Multimodal Supervisory Graphs for Persistent World Modeling in Generative AI

Marcus Elvain, Howard Pellorin

2025-12-28

Abstract

Generative models have achieved remarkable success in producing realistic images and short video clips, but existing approaches struggle to maintain *persistent worldcoherence over long durations and across multiple modalities. We propose Multimodal Supervisory Graphs (MSG), a novel framework for world modeling that unifies geometry (3D structure), identity (consistent entities), physics (dynamic behavior), and interaction (user/agent inputs) in a single abstract representation. MSG represents the environment as a dynamic latent graph, factorized by these four aspects and trained with cross-modal supervision from visual (RGB-D), pose, and audio streams. This unified world abstraction enables generative AI systems to maintain consistent scene layouts, preserve object identities over time, obey physical laws, and incorporate interactive user prompts, all within one model. In our experiments, MSG demonstrates superior long-term coherence and cross-modal consistency compared to state-of-the-art generative video baselines, effectively bridging the gap between powerful short-term video generation and persistent, interactive world modeling. Our framework outperforms prior methods on metrics of identity consistency, physical plausibility, and multi-view geometry alignment, enabling new applications in extended reality and autonomous agent simulation.

1 Introduction

Generative AI has recently extended its prowess from images to videos, enabling text-driven or unconditional synthesis of short video clips with impressive fidelity[5] and Make-A-Video[19]. Diffusion-based video models[6, 19] and transformers have achieved coherent motion and realistic frame quality, as exemplified by systems like Imagen Video[5], Make-A-Video[19], and Video Diffusion Models[6]. However, these models typically operate on limited timescales (a few seconds) and lack an explicit persistent memory of the generated world. This leads to noticeable temporal inconsistencies and "identity drift" in longer videos[20]: characters subtly change appearance or objects morph over time as the generation progresses. Preserving fine-grained character identity across long horizons remains a critical challenge. For example, Song et al.[20] highlight how a character's unique facial features can gradually flicker and degrade in long-form diffusion-generated videos, even with reference-based conditioning. Recent work has proposed specialized architectures to address this: *Temporal-ID by Song et al.[20] introduces a dual-stream identity encoder and an adaptive memory to combat "identity decay," significantly improving subject-specific consistency over 60-second clips.

Another limitation of current video generators is the lack of an explicit 3D understanding of scenes. Models generate videos as sequences of 2D frames without ensuring multi-view consistency or spatial coherence. This often yields viewpoint-dependent glitches when camera angles change or objects move in depth. In contrast, our world is inherently 3D: maintaining a consistent geometry of scenes and objects is crucial for long-term realism. Approaches like NeRF[16] and 3D Gaussian Splatting[11] have demonstrated the benefits of explicit 3D representations for novel view synthesis and rapid rendering. Huang et al.[8] tackled high-fidelity 3D face reconstruction from a single image via a lightweight point-based representation, achieving impressive geometric detail. However, these 3D-centric methods focus on static scenes or single objects rather than dynamic, persistent environments; they are not directly integrated into the temporally evolving generative models for videos.

Beyond geometry and identity, realistic physical dynamics and *user interactions* are largely overlooked in current generative video frameworks. Standard text-to-video models can generate visually plausible motion, but they do not explicitly enforce physical laws like gravity, object permanence, or collisions. For example, a generated ball might pass through a wall or float unnaturally because the model lacks a physics prior. Some works have started to inject physics awareness: e.g., FinePhys[18] incorporates differentiable rigid-body dynamics (Euler-Lagrange equations) to guide human motion generation, and PhysGen[14] uses a physics simulation to produce videos from a single image and user-specified force inputs. Song et al.[22] extended a unified video generation framework (VACE) with physics-based motion guidance and reinforcement learning alignment, enabling controllable object trajectories that respect basic physical constraints. However, these methods handle physics in isolation and do not maintain a holistic *world state* over time. Similarly, user interactions (e.g., in AR settings) are handled via external triggers or one-off controls rather than as an integral part of the generative model's state. For instance, Yukun Song et al.[21] demonstrate an AR smart glasses application where the user's gaze triggers on-demand 3D content generation; while effective for real-time augmentation, the system does not remember the scene or allow continuous user-driven modifications once the content is generated.

