Physics-Consistent World Models via Schrödinger-Bridge Optimal Transport for Computational Imaging and 3D-Consistent Video Generation

Abhiram Srivatsa Kadaba
Nanyang Technological University, Singapore
abhiram001@e.ntu.edu.sg

Abstract

Modern generative models often violate basic physical principles. Shadows drift, geometry becomes inconsistent across views, and measurement models are ignored, which limits trust in both video synthesis and computational imaging. We propose a finite time Schrödinger Bridge (SB) world model that formulates generation as entropy regularized optimal transport from a simple prior to a distribution that is consistent with both data and physics. Instead of applying consistency corrections only at the final output, the framework introduces geometric and physical structure directly along the generative path. For video, the model enforces multiview geometric constraints through reprojection and epipolar agreement, homographies, and depth guided warping. For imaging, it incorporates differentiable optical operators, including point spread function based defocus models and lightweight Fourier propagation for coherent and partially coherent settings. When camera poses are known, the model penalizes reprojection error and warp aligned photometric or feature inconsistencies. When poses are unknown, a compact motion or flow estimator encourages cycle consistent trajectories. A lightweight UNet or Vision Transformer backbone, together with a short SB horizon, maintains computational efficiency. Evaluation will measure three dimensional and temporal consistency, physics fidelity through forward simulation residuals, and overall generative quality and efficiency using FID, KID, and FVD. Comparisons will include modern video diffusion models, plug and play data consistency methods, and unconstrained SB variants. The central hypothesis is that constraining the entire generative trajectory, rather than only the final frame, can shorten sampling while improving cross view coherence and physical plausibility across diverse sensing modalities, including cameras, microscopes, and medical imaging systems.

Introduction

Modern generative models produce highly realistic content but often violate basic geometric and physical constraints. Temporal flicker, multiview drift, inconsistent shadows, and implausible motion remain common in state-of-the-art video diffusion systems, even at large scale (Melnik et al. 2024). In computational imaging, reconstructions frequently deviate from the underlying physics due to point spread function mismatch, diffraction errors, or geometric distortions (Gao et al. 2024; Zeng et al. 2025). A key reason is that diffusion models impose structure primarily as a post hoc correction over long denoising chains, rather than enforcing it throughout the generative trajectory.

Finite-time Schrödinger Bridge (SB) formulations (Léonard 2014; Peyré and Cuturi 2019) offer a complementary perspective by casting generation as entropy-regularized optimal transport over a short, path-consistent evolution. Instead of hundreds of noisy updates, SB models evolve from a simple prior to the data distribution in a fixed and relatively small number of steps, enabling geometric and physical constraints to be injected directly into the latent path. This project investigates whether finite-time SB models can generate video sequences consistent with a stable three-dimensional scene, with primary emphasis on cross-view and temporal coherence. Extensions to computational imaging with differentiable optical models are considered only if time permits.

Background

The Schrödinger Bridge (SB) problem seeks the most likely stochastic evolution connecting two distributions while remaining close to a reference diffusion (Léonard 2014). In machine learning, this yields generative models with explicit path structure, interpretable as entropy-regularized optimal transport between a prior and a data distribution (Peyré and Cuturi 2019). Neural SB solvers, including diffusion Schrödinger bridges, show that such models can be trained and sampled efficiently in finite time while retaining close connections to score-based diffusion (De Bortoli et al. 2021; Shi et al. 2023). Because the drift is explicitly parameterized at each step, SB models naturally support constraints on the entire trajectory rather than only the endpoint.

Multiview geometry provides complementary constraints on image evolution under camera motion. Epipolar geometry, depth reprojection, and homography-based transformations underpin view-consistent reconstruction and rendering (Hartley and Zisserman 2004). Datasets such as RealEstate10K, ScanNet, and TUM RGB-D provide camera trajectories and approximate depth, enabling quantitative evaluation of geometric consistency (Zhou et al. 2018; Dai et al. 2017; Sturm et al. 2012). In parallel, differentiable rendering and physics-based imaging introduce for-

ward models for light transport, shading, and wave propagation into deep learning pipelines, improving physical plausibility (Gao et al. 2024; Zeng et al. 2025). Together, these advances motivate path-aware generative models that impose geometric and physical structure at intermediate states, addressing key limitations of current video diffusion methods.

Prior Work by the Applicant

This project builds on three strands of my previous research. At ERI@N, I developed real-time perception and object detection pipelines for autonomous driving, where temporal stability, viewpoint consistency, and inference efficiency were all critical for deployment. At IDMxS, I worked on deep learning methods for partially coherent lensless imaging that used differentiable Fourier propagation and point spread function based optical operators, which strengthened my intuition for physical forward models and physics-consistent reconstructions. At I^{2}R A*STAR, I studied multimodal dense correspondence and geometric supervision for cross-view alignment, including epipolar geometry, warping functions, homography prediction, and robustness to sensor variation. These experiences provide practical familiarity with geometric consistency, differentiable physics, and sequence modeling, and they motivate the proposed SB-based approach.

