, a synthetic urban dataset and view interpolation pipeline, and a Virtual Lookahead Sink that continuously re-grounds each chunk to a retrieved future reference. SWM outperforms recent video world models across Seoul, Busan, and Ann Arbor while supporting diverse camera movements and text-prompted scenario variations.

## 1. Introduction

World models aim to learn internal representations of environments and predict their future states (Ha & Schmidhuber, 2018). With recent advances in video generation, such models have rapidly evolved toward video world simulation, where sequences of frames are generated conditioned on images, text prompts, and user actions (Agarwal et al., 2025; Team et al., 2026; Zhu et al., 2025b; He et al., 2025; HunyuanWorld, 2025; Mao et al., 2025; Chen et al., 2025; Dai et al., 2025; Zhu et al., 2025a; Li et al., 2025). 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. Such a real-world grounded simulation would enable urban planning visualization, autonomous driving scenario generation, and location-based exploration (Deng et al., 2024; Hu et al., 2023; Shang et al., 2024). Yet this direction remains unexplored: while large-scale 3D reconstruction systems model real cities (Liu et al., 2024; Tancik et al., 2022), they are fundamentally static and lack generative simulation capabilities.

As illustrated in Figure 1, we formalize this goal as **real-world grounded video world simulation** and instantiate it in Seoul, introducing **Seoul World Model (SWM)**. Our key observation is that widely available street-view photographs provide a scalable source of location-specific visual references. 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.

While retrieval-augmented grounding naturally anchors generation to real-world locations, it introduces three key challenges: temporal misalignment between street-view references and the dynamic simulated world, limited trajectory coverage and temporal sparsity from vehicle-mounted captures, and long-horizon error accumulation. We address these with cross-temporal pairing, synthetic urban data with an intermittent freeze-frame interpolation strategy, and a **virtual lookahead sink** that retrieves a nearby street-view

[*]Equal contribution. Note that the first author conducted this work as part of the NAVER Cloud Residency Program. [1]KAIST AI [2]NAVER AI Lab [3]SNU AIIS. Correspondence to: Seungryong Kim <seungryong.kim@kaist.ac.kr>, Jin-Hwa Kim <j1nhwa.kim@navercorp.com>.

*Accepted at the F2S Workshop at the $43^{rd}$ International Conference on Machine Learning (ICML)*, Seoul, South Korea, 2026. 


*Figure 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.


*Figure 2.* **Reference-target video pair.** A real training example pairs street-view references with a target video captured at the same location but at a different timestamp.

image at each generation chunk as a virtual future destination.

SWM demonstrates that world simulation can be faithfully grounded in real, physically existing environments at city scale. We evaluate SWM across Seoul, Busan, and Ann Arbor, where the latter two cities are absent from training, and show improved visual quality, camera adherence, temporal coherence, and structural fidelity to real locations.

## 2. Data Construction

For SWM training, we build aligned pairs between street-view references and target video sequences. Each reference carries camera pose and depth, providing geometric conditions that ground generation to real-world structure. We construct these pairs from Seoul street-view images, synthetic urban data, and a public driving video dataset (Sun et al., 2020). A real example of the reference-target video pair is shown in Figure 2.

**Street-view dataset.** We collect 1.2M panoramic images covering major urban areas of Seoul, with GPS coordinates and capture timestamps; license plates and pedestrians are blurred for de-identification. After processing, 440K images are used for training. We use Depth Anything V3 (Lin et al., 2025) to estimate per-keyframe depth maps and camera poses, aligned to real-world scale using GPS metadata. We caption all videos with Qwen2.5-VL-72B (Bai et al., 2025), augmented with predefined camera actions (straight, stop, left turn, right turn).

**Cross-temporal pairing.** A key design choice is that 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, giving the model no incentive to separate persistent structure from transient objects, so it 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. Figure 6 visualizes the resulting attention pattern.

