Ken Graham, 24 October 2025 Ken@maketech.com.au

Diffusion Forecasting White Paper v4

Table of Contents

Abstract

Reverse-time stochastic differential equation (SDE) diffusion models are emerging as a powerful approach for forecasting across vision, science, finance, and defense. These models learn to invert a noise-driven diffusion process to generate future outcomes – from video frames to weather maps – without traditional autoregression. This article provides a comprehensive overview of diffusion-based forecasting, defining key terms and tracing the technique's evolution from image generation to video prediction and time-series modeling. We synthesize recent experiments (2023–2025) and industrial prototypes by Google DeepMind, OpenAI (Sora), xAI, Stanford/MIT, Scale AI (Thunderforge), Runway, and others. A unifying theme is the proposal of a **unified latent-space** for perception, reasoning, and prediction: we discuss this philosophy and its implications. Cross-domain applications are examined, including robotics (physical trajectory prediction), meteorology (probabilistic nowcasting and 10-day forecasts), financial markets (regime-shift detection and scenario stress-testing), chemistry/biology (molecular and protein simulations), and military/geopolitical intelligence (agent-based simulations). We conclude with a forward-looking road-map of expected progress over the next 36 months – outlining milestones in real-time physics simulation, improved

weather forecasting skill, financial risk modeling, and AI-driven scenario planning – and we highlight open challenges and ethical considerations for this nascent field.

1. Introduction

Diffusion models – originally popularized for generating realistic images from noise – are now being harnessed to **predict future events** across diverse domains. Unlike classical predictive models, diffusion-based forecasters treat forecasting as a *generative modeling* problem: they learn a data-driven simulation of "what could happen next" by iteratively denoising random noise into plausible outcomes. This reverse-time SDE approach offers a unique capability: it naturally produces *probabilistic* forecasts (ensembles of many possible futures) rather than a single deterministic prediction. Recent advances suggest that a single latent generative model could integrate perception, reasoning, and prediction – essentially **simulating the world** in its hidden variables. AI researcher Ken has argued for a unified latent space that bridges sensory input and predictive imagination, allowing an AI to internally **"think" in latent space** and roll out future scenarios before deciding on actions. In this article, we explore that vision and survey the state of the art in diffusion-based forecasting. We maintain a neutral, encyclopedic tone appropriate for a Wikipedia-style overview, while highlighting exciting demos and research breakthroughs (2023–2025) that ground the discussion. A forward-looking section projects the likely progression in capabilities over the next 0–36 months, and we close with challenges and ethical considerations as this technology matures.

2. Key Terminology & Definitions

Term	Definition
reverse-time SDE	A stochastic differential equation integrated *backwards in time* to transform noise into data samples. In diffusion models, the reverse-time SDE describes how to gradually *remove* noise from an initially random signal, producing a coherent output (e.g. an image or forecast). This is the mathematical backbone of score-based generative modeling.
diffusion model	A generative model that learns to synthesize data by first adding noise (forward diffusion) and then training a network to remove noise stepwise (reverse diffusion). Notable for high-quality image generation, diffusion models produce samples by refining random noise through many iterations. They can capture complex data distributions without the mode collapse of GANs.
video diffusion model (VDM)	A diffusion model specialized for video generation. VDMs extend image diffusion to the time

	dimension, generating sequences of video frames. They often use 3D U-Nets or Transformers over space-time and can produce temporally coherent video clips. Ho et al.'s *Video Diffusion Models* (2022) demonstrated the first high-fidelity unconditional video synthesis using diffusion.
autoregression-free (AR-free) diffusion	A diffusion generation approach that avoids sequential token/frame generation. Traditional sequence models (e.g. language models) generate one step at a time autoregressively. In contrast, AR-free diffusion generates the entire output in parallel by iteratively refining noise. This enables faster and more coherent generation of long sequences (text or video) by treating them as a whole, at the cost of multiple refinement steps.
latent space	The hidden vector space in which data is represented in a compressed form. A latent space encodes high-dimensional inputs (images, text, sensor signals) into a lower-dimensional representation capturing essential features. Generative models (VAEs, diffusion models) operate in latent space to modify and sample data. A *unified latent space* refers to using one common representation for different modalities or tasks (vision, language, planning, etc.), enabling cross-modal understanding.
score network	The neural network in a diffusion model that estimates the *score* (gradient of the log-density) of the data distribution at a given noisy input. By predicting the noise residual or denoising direction, the score network guides the reverse diffusion process. Essentially, it tells how to tweak a noisy sample to make it more data-like at each step.
Frame-aware Video Diffusion (FVDM)	A novel video diffusion approach that assigns each frame its own noise schedule via a *vectorized timestep*. Introduced by Liu *et al.* (2024), FVDM lets different frames denoise at different rates, capturing fine-grained temporal dependencies beyond standard video diffusion. This mitigates issues like temporal blurriness or forgetfulness by treating frames individually during diffusion.
optical flow	The apparent motion of pixels between video frames, represented as a 2D vector field. Optical

	flow indicates how objects move from one frame to the next. It is often used as an intermediate representation or conditioning signal in video prediction and frame interpolation models, guiding diffusion models to maintain temporal consistency by aligning moving elements.
Slot Attention	A neural mechanism for unsupervised object-centric representation. Slot Attention (Locatello *et al.*, 2020) uses iterative attention to carve an input (image, scene) into a fixed number of "slots," each slot encoding an object or component. In the context of latent world models, Slot Attention can disentangle a scene into independent parts, which a diffusion model might then predict or manipulate separately, improving compositional generalization.
Perceiver IO	A transformer-based architecture (Jaegle *et al.*, 2021) that uses a latent array to **encode and decode** arbitrary inputs and outputs. Perceiver IO reads high-dimensional multimodal inputs (images, audio, etc.) into a fixed-size latent bottleneck via cross-attention, then generates task-specific outputs from that latent. It enables *cross-modal fusion* by mapping different modalities into one unified latent space for processing.
Sora	OpenAI's text-to-video diffusion model (first released 2024) capable of generating photorealistic videos from prompts. Sora uses a denoising latent diffusion process with a Transformer UNet as the denoiser. It generates videos in a compressed latent space by refining 3D spatiotemporal patches. While producing impressive results (up to 1-minute HD video), Sora still struggles with complex physics and logical consistency.
Gemini Diffusion	Google DeepMind's experimental **diffusion-based language model** (announced 2025) that generates text via diffusion instead of one word at a time. Gemini Diffusion produces whole blocks of tokens in parallel and iteratively refines them, allowing faster text generation with improved coherence and editability. Despite a smaller size, it achieved performance comparable to larger autoregressive models, pointing to diffusion as a promising alternative for language.

Thunderforge	A U.S. Department of Defense flagship AI program (led by Scale AI, started 2025) to integrate AI agents into military planning and operations. Thunderforge fuses large language models and multi-modal simulations (through partners like Anduril's Lattice) to provide decision support for military scenarios. While not a single diffusion model, it exemplifies applying generative AI (including possibly diffusion-based world models) to intelligence, surveillance, and reconnaissance (ISR) and wargaming in defense.
cross-modal fusion	The integration of multiple input data modalities (e.g. vision, text, audio, radar) into a single model or representation. In diffusion forecasting, cross-modal fusion might mean conditioning a model on different data sources – for example, using both satellite images and numerical data to generate a weather forecast. Architectures like Perceiver IO and multimodal Transformers enable such fusion by mapping all modalities into a joint latent space for unified processing.
latent probe	A hypothetical mechanism or tool for extracting or injecting information into a model's latent space. "Probing" a latent space means analyzing the hidden representation to understand what the model has learned (for example, using a simple classifier to detect if a certain concept is encoded). A *latent probe* could also refer to inserting a test signal into the latent state to see how the model's prediction changes – effectively querying the model's internal simulation without external outputs. This is useful for interpretability and for steering generation by nudging latent variables.
stochastic rollout	The simulation of a sequence of future states by repeatedly sampling from a probabilistic model. In reinforcement learning or planning, a rollout refers to generating a trajectory of states and actions. A *stochastic* rollout acknowledges uncertainty by sampling at each step (as opposed to taking the single most likely outcome). Diffusion models naturally produce stochastic rollouts when used iteratively (each step's output is sampled with noise), allowing many diverse futures to be rolled out from the same starting condition.

entropy funnel sampling	A strategy of progressively reducing uncertainty (entropy) during generation to converge on realistic outcomes. In diffusion, large noise (high entropy) at the start gradually reduces as the model refines the sample. “Entropy funnel” evocatively describes how a broad distribution of possibilities narrows to a focused set of plausible outcomes. By carefully scheduling noise levels or applying guidance (e.g. classifier guidance), one can “funnel” the random exploration toward higher-probability predictions. This ensures both diversity initially and accuracy finally – analogous to widening then narrowing a funnel.
world-model	An internal model that captures the dynamics and rules of an environment, allowing simulation of future trajectories. In robotics and AI, a world-model learns from past experience to predict what will happen next given the current state and potential actions. World-models can be explicit (physics-based) or learned (neural). A diffusion-based world-model would generate possible next states via reverse diffusion. The unified latent-space philosophy envisions a *single* world-model latent that encodes everything the AI observes and can imagine, serving as the substrate for both understanding the present and predicting the future.
now-casting	Ultra-short-range forecasting, typically covering the next few minutes to a few hours. Now-casting is heavily used in weather (e.g. predicting rain in the next 60 minutes) and often relies on latest sensor data like radar and satellite. Diffusion models have been applied to precipitation nowcasting by treating radar image sequences as data to denoise forward in time. These models can produce an ensemble of plausible high-resolution maps for the very near future, bridging the gap between instantaneous observation and longer-range forecasts.
edge inference	Running AI model predictions on edge devices (local hardware such as robots, smartphones, or IoT devices) rather than in a cloud data center. Edge inference for diffusion forecasting implies doing the compute-intensive diffusion sampling on-device, which can reduce latency and reliance on connectivity. Achieving real-time edge inference

	with diffusion models is challenging due to their iterative nature, but ongoing optimizations (model distillation, fewer diffusion steps, hardware acceleration) are pushing in this direction to enable on-the-fly predictions in the field.
cloud orchestration	Coordinating compute and data resources in the cloud to run large-scale models or pipelines. In a forecasting context, cloud orchestration might manage a network of diffusion models across a cluster – for example, a central server spawning multiple diffusion simulations in parallel (an ensemble) and aggregating the results. It ensures that massive predictive models (like global weather diffusion models) can run efficiently by scaling across GPUs/TPUs in the cloud, with scheduling that meets real-time requirements (such as delivering a forecast cone by a deadline).

