QUASAR: STRUCTURE-PRESERVING DENSITY-STATE DYNAMICS FOR RECURSIVE VIDEO PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual world models increasingly rely on structured latent dynamics, yet they are still evaluated mainly through decoded frames. In recursive prediction, latent states are fed back into later forecasts, so a model can produce accurate frames while its states leave the space its representation declares valid, a failure that standard forecasting metrics cannot detect. We introduce QUASAR, a video predictor whose local latent states are density matrices: symmetric, positive semidefinite, and unit trace. Motion-conditioned orthogonal transport preserves each state's spectrum, and learned Kraus channels, made complete by thin QR factorization, redistribute spectral mass. Every recurrent update is therefore completely positive and trace preserving (CPTP), so states remain admissible by construction rather than by post-hoc correction. Experiments show that forecast accuracy and state validity are distinct properties. On Human3.6M, an unconstrained transition matches QUASAR's four-frame mean squared error (MSE), yet its minimum eigenvalue reaches -1.36×10^{-3}, versus -8.79×10^{-9} for CPTP. Among matched transition controls under 16-frame recursion, CPTP gives the lowest mean MSE and highest mean structural similarity (SSIM) on both evaluation cohorts, and the lowest mean MSE at all three tested temporal rates. As a predictor, QUASAR reduces Human3.6M MSE by 17.95% relative to the strongest reproduced baseline and attains the highest SSIM on TaxiBJ and KITTI/Caltech, though not the lowest MSE on either. Added to PredFormer, the density-dynamics module reduces MSE by 14.69% with about 1.5% more parameters, although its gains vary across metrics. Density entropy is moderately associated with an occlusion proxy but only weakly with prediction error. These results indicate that admissible latent dynamics can be enforced without sacrificing forecast quality, and that latent states merit auditing alongside decoded frames.

1 INTRODUCTION

Visual world models increasingly encode scene evolution in structured latent states. Recent approaches separate appearance from dynamics, represent motion explicitly, and introduce transition laws for prediction, reasoning, and extrapolation Ren et al. (2026); Yang et al. (2026); Li et al. (2026); Radha & Goktas (2026). Convolutional, Transformer-based, and generative predictors have also advanced future-frame synthesis Gao et al. (2022); Tan et al. (2023); Tang et al. (2024); Ho et al. (2022). Together, these developments raise a fundamental question: how should a visual predictor be characterized beyond the accuracy of its decoded frames? Recursive prediction repeatedly feeds evolving internal states back into future forecasts. Standard metrics such as MSE, SSIM, PSNR, and LPIPS measure output quality but cannot determine whether the underlying predictive state remains admissible. Consequently, accurate decoded forecasts can coexist with invalid latent dynamics. We therefore ask whether structural validity can be enforced throughout recurrent prediction without sacrificing forecast quality, and whether this distinction becomes consequential under extended rollout.

We distinguish complementary aspects of predictive-state evaluation. **Observation fidelity** measures agreement between predicted and observed frames, while **state admissibility** measures membership in the representation's declared state space. We additionally probe **transition behavior** through controlled temporal-rate and numerical-composition variants under fixed conditioning. These measurements complement structured-dynamics approaches based on Koopman operators,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Hamiltonian systems, reversible computation, and unitary recurrence Lusch et al. (2018); Greydanus et al. (2019); Yu et al. (2020); Arjovsky et al. (2016), and characterize properties that observation-space metrics alone cannot establish. We introduce **QUASAR** as a constrained predictive-state framework for video forecasting. Rather than treating latent dynamics as unconstrained neural transitions, QUASAR defines an explicit admissible latent state space using density-matrix geometry and evolves these states through conditional CPTP maps that preserve validity by construction. Beyond spectrum-preserving evolution, learned Kraus channels allow the predictive state to redistribute spectral mass while remaining admissible; thin QR factorization enforces Kraus completeness, and a finite convex mixture controls transition strength. Each conditioned update is therefore completely positive and trace preserving (CPTP), keeping the state within the valid density space. In parallel, multiscale visual features are transported and modulated by the evolving density state through ρ-FiLM.

Experiments show that predictive accuracy and state validity need not coincide: unconstrained dynamics can remain competitive in observation space while violating the declared state geometry. QUASAR instead preserves admissibility by construction while retaining strong forecasting performance, with the benefits of structured evolution becoming more apparent under extended recursion. Our contributions are as follows.

- **Structure-preserving predictive states:** QUASAR represents local predictive states on an explicit density-matrix manifold and evolves them through motion-conditioned CPTP transitions, guaranteeing admissibility throughout recurrent forecasting.
- **Auditable recurrent dynamics:** We jointly evaluate decoded forecasts, state validity, and controlled transition variants, revealing properties of latent evolution that observation-space metrics cannot identify.
- **Cross-domain and cross-backbone validation:** Experiments across four datasets, 16-frame recursion, high-resolution forecasting, and PredFormer transfer establish the practical applicability of density-state conditioning beyond a single architecture or dataset.