**View interpolation.** City-scale street-view databases provide panoramas at sparse spatial intervals (5–20 m) rather than continuous video; abrupt jumps between distant viewpoints break the temporal continuity that pretrained video diffusion models expect. We synthesize $T$-frame videos from $N$ sparse keyframes ($T \gg N$). A straightforward option concatenates keyframe latents along the channel di-

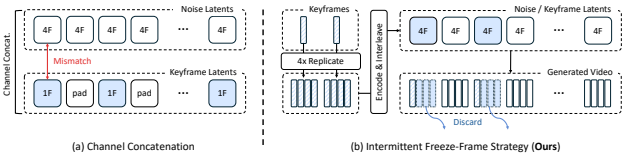


*Figure 3.* **View interpolation pipeline:** (a) keyframe conditioning via channel concatenation, and (b) intermittent freeze-frame strategy (**Ours**). 4F and 1F denote latents from four frames and one frame, respectively, before the 4× temporal compression of the 3D VAE.

mension at their timestamps (e.g. Wan2.1-FLF2V (Wan et al., 2025), Figure 3(a)), but we observe weak keyframe adherence: the 3D VAE compresses every 4 consecutive frames into one latent, so an isolated keyframe does not form a valid 4-frame group. We instead propose an *Intermittent Freeze-Frame* strategy (Figure 3(b)): each keyframe is repeated for 4 frames so it encodes to exactly one clean latent, motion between keyframes is generated normally, and the three duplicates are discarded after decoding. Quantitative comparisons are provided in our extended version.

**Synthetic dataset.** To complement driving-like street-view trajectories with diverse camera paths, we construct a synthetic dataset from CARLA (Dosovitskiy et al., 2017). We render 12.7K videos from 6 urban maps spanning approximately 431,500 $\mathrm{m}^2$ of city area, across pedestrian, vehicle, and free-camera trajectories, and render street-view references every 10 m along roads with eight directional views covering 360°. As in real data, references and target videos are rendered at different timestamps.

## 3. Model

As shown in Figure 4, SWM generates city-grounded videos through retrieval-augmented conditioning from a starting location, camera motion, and text prompt. We build on a pretrained Diffusion Transformer (Peebles & Xie, 2023; Agarwal et al., 2025) that operates in a latent space compressed from pixel-space frames $\mathbf{X} = \{x_t\}_{t=0}^{T-1}$ via a 3D VAE. Generation proceeds autoregressively in chunks: for the $i$-th chunk, the model takes a camera trajectory $\mathbf{C}^{(i)}$, a text prompt $P^{(i)}$, and noisy latents $\mathbf{Z}^{(i)} = \{z_l^{(i)}\}_{l=0}^{L-1}$ ($L$ latents per chunk) to produce target latents, additionally conditioning on $H$ history latents $\mathbf{Z}_{\text{hist}}^{(i)}$ from the tail of the preceding chunk for temporal continuity. We evaluate it under Teacher Forcing (Williams & Zipser, 1989) and Self-Forcing (Huang et al., 2025).

For each chunk, nearby street-view images are retrieved from a geo-indexed database of 1.2M Seoul panoramas, each rendered into 8 equi-angular pinhole views with metric-scale depth maps and 6-DoF poses estimated by Depth Anything V3 (Lin et al., 2025). Given the target trajectory, we retrieve in two stages: (1) nearest-neighbor search identifies candidate street-view locations, and (2) depth-based reprojection filtering retains references whose projected pixels exceed a coverage threshold in the nearest target view. This yields up to $K$ pinhole references $\mathbf{X}_{\text{ref}}^{(i)}$ with camera poses $\mathbf{C}_{\text{ref}}^{(i)}$ and depth estimates $\mathbf{D}_{\text{ref}}^{(i)}$.

**Virtual Lookahead Sink.** As contrasted in Figure 5, autoregressive generation accumulates errors across chunks. Prior work mitigates long-horizon degradation by maintaining an attention sink, typically the initial frame, as a fixed global context throughout generation (Liu et al., 2025; Shin et al., 2025). However, this static anchor becomes increasingly irrelevant as the camera moves farther from the starting point. We instead dynamically update the sink with a retrieved street-view image near the target trajectory endpoint. Because each chunk refreshes this virtual future destination, the anchor remains relevant to the region being generated.