3. Generative Video and RSDE in the Zeitgeist

Across public discussions of AI, diffusion-style, reverse-time SDE (RSDE) video models are increasingly framed not merely as “content generators,” but as emergent world-model learners that capture intuitive physics and thereby support near-term forecasting. In a recent conversation, Demis Hassabis emphasized that video models such as Veo 3 appear to “model enough of the dynamics” to generate coherent multi-second continuations; he characterized this as a form of understanding grounded in next-frame prediction and noted rapid empirical progress toward more consistent physics, lighting, materials, and liquids—an “intuitive physics” akin to what a human child acquires

4. Historical Background

Early generative diffusion models were developed for image synthesis. The foundational work by Sohl-Dickstein *et al.* (2015) introduced the idea of progressively adding noise to data and learning to reverse that process. This concept gained traction with **Denoising Diffusion Probabilistic Models (DDPMs)** in 2020 (Ho, Jain, and Abbeel) and **score-based generative modeling** (Song and Ermon, 2019–2021) which formalized the reverse-time SDE framework. These models demonstrated that diffusion could produce high-fidelity images competitive with GANs, without adversarial training. By late 2021, diffusion models were the state-of-the-art in image generation (exemplified by OpenAI’s GLIDE and StabilityAI’s Stable Diffusion).

The success in images naturally led researchers to ask: can diffusion generate *sequences* of data, not just independent pixels? The first explorations in video and time-series emerged in 2022. Jonathan Ho et al. extended DDPMs to video with a paper titled **"Video Diffusion Models"** (NeurIPS 2022), showing that a 3D U-Net can learn to generate coherent 16-frame video clips by treating time as an additional dimension in the diffusion process. Around the same time, Google's research on Imagen Video and Phenaki demonstrated diffusion-based video generation at higher resolutions and longer durations, using cascades of models. These early video diffusion models were *unconditional or text-conditional generation* – essentially the video analogues of image synthesis – rather than forecasting future frames of a given real video. Nonetheless, they proved the capability to model temporal data.

In parallel, diffusion ideas entered the time-series domain. Traditional forecasting methods (ARIMA, RNNs, Transformers) produce point estimates or require Monte Carlo simulation for uncertainty. By 2022, works like **Diffusion Models for Time-Series** (e.g. DiffWave for audio, TimeGrad for financial series) showed that diffusion models can natively generate probabilistic sequence forecasts. A prominent early example is **Diffuser** (Janner et al., 2022), which applied diffusion to planning in reinforcement learning: the model was trained on trajectories of states and actions, and could then *plan* by generating new trajectories via guided diffusion sampling. This was a form of *forecasting agent behavior* – given a start state and a goal, Diffuser could roll out a feasible path to achieve it by denoising a noise sequence into a trajectory that the agent might take. Such work hinted at the broad potential of diffusion beyond pixel-space.

Throughout 2023, diffusion-based forecasting gained momentum. Researchers began replacing specific parts of forecasting pipelines with diffusion models to leverage their strengths in uncertainty quantification. In weather prediction, for instance, Pathak et al. (2022) and Lam et al. (2022) showed that learned neural models (Fourier Neural Operators, Graph Neural Networks) could emulate numerical weather simulations orders-of-magnitude faster. Building on that, by 2023 teams started to incorporate diffusion for probabilistic weather forecasting. We see a convergence of the diffusion *framework* with classic state-space models: Song's **continuous-time SDE formulation** provided a principled way to treat the evolution of a system as a diffusion process, which is very natural for physical systems like atmosphere and oceans.

Also in 2023, the first **multimodal diffusion models** emerged. These handle complex data inputs and outputs – for example, conditioning video generation on text (as in Runway Gen-2's text-to-video system) or mixing image, audio, and other signals. The Perceiver IO architecture and similar latent-space Transformers offered a template for how a single model could take in different modalities, project them into a latent space, and generate forecasts. This set the stage for the unified latent space philosophy we discuss later.

By mid-2024, diffusion-based forecasting had transitioned from an experimental idea to initial real-world trials. OpenAI unveiled **Sora**, a diffusion Transformer model that can generate minutes of video conditioned on text, implicitly simulating physical events in those videos. DeepMind (Google) announced **Gemini Diffusion**, a text-generation model that breaks with the autoregressive tradition and generates full passages via diffusion. In weather, DeepMind also revealed **GenCast**, a diffusion-based global forecasting model that produces an *ensemble* of 15-day predictions faster than the leading numerical ensemble – a milestone in AI now-casting and forecasting. These developments indicated that diffusion models were no longer confined to toy problems; they were tackling high-value, real-world forecasting tasks.

In summary, the field progressed from basic research on image diffusions (2015–2020), to early extensions in video and time-series (2021–2022), to a proliferation of domain-specific diffusion forecasters

(2023–2024). This historical trajectory sets the stage for the detailed examples and unifying concepts we explore in the following sections.

5. Recent Experiments & Industrial Prototypes (2023–2025)

Recent years have seen a flurry of prototypes that apply reverse-time diffusion to forecasting problems:

- **OpenAI's Sora (2024)** – A landmark text-to-video diffusion model, Sora was trained on both images and videos, enabling it to generate up to 60-second videos from prompts. Sora operates in a compressed latent video space using a Transformer as the "denoiser." Impressively, it can simulate complex scenes (e.g. a tiger running through a forest) with evolving backgrounds and camera angles, suggesting an internal physics intuition. OpenAI described Sora as a step toward "**world simulators**" – models that can imagine plausible physical sequences unfolding. While primarily generative, Sora's ability to extend and alter existing videos also makes it a forecasting tool (e.g. one demo shows it extending a video clip forward in time with logically consistent outcomes). Its development highlighted challenges like temporal coherence and object permanence, partially addressed by latent patches and high-capacity temporal attention.

- **Google DeepMind's Gemini Diffusion (2025)** – Unveiled at Google I/O 2025, Gemini Diffusion is an experimental **diffusion-based large language model**. Instead of predicting words one by one, it generates entire passages by denoising in token-space. This allows *autoregression-free text generation*, yielding notable speed-ups and coherence gains. For example, Gemini Diffusion can produce a paragraph in a few large refinement steps, rather than sampling 100+ individual tokens. Internally, it can "think" by iteratively refining its output – an editing-like process that reduces errors. Benchmarks showed Gemini Diffusion achieved comparable performance to a standard autoregressive model several times its size, while generating **~1,500 tokens/second**. This prototype demonstrated that diffusion is not limited to visual data – it can competently handle structured, semantic sequences like code and language, potentially opening the door to models that unify vision and language forecasting.

- **Weather Forecasting Models (2024–25)** – The meteorology domain has embraced diffusion for uncertainty-aware forecasts. *DeepMind's GenCast* system is a prime example: It frames 15-day global weather prediction as a conditional generation task. Given the last two time steps of atmospheric state, GenCast's diffusion model samples the next state by iterative refinement, then repeats sliding forward in time. Price *et al.* (2025) report GenCast outperforms the ECMWF's operational ensemble (ENS) on most metrics, while producing 20 probabilistic forecasts in only 8 minutes – far faster than traditional ensembles. Another example is *CoDiCast* (Shi *et al.*, 2025), a diffusion now-caster that can generate 6-day global forecasts with quantified uncertainty. By drawing multiple noise samples, CoDiCast naturally creates an *ensemble of outcomes*, capturing the cone of uncertainty of, say, a hurricane's track. Meanwhile, several 2024 works addressed high-resolution nowcasting: e.g. **DiffCast** (CVPR 2024) for radar-based rain nowcasts and **EnsDiff** (2024) which uses diffusion to produce ensemble predictions for imminent precipitation. These prototypes show diffusion models matching or exceeding physics-based models in accuracy, and providing calibrated

probabilistic outputs “for free” via sampling.

