

3D Point Cloud Processing: Driving Generative AI Toward Physical Realism

Junwen Wang

*College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
wjw24@hnu.edu*

Abstract. Generative Artificial Intelligence (Generative AI) aims to improve world generation beyond simply visual synthesis. Evaluation criteria have expanded from concerns about the fidelity of generated images to multiple others, including the persistence of generated objects, the consistency of structure across generated shots, and the presence of physical causality among the generated elements. This paper explores the Seedance and Sora pathways to show the impact of growing long-range coherence on the generation of 3D worlds using 2D observational signals. This research shows that 3D point clouds exhibit the central pivot of the synthesis of 2D content coupled with the modeling of 3D physical worlds of the generation pipeline. As an openly stated world state, point clouds spatially alleviate simmered tensions of multiple modalities, persistence, and relationships that exist in space and time. Long-range coherence demonstrates the relationships that rationalize persistence in relationships and narratives. This study also demonstrates the intrinsic value of point clouds in the areas of geometric supervision, cross-modal consistency, and spatio-temporal organized space, showing the capacity for evolving generation models from content tools to spatial agents.

Keywords: 3D point clouds, video generation, world simulator, physical realism, spatial consistency

1. Introduction

Generative Artificial Intelligence (Generative AI) is in transition from simple image and video generation to world building. As a result, the criteria for evaluating the quality have shifted from realism to world realism. World realism demands characters retain identities across shots, scenes remain coherent as viewpoints change, and high audio-video unification in terms of lip-sync, sound sources and narrative action. Sequence lengths must communicate the presence of a single world.

In this evolution, Seedance and Sora are two technical approaches. Seedance aims for industrial application and artistic control, and has evolved to become a multi-modal storytelling system with joint multi-modal input, sophisticated camera scheduling and on-board audio-visual generation. Sora proposes a world simulator, where the model learns both spatial and temporal structures and causal relationships via large-scale pre-training on video data. Despite the different designs, Seedance and Sora both share the same paradox: they use 2D observational signals to do 3D dynamic world

simulation. When expanding to long video, multiple shots, or jointly generating audio and video, problems such as character drift and scene consistency quickly arise.

3D point clouds are the answer. They explicitly preserve world structure by keeping object and scene layout invariant, and connecting audio and visual events. Point clouds as a world state enable models to transition from observation fitting to world change maintaining, and approach the world's ontology. This paper examines the bottlenecks in world generation, and shows that 3D point clouds as a structural pivot transform generative models into world agents.

2. From controllable generation to world simulation

In recent years, the progress of generative models has been accelerating rapidly, mainly thanks to the strong modeling ability at the observational level. Through fitting massive distributions, diffusion models and their extensions have achieved visual quality approximating real distributions [1]. But with more complex tasks, the model must not only be observationally plausible, but also consistent and causally rational across time, view, and modalities.