We encode the retrieved image into a single latent $z_{\text{VL}}^{(i)}$ and assign it a RoPE (Su et al., 2024) temporal position beyond the current chunk. The latent sequence $\mathbf{Z}_{\text{seq}}^{(i)}$ fed to the model and its RoPE positions $\mathbf{p}_{\text{seq}}^{(i)}$ are

$$\mathbf{Z}_{\text{seq}}^{(i)} = \left[\mathbf{Z}_{\text{hist}}^{(i)};\, \mathbf{Z}^{(i)};\, z_{\text{VL}}^{(i)}\right], \text{ and}$$
$$\mathbf{p}_{\text{seq}}^{(i)} = \big[\underbrace{1,\ldots,H}_{\text{history}};\, \underbrace{H{+}1,\ldots,H{+}L}_{\text{target}};\, \underbrace{H{+}L{+}\Delta_{\text{VL}}}_{\text{sink}}\big], \tag{1}$$

where $\Delta_{\text{VL}}$ is a temporal offset. During training, a ground-truth future frame is sampled at a random offset, exposing the model to varying lookahead distances (Seo et al., 2025); at inference, $\Delta_{\text{VL}}$ is fixed and the frame is replaced by a retrieved street-view image.

**Geometric and semantic referencing.** Since each retrieved reference and the target observe the same scene from known camera poses, their geometric relationship enables two complementary pathways: geometric referencing for spatial layout and semantic referencing for appearance detail. Geometric referencing reprojects the spatially nearest reference $x_{\text{ref},j}^{(i)}$ into the target view via depth-based forward splatting (Ren et al., 2025; Li et al., 2025; Wu et al., 2025; Seo et al., 2024):

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

using only the single nearest reference per frame to avoid the artifacts of multi-image splatting; the warped video is encoded and channel-wise concatenated with the noisy target latent. Semantic referencing instead injects each reference latent into the transformer's latent sequence at RoPE position $p_{\text{ref},k}^{(i)} = H{+}L{+}G{+}k\Delta_{\text{ref}}$, where $G$ is a large gap separating references from the generation window and $\Delta_{\text{ref}}$ is the inter-reference spacing, letting each target latent attend to all $K$ references. Camera poses for target, reference, and sink tokens are encoded via Plücker ray embeddings.