- **Robotics & Trajectory Planning** – Building on Diffuser (2022), new experiments integrate diffusion into robot planning and state estimation. One 2023 prototype from Stanford generates diverse possible trajectories for autonomous vehicles using a diffusion model trained on driving data. Given the current scene and a goal (e.g. turn right at intersection), the model denoises multiple random trajectories into physically plausible vehicle motions avoiding obstacles. The diversity of outputs helps identify rare but critical scenarios (like a pedestrian running into the road) that deterministic planners might miss. Another example is **SafeDiffuser (2023)**, which adds safety constraints (via control barrier functions) to diffusion planning – ensuring the sampled trajectories respect safety limits (such as staying within lanes or maintaining safe distances). In state estimation, diffusion has been used for smoothing and filtering; e.g., a diffusion model can “fill in” missing sensor signals or video frames by treating them as an image inpainting task, leveraging the model’s learned physics to infer likely values.

- **Finance and Economics** – Financial firms have begun prototyping diffusion-driven simulators for markets. One recent preprint (Lesniewski & Trigila, 2024) introduced a diffusion model that learns joint distributions of asset returns and can generate **synthetic market scenarios** indistinguishable from real data. The authors highlight that large batches of generated scenarios have well-conditioned covariance matrices and realistic heavy tails, useful for stress-testing portfolios. Another prototype by researchers at J.P. Morgan applies diffusion to limit order book data (the stream of buy/sell orders), enabling simulation of extreme market conditions. Compared to earlier GAN-based approaches, the diffusion model captured subtle statistical properties and regime shifts (e.g. volatility clustering) more faithfully. The **TRADES** system (Berti *et al.*, 2025) combines Transformers with diffusion to simulate order book evolution, and can generate entire future *order book trajectories in parallel* rather than step-by-step, improving long-horizon stability. These financial demos remain mostly in research phase, but signal a trend toward using generative models for risk management.

- **Scale AI’s Thunderforge (2025)** – Though Thunderforge is an integrated program rather than a single model, it provides a window into industrial adoption of AI scenario simulation. Announced in March 2025, Thunderforge brings AI decision-support to the U.S. Department of Defense planning and wargaming. It will incorporate multiple AI models (LLMs, presumably simulation engines) across security domains. For example, an intel analyst might use Thunderforge’s system to ask “What are plausible developments in region X over the next 30 days given current data?” and get a range of simulated scenarios with probabilities. While details are limited, one can imagine diffusion-like generative models being used to produce these multi-factor geopolitical scenarios (integrating satellite imagery, reports, economic data, etc.). Thunderforge underscores the demand for *agentic simulators*: AI that can play out “what-if” stories in a controlled, probabilistic manner to aid human decisions. It is effectively applying unified latent world-models to real-world complexity, under human oversight.

- **Others** – *Runway Gen-2* (2023) deserves mention as an industry system for text-conditioned video generation available to creators. It uses latent diffusion (similar to Stable Diffusion) extended to video frames, and though oriented to creative content, the underlying tech (multi-frame diffusion with temporal attention) is applicable to forecasting future video frames. *Pika Labs* has a text-to-video service as well, indicating how quickly diffusion video generation has become accessible. In academic labs, researchers have demonstrated diffusion models for **molecular dynamics**

(predicting atomic trajectories), for **traffic flow prediction** on road networks, and even for **medical prognosis** (e.g. diffusing patient health indicator trajectories to predict possible future health states with uncertainty bounds). Each of these prototypes tackles domain-specific challenges (respectively: physical conservation laws, spatial correlations, and personalized risk profiles) by leveraging the intrinsic strengths of diffusion – flexibility in data distribution modeling and the ability to generate many outcome samples for analysis.

Collectively, these experiments across visual, textual, scientific, and strategic domains illustrate the versatility of diffusion-based forecasting. They also provide early validation that a *single paradigm* can span very different applications, lending credence to the idea of a unified latent model that underlies them all.

6. Unified Latent-Space Philosophy (thread synthesis)

At the heart of the discussion is the vision of a **unified latent space** for AI – a single representational space where an AI system can simultaneously understand what it perceives, reason about it, and imagine future possibilities. The author Ken posits that rather than splitting AI into separate modules (vision module, planning module, language module, etc.), we should train one *world-model* that encodes all modalities and cognitive functions in one latent representation. In this latent space, the distinctions between "perception", "reasoning", and "prediction" blur: reasoning becomes an *internal transformation* of the latent state, and prediction is just rolling that latent state forward in a learned simulation.

This philosophy is inspired by how humans seem to think – we don't always verbalize our chain-of-thought; much happens in a subconscious mental model of the world (a "latent" space in the brain). Ken argues an AI should do the same: for example, when an autonomous robot observes a scene, it projects the visual input into its latent space (perception); it can then query or adjust that latent (reasoning) to test various outcomes (imagination); finally, it samples an evolved latent representing the predicted next state (forecasting). All these steps occur with the same distributed representation, allowing fluid movement between understanding the present and predicting the future. Notably, OpenAI's technical report on Sora hints at this approach by converting **all visual input into a unified latent patch representation** before modeling, drawing an analogy to how large language models tokenize everything in text. In Sora's case, video frames are compressed into latent codes, and the diffusion model operates on those – essentially treating perception and generation in one space.

The unified latent view aligns with concepts like **Perceiver IO** and **Slot Attention**. Perceiver IO provides a way to map diverse inputs into one latent bottleneck and derive outputs, suggesting how an AI might fuse vision, language, and other data into one space for joint processing. Slot Attention, on the other hand, gives a mechanism for the latent space to be structured into meaningful parts (slots could represent objects or concepts) which the model can manipulate. If each slot can hold an object's state, the model's latent dynamics could simulate object interactions over time – a form of reasoning and prediction unified. Recent research on "latent reasoning" in language models (e.g. Huginn, 2025) also plays into this. Huginn showed a smaller language model can outperform larger ones by **pondering in latent space** – it recurrently updates its hidden state (without producing output tokens) to work through a problem, akin to mental

rehearsal. This is exactly what Ken advocates: let the model internally figure things out in a continuous latent space, then produce a result when ready, rather than exposing every reasoning step.

Another influence on the unified latent philosophy is the success of **world models in reinforcement learning**. For instance, Ha and Schmidhuber's "World Models" (2018) trained a variational autoencoder (VAE) to encode game images into a latent state, and a recurrent network to predict the next latent state given actions. The agent could then plan by imagining sequences of latent states. Modern diffusion world-models can do similar planning via *stochastic rollout* in latent space, but with much richer representations and data. Imagine an AI that encodes not just images but also text, audio, and symbolic data into one state; it could simulate futures in that abstract space (using diffusion to sample different possible evolutions) and decode those futures into human-interpretable form (text explanations, visualizations, etc.).

Ken's argument also touches on **cross-modal coherence**: a unified latent model can ensure that predictions across modalities remain consistent. For example, suppose an AI is predicting a political scenario. In a siloed approach, one model might predict a graph of alliances, another generates a textual news headline, another predicts economic indicators – and one must hope they align. In a unified latent approach, the model would carry all these aspects in one state vector; when it samples a future latent, that latent simultaneously determines the alliances graph, the likely news headlines, and the economic indicators, ensuring they reflect one coherent scenario. This is analogous to how humans maintain a single mental model of a situation that can answer different questions consistently.

To make this concrete, consider a **self-driving car AI**. With a unified latent world-model, the car's cameras, LiDAR, maps, and even high-level route plans would all feed into one latent representation of "what's happening now". The diffusion model can then simulate that latent a few seconds forward to see "what might happen next" (for instance, a pedestrian might step into the road or the traffic light might change). Because it's one model, it inherently combines visual cues, physics, and common-sense (learned) to produce each simulated outcome. It might run 100 such simulations in parallel (via random latent perturbations) to build a probabilistic picture of the next few seconds. The car then makes a decision based on this forecast ensemble. All of this – perception, imagination, planning – occurs within the weights of a single neural model (or tightly integrated set of models), rather than discrete modules passing simplified messages.

Ken acknowledges that this approach is ambitious and there are challenges: training a single model to do "all of the above" requires vast data and careful architecture design to avoid interference between tasks. However, recent large models provide evidence it's feasible – e.g. DeepMind's Gato (2022) was a single Transformer that could play video games, caption images, chat, and control a robot arm, by mapping all tasks into a common embedding space. **Diffusion models** add the ability to *predictably diverge* into the future within such a space. They can take a latent that encodes the current world, inject noise (i.e. explore uncertainty), and then refine it to a plausible new world state. By chaining this, the unified model "dreams" forward, not unlike a physicist mentally simulating how a scene will evolve.

In summary, the unified latent-space philosophy is about breaking down barriers between different cognitive functions of AI. It envisions an AI agent with a single brain (latent space) where it stores what it sees, thinks with it, and uses it to predict. Reverse-time SDE diffusion is a key enabling mechanism for the prediction part of that brain, allowing the agent to draw multiple future trajectories from its latent state. The benefit, proponents argue, is a more **holistic intelligence** that can leverage the full knowledge of the system for any sub-task – producing more coherent, context-aware, and reliable predictions. In the next section, we

will see how this vision plays out in specific domains, tying together the concepts through concrete application scenarios.