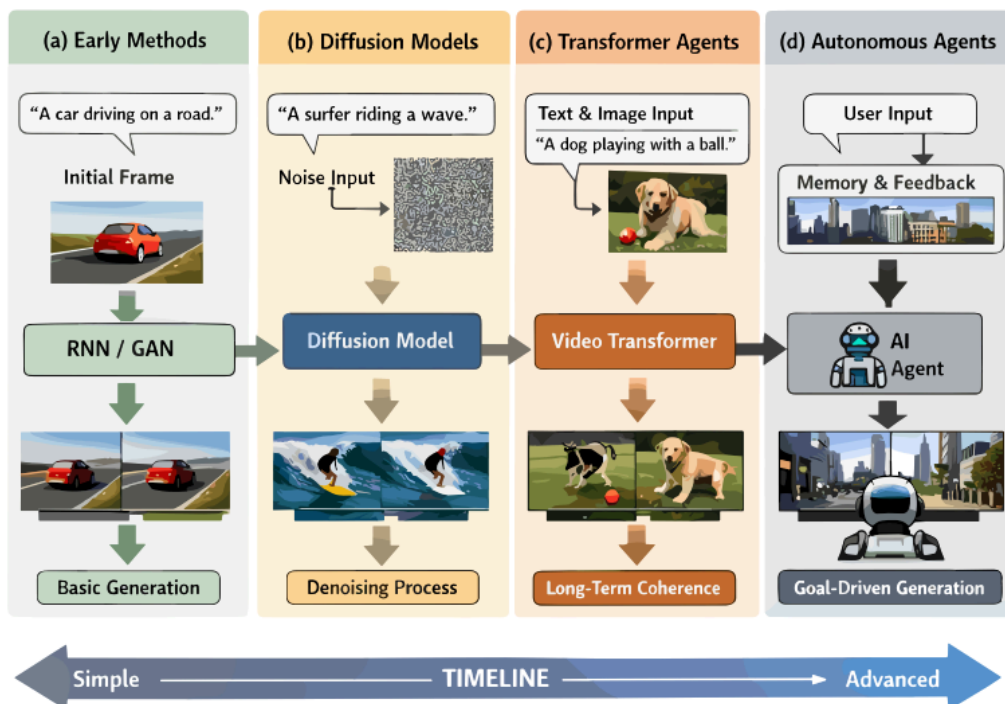


Figure 1. Evolution of video generation agents [2-4]

Figure 1 demonstrates video generative models which have evolved from early recurrent neural networks and generative adversarial networks (GANs) to diffusion models and Transformer-based generative agents. This progress represents an upgraded ability in temporal modelling; however, its core remains at the sequence of observations.

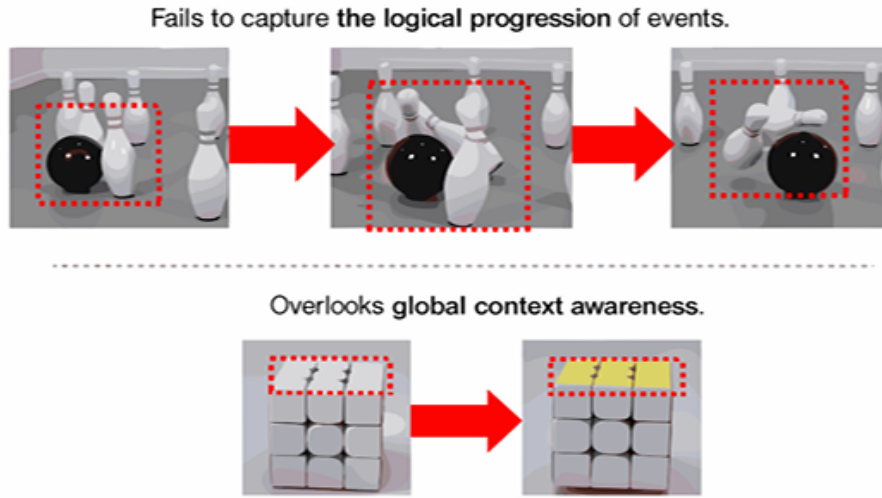


Figure 2. Examples highlighting failures in causal reasoning during multi-step generation(picture credit: original)

This is especially problematic in complex tasks. Figure 2 show the common error types: temporal models cannot establish logical event continuity, resulting in sudden changes in the number of objects or physical inconsistencies. Spatially, the lack of consideration for the global structure leads to mismatches between local changes and global consistency. Although some research takes into account 3D structure or regularize physical properties to address these problems, they still fail to effectively solve the instability problem caused by absent explicit state [5].