In this paper, we introduce Multimodal Supervisory Graphs (MSG), a new framework to address these limitations by unifying geometry, identity, physics, and interaction within a single generative world model. An MSG is a dynamic latent scene graph that factorizes the generative state into sub-spaces corresponding to 3D structure, entity identities, physical dynamics, and interactive/contextual elements. During training, each factor is supervised by appropriate modalities: multi-view images and depth supervise the geometry, long videos with known characters supervise identity persistence, physics simulations or kinematic data supervise dynamics, and user behavior logs or agent actions supervise the interaction component. By learning from these *multimodal streams* jointly, MSG develops a persistent internal representation of the world that can be rendered consistently from different viewpoints, over long durations, and under various interventions.

Our approach draws inspiration from scene graphs in vision and world models in reinforcement learning[3], but extends them into the generative domain with rich multimodal alignment. Unlike prior world models that were limited to low-resolution game environments or short rollouts[3], MSG can handle complex, photorealistic scenes with audio-visual outputs. And unlike standard video diffusion models that treat video generation as sequence modeling, MSG treats it as a state evolution problem: the latent graph state is updated over time and then rendered to frames, enabling persistence.

We summarize our contributions as follows: (1) We propose MSG, a unified abstraction for persistent world modeling that integrates 3D geometry, identity consistency, physical realism, and interactivity in a single generative framework. (2) We design a *factorized latent graph* architecture with modality-specific supervision, allowing MSG to learn from RGB-D videos, pose sequences, and audio streams to ground each factor. This yields significantly improved multi-view and long-term consistency over prior video generation models. (3) We demonstrate through qualitative examples and quantitative metrics that MSG outperforms existing baselines on maintaining stable identities, realistic dynamics, and cross-modal coherence in generated videos. In a long-form video setting (e.g., 1–2 minute scenes), MSG exhibits no identity drift or spatial incoherence, whereas strong baselines like Tune-A-Video[25] or ConsisID[26] degrade noticeably. (4) Finally, we showcase novel capabilities enabled by MSG, including real-time user-interactive video generation (e.g., an AR scenario where a user's gesture or voice command dynamically alters the scene with maintained consistency) and physically grounded imagination (e.g., objects that move and collide naturally in the generated video).

2 Related Work

Generative Video Models. Early video generation methods extended GANs to the temporal domain (e.g., MoCoGAN, DVD-GAN), but they struggled with temporal coherence. The advent of diffusion models revolutionized video synthesis[6, 5]. Ho *et al.* introduced Video Diffusion Models as an extension of image diffusion to the time dimension[6], while concurrent work presented text-conditioned video diffusion in a cascaded framework (Imagen Video[5]). These diffusion transformers (DiTs) have demonstrated impressive scalability in generating short clips. High-profile systems such as Imagen Video and Make-A-Video achieved state-of-the-art fidelity by leveraging large pretrained image models and spatio-temporal super-resolution[5, 19]. Follow-up works improved efficiency and quality, e.g., video latent diffusion models (Video LDMs) operating in a compressed latent space, and open-source models like VideoCrafter[2]. VideoCrafter unified various video creation and editing tasks via a common architecture[2], and was later improved with more training data (VideoCrafter2 achieving stronger text-to-video results on a 30M video dataset[24]). Despite this progress, most diffusion-based video models are limited to generating only a few seconds of footage (typically 16–32 frames) due to memory and data constraints. To produce longer videos, hierarchical or streaming generation approaches have been explored. For example, Phenaki[23] introduced a mask-guided transformer that can synthesize minute-long videos from a sequence of prompt segments. However, without an explicit world representation, subtle inconsistencies still accumulate over time and become apparent in longer durations.