2 Related Work

Visual prediction and latent world models: Video prediction spans recurrent architectures such as PredRNN, PredRNN++, and MIM, convolutional and Transformer-based predictors including SimVP, TAU, and PredFormer, and diffusion-based generation (Wang et al., 2017; 2018; 2019; Gao et al., 2022; Tan et al., 2023; Tang et al., 2024; Ho et al., 2022). VideoWorld 2 decouples task-relevant dynamics from appearance, while Chain of World represents embodied evolution through latent motion chains (Ren et al., 2026; Yang et al., 2026). QUASAR complements predictive and downstream evaluations with measurements of state admissibility and transition behavior.

Structured latent dynamics: Structured models constrain evolution through Koopman linearization, Hamiltonian dynamics, unitary recurrence, reversible computation, and structured state-space models (Lusch et al., 2018; Greydanus et al., 2019; Arjovsky et al., 2016; Yu et al., 2020; Gu et al., 2022). PhyDNet separates physics-oriented dynamics from residual appearance, while LDR performs kinematic latent integration for extrapolation (Le Guen & Thome, 2020; Li et al., 2026). QUASAR combines spectrum-preserving orthogonal conjugation with channels that can redistribute spectral mass, and audits temporal composition using Lie and Strang splitting (Strang, 1968).

Recursive belief-state estimation: Classical Kalman filtering propagates Gaussian means and covariances; deep variants introduce learned stochastic or locally linear dynamics (Kalman, 1960; Krishnan et al., 2015; Becker et al., 2019). Recurrent Kalman Networks retain explicit uncertainty under locally linear propagation. QUASAR maintains local positive-semidefinite, unit-trace states through conditional CPTP transitions, with numerical audits of state membership during rollout.

Density-matrix and quantum-structured representations: Density matrices support classical estimation and probabilistic modeling (González et al., 2022; 2025). UWM-JEPA uses unitary density-state prediction, preserving the joint-state spectrum, while QSWM studies complex-valued and density-matrix-like representations on elementary cellular automata (Radha & Goktas, 2026; Jiang et al., 2026). Unlike UWM-JEPA, which preserves the state spectrum through unitary evolution, QUASAR permits learned spectral redistribution through conditional CPTP evolution while maintaining state admissibility. QUASAR further operates on spatially local density states for

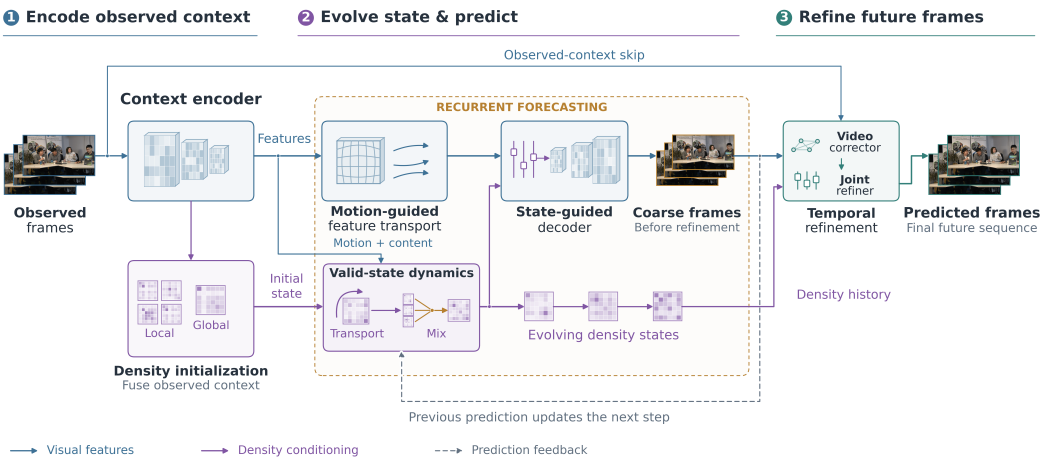

Figure 1: Overview of QUASAR. Observed frames are encoded into visual features and density states. During recurrent forecasting, motion-guided feature transport and valid-state dynamics jointly produce coarse future frames, which are fed back to the next step. A temporal refiner then combines the coarse sequence, observed context, and density history to generate the final predictions.

natural-video forecasting and jointly evaluates forecast quality, state admissibility, and transition behavior.

3 METHOD

3.1 OVERVIEW

Given observed frames $x_{1:T}$, QUASAR predicts the next P frames through the three-stage pipeline in Fig. 1. First, the observed sequence is encoded into multi-scale visual features and valid density states. Second, future states and visual features are evolved recurrently: motion guides feature transport, while density states undergo structure-preserving dynamics and condition frame decoding through ρ-FiLM. Each coarse prediction is fed back to condition the next forecasting step. Finally, a temporal refiner jointly processes the observed context, coarse forecasts, and density history to produce the final future sequence.