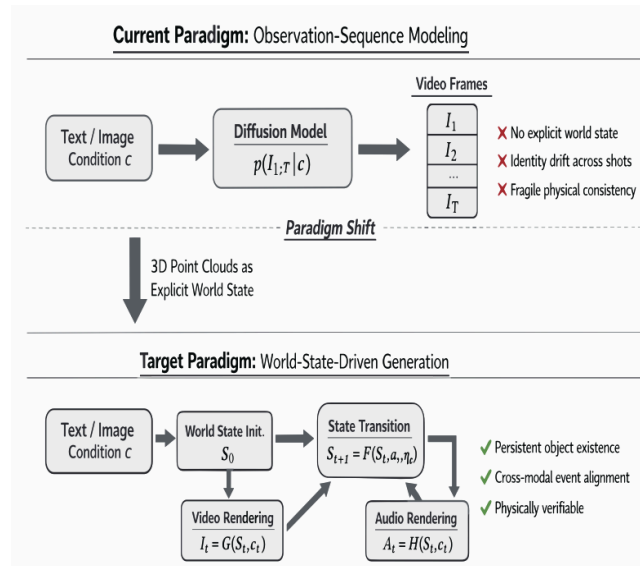


Figure 3. Paradigm shift from observation-sequence modeling to world-state-driven generation. Point clouds serve as the structural bridge enabling this transition [3, 5]

The current dominant approaches are defined as observational sequence modeling (Figure 3, top). Models use diffusion processes to generate video from textual or image conditions but do not explicitly model the world state. This results in problems like identity drift, weak physical consistency and multi-modal alignment [1, 6, 7].

In this regard, Seedance and Sora take different paths (Figure 4). The former focuses on multi-modal integration and controllability [8], while Sora brings in physical priors, viewing video generation as latent world simulation [3, 9]. These approaches point to a general issue of controllability and consistency without an explicit notion of the world state: it remains approximate fitting of an observational model.

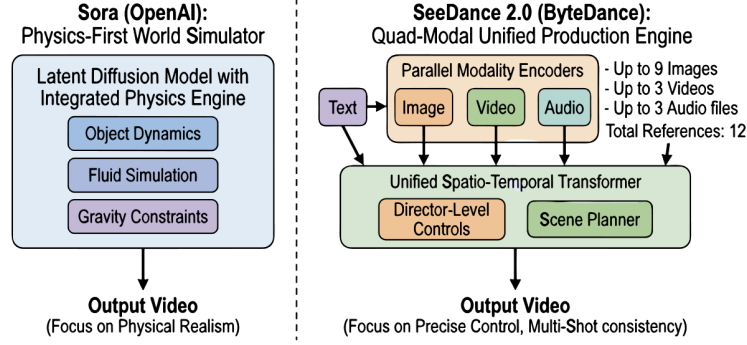


Figure 4. Comparative analysis of core architectures between Sora and SeeDance 2.0 [8, 9]

Theoretically, the real world is best described as a state-driven system. Let S_t denote the world state at time t ; its evolution follows:

$$S_{t+1} = F(S_t, a_t, \eta_t) \quad (1)$$

$$I_t = G(S_t, c_t) \quad (2)$$

That is, the world is in a state space and then observed through imaging. This is consistent with recent works interpreting generative models as world models [3, 9]. But existing models skip state modeling, and learn distributions of observations, effectively compressing the state transitions into high-dimensional pixel distributions. While some approaches approximate state space using point clouds or 3D models [6, 10, 11], this is yet to be unified. This is adequate for simple tasks but insufficient for object persistence, spatial consistency and cross-modal alignment.

Therefore, the bottlenecks are paradigmatic. Although modeling 2D observations is expressive, it does not provide support for the consistency of rich worlds. Figure 3 (bottom) shows a possible transition: to explicit world states for state transitions and multi-modal rendering. Video and audio are no longer the targets but observations of a single state. This transition from observation to state modeling makes generative models evolve from content to world modeling systems.

3. Point clouds: the crucial intermediate layer for physical realism

If the fundamental issue is the absence of a world state, then the state representation is pivotal. The choice of representation impacts training efficiency, editability and scalability [12]. The representation of video and audio should allow spatial representation, temporal updates and multi-modal fusion.

Here, point clouds are an important state representation. Their strength is offering a spatial representation independent of viewpoints. Point clouds represent points in 3D space. They represent

the true distribution of objects. This information can be reused over views and time. The model changes from matching shapes to representing scenes. This results in more persistent objects and scenes [6, 10, 11].