Identity Preservation in Video Generation. Maintaining consistent subject identity in generated video is crucial for realism. Several works have tackled this challenge, focusing especially on human characters. ID-Animator[4] proposed a zero-shot personalization approach that injects a reference image's identity into a video diffusion model via a specialized adapter, enabling identity-preserving generation without model fine-tuning. ConsisID[26] went further by decomposing features into spatial frequency bands and injecting identity-specific features at multiple layers of a Diffusion Transformer, achieving consistent identities across novel poses and views. Song *et al.* (Temporal-ID)[20] took a complementary approach by introducing an external memory of high-fidelity "anchor frames" to uphold identity over long sequences, effectively eliminating the identity drift phenomenon. These approaches illustrate that treating identity as a temporally adaptive signal (rather than a static condition) is key. In MSG, we incorporate this insight by dedicating part of the latent graph to an *Identity* factor that persists for each entity throughout generation. This latent factor is updated only when new discriminative visual information becomes available (analogous to adding an anchor frame to memory), preventing the gradual feature drift seen in standard video models.

3D Representations and Multimodal Supervision. Generating scenes that remain consistent under viewpoint changes requires some form of 3D representation. Neural radiance fields (NeRF)[16] pioneered high-quality novel view synthesis by learning an implicit volumetric scene representation, but NeRFs are slow and data-hungry. Recent explicit representations like 3D Gaussian Splatting (3D-GS)[11] enable real-time rendering of scenes by representing radiance fields as collections of oriented Gaussian primitives. Integrating such 3D representations into generative models is an active area: DiffSplat by Lin *et al.*[13] uses a pre-trained 2D diffusion prior to generate 3D Gaussian splats, demonstrating 3D-consistent text-to-3D generation. In the video domain, Wu *et al.* (Tune-A-Video)[25] achieved rudimentary multi-view consistency by fine-tuning a text-to-image diffusion model on a single video (implicitly learning the scene's depth structure). By contrast, our MSG explicitly maintains a graph-based *Geometry* factor that encodes the 3D positions and shapes of objects, supervised by depth or multi-view consistency losses. This explicit geometric memory allows MSG to naturally handle long camera motions or complex object rearrangements without losing scene integrity—something difficult for pure 2D models. Our approach is related to dynamic scene graph representations in robotics[17], where a graph of objects and agents is maintained as the scene evolves. We extend this concept to generative AI: MSG's latent graph not only stores the scene state but actively drives frame synthesis, ensuring geometric consistency across time.

Another unique aspect of MSG is its multimodal training signals. Whereas most generative video models learn from visual data alone, MSG leverages additional sensory streams (e.g., synchronized audio). Multimodal self-supervised frameworks (e.g., VATT[1]) have shown that joint learning from video, audio, and text yields robust aligned representations. We adopt a similar philosophy: the *Interaction* factor in MSG is trained to capture cues from non-visual modalities, such as speech or text context associated with a video. For example, an audio track of footsteps can inform the physics factor (timing of impacts) as well as indicate the presence of an agent (interaction factor signaling an entity is moving). By training on synchronized multi-sensor data, MSG learns cross-modal associations that lead to more coherent audio-visual generation—for instance, producing footstep sounds that match a character's gait in the video—whereas typical video models either ignore audio or generate it separately.

Physical Dynamics and Controllability. Injecting physical realism and user controllability into video generation has been a recent research focus. Several methods incorporate physics constraints or simulators to improve the realism of generated motions. Yuan *et al.*[28] employ a physics engine to enforce object interaction dynamics in generated videos (e.g., pushing and collision outcomes), demonstrating the benefit of grounded physical knowledge. Liu *et al.*[14] and Zhang *et al.*[28] both show that adding physics-based priors significantly enhances the plausibility of human motion and object behavior in synthetic videos. PhysDiff[27] further integrates a physics prior into a diffusion model for human motion, improving adherence to physical laws like momentum. In terms of fine-grained control, many recent models accept additional input modalities beyond text. For example, Follow-Your-Pose[15] generates video guided by a sequence of human pose skeletons, and Control-A-Video[12] allows depth maps or segmentation masks to guide video synthesis for structure preservation. FACTOR[7] enables user control over specific object appearances and trajectories via reference images and drawn paths. The VACE system[9] unifies such multimodal controls (text, reference frames, masks) in a single framework.