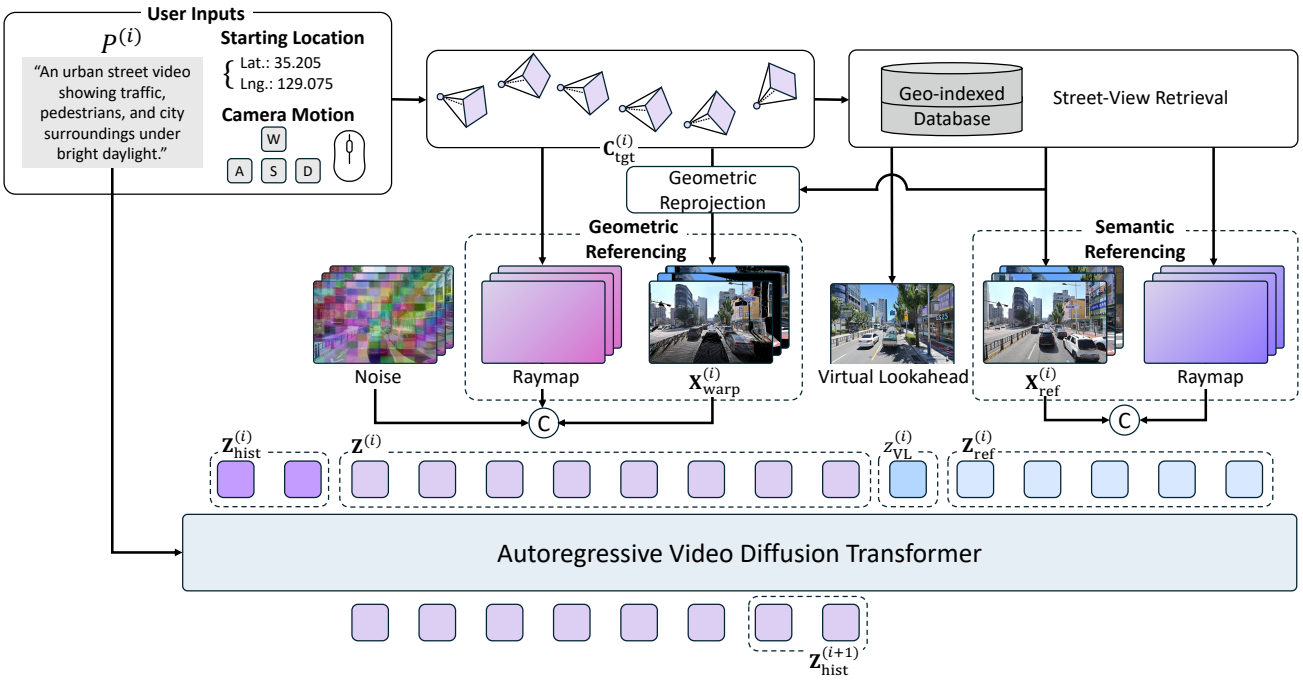


*Figure 4.* **Model overview.** Given a start location, SWM autoregressively generates video grounded in a real city based on text prompt $P^{(i)}$ and target camera trajectory $\mathbf{C}^{(i)}$, retrieving relevant street-view images from a geo-indexed database.

*Table 1.* **Quantitative comparison with other methods.** Values are reported as Busan-City-Bench / Ann-Arbor-City-Bench.

| Method | FID↓ | FVD↓ | Img.Q.↑ | RotErr↓ | TransErr↓ | mPSNR↑ | mLPIPS↓ |
|---|---|---|---|---|---|---|---|
| Aether (Zhu et al., 2025a) | 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 (Chen et al., 2025) | 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 (Mao et al., 2025) | 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 (HunyuanWorld, 2025) | 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 (Dai et al., 2025) | 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 (Team et al., 2026) | 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**/56.61 | **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)** | 32.50/**43.97** | 325.87/779.94 | 0.77/0.57 | 0.028/0.217 | 0.033/0.208 | 13.52/14.20 | 0.478/0.573 |

*Table 2.* **Ablation on Busan-City-Bench.**

| Variant | FID↓ | FVD↓ | Img.Q.↑ | RotErr↓ | TransErr↓ | mPSNR↑ | mLPIPS↓ |
|---|---|---|---|---|---|---|---|
| Full model | 28.43 | **301.76** | 0.78 | 0.020 | **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 | 0.78 | 0.021 | 0.020 | 13.52 | 0.427 |
| w/o geometric referencing | 33.01 | 398.74 | **0.79** | 0.036 | 0.051 | 12.33 | 0.525 |
| w/o semantic referencing | 30.27 | 326.18 | 0.78 | 0.032 | 0.022 | 14.08 | 0.442 |
| w/o any attention sink | 33.06 | 342.81 | 0.78 | 0.021 | 0.016 | 14.16 | 0.406 |
| w/ first frame attention sink | 32.71 | 378.92 | 0.78 | **0.018** | 0.018 | 14.25 | 0.388 |
| w/ first position attention sink | 32.41 | 354.61 | 0.78 | 0.026 | 0.027 | 14.35 | **0.379** |

# 4. Experiments

## 4.1. Setup

**Implementation details.** SWM fine-tunes Cosmos-Predict2.5-2B (Agarwal et al., 2025) with AdamW (Loshchilov & Hutter, 2019) (learning rate 4.8e−5) for 10K iterations at a total batch size of 48 on 24 H100 GPUs. We train a Teacher-Forcing (TF) model and derive a faster Self-Forcing (SF) variant (Huang et al., 2025) from the TF checkpoint via ODE initialization followed by 10K iterations of fine-tuning. TF uses $T$=77-frame chunks with $H$=5 history latents, $K$=5 references, and gap $G$=50; SF uses $H$=3, 12-frame chunks, and $K$=1, reaching 15.2 fps on a single H100. Both apply the Virtual Lookahead Sink with $\Delta_{\text{VL}}$=5.