3.2 OBSERVED-CONTEXT ENCODING AND STATE INITIALIZATION

The context encoder first extracts visual features from the observed sequence. In parallel, at each observed time t and bottleneck location g, it predicts M vectors $a_{t,g,m} \in \mathbb{R}^d$ and logits $\ell_{t,g,m}$. Their normalized weights and directions are

$$p_{t,g,m} = \frac{\exp(\ell_{t,g,m})}{\sum_{j=1}^{M} \exp(\ell_{t,g,j})}, \tag{1}$$

$$\psi_{t,g,m} = \frac{a_{t,g,m}}{\max(\|a_{t,g,m}\|_2, \epsilon_a)}. \tag{2}$$

Here d is the state dimension and $\epsilon_a > 0$ stabilizes normalization. These components are collected into

$$F_{t,g} = \left[\sqrt{p_{t,g,1}}\psi_{t,g,1}, \ldots, \sqrt{p_{t,g,M}}\psi_{t,g,M}\right], \tag{3}$$

$$\rho_{t,g} = \frac{F_{t,g}F_{t,g}^{\top} + \varepsilon I_d}{\mathrm{Tr}(F_{t,g}F_{t,g}^{\top}) + \varepsilon d}. \tag{4}$$

where $F_{t,g} \in \mathbb{R}^{d\times M}$ and $\varepsilon > 0$ is a small jitter term. This construction makes every state symmetric, positive semidefinite, and unit trace:

$$\mathcal{D}_d = \left\{\rho \in \mathbb{R}^{d\times d} : \rho = \rho^\top,\ \rho \succeq 0,\ \mathrm{Tr}(\rho) = 1\right\}. \tag{5}$$

To initialize forecasting, density-aware temporal attention summarizes the observed context and predicts a skew-symmetric context matrix A_c together with location-dependent gates $\alpha_g \in [0, 1]$. Setting $U_c = \exp(A_c)$, the final observed state is corrected as

$$\rho_{0,g} = (1 - \alpha_g)\rho_{T,g} + \alpha_g U_c \rho_{T,g} U_c^\top. \tag{6}$$

Because U_c is orthogonal and the update is a convex combination, $\rho_{0,g}$ remains in $\mathcal{D}_d$. This provides the initial density state used by the recurrent forecasting stage.

3.3 Recurrent Valid-State Forecasting

For every future step $h \in \{1, \ldots, P\}$, QUASAR updates the density state and visual prediction together. A motion encoder first computes a motion representation $\mathbf{m}$ from the final two observed bottleneck features. This representation controls both state evolution and visual feature transport.

3.3.1 Motion-conditioned state transport

The motion embedding parameterizes a skew-symmetric matrix $A_h^\top = -A_h$, giving

$$U_h = \exp(-\Delta t A_h), \tag{7}$$

$$\rho_{h,g}^{\mathrm{rev}} = U_h \rho_{h-1,g} U_h^\top. \tag{8}$$

Here $\Delta t \geq 0$ denotes the transition step. Since U_h is orthogonal, this transport preserves positivity, trace, and spectrum while moving the previous predictive state forward in time.

3.3.2 Conditional CPTP Evolution

After orthogonal transport, the motion embedding conditions two learned families of operator blocks. Let

$$D(\mathbf{m}) = \mathrm{reshape}_{L\times d\times d\times 2}\left[W_2\,\mathrm{GELU}(W_1\mathbf{m} + b_1) + b_2\right],$$

$$A_\ell(\mathbf{m}) = \bar{A}_\ell + 0.1 D_{\ell,:,:,0}(\mathbf{m}), \tag{9}$$

$$C_\ell(\mathbf{m}) = \bar{C}_\ell + 0.1 D_{\ell,:,:,1}(\mathbf{m}), \tag{10}$$

where $\bar{A}_\ell$ and $\bar{C}_\ell$ are learned base matrices. Together with an identity reference, they form

$$B_h = \begin{bmatrix} I_d \\ A_1(\mathbf{m}) \\ \vdots \\ A_L(\mathbf{m}) \\ C_1(\mathbf{m}) \\ \vdots \\ C_L(\mathbf{m}) \end{bmatrix} \in \mathbb{R}^{Jd\times d}, \qquad J = 2L + 1. \tag{11}$$

Thin QR gives

$$B_h = V_h R_h, \tag{12}$$

$$V_h = [K_{h,1}^\top, \ldots, K_{h,J}^\top]^\top, \tag{13}$$

where the identity reference ensures full rank. Since $V_h^\top V_h = I_d$,

$$\sum_{j=1}^{J} K_{h,j}^\top K_{h,j} = I_d, \tag{14}$$

$$\mathcal{E}_h(\rho) = \sum_{j=1}^{J} K_{h,j}\rho K_{h,j}^\top, \tag{15}$$

which defines a CPTP channel. In the native rollout, the motion embedding is fixed across forecast steps, so this operator set is shared across steps and spatial locations. The operators are shared spatially, while the transition intensity is modulated by local visual content and prediction feedback. Let b_T denote the final observed bottleneck feature, and define $p_0 = x_T$ and $p_{h-1} = \bar{x}_{T+h-1}$ for $h > 1$. The feedback multiplier and nonnegative local rates are

$$\begin{aligned} \omega_{h,\ell} &= 2\sigma(f_\ell(p_{h-1}))\,, \\ \gamma_{h,\ell,g} &= 2\ \mathrm{softplus}(a_\ell)\,\sigma(s_\ell(b_T)_g)\ \omega_{h,\ell}, \end{aligned} \tag{16}$$

where a_ℓ is a learned log-rate, s is the spatial-content head, and f is the prediction-feedback head. The resulting channel intensity is

$$\mu_{h,g} = \Delta t\ \max\left(0, \sum_{\ell=1}^{L} \gamma_{h,\ell,g}\right). \tag{17}$$