7. Cross-Domain Applications

7.1 Robotics & Autonomous Vehicles

Modern robots and self-driving vehicles must **predict physical dynamics** to operate safely. Diffusion-based world models offer a promising route to faster-than-real-time physics prediction for these systems. Instead of running computationally expensive physics simulators or handcrafted motion models, a robot can learn a diffusion model of its environment's dynamics from experience. This model can then imagine many possible futures in parallel by sampling the reverse diffusion process.

Trajectory Forecasting: One application is forecasting the trajectories of objects (or the robot itself) in real time. For example, an autonomous car needs to anticipate the next few seconds of motion for nearby vehicles and pedestrians. A diffusion model can take the current state of the traffic scene (positions, velocities, etc. embedded in a latent) and generate dozens of possible future trajectories for each actor via stochastic rollout. Because diffusion generation is parallel, the model can produce an entire multi-agent trajectory sequence in one go, rather than simulating each timestep sequentially. Janner et al.'s *Diffuser* system demonstrated this by denoising complete **trajectory sequences** in a single process. The benefit is that long-term dependencies (like where a pedestrian will be 5 seconds from now) are handled with global coherence in the sequence, reducing the compounding errors that often plague autoregressive predictions. This AR-free sequence generation was noted to potentially *speed up inference for longer horizons*, since we don't need to iterate step-by-step (though diffusion itself involves many internal steps).

Physics and Kinematics: Robots with manipulators (arms, legs) are using diffusion models as internal physics simulators. Consider a bipedal robot walking on uncertain terrain. A diffusion model can learn the distribution of next joint states given the current state and desired foot placement. During operation, the robot can sample many potential outcomes of its next footstep (slip, firm plant, etc.) by injecting noise and denoising, effectively doing a Monte Carlo roll-forward of its physics. The *central advantage* is that the model is learned from real data, so it can capture complex contacts and friction effects that are hard to model analytically. Over a 0–6 month horizon, we expect lab robots to begin demonstrating diffusion-based balance and gait prediction that anticipates falls faster than real time (e.g. predicting a slip 0.5s before it completes and adjusting accordingly). By 12–24 months, such learned predictive models could be integrated into control loops of quadrupeds and drones, giving them an "intuition" for dynamics that extends classical model-predictive control.

Intention and Multi-Agent Interaction: Beyond physics, robots often need to infer *intentions* – e.g., is that oncoming car planning to turn? Diffusion models can help here by generating multimodal predictions. A single scenario can branch into multiple possible futures, and diffusion can naturally represent that uncertainty. For instance, a diffusion model controlling a social robot in a crowd might simulate one latent future where the person ahead moves left, and another where they move right, etc., covering all plausible intent interpretations. Each sample is a complete trajectory consistent in itself. This aligns with the concept of an *entropy funnel*: early in the diffusion process the model explores a wide range of intents (high entropy), but by the final denoising steps, each sample converges to a concrete hypothesis (low entropy

outcome). The robot can then use this ensemble of hypotheses for robust planning (worst-case avoidance, average-case efficiency, etc.).

Edge vs. Cloud: In robotics, latency is crucial. Edge inference of diffusion models (onboard the robot or vehicle) is challenging due to the iterative nature of diffusion, but progress in model distillation and efficient solvers is narrowing the gap. We anticipate a hybrid **edge-cloud orchestration**: the most immediate predictions (next second or two) might be handled by a small distilled diffusion model on the edge, providing reflexes faster than real-time. Meanwhile, cloud-based larger diffusion models could run slightly longer-term predictions (next 5–10 seconds or more) in parallel and feed those back to the robot for strategic planning. Over a 6–12 month timeline, specialized accelerator hardware (TPUs, neural engines) and algorithmic tricks (fewer diffusion steps via improved samplers like DPM-Solver++) will likely make real-time diffusion inference feasible on vehicles. By 24–36 months, it's conceivable that autonomous cars will routinely run learned world-model simulations at 2× real-time speed or more – effectively *fast-forwarding reality* internally – to foresee hazards and coordinate maneuvers accordingly.

Validation and Safety: An open challenge in this domain is guaranteeing safety with learned predictions. Efforts like *SafeDiffuser* (2023) incorporate safety constraints so that even the imagined trajectories obey certain rules. This will be an ongoing area of research. Regulatory bodies may eventually require demonstrating that generative trajectory models won't produce unsafe outputs that mislead the robot. Nonetheless, early prototypes indicate that diffusion models, with their ability to produce *diverse, realistic scenarios*, could significantly enhance the foresight of robotic and autonomous systems.

7.2 Weather & Climate

Weather forecasting is a natural fit for diffusion methods because of the inherently uncertain and chaotic nature of atmospheric dynamics. Traditional numerical weather prediction (NWP) produces a single deterministic forecast (or a small ensemble of runs for uncertainty). Diffusion-based forecasting turns this on its head by effortlessly generating *hundreds of plausible weather scenarios* consistent with recent observations, effectively modeling the probability distribution of future weather.

Now-casting (0–6 hours): Now-casting focuses on short-term, high-resolution forecasts like predicting where thunderstorms will form or how a rain band will move in the next hour. Data-driven nowcasting systems such as Google's MetNet (2020) started this trend using deep learning. In 2024, diffusion models took it further – for instance, **DiffCast** (2024) unified radar-based rain prediction as a diffusion problem. It inputs the latest radar and satellite frames, encodes them into a latent (possibly using optical flow to inform motion), and then generates the next frames by denoising. The result is an ensemble of high-resolution rainfall maps 1-3 hours ahead, with realistic variability (capturing uncertainties in storm initiation or dissipation). Physical consistency can be built-in by conditioning the diffusion on known physics (e.g. large-scale wind fields from NWP). In the next 0–6 months, we expect to see national weather services test such diffusion nowcasts, comparing them to extrapolation-based methods. A key milestone will be demonstrating skill improvement (accuracy gains) over legacy nowcasting for extreme events (flash floods, convective storms) thanks to the generative approach's ability to *imagine rare but possible outcomes* (something deterministic models struggle with).

Mid-range Forecasts (1–10 days): This is where models like **GraphCast** and **GenCast** have shined. GraphCast (DeepMind 2022) was a GNN that predicted global weather quickly but deterministically. **GenCast (2024)** upgraded this by using a diffusion model to produce a *probabilistic ensemble* that beat the accuracy of ECMWF's 51-member ensemble. In GenCast, each diffusion sample is essentially one

plausible simulation of the atmosphere's evolution, conditioned on the current state. Because diffusion steps are iteratively correcting errors, the model can incorporate physical constraints learned from data (conservation laws, etc.) at each step, maintaining realism. Over 6–12 months, we expect diffusion models to start getting integrated into operational forecasting workflows. For example, NOAA's researchers are exploring diffusion for regional rapid refresh models (as hinted by *Operationalizing diffusion models for regional weather* at the NOAA AI Workshop 2025). A likely milestone: a **10-day forecast cone** generated by a diffusion model that forecasters use alongside traditional ensembles by late 2025. This cone would give probabilities of extreme deviations (like heatwave intensity or hurricane track shifts) with a level of detail (spatial and distributional) that classical ensembles find hard to match.

Climate and Long-term Projections: While weather prediction deals with day-to-day variability, climate modeling looks at statistical trends over decades. Diffusion models can serve as emulators for expensive climate simulators. For instance, a diffusion model could learn to generate yearly outcomes for temperature and precipitation fields given some forcing conditions (like greenhouse gas levels). The **DiffESM (2023)** and *ClimateDiffuse* projects attempt to replace or augment parts of climate models with diffusion components. Over 12–24 months, we might see diffusion-driven downscaling tools – taking coarse climate model output and generating high-resolution local projections – become available for climate researchers. Additionally, diffusion can help explore *tipping point scenarios* by sampling low-probability trajectories in climate variables that standard simulators might not focus on.

Uncertainty Quantification: A major advantage confirmed by recent papers is the quality of uncertainty quantification diffusion provides. *Continuous Ranked Probability Score* (CRPS) metrics in studies (e.g. EnsDiff 2024) show diffusion ensembles are well-calibrated – the spread of the ensemble correctly reflects actual forecast errors in a statistically reliable way. This is partly because diffusion doesn't just perturb initial conditions (like traditional ensembles) but actually learns the error growth patterns from data. We expect further improvements in calibration with techniques like **entropy adjustment** – adding just enough noise at each step to maintain the correct spread.

Integration with NWP: Rather than completely replacing physical models, the near-term trend is hybrid systems. For example, one could take a high-quality deterministic forecast from an NWP model and use a conditional diffusion model to generate perturbations around it, creating a super-ensemble (this could be seen as learning a *forecast error model* via diffusion). NOAA's mention of **"Diffusion models for data assimilation"** also points to using diffusion to intelligently add perturbations during the data assimilation step, potentially improving initial conditions for NWP. Over 24–36 months, if diffusion models continue to prove their worth, we might even see the first instances of *AI-first forecasting systems* – where a diffusion model produces the primary forecast and the physics model is used as a sanity check or for guidance (a reversal of today's paradigm).

One concrete milestone would be **probabilistic hurricane forecasting** via diffusion. Currently, hurricane track and intensity cones are made from a few deterministic runs and statistical models. A diffusion model fed with satellite imagery, ocean state, and previous track could instantly generate thousands of plausible track and intensity scenarios, better capturing uncertainty in rapid intensification or curving tracks. If by the 2025–2026 hurricane seasons such a system can demonstrate equal or better performance than NHC's statistical methods, it would mark a significant breakthrough – delivering richer information (like a full probability distribution of landfall locations) to decision-makers and the public.