**Benchmarks.** Since SWM is trained on Seoul data, we

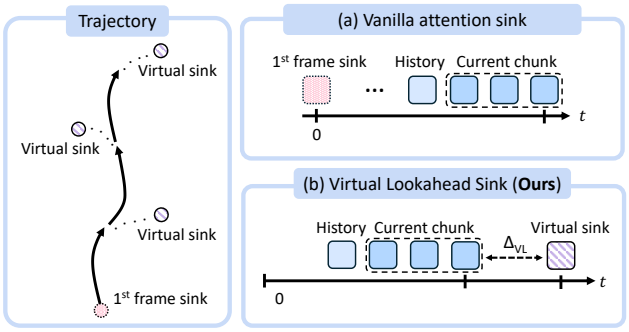


*Figure 5.* **Virtual Lookahead Sink:** (a) vanilla attention sink and (b) virtual lookahead sink (**Ours**).

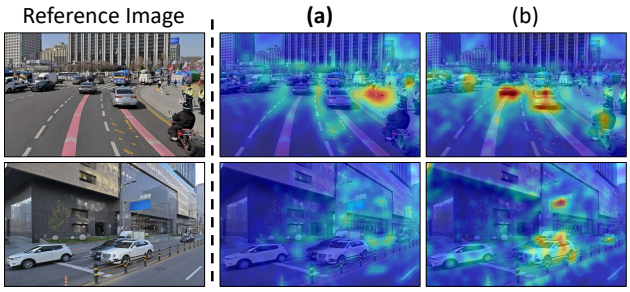


*Figure 6.* **Attention scores on references.** Cross-temporal pairing makes the model less attentive to dynamic objects at the reference.

evaluate cross-city generalization on Busan-City-Bench and Ann-Arbor-City-Bench (from the MARS dataset (Li et al., 2024)). Each benchmark contains 30 test sequences of 365 frames (about 100 m each); references are retrieved from nearby locations but exclude any street-view image of the test sequence itself.

**Baselines.** As real-world grounded world simulation is a new task requiring inputs not fully supported by existing models, we adapt six recent video world models by providing each with its supported inputs: Aether (Zhu et al., 2025a), DeepVerse (Chen et al., 2025), Yume1.5 (Mao et al., 2025), HY-World1.5 (HunyuanWorld, 2025), Fantasy-World (Dai et al., 2025), and Lingbot (Team et al., 2026).

**Metrics.** We assess three aspects. Visual and temporal quality: FID (Heusel et al., 2017), FVD (Unterthiner et al., 2018), and VBench Image Quality (Huang et al., 2024). Camera-following accuracy: Rotation Error (RotErr) and Translation Error (TransErr). 3D adherence: masked PSNR and LPIPS (Zhang et al., 2018) computed only on static regions, using SAM3 (Carion et al., 2025) to segment moving objects whose dynamics need not match the ground truth.

### 4.2. Results

As shown in Table 1, SWM achieves the best performance on both benchmarks 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 camera motion and scene structure and resulting in blurred videos, reduced motion, or complete collapse. By leveraging retrieved images, SWM remains anchored to the real-world scene layout and preserves alignment with the target trajectory. **Video qualitative results can be found in the project page.**

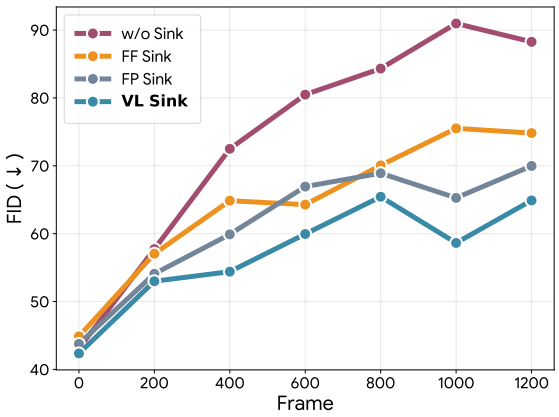


*Figure 7.* **Performance over time.** Sliding-window FID with a 200-frame window for different attention sink strategies.