Approach

The goal is to construct a finite-time SB generator for video synthesis in which geometry and physics guide the model throughout the generative path, not just at the final output. A video autoencoder encodes real video sequences into a terminal latent distribution $p_T(z)$, while a simple Gaussian prior defines the initial distribution $p_0(z)$. The SB objective seeks a latent process $\{z_t\}_{t=0}^{K}$ that minimizes

$$\begin{aligned} \min_{P} \quad & \mathrm{KL}(P \,\|\, Q) \\ \text{s.t.} \quad & z_0 \sim p_0, \\ & z_K \sim p_T, \end{aligned}$$

where Q is a reference diffusion. The forward dynamics are parameterized as

$$z_{t+1} = z_t + f_\theta(z_t, t) + \sigma \epsilon_t.$$

with a corresponding backward update network to enforce consistency at the terminal time. A short horizon (around eight to twelve steps) keeps sampling efficient while providing enough temporal resolution to impose meaningful constraints along the path.

Each latent state z_t is decoded into an intermediate frame $\hat{x}_t$. A depth prediction module produces a depth map d_t, and a pose estimator predicts the relative transform $T_{t\to t+1}$ between consecutive frames. When camera intrinsics and poses are available, the model measures agreement between $\hat{x}_{t+1}$ and a rendering obtained by reprojecting $\hat{x}_t$ using d_t and $T_{t\to t+1}$. This reprojection consistency, together with penalties on epipolar line deviations and forward–backward pose composition, encourages stable cross-frame geometry and reduces 3D drift. When ground truth poses are unavailable, a compact motion or optical flow head predicts trajectories shared across SB steps and is regularized through forward–backward agreement and simple cycle consistency. The generator will use a lightweight architecture (for example, a compact UNet or small Vision Transformer with shared weights across SB time), mixed-precision training, and moderate spatial resolution to stay computationally feasible.

If time permits, a small imaging extension will be explored. Latent scenes will be decoded and passed through differentiable optical operators, such as defocus point spread function convolution for incoherent imaging or Fourier propagation for coherent setups (Gao et al. 2024; Zeng et al. 2025). Deviations between measured and simulated intensities will then penalize physically inconsistent generations. This extension is deliberately scoped to remain secondary to the main video generation contribution.

Evaluation

Evaluation is conducted along three axes. Geometric consistency is measured using reprojection error, warp-aligned photometric or feature discrepancies, and depth agreement on multiview datasets such as RealEstate10K, ScanNet, and TUM RGB-D (Zhou et al. 2018; Dai et al. 2017; Sturm et al. 2012), assessing whether generated sequences correspond to a coherent 3D scene. Generative quality and efficiency are evaluated using Fréchet Video Distance for temporal coherence and frame-level metrics (FID/KID), comparing the SB model against modern video diffusion baselines at matched wall-clock time to determine whether finite-time SB models achieve comparable or superior quality with fewer sampling steps. Finally, if conducted, physics fidelity is evaluated through forward-model residuals and spectral plausibility checks on imaging benchmarks such as NTIRE or DIV2K (Agustsson and Timofte 2017), and is included only if time permits beyond the core video experiments.

Discussion

Constraining the latent trajectory rather than only the final frame yields more stable temporal dynamics, stronger parallax, and reduced multiview drift than conventional diffusion models. By integrating Schrödinger Bridge–based optimal transport with geometry-aware and physics-inspired supervision, the model aims to generate video sequences consistent with a plausible 3D world. If validated empirically, these results would support SB methods as a foundation for physically grounded world models, benefiting robotics, simulation, and AR/VR, and enabling more trustworthy generative AI through structural consistency checks.

Conclusion

This project proposes a finite-time Schrödinger Bridge framework for physically and geometrically consistent video generation. By merging entropy-regularized optimal transport with multiview geometry and, where feasible, differentiable optical models, the method aims to produce video sequences whose evolution follows plausible 3D structure.

References

Agustsson, E.; and Timofte, R. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1122–1131. NTIRE / DIV2K series.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5828–5839.

De Bortoli, V.; Thornton, J.; Heng, J.; and Doucet, A. 2021. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Advances in Neural Information Processing Systems*, volume 34, 17695–17709.

Gao, Y.; Zhang, L.; Wang, P.; et al. 2024. A Brief Review on Differentiable Rendering. *Computer Graphics Forum*. Early access.

Hartley, R.; and Zisserman, A. 2004. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition.

Léonard, C. 2014. A Survey of the Schrödinger Problem and Its Connections with Optimal Transport. *Discrete and Continuous Dynamical Systems A*, 34(4): 1533–1574.

Melnik, A.; et al. 2024. Video Diffusion Models: A Survey. *arXiv preprint arXiv:2403.13038*.

Peyré, G.; and Cuturi, M. 2019. *Computational Optimal Transport*, volume 11 of *Foundations and Trends in Machine Learning*. Now Publishers.

Shi, Y.; Minner, R.; Vanden-Eijnden, E.; et al. 2023. Diffusion Schrödinger Bridge Matching. In *Proceedings of the 40th International Conference on Machine Learning*, ICML.

Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; and Cremers, D. 2012. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 573–580.

Zeng, H.; Li, M.; Xu, R.; et al. 2025. A Survey on Physics-Based Differentiable Rendering. *ACM Transactions on Graphics*. To appear.

Zhou, T.; Tucker, R.; Flynn, J.; and Snavely, N. 2018. RealEstate10K: Learning Image-Based Rendering from Large Video Collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1578–1587.