7.3 Financial Markets

Financial markets are notoriously hard to predict – they're noisy, influenced by myriad factors, and prone to abrupt regime changes. Diffusion models, with their ability to model complex distributions and generate many sample paths, are well suited to **financial forecasting and stress-testing**. Rather than point predictions of, say, stock prices, diffusion-based approaches aim to simulate the many ways the market could evolve, especially the extreme but plausible scenarios.

Market Regime-Shift Detection: A regime shift (e.g. from low volatility to high volatility, or bull to bear market) often manifests as a distributional change in time series data. Diffusion models can aid in detecting and even *anticipating* such shifts in a few ways. First, a diffusion model trained on historical data can serve as a generative reference: if the actual market data suddenly has very low likelihood under the model (i.e. the model's samples no longer resemble reality), that's a signal of a regime change. For instance, a diffusion model might be trained to generate typical daily return patterns for equities. If a cluster of actual days occurs that the model finds "surprising" (statistically far in the tail of its learned distribution), analysts are alerted that "something fundamental has changed." Additionally, one can condition diffusion models on macro indicators or latent regime variables and use Bayesian inference to update the probability of being in a certain regime as new data comes in. Over 0–6 months, we expect quant researchers to experiment with diffusion models as unsupervised anomaly detectors for market data – basically flagging when the data distribution is shifting away from the training distribution.

Scenario Generation for Stress-Testing: Regulators and banks conduct stress tests to ensure financial institutions can withstand adverse market conditions. Traditionally, stress scenarios are handcrafted (e.g. "what if unemployment rises by 5% and stock market falls 30%?"). Diffusion generative models can provide a data-driven complement: generate hundreds of plausible crisis scenarios drawn from the tail of historical data distributions. For example, *Lesniewski & Trigila (2024)* generated synthetic multi-asset price paths that preserved realistic cross-asset correlations and tail behaviors. Because diffusion models can incorporate known constraints (no arbitrage conditions, etc.) through how they're trained or via conditioning, the scenarios can be financially sane yet diverse. Over 6–12 months, we might see initial adoption of diffusion-based scenario sets in portfolio risk software – e.g. a risk manager can push a button to generate 1,000 1-year market paths for a given portfolio, then compute losses distribution. One key milestone would be a major bank or asset manager publicly validating that a diffusion-generated stress test predicted a vulnerability that traditional methods missed (for instance, uncovering a nonlinear correlation spike that could occur in extreme conditions).

Portfolio Optimization & Hedging: The outputs of diffusion models can feed into downstream decision models. For example, if a diffusion market simulator produces an ensemble of future yield curves, a bank can optimize its bond portfolio to minimize risk across that ensemble (robust optimization). Similarly, hedging strategies could be tested against a wide range of generated scenarios, ensuring they perform not just on average but in worst-cases. By 12–24 months, as confidence in these models grows, we might see them used for real-time risk monitoring – a sort of "risk radar" that constantly generates near-future market movements (over days/weeks) and flags if a portfolio would incur heavy losses in a significant fraction of those. Because the diffusion model inherently produces *heavy-tail events* with the correct frequency, this approach could better prepare institutions for market crashes or spikes.

High-Frequency and Limit Order Books: Another area is intraday trading and market microstructure. The *TRADES* model (2025) and others have applied diffusion to simulate limit order book dynamics. An advantage noted is the "*parallel generation of future timesteps*" – diffusion can treat an entire sequence of order flow in one pass, which helps capture long-range dependencies such as order book pressure building up. High-frequency trading firms might use such models to stress-test their algorithms: simulate a flash

crash scenario via the generative model and see how their algorithm behaves. If diffusion models prove accurate in microstructure (a big “if,” given how noisy and reflexive those systems are), by 24–36 months they could be part of exchange surveillance tools that generate synthetic illegitimate trading patterns (spoofing, manipulation) to test detection systems, or conversely generate benign patterns to reduce false alarms.

Challenges: Finance is an especially high stakes domain for predictive models, so trust is a major issue. Diffusion models will have to earn trust by demonstrating reliable performance over time. They also must deal with non-stationarity (markets evolving – a model trained on one decade may not apply to the next). Continual learning or adaptive retraining will likely be needed, perhaps with techniques to avoid “catastrophic forgetting” of rare crises when they are not in recent data. Another consideration is interpretability: regulators might require explanations for scenarios. Some work in 2025 has looked at “latent probes” on financial diffusion models – essentially trying to tie latent dimensions to economic factors so that scenarios can be described (“this scenario corresponds to a rapid interest rate hike shock”). By 36 months out, if diffusion models are mainstream in finance, we should expect a suite of interpretability and validation tools accompanying them, ensuring they are used responsibly and do not inadvertently encourage reckless risk-taking due to misplaced confidence in AI-generated scenarios.

7.4 Chemistry & Biology

In chemistry and biology, many problems involve exploring a vast space of possibilities – be it molecular configurations, drug designs, or protein structures. Diffusion models have recently made inroads here as **generative forecasters of molecular outcomes**. While not “forecasting” in a temporal sense, they predict *what structures or reactions are likely* given certain starting conditions, essentially mapping out future states in a chemical or biological process.

Drug Molecule Generation: One of the hottest areas is using diffusion models to generate novel molecular structures with desired properties. Traditionally, drug discovery involved predicting how a molecule might bind to a target protein (docking) or doing virtual screening through millions of candidates. Diffusion models can learn the distribution of stable, synthesizable molecules and then generate new candidates *conditioned* on a target’s characteristics. For example, *DrugDiff (2023)* and similar approaches use a latent diffusion model guided by property predictors to create molecules optimized for specific properties (like activity or solubility). This is akin to “forecasting” a molecule that will succeed in a medicinal role, from the space of all possibilities. Over the next 0–6 months, we anticipate diffusion-generative models being integrated into medicinal chemistry pipelines; companies like Insilico Medicine and Generative AI startups are actively exploring these. A milestone will be a diffusion-designed compound advancing to pre-clinical testing, demonstrating the model’s ability to foresee a viable drug molecule.

Protein Structure and Function Design: A groundbreaking development was **RoseTTAFold Diffusion (RFdiffusion)** in 2023, which applied diffusion to protein structure design. By fine-tuning a structure prediction network as a denoiser, RFdiffusion could generate new protein backbones achieving *state-of-the-art* success on multiple design tasks. Essentially, given a rough specification (like “protein that binds X” or “protein with Y symmetry”), the model diffuses random backbones into functional ones meeting the criteria. This is forecasting in a structural sense: it predicts a protein configuration that would result from evolution or design pressures. Within 6–12 months, we expect at least a few instances of novel enzymes or therapeutics designed by diffusion models being validated in labs (some are already in testing as per RFdiffusion’s results). A concrete milestone: design a new enzyme for a biochemical reaction via diffusion modeling the catalytic site formation – and experimentally show it works. Because diffusion can generate a

diverse set of solutions, it might find non-intuitive protein folds that human designers wouldn't think of, potentially opening new areas in protein engineering.

Chemical Reaction Prediction: Another aspect is reaction outcome prediction – given reactants and conditions, what products form (and in what yield)? Diffusion models can be trained on known reaction databases to generate product distributions. One could encode the molecular graphs of reactants and then diffuse to predict the product graph. If conditioned on temperature, solvent, etc., the model can also forecast side products or failure modes (important for chemical process planning). Over 12–24 months, diffusion-based reaction models might become a tool for chemists to explore novel synthetic routes: the model could generate possible reaction pathways step by step (like a multi-step synthetic plan), with each step a diffusion prediction of products from given reagents. While still early, this could revolutionize computer-aided synthesis by suggesting plausible new reactions or highlighting that a certain route might produce a dangerous byproduct with some probability (thus warning the chemist).

Molecular Dynamics and Folding Pathways: Classical molecular dynamics (MD) simulations are computationally costly (simulate femtoseconds to see nanoseconds of motion). Diffusion models offer a data-driven shortcut: train on lots of MD trajectories and then generate likely trajectories at a higher level of abstraction. For instance, a diffusion model could model the conformational changes of a protein or RNA molecule over time at a coarse level (like alpha-carbon trace) much faster than physics-based MD. It samples possible folding pathways or conformational transitions (like a channel protein opening/closing). By 24–36 months, such models could help in understanding complex biological processes – e.g., generating hypotheses of how a protein might misfold (which is relevant to diseases like Alzheimer's) that experimentalists can then verify.

A specific milestone in this vein could be **binding affinity prediction** by diffusion: Imagine generating many poses of a drug molecule in a protein binding site via diffusion (essentially predicting how the molecule could fit and what conformations the protein might adopt), then evaluating those for binding strength. This could outperform single "best pose" docking by exploring the rugged landscape of binding configurations more thoroughly.

Challenges: Chemical space is huge, and biological systems are extremely intricate. Ensuring validity of generated molecules (e.g. no radicals or disallowed substructures) is non-trivial – diffusion models sometimes generate chemically invalid structures if not properly constrained. Researchers address this by incorporating chemical rules into the model or filtering outputs. Another challenge is data: for some tasks (like novel reaction prediction), available data is limited. One strategy is to use **transfer learning in latent space** – e.g., pretrain a diffusion model on generic molecules, then fine-tune on the smaller reaction dataset to bias it towards chemically feasible transformations. Ethical considerations also arise: generative models could be misused to design harmful substances (toxins or chemical weapons). Monitoring and putting proper safeguards (like having the model also predict toxicity and flagging dangerous outputs) will be important.