This affects generative reasoning. Objects become stable in 3D space. Modifications merely reflect pose or view changes. Camera control is now pose variation. Temporal audio-visual sync becomes spatial and motion sync. So point clouds modify the generation problem.

Moreover, point clouds are updateable. This is crucial for long-term generation and spatial memory [2, 6]. Their view independence supports consistency. Research indicates point clouds can be generated from video for pre-training. This gives models an inductive bias [10].

Put simply, point clouds aren't just geometric. They offer a world state. Learning and inference can focus on a single object. Point clouds are the state of the art for carrying state.

4. Restructuring generative models via point clouds: geometric supervision, multi-modal alignment, and world constraints

With the implementation of points in the cloud, the initial adjustment doesn't pertain to the quality of the rendering, but takes form in the kind of problem that the model has to address. An example of this, as demonstrated in Figure 5, is the fact that the majority of traditional video diffusion models focus their optimizations in the observation domain. One of the primary purposes of these models is to minimize the error in the predicted noise:

$$\mathcal{L}_{2D} = \mathbb{E}_{x, \epsilon, t} \left[\|\epsilon - \epsilon_{\theta}(x_t, t, c)\|_2^2 \right] \quad (3)$$

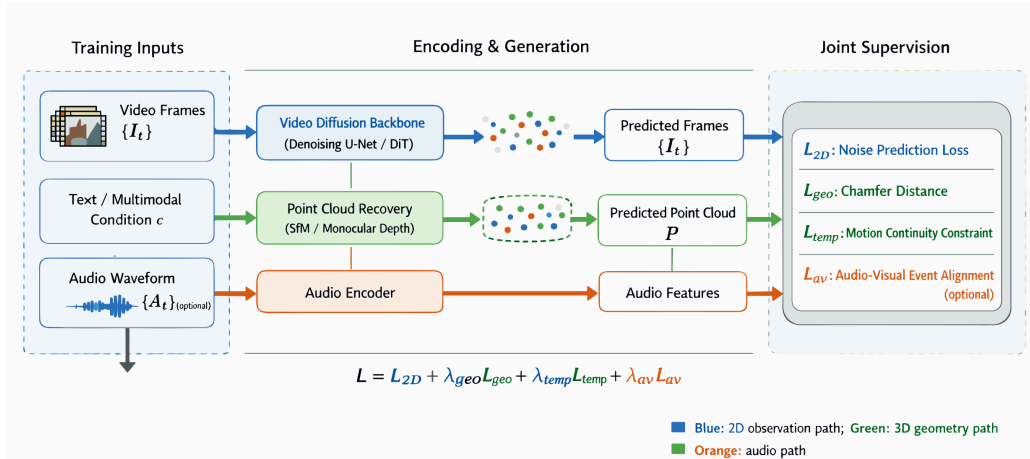


Figure 5. Point cloud-enhanced joint training framework. The model applies multi-level supervision to achieve multi-modal fusion [10, 11, 13]

This target is suitable for the model to learn the different visual styles, local textures, and the short-term and transitional motions that are commonly seen in videos. Even though the optimization improves the degree of coherence and the quality of the motion, the results don't correlate and belong to the same 3D environment. Most of the time, the optimization of these models is only guided by the quality of the rendered motion. These models can hide severe structural flaws in the output. For instance, the image results can produce odd or anomalous proportions for the actors.

Models output inter-penetrations or seams, and temporal inconsistencies can be seen. Optically, these flaws are usually hardly noticeable in the final output, but in practice, they completely invalidate the model of the world.