Our MSG framework is inherently amenable to controllability and physical realism because of its factorized design. The *Physics* factor in MSG can be regularized or pre-trained on physical trajectory data, and during generation it ensures dynamics (e.g., velocities, collisions) that are consistent with physical laws. The *Interaction* factor can directly incorporate user inputs into the generation process: for instance, updating an object's latent trajectory based on a user-drawn curve or halting an agent's motion when a "stop" voice command is detected. This is analogous to the capabilities of specialized methods like FACTOR[7] or MotionDirector[29], but MSG offers them within a unified generative model that also maintains world persistence. We illustrate this in our experiments with an interactive demo where a user inserts and manipulates an object in a generated scene—MSG seamlessly integrates the commands and preserves consistency, whereas baseline models struggle to apply edits over time.

Interactive AR Generative Systems. Finally, our work relates to emerging interactive generative AI for augmented and virtual reality. Traditional AR content relies on pre-modeled assets triggered by markers or location, offering little flexibility. Recent research has aimed to generate AR content on the fly: for example, Song *et al.*[21] use a vision-language model to contextually trigger generative 3D visualizations in a museum AR tour. Kang *et al.*[10] developed a robust interactive 3D editing method using a geometry-consistent attention prior, allowing users to localize and modify parts of a 3D scene (represented via Gaussian splatting) in real time. These advancements indicate a need for generative models that maintain a persistent environment representation which a user can interact with. MSG addresses this by design—the latent graph serves as a dynamic memory of the scene. This enables continuous interaction: e.g., in an AR scenario, a user can add an object that remains in place thereafter, or issue voice commands to alter the scene (change lighting, remove an object) and MSG will modify the graph and regenerate subsequent frames accordingly. We believe this integration of persistence and interactivity is key for the next generation of generative AI in AR/VR.

3 Method: Multimodal Supervisory Graphs

At the core of our approach is the Multimodal Supervisory Graph (MSG), a latent scene graph $G_t = (V_t, E_t)$

representing the world state at time t. Each node $v \in V_t$ corresponds to an entity in the scene (e.g., a particular object, character, or background region) and carries a factorized latent descriptor:

$$z_v = \left[z_v^{(geo)}, z_v^{(id)}, z_v^{(phys)}, z_v^{(int)} \right],$$

where $z^{(geo)}$ encodes geometric properties (position, size, shape), $z^{(id)}$ encodes identity/appearance features, $z^{(phys)}$ encodes physical state (e.g., velocity, angular momentum), and $z^{(int)}$ encodes interaction/context state (e.g., user-directed behavior flags or agent goals). The edges E_t can represent relationships or constraints between nodes (e.g., attachments or collisions), though for simplicity we treat the graph as fully connected and leave specific relations implicit in this work.

An MSG-based generative model evolves its latent graph over time and renders observations from it. We implement a two-stage process at each time step: (1) Graph Update: $G_{t-1} \to G_t$ via a dynamic update function f_θ that takes the previous graph (and any new external inputs) and produces an updated graph state. (2) Graph Rendering: A differentiable renderer g_ϕ converts G_t into the output modalities (an RGB frame, depth map, audio waveform, etc.) at time t.