In summary, diffusion models in chemistry/biology act like **hypothesis generators** for the future of molecular systems – "What if I mix these chemicals?", "What protein shape could perform this function?", "What drug might bind here and stay stable?". By providing a breadth of plausible answers, they accelerate the scientific process of narrowing down to the most promising hypotheses that can then be tested in the lab.

7.5 Military & Geopolitical Forecasting

National security and geopolitical analysis involve high stakes forecasting where scenarios are complex, multi-factorial, and often there is very little historical data for unprecedented events. Generative AI, including diffusion models, is poised to become a critical tool in this arena by **simulating conflict and political scenarios** in a way that allows strategists to explore "what if" questions with AI-powered imagination.

Intelligence & Surveillance (ISR) Simulation: Military planning often uses wargames – essentially human-driven simulations of conflicts or crises. AI agents can augment this by playing out scenarios faster and suggesting novel strategies. A diffusion-based approach could encode the state of a conflict (force positions, readiness levels, political climate) in a latent vector and then generate possible next-week developments by denoising that latent. The stochasticity would capture the fog of war and unpredictability of adversary actions. Scale AI's **Thunderforge** initiative indicates a move in this direction: integrating AI into workflows to provide decision advantage. We can imagine Thunderforge deploying diffusion models to simulate, say, the range of outcomes of a naval standoff in the South China Sea – from peaceful resolution to escalation – each outcome with a likelihood and key indicators to watch (the model might highlight that certain patterns of adversary deployment often precede a de-escalation, gleaned from training on historical analogues or war game data). Over 0–6 months, initial prototypes might focus on narrow questions (e.g., "If surveillance drones are removed from area X, what's likely to happen?") using diffusion conditioned on variables like presence/absence of intel, readiness, etc.

Strategic Scenario Modeling: Geopolitical forecasting can benefit from cross-modal diffusion models that combine economic data, social media signals, satellite imagery, etc., to predict events like coups, migrations, or treaty formations. For instance, a model could take a latent state representing a country's economic indicators, internal unrest level, neighboring countries' stances, and then generate a distribution of outcomes 6 months out (stable, protests, regime change, etc.). While this sounds almost like science fiction, early steps are being taken: projects in academia and companies like Palantir are exploring "AI red teaming" – using AI to generate adversarial scenarios that stress-test policies or strategies.

One concrete application is **ISR data hallucination**: filling in missing information. If satellite coverage or recon is missing in an area, a diffusion model could generate likely enemy force distributions based on what it has seen elsewhere (with uncertainty). This is dangerous if used blindly, but as a tool it could help analysts consider possibilities rather than assume unknown areas are benign. Over 6–12 months, analysts might start using AI-generated scenario narratives as part of their intelligence briefings – e.g. an AI might output: "There is a 20% chance that over the next year country Y will attempt a blockade of strait Z, likely preceded by naval exercises and trade rhetoric escalations," with the narrative drawn from patterns it learned. These narratives could be considered alongside human judgment.

Wargaming and Training: The military often conducts exercises to train decision-makers for unexpected situations. AI-driven simulation could create a richer set of training scenarios. For example, generative models can produce plausible but fictitious political crises for war game exercises ("Country A suffers a cyber attack on its grid amid election turmoil"). A diffusion model might be tasked with generating consistent multi-domain conditions (political, cyber, kinetic, economic) that fit a scenario prompt. Because it's generative, each exercise could be slightly different, preventing participants from gaming known scripts. By 12–24 months, we might see AI-simulated exercises being tested in military education, with feedback loops where human decisions are fed back into the model to generate the next developments (making it an interactive narrative generation problem, which diffusion models can do with iterative conditioning).

Global Politics and Economy Interplay: Geopolitical outcomes often depend on many interconnected systems. A unified latent model could, for instance, encode the state of global oil markets, alliances, and

conflict flashpoints, then predict how a shock (like a sudden sanction or conflict) cascades. Think of it as a "world sandbox" powered by AI – policy analysts could ask, "What if country X defaults on its debt?" and the model, having learned from analogous events, could generate a spectrum of outcomes: perhaps predicting increased unrest in region Y or shifts in global currency use. While such a tool would not be perfectly accurate, it offers a way to explore complex interdependencies quantitatively. Over 24–36 months, if progress continues, organizations like the UN or World Bank might use AI scenario models to aid in planning humanitarian responses or economic interventions by foreseeing possible consequences of various actions in unstable regions.

Ethical and Reliability Concerns: Using AI in military/political forecasting raises heavy ethical questions. There's a risk of the "oracle problem" – leaders might over-rely on AI predictions, which could be wrong and lead to self-fulfilling prophecies or unnecessary conflict. Ensuring a human-in-the-loop is essential: AI should suggest possibilities, not make decisions. To address this, Thunderforge and similar efforts emphasize human oversight and interpretability (e.g. why does the model think a blockade is likely? It might point to troop movements and past patterns as justification). Another risk is adversarial: if one side's AI is used in planning, an opponent could try to spoof it with disinformation (causing the AI to predict a false outcome). This means any generative model has to be robust to intentional data manipulation – an active area of research (adversarial robustness).

In summary, diffusion models and allied generative techniques could become a **strategist's assistant**, offering a richer imagination of future geopolitical trajectories. By generating full-spectrum scenarios (diplomatic, informational, military, economic) and their probabilities, they enhance human planners' ability to anticipate and thus prevent or mitigate adverse outcomes. Achieving this will require careful validation, trust-building, and ethical guardrails, likely over the coming 2–3 years of incremental deployment and testing in closed settings before any high-stakes reliance.

7.6 Document Ingestion & Long-Context Compression via Vision-Token OCR (DeepSeek-OCR)

Summary. DeepSeek-OCR introduces an "optical 2D mapping" pipeline that converts full pages into **compact vision tokens**, then reconstructs text and layout with a Mixture-of-Experts (MoE) decoder. The approach targets the long-context bottleneck by replacing thousands of text tokens per page with on the order of **10^2–10^3** visual tokens while preserving most semantics. Recent reports claim **≈10× compression at ~97% OCR precision**, with accuracy degrading toward **~60% at ≈20×**; the system is open-sourced and reportedly processes **>200k pages/day on a single A100-40G**, scaling linearly across nodes. (eWeek)

Mechanism (Encoder → Decoder). A high-resolution page is rasterised to a tensor, partitioned into **patches** (e.g., 16×16), embedded, and **pooled/down-sampled** to a small set of **vision tokens** by the *DeepEncoder* (ViT-style blocks plus convolutional/token pooling). A **MoE text decoder** then reconstructs text and layout from these tokens, specialising experts for tables, math, multilingual text, etc. Architectural details and naming (DeepEncoder; DeepSeek-3B-MoE) are outlined in the paper. (PixelsTech)

Reported Capability & Throughput. Public write-ups summarise **~97% OCR precision at <10× compression** and **~60% at ~20×**, positioning DeepSeek-OCR as a token-budget reducer rather than a lossless parser. Throughput claims include **~200k pages/day on one A100** and **tens of millions/day on small clusters**, with open-source code on GitHub and a model card on Hugging Face. (eWeek) These align with our internal summary of the encoder/decoder pipeline and token-compression rationale.

Why this matters for diffusion-forecasting systems. Our white paper argues for a **unified latent space** that couples perception, reasoning, and prediction; vision-token OCR is a practical *front-end compressor* that makes extremely long contexts feasible within fixed attention budgets. Shorter sequences lower memory and step-time while retaining layout cues that are vital for downstream reasoning (tables, captions, figure references). In effect, DeepSeek-OCR supplies a **layout-aware latent preamble** that our diffusion/RSDE back-ends can condition on, enabling larger retrieval windows and cheaper ensemble rollouts over document corpora.

Integration sketch (production).

1. **Pre-ingest:** Rasterise PDFs/pages → **DeepEncoder** → persist **vision tokens** alongside canonical text.

2. **Indexing/RAG:** Build dual indices (text + vision-latent) so retrieval can hit layout-dense segments (tables/figures).

3. **Reasoning:** Feed compact page-latents as **context priors** into diffusion-based planners/forecasters; expand only the cited spans to text when emitting final outputs.

4. **Ops:** Target A100-40G or equivalent; batch PDF paths via vLLM per project guidance in the repo. (GitHub)

Limitations & risk controls. Accuracy drops beyond ~10× compression; high-stakes domains (legal/medical/QA datasets) should cap compression and run **A/B sanity checks** against text-layer OCR. Complex handwriting/exotic layouts remain failure modes; gate deployment with **layout coverage tests** and fall back to standard OCR where confidence is low. See concurrent coverage for broader context and adoption signals. (South China Morning Post)

Current news context (Oct 2025). Multiple outlets report DeepSeek-OCR as **open source** and highlight the **long-context** angle (visual tokens vs. text tokens). Headlines emphasise the 10× token shrink with high precision and A100-class throughput; GitHub and Hugging Face releases are live as of 21–23 Oct 2025. (eWeek)

Cross-reference: This section extends §6 (*Unified Latent-Space Philosophy*) and §7 (*Cross-Domain Applications*) by providing a concrete **document-compression front-end** that slots before forecasting modules.