With the introduction of point cloud representations, the training objective expands from simple observation consistency to both geometric and temporal consistency. As shown by the green path in Figure 5, the model no longer just predicts video frames; it must also generate point cloud structures simultaneously. Let $\mathcal{P}$ denote the ground-truth point cloud and $\widehat{\mathcal{P}}$ the predicted point cloud. Geometric consistency loss can then be implemented using the Chamfer Distance:

$$\mathcal{L}_{geo} = \sum_{p \in \mathcal{P}} \min_{q \in \widehat{\mathcal{P}}} \|p - q\|_2^2 + \sum_{q \in \widehat{\mathcal{P}}} \min_{p \in \mathcal{P}} \|q - p\|_2^2 \quad (4)$$

Furthermore, to ensure the continuity of temporal evolution, motion constraints can be introduced on point-level trajectories:

$$\mathcal{L}_{temp} = \sum_i \| (p_i^{t+1} - p_i^t) - v_i^t \|_2^2 \quad (5)$$

This results in a joint optimization objective:

$$L = \mathcal{L}_{2D} + \lambda_{geo} \mathcal{L}_{geo} + \lambda_{temp} \mathcal{L}_{temp} \quad (6)$$

The joint supervision module (Figure 5, right) constrains the model in three ways: observation, geometry and time. This means the goal of generation is no longer just to simulate the observation. It must strive to be independent in the sense of structure and evolution.

This evolution from data-driven to structure-constrained generative models makes generative models world modelers. The model has a persistent spatial structure. Objects remain stable across time and viewpoints as objects, not just visual proxies. This model transitions the generative process from 2D surface properties to 3D spatial properties. Character consistency is now achieved via persistence of 3D objects. Camera motion is expressed in terms of pose variations. And explicit geometry constrains complex interactions for physical consistency. These developments eliminate pixel-based heuristics in favour of geometry.

This is essential for physics. Observational statistics cannot ensure stable position and motion laws. Point clouds explicitly contain spatial information. This enables a physical factor to be included in the loss function to be optimized. Physical consistency thus becomes an objective structural property. This enables explainable modeling.

This structural representation is also crucial in multi-modal tasks. The audio and visual pathways meet at the feature layer (Figure 5). Joint supervision then provides audio-visual constraints. Conventional approaches align audio and lip movements in time. This spatialisation extends to point clouds. Audio sources, humans and poses are connected via a shared spatial coordinate system. This elevates audio-visual synchronization from temporal to spatial and event consistency. This is crucial for multi-person engagement and environmental response.

Table 1. Role of point clouds across key generation tasks

Task	Failure w/o PC	Mechanism	Constraint / Loss
Identity Consistency	Appearance drift	3D identity anchoring	Chamfer distance
Camera Control	Unstable viewpoint	Pose as 3D transform	Projection consistency
Occlusion	Wrong occlusion	Depth & contact reasoning	Geometry constraints
Temporal Consistency	Scene drift	Global 3D anchor	L_temp
Multi-view	Inconsistent identity	Shared point cloud	Reprojection error
Audio-Visual Sync	Misalignment	Audio bound in 3D	Spatial audio loss
Physical Realism	Unrealistic motion	Physical structure encoding	Physical reg. [10]

Based on the above analysis, this transition can be further categorized at the task level. As shown in Table 1, the absence of point clouds in various generation tasks often leads to identity drift, viewpoint instability, and incorrect occlusion relationships. The root cause is that models only fit the observation domain and lack the ability to model 3D consistency.

Point clouds resolve inconsistencies through structured mechanisms. 3D Identity Anchoring ensures object persistence. Pose Modeling treats viewpoints as spatial transformations to stabilize perspectives. Finally, depth and contact reasoning improve occlusion and interaction logic. These tools replace visual estimation with geometric precision.

By incorporating loss functions such as Chamfer Distance, projection consistency, and geometric constraints, the system applies continuous structural pressure during training. This "Task Defect—Mechanism—Quantitative Constraint" mapping confirms the core role of point clouds. They do more than improve stability; they fundamentally push models from observation fitting toward global modeling of spatial structure, temporal evolution, and cross-modal correlation.