Thus, the preceding prediction conditions the intensity of the following state transition. Using Poisson weights

$$w_n(\mu) = e^{-\mu}\frac{\mu^n}{n!}, \tag{18}$$

and $\mathcal{E}_h^0 = \mathcal{I}$, we define

$$\begin{aligned} \mathcal{C}_{h,g}^{(Q)} = &\sum_{n=0}^{Q-1} w_n(\mu_{h,g})\mathcal{E}_h^n \\ &+ \left(1 - \sum_{n=0}^{Q-1} w_n(\mu_{h,g})\right)\mathcal{E}_h^Q, \end{aligned} \tag{19}$$

where the final term collects the remaining Poisson mass, preserving nonnegative normalized weights. The recurrent update is

$$\rho_{h,g} = \mathcal{C}_{h,g}^{(Q)}(\rho_{h,g}^{\mathrm{rev}}) = \Phi_{h,g}(\rho_{h-1,g}), \tag{20}$$

so, for fixed conditioning,

$$\rho_{h-1,g} \in \mathcal{D}_d \quad \Longrightarrow \quad \rho_{h,g} \in \mathcal{D}_d. \tag{21}$$

Thus, admissibility is preserved at every recurrent step. In FP32, only symmetrization and scalar trace normalization are used to correct numerical drift, without eigenvalue clipping.

Complexity: Density storage is $O(BGd^2)$, while channel evolution costs $O(PBGQJd^3)$. QR and matrix exponentials are computed once per forecast step. We use $d = 8$, $M = 4$, $J = 17$, and $Q = 2$.

3.4 State-Guided Prediction and Refinement

At scale s, motion-guided transport propagates the final observed feature,

$$z_h^{(s)} = \mathcal{D}_h^{(s)}\left(e_T^{(s)}; \mathbf{m}, q_h^{(s)}\right), \tag{22}$$

where $\mathbf{m}$ and $q_h^{(s)}$ encode motion and step conditioning. The evolved density state then modulates the feature via ρ-FiLM,

$$\tilde{z}_h^{(s)} = \left(1 + \Gamma_h^{(s)}(\rho_h)\right) \odot z_h^{(s)} + \beta_h^{(s)}(\rho_h). \tag{23}$$

A decoder predicts a bounded residual δ_h and mask s_h,

$$\bar{x}_{T+h} = \mathrm{clip}_{[0,1]}\left[\bar{x}_{T+h-1} + s_h \odot \delta_h\right]. \tag{24}$$

Each coarse prediction conditions the next recurrent update. After P steps, a temporal refiner combines the observed frames, coarse predictions, and detached pooled density features to produce $\hat{x}_{T+1:T+P}$, with an observed-context skip preserving source information.

3.5 Training Objective

QUASAR is trained end-to-end using pixel, structural, perceptual, temporal, and state-regularization terms:

$$\begin{aligned}\mathcal{L} = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{vgg}} \\ + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}}.\end{aligned} \tag{25}$$

Here $\mathcal{L}_{\text{mse}}$ and $\mathcal{L}_{\text{ssim}}$ measure pixel and structural fidelity, $\mathcal{L}_{\text{vgg}}$ and $\mathcal{L}_{\text{lpips}}$ provide perceptual supervision, $\mathcal{L}_{\text{temp}}$ encourages temporal consistency, and $\mathcal{L}_{\text{ent}}$ regularizes the predictive density states. The corresponding Human3.6M weights are reported in Table 1.

For $P > 1$, entropy is encouraged to remain within a target range while changing smoothly over time:

$$\begin{aligned}\mathcal{L}_{\text{ent}} = \frac{1}{P}\sum_{h=1}^{P}\left([s_{\min} - \bar{S}_h]_+^2 + 10[\bar{S}_h - s_{\max}]_+^2\right) \\ + \frac{0.05}{P-1}\sum_{h=2}^{P}(\bar{S}_h - \bar{S}_{h-1})^2.\end{aligned} \tag{26}$$

Here $\bar{S}_h$ is the mean state entropy, $s_{\min}$ and $s_{\max}$ define its target range, and $[u]_+ = \max(u, 0)$. For Human3.6M, we use $s_{\min} = 0.3$, $s_{\max} = 1.5$, and $\lambda_{\text{ent}} = 0.1$.

4 Experiments

4.1 Setup

Datasets and protocols: We use Human3.6M ($4 \to 4$ frames at 256×256; implementation details in Table 1), TaxiBJ ($4 \to 4$ at 32×32), and KITTI/Caltech ($10 \to 1$ at 128×128). Five JVET CTC sequences Bossen et al. (2020) at up to 3840×2160 serve as high-resolution stress tests.

Metrics: Observation fidelity is measured by MSE, MAE, SSIM, PSNR, and LPIPS; LPIPS is not reported for TaxiBJ. State admissibility is measured by the minimum eigenvalue $\lambda_{\min}$ of the density states, the maximum trace error $|\operatorname{Tr}(\rho) - 1|$, and the deviation of $\sum_j K_j^\top K_j$ from I_d (Kraus-completeness error).

Baselines: Forecasting baselines are E3D-LSTM, MAU, SimVP variants, PredFormer, GMG, and SVQ (Table 2).

Transition controls: All controls share the visual architecture, objective, and optimization schedule, and differ only in the density-state update. Unconstrained applies a learned update without structural constraints; Trace normalized rescales that update to unit trace; Gram repair and Eigenvalue projection restore admissibility after the update by re-forming a normalized Gram matrix and by clipping negative eigenvalues, respectively; Identity keeps the density state fixed. Cohorts, seeds, and temporal rates are given in the table captions.