8. Road-Map: Expected Progression (0–36 Months)

Predicting the progress of AI is itself a speculative task, but based on current trends, we can outline key milestones for diffusion-based forecasting in the near future. We break it down by time frames:

0–6 Months (Late 2025 – Early 2026): In this period, we expect *pilot projects and proofs-of-concept* to flourish. Many will be extensions of the prototypes discussed:

- **Robotics:** At least one notable demo of an autonomous vehicle or robot using a diffusion model in its prediction pipeline will surface. This might be a self-driving car that visualizes multiple possible futures on a heads-up display for the safety driver, or a drone that uses diffusion to anticipate wind gusts. Real-time capability will still be limited (perhaps using 5–10 diffusion steps for a short horizon), but the concept of generative foresight on the edge will be proven. Look for a research paper or tech blog from an AV company showcasing this.

- **Weather:** National weather agencies (e.g. NOAA, ECMWF) will announce internal trials comparing diffusion-based ensemble forecasts with traditional ensembles. A milestone could be NOAA's GraphCast-Diffusion upgrade – incorporating a diffusion nowcast module into their GraphCast framework to improve regional severe weather warnings. By early 2026, a diffusion model may be running quasi-operationally in parallel with existing systems for things like precipitation nowcasting, providing forecasters with additional guidance (especially for extreme rainfall events).

- **Markets:** Financial firms will publish whitepapers on using diffusion models for risk. Possibly a collaboration between an AI lab and a bank will demonstrate that a diffusion stress test could have flagged a vulnerability in a historical event (for instance, showing that prior to the 202X flash crash, a diffusion model trained on 1990s data would have generated a similar crash scenario with non-negligible probability). Expect initial skepticism from conservative finance folks, but enough promise that more investment flows into fintech AI startups focusing on generative market modeling.

- **Military:** The first phase of Thunderforge will likely be in planning and development. We might not see public demos, but there could be mentions in defense AI conferences of using large language models for scenario planning. If an AI is used in a real-world crisis simulation exercise, that would be a significant benchmark – e.g., an exercise where AI-generated scenario injects catch participants off guard (as a test of unpredictability).

- **Unified Models:** We may see a publication or tech demo of a **multi-domain diffusion model** – one that can, say, take satellite images and economic indicators and generate a short text forecast of a humanitarian situation. This could be from a university lab showing the feasibility of unified latent representations across modalities. The quality will be rudimentary, but it will light the path forward.

6–12 Months (Mid/Late 2026): In this window, *early adoption and integration* starts:

- **Robotics:** More capable hardware (new GPU releases, edge AI chips) and algorithm optimizations (perhaps diffusion models with fewer steps or clever partial updates) will allow near real-time performance. We might hear of an autonomy stack using a diffusion model to predict pedestrian intent two seconds ahead, updating 5 times a second – effectively beating real-time (2s prediction in <200ms). A milestone might be a self-driving truck system that drives safely through a complex scenario that a previous system struggled with, explicitly thanks to improved prediction by a learned model. Regulators might not yet allow AI-only systems on public roads, but pilot operations in controlled environments (like automated forklifts in warehouses or mining trucks) could employ this tech.

- **Weather:** By mid-2026, one or more major forecasting centers could officially incorporate diffusion models in generating products. For example, **probabilistic 10-day forecasts** to the public might include diffusion-based uncertainty cones (maybe labeled as “experimental probabilistic guidance”). If GraphCast-Diffusion performed well, NOAA might replace or augment some of its Global Ensemble Forecast System (GEFS) with an AI ensemble for certain variables. A major milestone: an AI-forecast predicted a weather extreme (e.g. an abrupt rapid intensification of a hurricane, or an unusual track) that the human consensus missed, and it turned out correct – this would be widely noted and likely accelerate adoption.

- **Markets:** Some hedge funds likely already quietly use these techniques; by late 2026, we may see diffusion models used in real-time trading as scenario generators for reinforcement learning-based trading agents. Also, regulators might start paying attention – possibly requiring model risk management for generative models if banks use them. A concrete milestone could be an IMF or Federal Reserve research paper on using diffusion models to model macroeconomic risk scenarios (e.g., multiple central banks exploring it for systemic risk assessment).

- **Chemistry/Biology:** In pharma, if a drug designed with the aid of diffusion AI enters Phase I trials, that’s a huge validation. 6–12 months might be too soon for clinical, but at least lab results on AI-designed molecules or proteins with therapeutic function will be out. We may also see diffusion models integrated into electronic lab notebooks or cloud lab services, offering chemists on-the-fly reaction outcome predictions or peptide designs. The community might establish a benchmark suite for generative chemistry models (to compare diffusion vs. GAN vs. language model approaches on molecule generation tasks).

- **Unified Latent AI:** Perhaps a notable open-source project will appear (like a “WorldModelGPT” or so) that tries to combine modalities and tasks – maybe funded by a billionaire’s AI venture aiming at AGI (xAI?). This model might not surpass specialized models in each domain, but it will show intriguing cross-capabilities (e.g., feed it a short video of an event and it outputs a plausible continuation in text form – a cross-modal prediction). This might spur debate on the merits of specialization vs. generalization in AI forecasting.

12–24 Months (2027 into early 2028): This could be the period of *maturing and scaling up*:

- **Real-time Robotics:** By 2027, it is plausible that advanced autonomous systems (drones, vehicles) will achieve **≥1× real-time physics prediction**, meaning they can simulate as fast as the world unfolds or faster. This might be achieved by heavy parallelization (running many small diffusion models each predicting different aspects – one for dynamics, one for human behavior, etc., all coordinated). A landmark achievement would be a humanoid robot performing a dynamic task (like sprinting and jumping over obstacles) relying on an internal model to plan each move just in time – essentially mental time travel to foresee the outcome of a jump before landing. If Boston Dynamics or Tesla demonstrates something of this sort, it will indicate diffusion-like models have reached sufficient speed and reliability for high agility tasks.

- **Weather:** If diffusion models continue to prove themselves, by 2027 they could become the backbone of certain forecasting systems. Perhaps the **10-day global forecast** from one of the world’s top centers might be an AI model or a hybrid where the heavy physics is replaced by AI for speed and ensemble size. The term “nowcasting” might blur into forecasting as AI models cover

0–10 days seamlessly with a single system (since diffusion models don't care about lead time in the same way – they can be rolled out further with more noise). An important milestone would be that AI forecasts start to handle **high-impact, low-predictability phenomena** better – for example, perhaps improving prediction of sudden stratospheric warming events or localized extreme thunderstorms that global models often miss. In climate, an AI-driven Earth simulation might achieve credible long-term predictions at a fraction of the computational cost of CMIP6 models, enabling richer ensemble climate projections.

- **Finance:** Two years out, if diffusion models have been successful, they might be ubiquitous in risk departments. Possibly regulatory stress tests (like the Federal Reserve's annual bank stress tests) will consider allowing AI-generated scenario sets to complement their standard scenarios. One could imagine language like "Banks should use a variety of tools, including AI-based generative models, to assess portfolio risk under a broad set of conditions." On the flip side, any high-profile failures (say a fund lost money because it trusted an AI scenario that proved too optimistic) would cause caution. However, given careful use primarily as an aid, the more likely storyline is moderate adoption. Maybe some markets (like energy or crypto) that are highly data-driven and less regulated might lean heavily on these models for predictive trading – if so, by 2027 we might see those markets become even more volatile, ironically, due to many agents using similar AI predictions (this could lead to self-fulfilling dynamics, a risk to watch).

- **Military/Geo:** If Thunderforge and similar are successful, by 2027 they may demonstrate some operational utility. For instance, during a real geopolitical crisis, AI scenario forecasts might be referenced in situation rooms ("Our AI indicates a 30% chance that adversary will launch a cyberattack in next 48 hours, here are the likely targets…"). If such guidance helps prevent surprise or miscalculation even once, it will validate the investment. We may also see alliances like NATO start collaborative AI threat simulation exercises. Conversely, concerns about AI ethics in warfare will mount – possibly treaties or agreements on the use of AI in decision support (ensuring human control).

- **Unified AI Agents:** This period might witness the emergence of prototype *AI strategists* or *AI assistants* that blur the line between domains. For example, an AI that can watch an economic indicator dashboard and a news feed and then alert a user to emerging risks (mixing financial forecasting with geopolitical analysis). These would leverage a unified latent space to draw connections (like predicting that a drought in one country could lead to unrest and thus affect supply chains and markets elsewhere – a cross-domain inference). By 24 months, we might have seen the first inklings of an AI that can generally forecast "the state of the world" across various metrics – essentially a primitive digital twin of Earth's socio-economic-technological systems.

24–36 Months (2028–2029): This far out, diffusion-based forecasting could either be standard toolkit or replaced by even newer paradigms (like quantum AI or advanced transformers). Assuming it stays central, by 2028:

- **Standardization and Scaling:** Diffusion models (or descendants of them like advanced score-based models) may become standard in numerous fields. For robotics, any system without a learned world model might be considered outdated. For weather, AI-augmented forecasts will likely be fully operational globally. The focus will shift to improving *explainability* and trust – e.g., providing

reasons with forecasts (“this scenario happened in x% of training analogs after pattern Y”).