5. Point clouds and spatial memory: internal world states as the foundation for world simulation

In long-term storytelling and multi-modal generation, consistency stems from a maintainable internal world state, rather than the passive compression of instantaneous latent vectors. The updatable spatial memory proposed by Spatia can be formalized as [2]:

$$\mathcal{M}_t = \{\mathcal{P}_t, \mathcal{I}_t, \mathcal{R}_t\} \quad (7)$$

Where $\mathcal{P}_t$ represents the scene point cloud, $\mathcal{I}_t$ denotes object semantics and identity information, and $\mathcal{R}_t$ describes the relationships and events between objects. When new observations o_{t+1} and control conditions u_{t+1} arrive, the state undergoes an incremental update:

$$\mathcal{M}_{t+1} = f(\mathcal{M}_t, o_{t+1}, u_{t+1}) \quad (8)$$

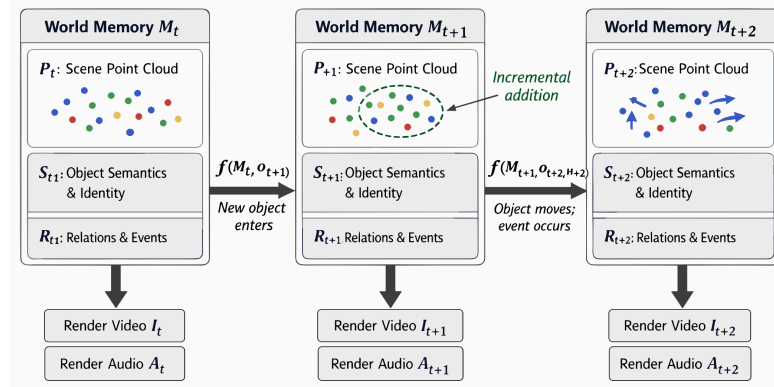


Figure 6. Illustrates the temporal flow of point cloud - based spatial memory update [2, 6]

Figure 6 illustrates the timeline of this mechanism. When a new object enters or moves, it only triggers local point additions, semantic binding, and pose corrections within the point cloud. Video frames and audio are then rendered from this unified world state. This ensures cross-frame stability, character identity continuity, and audio-visual synchronization.

But point cloud memory needs to be more structure-aware to better support world simulation. Xia et al.'s point cloud prior in Points-to-3D is an effective approach [13]. This research employs a point cloud prior. Through the injection of geometric skeletons, part relationships, and physics constraints at the beginning of the generation process, it improves long-term consistency greatly.

Table 2. Technical road-map from observation modeling to structure-aware world simulation

Tier	Function	Representative Work	Core Contribution
Digitization	Obtain 3D scene from vision	Yamada et al. [10]	Scalable point cloud pre-training from video
Representation Learning	Generate point cloud representations	Mo et al. [4], Huang et al. [11]	Learn structured feature codes from sparse points
Geometric Supervision	Distill geometry with regularization	Y. Chen et al. [5]	Supervise 3D point transformations and consistency
Structure-Aware Generation	Use point cloud priors for structural control	Xia et al. (Points-to-3D) [13]	Embed point cloud priors early for geometry, part relations and long-term consistency
Temporal World State	Maintain persistent 3D world memory	Gu et al. [1], Zhao et al. (Spatia) [2]	Enable incremental updates and cross-modal alignment
World Simulation Frameworks	Integrate into full world simulators	Brooks et al. (Sora) [3], He et al. [9]	Build end-to-end systems with explicit world state

Table 2 shows the evolution of generative AI from observation to structure-aware world simulation. It categorizes the representative works by hierarchy and contributions. It goes from primitive 3D digitization and point cloud learning to geometric supervision. It continues to structure-aware generation and world state representations. It concludes with unified world simulation. This transforms point clouds from a simple geometry to the main representation of the world state.