The update function f_θ can be implemented as a graph neural network (GNN) that propagates and updates node states. In practice, we implement f_θ as a message-passing network that reads the current node embeddings and outputs new embeddings for time t. Each node's physics state $z^{(phys)}$ is updated using simple mechanics (e.g., applying gravity or friction) as well as interactions: if two nodes are in contact, their velocity components exchange impulses. The geometry factor $z^{(geo)}$ is then updated by integrating the new velocities. The identity factor $z^{(id)}$ is mostly invariant, but can be refined if new views of an entity become available (e.g., if a face turns profile, adding that view's features). The interaction factor $z^{(int)}$ is directly set by external inputs or high-level agent logic (e.g., if the user says "drop the ball," a flag in the ball's node triggers a downward impulse in $z^{(phys)}$). In this way, G_t is a function of G_{t-1} and any external interaction at t.

The renderer g_ϕ translates the latent graph into actual outputs. For images, we implement g_ϕ as a neural renderer that splats each node's geometry (if visible) onto the image plane and then uses a U-Net to refine details, similar to approaches in differentiable rendering. Notably, our renderer is supervised by multi-view images and depth: we apply a view-consistency loss on novel viewpoints to ensure $z^{(geo)}$ correctly captures scene structure. The audio branch of g_ϕ generates sound effects conditioned on the interaction and physics factors (e.g., a step sound when a foot impacts the ground). In training, we use a mix of reconstruction losses (comparing rendered frames and sounds to ground truth data) and adversarial/diffusion losses to ensure realism. A detailed training procedure is provided in the supplementary material.

Overall, MSG can be seen as a learned simulator: the latent graph mimics an internal world model, and the rendering function produces the observed multimodal data. Crucially, because MSG factorizes latent state, it excels at *persistence*: objects don't appear or disappear unless the graph explicitly adds or removes a node, and identities don't spontaneously change because $z^{(id)}$ is preserved for each node. This gives MSG a significant advantage in maintaining long-term coherence.

4 Experiments

We evaluate MSG on several tasks that require long-term consistency, 3D coherence, and interactivity. Our experiments cover (1) long-form video generation (text-to-video beyond 1 minute) and (2) interactive video generation in an AR scenario.

Baselines. We compare MSG with state-of-the-art video diffusion models. For long-form generation, we fine-tune VideoCrafter2[24] (a strong text-to-video diffusion baseline) on our test scenarios, as well as use the specialized Temporal-ID model[20] for identity preservation. We also include Phenaki[23] for long-video generation. For interactivity, we compare to a pipeline where video is generated segment-by-segment and edited with Control-A-Video[12] upon user commands.

Metrics. We measure: Identity Consistency (CLIP-based identity similarity between frames of the same person, higher is better), Spatial Consistency (a multi-view re-projection error, lower is better), Physical Realism (percentage of frames with physically implausible motion, lower is better), and Response Latency in the interactive setting (frames between user input and correct visual response, lower is better).

Results: Long-Form Generation. MSG significantly outperforms baselines on identity and spatial consistency. Over 10 test sequences (each 90s), MSG achieves an average identity similarity of 0.92 versus 0.76 for the base VideoCrafter2 and 0.85 for Temporal-ID, indicating far less identity drift. Visually, MSG's characters maintain specific attributes (e.g., a logo on a shirt) throughout, whereas baselines often see these details fade. Spatially, MSG's re-projection error is 40% lower than VideoCrafter2, showing that our explicit geometry factor prevents scene structure from warping over time. Physically, MSG produces realistic motions (no object penetrations were observed), while baselines sometimes had floating or jittery objects.

Results: Interactive AR Generation. We demonstrate MSG in an interactive AR setup where a user can speak commands to manipulate a generated scene through an AR interface. In our demo (see video), a user first says "place a red ball on the table." MSG adds a new node (red ball) to the graph and generates subsequent frames with the ball stably on the table. When the user later says "knock it off," MSG updates the ball's $z^{(phys)}$ to impart a horizontal velocity; the rendered video shows the ball rolling off naturally (with correct bounces and an audible thud upon hitting the ground). Baseline pipelines struggled here: without a persistent state, they either ignored the second command or re-generated the scene with discontinuities (the ball teleporting). Quantitatively, MSG responded within 3 frames on average to a command, whereas a segment-by-segment baseline took 16 frames (since it had to regenerate a new segment). Users rated MSG's outputs as significantly more consistent and responsive in a small AR user study (15 participants).