### 4.3. Ablation Study

Table 2 summarizes ablations on three components of SWM. **Dataset construction.** Removing cross-temporal pairing causes the largest degradation across metrics, indicating that the model fails to disregard dynamic objects mismatched between references and generated frames; removing synthetic data slightly improves FID but substantially harms camera-following accuracy and 3D adherence, as the model no longer learns diverse trajectories. **Referencing.** Geometric and semantic referencing play complementary roles: geometric referencing supports camera alignment and static structural consistency, while semantic referencing improves appearance fidelity, so removing either degrades quality. **Attention sink.** Among the Virtual Lookahead (VL) Sink, no sink, a first-frame sink, and a first-position sink, removing the sink causes drift while the first-frame and first-position sinks remain limited as the camera moves far from the anchor; the VL Sink achieves the best FVD and the lowest sliding-window FID over time (Figure 7).

## 5. 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.

## References


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.

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.

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.

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.

Dai, Y., Jiang, F., Wang, C., Xu, M., and Qi, Y. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. *arXiv preprint arXiv:2509.21657*, 2025.

Deng, B., Tucker, R., Li, Z., Guibas, L., Snavely, N., and Wetzstein, G. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *SIGGRAPH*, 2024.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. Carla: An open urban driving simulator. In *CoRL*, 2017.

Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2018.

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.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.

Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., and Corrado, G. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Huang, X., Li, Z., He, G., Zhou, M., and Shechtman, E. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

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., and Liu, Z. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

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

Li, G., Zheng, S., Xu, S., Chen, J., Li, B., Hu, X., Zhao, L., and Jiang, P.-T. Magicworld: Interactive geometry-driven video world exploration. *arXiv preprint arXiv:2511.18886*, 2025.

Li, Y., Li, Z., Chen, N., Gong, M., Lyu, Z., Wang, Z., Jiang, P., and Feng, C. Multiagent multitraversal multimodal self-driving: Open mars dataset. In *CVPR*, 2024.

Lin, H., Chen, S., Liew, J., Chen, D. Y., Li, Z., Shi, G., Feng, J., and Kang, B. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Liu, K., Hu, W., Xu, J., Shan, Y., and Lu, S. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Liu, Y., Luo, C., Fan, L., Wang, N., Peng, J., and Zhang, Z. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In *ECCV*, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2019.

Mao, X., Li, Z., Li, C., Xu, X., Ying, K., He, T., Pang, J., Qiao, Y., and Zhang, K. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *ICCV*, 2023.

Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., and Gao, J. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, 2025.

Seo, J., Fukuda, K., Shibuya, T., Narihira, T., Murata, N., Hu, S., Lai, C.-H., Kim, S., and Mitsufuji, Y. Genwarp: Single image to novel views with semantic-preserving generative warping. *NeurIPS*, 2024.

Seo, J., Mira, R., Haliassos, A., Bounareli, S., Chen, H., Tran, L., Kim, S., Landgraf, Z., and Shen, J. Lookahead anchoring: Preserving character identity in audio-driven human animation. *arXiv preprint arXiv:2510.23581*, 2025.

Shang, Y., Lin, Y., Zheng, Y., Fan, H., Ding, J., Feng, J., Chen, J., Tian, L., and Li, Y. Urbanworld: An urban world model for 3d city generation. *arXiv preprint arXiv:2407.11965*, 2024.

Shin, J., Li, Z., Zhang, R., Zhu, J.-Y., Park, J., Shechtman, E., and Huang, X. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

Sun, P., Kretzschmar, 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., and Anguelov, D. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020.

Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P. P., Barron, J. T., and Kretzschmar, H. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, 2022.

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.

Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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.

Williams, R. J. and Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1989.

Wu, T., Yang, S., Po, R., Xu, Y., Liu, Z., Lin, D., and Wetzstein, G. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and 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.

Zhu, H., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Chen, J., Shen, C., Pang, J., and He, T. Aether: Geometric-aware unified world modeling. In *ICCV*, 2025a.

Zhu, Y., Feng, J., Zheng, W., Gao, Y., Tao, X., Wan, P., Zhou, J., and Lu, J. Astra: General interactive world model with autoregressive denoising. *arXiv preprint arXiv:2512.08931*, 2025b.