4.2 Forecasting Results

On Human3.6M, QUASAR attains the best MSE, MAE, SSIM, and PSNR, reducing MSE by 17.95% relative to SimVP+IncepU, while SVQ has the lowest LPIPS (Table 2). On TaxiBJ, QUASAR attains the best SSIM and PSNR, whereas GMG and SVQ achieve lower MSE and MAE. On KITTI/Caltech, widened SimVP achieves lower MSE and MAE and higher PSNR, while QUASAR attains the best SSIM and LPIPS.

High-Resolution Performance: On the JVET sequences (FourPeople 1280×720, MarketPlace and RitualDance 1920×1080, DaylightRoad2 and Tango2 3840×2160; $4 \to 4$ clips), a Repeat Last baseline quantifies the gain beyond copying the final observed frame. QUASAR outperforms both baselines on every sequence (Table 3); qualitative comparisons on FourPeople and MarketPlace appear in Appendix A.

Table 1: Key implementation details for Human3.6M.

Setting	Configuration
Optimization	AdamW; LR 2×10^{-4}; 5-epoch warm-up + cosine decay; 50 epochs
Batching	Batch size 1; 500 minibatches/epoch; EMA 0.9995
Data / protocol	$4 \rightarrow 4$, 256×256, interval 5; 71,378 train / 2,026 val windows
Density state	32×32 grid; $d = 8$, $M = 4$, $J = 17$, $Q = 2$, $\Delta t = 0.15$
Loss weights	MSE / SSIM / VGG / LPIPS / temporal / entropy = $0.30/1.0/0.05/0.01/0.05/0.10$
Training setup	RTX 3090 (24 GB)

Table 2: Native forecasting results. Parameters are in millions; bold indicates the best mean within each dataset.

Model	Params	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Human3.6M: $4 \rightarrow 4$, 256×256						
E3D-LSTM	60.850	0.0035147	0.020799	0.848524	24.772	0.393745
PredFormer	12.731	0.0015056	0.012919	0.935030	29.417	0.063367
SVQ	30.702	0.0012580	0.009962	0.949581	30.477	**0.031229**
SimVP+IncepU	41.156	0.0012109	0.011130	0.946515	30.514	0.045738
QUASAR	46.868	**0.0009935**	**0.008354**	**0.964006**	**32.230**	0.033918
TaxiBJ: $4 \rightarrow 4$, 32×32						
E3D-LSTM	50.987	0.0004579	0.011747	0.956392	35.744	n/a
MAU	4.410	0.0002211	0.008460	0.974902	38.549	n/a
SimVP+IncepU	13.795	0.0002170	0.008321	0.976387	38.755	n/a
SimVP+gSTA-S	9.961	0.0002125	0.008184	0.977154	38.895	n/a
PredFormer	6.333	0.0002119	0.008225	0.975663	38.729	n/a
GMG	31.223	**0.0002093**	0.008198	0.976554	38.791	n/a
SVQ	11.575	0.0002102	**0.008055**	0.977900	39.010	n/a
QUASAR	46.603	0.0002149	0.008110	**0.980212**	**39.162**	n/a
KITTI/Caltech: $10 \rightarrow 1$, 128×128						
SimVP	15.634	0.0040836	0.034379	0.830820	24.558	0.081540
SimVP+IncepU	12.154	0.0045692	0.037662	0.806159	23.915	0.113542
SimVP, widened	46.806	**0.0038617**	**0.033541**	0.836355	**24.780**	0.078023
QUASAR	46.908	0.0041807	0.033788	**0.846451**	24.641	**0.062698**

Table 3: High-resolution forecasting results. Entries report MSE $\times 10^3$ / SSIM / PSNR. Bold indicates the best result for each sequence.

Sequence	Repeat last	SimVP+IncepU	QUASAR
FourPeople	0.8310 / .9306 / 31.25	0.9245 / .8770 / 30.34	**0.6670 / .9417 / 32.28**
MarketPlace	4.7910 / .6959 / 25.94	4.6320 / .7115 / 26.08	**4.5260 / .7248 / 26.21**
RitualDance	14.4824 / .7649 / 19.46	13.8240 / .7751 / 19.78	**13.2927 / .7886 / 20.00**
DaylightRoad2	8.5183 / .5360 / 20.95	8.0145 / .5820 / 21.18	**7.6048 / .6215 / 21.46**
Tango2	16.6903 / .6803 / 18.28	16.3210 / .6982 / 18.40	**16.0189 / .7147 / 18.51**

4.3 Ablations and State Analysis

4.3.1 Transition controls and backbone transfer

All transition controls use a four-frame training horizon. Each recursive call re-encodes the latest four frames. Identity keeps the density state fixed within that block while CPTP gives the lowest mean 16-frame MSE and highest mean SSIM on both cohorts (Table 4). The PredFormer study fixes the backbone, FiLM location, recipe, split, and seed pair. CPTP reduces MSE by 14.69%, Gram repair by 12.08%, and generic recurrence by 6.13%. CPTP also attains the highest PSNR, while Generic conditioning obtains the highest SSIM, indicating that the improvements are metric-