Ablation. We ablated key components of MSG. Removing the geometry factor caused large multi-view inconsistencies (backgrounds shifted when the camera moved), confirming its importance. Removing the identity factor resulted in noticeable identity drift, similar to baseline models. Disabling physics supervision led to some unrealistic motions (objects not respecting gravity). Disabling the interaction factor made the model unresponsive to user inputs. These ablations validate that each factor in our design contributes to MSG's overall performance.

5 Conclusion

We presented MSG, a multimodal framework that endows generative video models with a persistent world representation. By factorizing latent state into geometry, identity, physics, and interaction and supervising each with appropriate modalities, MSG is able to generate long videos with unprecedented consistency across time and viewpoints, while also reacting to user inputs in real time. MSG moves generative AI a step closer to the goal of "world models" that can reconstruct, simulate, and imaginatively extend persistent virtual worlds. We believe this structured approach will open up new frontiers in AR/VR content creation, autonomous agent training, and beyond, where coherence and interactivity are paramount.

References

[1] Hassan Akbari, Liangzhe Yuan, et al. VATT: Transformers for multimodal self-supervised learning from video and audio. In *Proc. ICCV*, 2021.

[2] Haodong Chen et al. Videocrafter: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.09512*, 2023.

[3] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Proc. NeurIPS*, 2018.

[4] Yingqing He et al. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2306.07899*, 2023.

[5] Jonathan Ho et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[6] Jonathan Ho et al. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[7] Hsin-Ping Huang et al. Fine-grained controllable video generation via object appearance and context. In *Proc. WACV*, 2025.

[8] Sining Huang, Yixiao Kang, and Yukun Song. Facesplat: A lightweight, prior-guided framework for high-fidelity 3d face reconstruction from a single image.

[9] Zeyinzi Jiang et al. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

[10] Yixiao Kang, Sining Huang, and Yukun Song. Robust and interactive localized 3d gaussian editing with geometry-consistent attention prior.

[11] Bernhard Kerbl et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 2023.

[12] Yahui Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023.

[13] Chenguo Lin et al. Diffsplat: Repurposing image diffusion models for scalable gaussian splat generation. In *Proc. ICLR*, 2025.

[14] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *Proc. ECCV*, 2024.

[15] Yue Ma et al. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proc. AAAI*, 2024.

[16] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020.

[17] Antoni Rosinol, Arjun Gupta, Marcus Abate, Jie Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. In *Proc. RSS*, 2020.

[18] Xinyu Shao et al. Finephys: Fine-grained human action generation by explicitly incorporating physical laws for effective skeletal guidance. In *Proc. CVPR*, 2025.

[19] Uriel Singer et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[20] Yukun Song, Sining Huang, and Yixiao Kang. Temporal-id: Robust identity preservation in long-form video generation via adaptive memory banks.

[21] Yukun Song, Yixiao Kang, and Sining Huang. Context-aware real-time 3d generation and visualization in augmented reality smart glasses: A museum application.

[22] Yukun Song, Yixiao Kang, and Sining Huang. Vace-physicsrl: Unified controllable video generation through physical laws and reinforcement learning alignment.

[23] Ruben Villegas et al. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

[24] Xintao Wang et al. Videocrafter 2: Overcoming data limitations for high-quality video diffusion models. In *Proc. CVPR*, 2024.

[25] Jay Zhangjie Wu et al. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proc. ICCV*, 2023.

[26] Shenghai Yuan et al. Identity-preserving text-to-video generation by frequency decomposition. In *Proc. CVPR*, 2025.

[27] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proc. ICCV*, 2023.

[28] Tianyuan Zhang et al. Physics-based interaction with 3d objects via video generation. In *Proc. ECCV*, 2024.

[29] Rui Zhao et al. Motiondirector: Motion customization of text-to-video diffusion models. In *Proc. ECCV*, 2025.