So, point clouds and spatial memory form the world state. This is the basis for simulation in generative AI. Point clouds are no longer geometric representations. They are a world skeleton that is structured, persistent and multimodal. This makes storytelling, multi-camera, and physics-based simulations possible. It's a paradigm shift. Generative AI is transitioning from 2D synthesis to 3D

world modeling with spatial intelligence. Only through modelling a persistent and explicit world state can models simulate and comprehend the physical world.

6. Conclusion

AI systems on the frontier of generative processes are undergoing a deep integration towards new world representations. Their generative scope is also advancing beyond statistical linking. They begin to track identity in complex spatio-temporal frameworks beyond a single camera or a single modality. Production Control and world simulation are two different pathways pursued by Seedance and Sora respectively. However, they equally posit that for realism and consistency to be maintained in the simulacrum of a multi-camera system, there has to be a persistent world state.

Innovations in the spatial representation of 3D point clouds indexed to a spatio-temporal world is one such semiotic disruption in the current state of generative AI. 3D point clouds not only include 3D coordinates. They also allow CAIS systems to begin tracking spatial worlds of complex systems and subsystems of different degrees of phenomenality and spatial-temporal existence. 3D point clouds further allow different phenomenological agents to integrate into a coherent simulacrum, which, in the case of generative AI, precedes the current state of the world to produce new content. 3D point clouds will not be the only sign of progress to come, but of all the innovations along the generative AI continuum, they will be the most decisive.

References

- [1] Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., ... & Liu, Y. (2025). Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (pp. 1-12).
- [2] Zhao, J., Wei, F., Liu, Z., Zhang, H., Xu, C., & Lu, Y. (2025). Spatia: Video Generation with Updatable Spatial Memory. arXiv preprint arXiv: 2512.15716.
- [3] Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., ... & Ramesh, A. (2024). Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>, 3(1), 3.
- [4] Mo, S., Xie, E., Wu, Y., Chen, J., Nießner, M., & Li, Z. (2024). Fast training of diffusion transformer with extreme masking for 3d point clouds generation. In European Conference on Computer Vision (pp. 354-370). Cham: Springer Nature Switzerland.
- [5] Chen, Y., Cao, J., Goel, V., Korolev, S., Jiang, C., Ren, J., ... & Kag, A. (2025). Towards physical understanding in video generation: A 3d point regularization approach. arXiv preprint arXiv: 2502.03639..
- [6] Chen, L., Zhou, Z., Zhao, M., Wang, Y., Zhang, G., Huang, W., ... & Li, C. FlexWorld: Progressively Expanding 3D Scenes for Flexible-View Exploration. In The Thirty-ninth Annual Conference on Neural Information Processing Systems.
- [7] Han, J., Kokkinos, F., & Torr, P. (2024). Vfusion3d: Learning scalable 3d generative models from video diffusion models. In European Conference on Computer Vision (pp. 333-350). Cham: Springer Nature Switzerland.
- [8] Team, B. S. (2025). Seedance 1.5 pro: A native audio-visual joint generation foundation model. arXiv preprint arXiv: 2512.13507.
- [9] He, H., Zhang, Y., Lin, L., Xu, Z., & Pan, L. (2026). Pre-trained video generative models as world simulators. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 6, pp. 4645-4653).
- [10] Yamada, R., Ide, K., Fukuhara, Y., Kataoka, H., Puy, G., Bursuc, A., & Asano, Y. M. (2025). 3D sans 3D Scans: Scalable Pre-training from Video-Generated Point Clouds. arXiv preprint arXiv: 2512.23042.
- [11] Huang, Z., Johnson, J., Debnath, S., Rehg, J. M., & Wu, C. Y. (2024). Pointinfinity: Resolution-invariant point diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10050-10060).
- [12] Wen, B., Xie, H., Chen, Z., Hong, F., & Liu, Z. (2025). 3d scene generation: A survey. arXiv preprint arXiv: 2505.05474.

- [13] Xia, J., Duan, Z., Hengel, A. V. D., & Liu, L. (2026). Points-to-3D: Structure-Aware 3D Generation with Point Cloud Priors. arXiv preprint arXiv: 2603.18782.