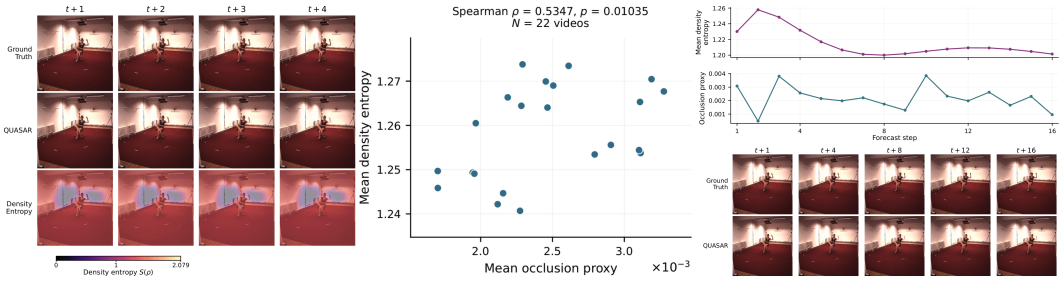

Figure 2: **Density entropy and visibility.** Left: predictions and entropy maps; center: entropy–occlusion association ($\rho = 0.5347$, $p = 0.01035$); right: 16-frame evolution.

dependent rather than uniformly dominant (Table 5). The added modules use approximately 1.5% of baseline parameters.

Table 4: Sixteen-frame recursion after four-frame training. Each cohort has 512 windows. Values are means $\pm$ sample standard error over seeds $\{0, 43, 44, 45\}$; MSE is multiplied by 10^3.

Transition	MSE ↓	SSIM ↑
Original cohort: 7 source videos		
Unconstrained	2.62468 ± 0.04700	0.945845 ± 0.001510
Gram repair	2.55689 ± 0.01711	0.947979 ± 0.000426
Eigenvalue projection	2.57500 ± 0.02938	0.947906 ± 0.000934
Identity	2.60033 ± 0.09115	0.947101 ± 0.002507
QUASAR (CPTP)	$\mathbf{2.51680 \pm 0.01040}$	$\mathbf{0.949562 \pm 0.000125}$
Balanced cohort: 22 source videos		
Unconstrained	2.84126 ± 0.05208	0.947155 ± 0.001414
Gram repair	2.79412 ± 0.02115	0.949124 ± 0.000461
Eigenvalue projection	2.79332 ± 0.02981	0.949050 ± 0.000873
Identity	2.80951 ± 0.08550	0.948373 ± 0.002379
QUASAR (CPTP)	$\mathbf{2.73481 \pm 0.01524}$	$\mathbf{0.950691 \pm 0.000128}$

Table 5: PredFormer conditioning at a fixed FiLM location. Means $\pm$ sample standard deviations use seeds 1 and 2. Metric scaling applies to both means and deviations.

Metric	None	Generic	Gram	CPTP
MSE $\times 10^3$ ↓	1.50558 ± 0.02546	1.41327 ± 0.00656	1.32367 ± 0.02476	$\mathbf{1.28446 \pm 0.02192}$
SSIM $\times 10^2$ ↑	93.5030 ± 0.1618	$\mathbf{93.9323 \pm 0.0328}$	93.7780 ± 0.1136	93.6161 ± 0.1375
PSNR ↑	29.417 ± 0.115	29.851 ± 0.018	29.862 ± 0.102	$\mathbf{29.994 \pm 0.084}$

4.3.2 State validity, persistence, and entropy

Table 6 shows that unconstrained and trace-normalized updates violate the density-state geometry, whereas Gram, projection, and CPTP remain near numerical tolerance. QUASAR preserves admissibility by construction while retaining competitive short-horizon accuracy. Under 16-frame recursion, CPTP achieves the best observed mean MSE and SSIM on both cohorts (Table 4); persistent propagation retains nearly identical accuracy with validity diagnostics near numerical tolerance (Appendix B). Density entropy is moderately associated with an independent occlusion proxy (Spearman $\rho = 0.5347$, $p = 0.01035$) but only weakly with prediction error, supporting its interpretation as visibility-related rather than calibrated predictive uncertainty.

Temporal rate and composition: CPTP achieves the lowest observed mean MSE at all tested temporal rates (Table 7), while Strang splitting remains close, indicating limited forecast sensitivity to this alternative composition; these results do not directly measure composition error.

Table 6: Human3.6M four-frame state audit under the protocol of Table 2. Bold indicates the best forecast means and least-violating $\lambda_{\min}$.

Transition	MSE ↓	SSIM ↑	PSNR ↑	$\lambda_{\min}$	Max. trace error
Unconstrained	0.0009936	0.963815	32.207	-1.36×10^{-3}	9.06×10^{-6}
Trace normalized	0.0010375	0.962110	31.980	-2.85×10^{-3}	2.38×10^{-7}
Gram repair	0.0010119	0.962822	32.119	-6.56×10^{-8}	2.38×10^{-7}
Eigenvalue projection	**0.0009693**	0.964090	**32.305**	-2.21×10^{-8}	2.38×10^{-7}
QUASAR (CPTP)	0.0009935	**0.964006**	32.230	$\mathbf{-8.79 \times 10^{-9}}$	2.38×10^{-7}

Table 7: Temporal-rate study on 264 clips from 22 videos. Intervals $1\times, 2\times, 4\times$ use $\Delta t = 0.15, 0.30, 0.60$; MSE $\times 10^3$, mean $\pm$ std. over four seeds.

Transition	$1\times$	$2\times$	$4\times$
Identity	0.98541 ± 0.02108	1.94618 ± 0.01998	3.24317 ± 0.01612
Unconstrained	1.01143 ± 0.04050	1.96876 ± 0.03830	3.26131 ± 0.05376
Gram repair	0.98255 ± 0.01151	1.94659 ± 0.00869	3.23814 ± 0.01754
Eigenvalue projection	0.98775 ± 0.03020	1.94859 ± 0.01897	3.24134 ± 0.03552
QUASAR (CPTP)	$\mathbf{0.97765 \pm 0.00631}$	$\mathbf{1.93988 \pm 0.00853}$	$\mathbf{3.23158 \pm 0.03862}$
CPTP + Strang	0.98552 ± 0.02121	1.94659 ± 0.02057	3.24415 ± 0.01709

5 Discussion

Our results reveal a distinction that observation-space evaluation alone cannot capture. Good forecasts do not necessarily imply valid latent dynamics. Unconstrained transitions remain competitive over short horizons despite leaving the declared density-state space, whereas CPTP evolution preserves admissibility by construction and achieves the strongest observed performance under extended recursion.

5.1 Limitations and Future Work

Scalability: Matrix exponentials, QR factorization, and Kraus operations incur $O(d^3)$ cost; low-rank, sparse, or block-structured CPTP parameterizations could enable larger states and finer grids. **State semantics:** Density states are mathematically interpretable but are not explicitly aligned with semantic scene factors; future work could connect their structure to objects, motion, and physical attributes. **Spatial dynamics:** Shared transition operators improve efficiency and auditability but may limit spatial specialization; region-conditioned Kraus channels could better model independently evolving objects and agents.

6 Conclusion

We presented QUASAR, a video prediction framework that directly validates latent dynamics rather than inferring their quality solely from decoded forecasts. QUASAR represents predictive states as density matrices and evolves them through motion-conditioned CPTP transitions, guaranteeing admissibility by construction rather than post-hoc correction. Experiments show that forecast quality and state validity are distinct: unconstrained dynamics can produce competitive predictions while violating the declared state geometry, whereas CPTP evolution preserves validity and achieves strong recursive forecasting across cohorts and temporal rates. The formulation also transfers to PredFormer with minimal parameter overhead. Overall, QUASAR demonstrates that structurally valid, auditable latent dynamics can be enforced without sacrificing predictive performance, providing a foundation for visual world models whose internal evolution can be inspected directly.

AI USE STATEMENT

Generative AI tools were used solely for copy-editing and language polishing, including improving grammar, clarity, readability, and stylistic flow of the manuscript. Generative AI was not used for developing research ideas, conducting literature searches, collecting or analyzing data, deriving mathematical formulations, designing experiments, interpreting results, or drawing scientific conclusions. All AI-assisted edits were carefully reviewed by the authors to ensure that the original scientific meaning and intent were preserved. The authors take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work focuses on methodological research in video prediction and structured latent dynamics. We did not recruit human participants, conduct human-subject experiments, or collect new personally identifiable or sensitive data. All experiments were performed using established research datasets and benchmark sequences for the purpose of model development and evaluation. The proposed method does not involve biometric identification, surveillance, decision-making about individuals, or other high-risk human-facing applications. We are not aware of any direct ethical risks beyond those generally associated with video prediction technologies, including the possibility of downstream misuse when applied to real-world visual data. The authors are responsible for the integrity of the experiments, analyses, and reported results, and no undisclosed conflicts of interest affected the conduct or interpretation of this research.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility of our results. The complete QUASAR architecture, predictive-state construction, motion-conditioned orthogonal transport, CPTP transition parameterization, density-conditioned visual prediction mechanism, and training objective are specified mathematically in Section 3. The experimental section reports the datasets, prediction horizons, evaluation protocols, model parameter counts, comparison settings, and the number of training seeds used for the reported results. Controlled transition experiments use matched architectures, objectives, optimization settings, and training horizons to enable direct comparison between transition mechanisms. We additionally report state-validity diagnostics, persistent-rollout evaluations, temporal-rate experiments, and numerical audits of trace preservation, minimum eigenvalues, and Kraus completeness. Together, these methodological and experimental details are provided to support independent implementation, reproduction, and verification of the reported findings.

REFERENCES

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 1120–1128. PMLR, 2016.

Philipp Becker, Harit Pandya, Gregor Gebhardt, Cheng Zhao, C. James Taylor, and Gerhard Neumann. Recurrent kalman networks: Factorized inference in high-dimensional deep feature spaces. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 544–552, 2019.

Frank Bossen, Jill Boyce, Xiang Li, Vadim Seregin, and Karsten Sühring. Vtm common test conditions and software reference configurations for sdr video. Technical Report JVET-T2010, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, October 2020.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. SimVP: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3170–3180, 2022.

Fabio A. González, Alejandro Gallego, Santiago Toledo-Cortés, and Vladimir Vargas-Calderón. Learning with density matrices and random features. *Quantum Machine Intelligence*, 4:23, 2022. doi: 10.1007/s42484-022-00079-9.

Fabio A. González, Raúl Ramos-Pollán, and Joseph Gallego. Kernel density matrices for probabilistic deep learning. *Quantum Machine Intelligence*, 7:94, 2025. doi: 10.1007/s42484-025-00299-9.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Hailong Jiang, Emran Hossain, Feng Yu, Jianfeng Zhu, Guilin Zhang, and Wulan Guo. Quantum-structured world models (QSWMs) for predictive latent dynamics. *arXiv preprint arXiv:2608.05371*, 2026.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11474–11484, 2020.

Haodong Li, Shaoteng Liu, Tianyu Wang, Chongjian Ge, Sihui Ji, Jiahan Zhang, Xin Lin, Haolin Lu, Zhe Lin, and Manmohan Chandraker. Learning how the world evolves: Extrapolative video world models via latent dynamics reasoning. *arXiv preprint arXiv:2608.09926*, 2026.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018. doi: 10.1038/s41467-018-07210-0.

Santosh Kumar Radha and Oktay Goktas. UWM-JEPA: Predictive world models that imagine in belief space. *arXiv preprint arXiv:2605.25313*, 2026.

Zhongwei Ren, Yunchao Wei, Xiao Yu, Guixun Luo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. VideoWorld 2: Learning transferable knowledge from real-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40569–40580, 2026.

Gilbert Strang. On the construction and comparison of difference schemes. *SIAM Journal on Numerical Analysis*, 5(3):506–517, 1968.

Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z. Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18770–18782, 2023.

Yujin Tang, Lu Qi, Fei Xie, Xiangtai Li, Chao Ma, and Ming-Hsuan Yang. PredFormer: Transformers are effective spatial-temporal predictive learners. *arXiv preprint arXiv:2410.04733*, 2024.

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 5123–5132. PMLR, 2018.

Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9154–9162, 2019.

Fuxiang Yang, Donglin Di, Lulu Tang, Xuancheng Zhang, Lei Fan, Hao Li, Wei Chen, Tonghua Su, and Baorui Ma. Chain of world: World model thinking in latent motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6675–6684, 2026.

Wei Yu, Yichao Lu, Steve Easterbrook, and Sanja Fidler. Efficient and information-preserving future frame prediction and beyond. In *International Conference on Learning Representations*, 2020.

A Additional Qualitative Results

Figures 3 and 4 provide high-resolution qualitative comparisons on the FourPeople and MarketPlace sequences. QUASAR preserves fine text, facial details, clothing boundaries, and the structure of moving people across the predicted frames.

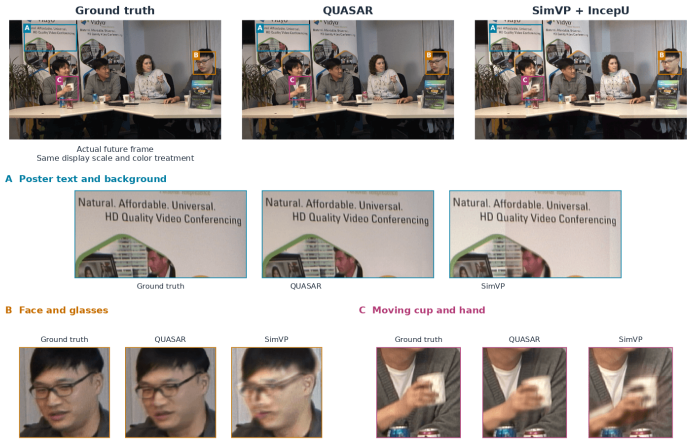

Figure 3: Qualitative comparison on FourPeople. QUASAR better preserves fine text, facial details, and moving-object structure than SimVP+IncepU.

B Persistent-State and Entropy Analysis

We provide additional results for persistent density-state propagation, numerical validity, and entropy associations. Persistent propagation produces nearly identical forecasting performance to block-recursive prediction, while numerical state diagnostics remain near floating-point tolerance. We additionally report window- and video-level associations between density entropy, occlusion, and prediction error.



Figure 4: Qualitative comparison on the MarketPlace sequence (1920×1080). QUASAR better preserves facial details, clothing boundaries, and the structure of the moving person compared with SimVP.

Table 8: Persistent evolution and state statistics. Panel A compares block-recursive and persistent-state propagation over 16 predicted frames on 512 windows. Panel B audits numerical validity during persistent evolution, and Panel C reports associations between density entropy, occlusion, and prediction error

A. Persistent rollout: 16 predicted frames		
Rollout mode	MSE ↓	SSIM ↑
Block-recursive	0.00248950	0.951200
Persistent state	0.00248910	0.951240
Persistent + adaptive conditioning	0.00247650	0.951890
B. Persistent-state numerical audit		
Diagnostic	Extremum	Value
Post-cleanup trace deviation	Maximum	3.05×10^{-7}
Post-cleanup eigenvalue	Minimum	-1.14×10^{-9}
Kraus-completeness error	Maximum	5.12×10^{-7}
C. Density-entropy associations		
Comparison	Spearman ρ	Statistical summary
Entropy–occlusion (video)	0.5347	$p = 0.010$
Entropy–occlusion (window)	0.3100	95% CI $[0.02, 0.54]$
Entropy–prediction error (video)	0.06	–