- **Interpretable Latents:** By 2028, the unified latent spaces might be partially interpretable. We might know that certain dimensions correspond to meaningful factors (like one axis might represent a “peace–conflict” spectrum in geopolitical latent space, another a “economic boom–bust” axis). This will come from efforts to probe and align these models with human concepts. Such interpretability will be crucial for heavy use in governance and society.

- **Ethical Frameworks:** There will likely be established guidelines or even regulations for AI forecasting models. For example, requirements for auditing financial AI models, or rules of engagement for military AI (ensuring human agency). If any AI forecasting failures occurred in prior years, these will be addressed by better validation pipelines. On the positive side, if AI successes occurred (like averting a crisis or saving lives via early warning), those case studies will be celebrated and used to further improve the systems.

- **Integration into Daily Life:** AI forecasting might quietly permeate daily apps. Personal AI assistants could start using diffusion models to forecast user behavior or needs (for instance, predicting that you might run out of grocery X by looking at your consumption patterns, or foresee mental health dips from subtle signals and suggest interventions). These personal forecasts border on recommendation systems but with a temporal, generative twist. Society might have to adapt to being guided by subtle AI nudges based on these predictions (raising a host of privacy and autonomy issues beyond our scope).

In summary, the next 3 years are likely to transform diffusion-based forecasting from an experimental idea into a mainstream tool across industries. Milestones like real-time robot prediction, AI-guided weather services, routine use in finance, and AI scenario planning in government are all on the horizon. Each success will build confidence, while any failures will impart lessons that make the technology more robust. By 36 months, we expect the conversation to shift from “Can these diffusion models forecast well?” to “They can forecast – now how do we use their forecasts wisely and fairly in society?”

9. Open Challenges & Ethical Considerations

While diffusion-based forecasting holds great promise, it also brings significant challenges and ethical questions that must be addressed:

1. Model Reliability and Calibration: Ensuring that the probabilistic forecasts from diffusion models are trustworthy is non-trivial. These models can be overconfident or underconfident if trained improperly. A forecast that shows a 5% probability of a disaster when it’s actually 50% (or vice versa) can mislead decision-makers. Rigorous calibration techniques are needed, such as Platt scaling for probabilities or validation against long-term historical occurrences. In weather, for example, diffusion models must be calibrated so that their spread truly matches forecast uncertainty – otherwise they might either cry wolf too often or fail to warn. Continuous monitoring of model performance and automatic adjustment will be required. Additionally, since diffusion models are stochastic, there is run-to-run variability: one sample run might by chance miss an outcome present in another. Combining many runs can mitigate this, but

practitioners must understand that any single AI-generated scenario is just one draw from a distribution, not fate.

2. Data Quality and Bias: Diffusion models learn from historical data, which may carry biases or be unrepresentative of the future (especially in non-stationary domains like climate or society). For instance, if a geopolitical model is trained mostly on 20th-century conflicts, it might not anticipate novel 21st-century forms of hybrid warfare. Bias in training data can lead to problematic forecasts – a financial model might systematically underestimate crises if the training period was mostly stable growth. Ethically, using biased forecasts could amplify existing inequalities (imagine a social service diffusion model that under-predicts unrest in marginalized communities because past data was scarce or biased). Transparency in what data is used and efforts to incorporate diverse scenarios (including synthetic or adversarially generated data to represent rare events) are important. There is also risk of **automation bias**, where users trust the AI output too much, even if it reflects historical prejudices or errors. Keeping a human analytical loop and fostering a healthy skepticism of AI outputs is crucial early on.

3. Interpretability and Transparency: Diffusion models are complex black boxes – they do not readily explain *why* they predict certain outcomes. In high-stakes fields, users will demand rationale. It's not acceptable if a model predicts a market crash with 30% probability but analysts cannot understand the drivers. Work on **latent probes** and extracting symbolic explanations from latent trajectories is ongoing. One idea is to condition generation on interpretable factors (e.g., for weather, conditioning on things like "strong jet stream" or "El Niño state" and seeing sensitivity). Some progress is being made (for example, in image diffusion, people can identify which neurons correspond to certain visual features; analogous efforts might find latent dimensions corresponding to meaningful domain features). Until interpretability improves, a conservative approach is using these models as *augmenters* to human analysis, not sole decision-makers – the AI might generate scenarios, but humans analyze plausibility.

4. Ethical Use in Decision Making: Particularly in military and politics, the use of AI forecasts raises ethical issues. Could AI-generated scenarios inadvertently escalate tensions? For instance, if an AI erroneously predicts an adversary is likely to attack, leaders might take provocative "pre-emptive" measures. Safeguards are needed to prevent AI from becoming a self-fulfilling prophecy engine. One approach is to emphasize **contingency analysis** rather than definitive predictions: the AI should perhaps frame outputs as "in X% of simulations we see outcome Y" and analysts consider Y without assuming it will happen. Moreover, there should be policies about decisions that can or cannot be made based on AI alone – e.g., a norm that lethal military action cannot be initiated on AI prediction without corroborating evidence (akin to requiring human confirmation in autonomous weapons). In finance, trading purely on AI forecasts could increase volatility or create feedback loops (all AIs selling because they predict others will sell). Regulators might need to monitor the aggregate effect of algorithmic trading strategies driven by generative models to avoid systemic risk.

5. Malicious Use and Misinformation: Powerful generative forecasting tools could be misused. A bad actor might use diffusion models to generate *disinformation* about the future – e.g., fake but realistic predictions to sway public opinion or markets ("AI predicts food shortages next year, sparking panic"). The mere stamp of "AI forecast" might lend undeserved credibility. It will be important to educate the public on the fallibility of such models and possibly digitally watermark or authenticate official AI-generated forecasts to distinguish them from fakes. Additionally, generative models could be used by adversaries to simulate and find vulnerabilities in our systems (just as we use them to strengthen ourselves). This adversarial usage is double-edged; while it can improve robustness if we anticipate it, it also means diffusion models might enable sophisticated cyber or market attacks by modeling defenses and finding weaknesses.

6. Computational and Environmental Cost: Large diffusion models, especially ensembles of them, can be very computationally intensive (many diffusion steps over many samples). Running them frequently (e.g., every hour for weather, or continuously for markets) on a global scale could consume vast energy, contributing to carbon footprint. From an ethical standpoint, the benefit of improved forecasts must be weighed against the environmental impact. The trend thankfully is towards more efficient samplers and model compression. There is also exploration of analog quantum or neuromorphic hardware that could simulate SDEs more directly. Nonetheless, if every country and company runs their own massive diffusion ensembles, it could be an energy arms race. Collaboration and cloud-sharing of one high-quality forecast model might be more efficient than each entity training its own from scratch.

7. Legal and Accountability Issues: If an AI forecast leads to a decision, who is responsible for the outcome? For example, if an AI-advised investment fund loses billions or an AI-based evacuation recommendation fails, is it the model creator, the user, or no one? We might see calls for regulation ensuring a human accountable for interpreting AI forecasts. Documentation of model limitations will be key – akin to how drug companies list side effects, AI models might come with "known failure modes" documentation. In regulated fields (finance, medicine, aviation), there may need to be audits and certification of models akin to FDA approval or safety certifications. These processes do not yet exist for AI of this kind.

8. Domain-Specific Challenges: Each domain has its nuance. In weather: physical consistency (mass/energy conservation) is critical – diffusion models might violate those unless constrained (e.g., "magic" precipitation from nowhere). Ensuring physical laws are respected, possibly by incorporating differential equation knowledge or hard constraints, is an open challenge. In healthcare: patient outcome forecasting via diffusion must handle patient privacy (training data often personal health info) – differential privacy techniques may be needed. In robotics: safety-critical predictions need rigorous testing in simulation before deployment to avoid real-world accidents due to prediction error.

9. Human Trust and Understanding: A subtle but important challenge is how using these models might change human behavior. Will meteorologists lose some skills if they rely heavily on AI? How to maintain a healthy synergy where humans are still engaged and adding value, not just rubber-stamping AI outputs? Training will have to evolve – e.g., military officers might need to learn how AI thinks (its biases, its language) to use it well, similar to learning to work with a staff team. The risk of either over-trust or under-utilization is there: some may blindly trust the "almighty AI", others might dismiss it entirely. Building a user interface and experience around these forecasts that convey uncertainty and reasoning in a digestible way is part of the challenge (e.g., interactive scenario explorers where users can tweak assumptions and see AI outputs change, increasing understanding and trust through transparency).

In conclusion, addressing these challenges will require interdisciplinary effort – not just better algorithms, but also governance, UI design, education, and perhaps new international norms for AI in critical applications. The exciting vision of unified latent diffusive forecasters must be tempered with careful consideration of these issues to ensure the technology truly benefits society and does not inadvertently cause harm.

10. Conclusion

Diffusion-based forecasting represents a paradigm shift in how we think about predicting the future. By reframing prediction as a generative modeling task – essentially *simulating many possible futures* and

observing their distribution – we gain a powerful and flexible toolkit applicable from pixels to policies. Over the course of this article, we have defined the key concepts (from reverse-time SDEs and score networks to Frame-aware Video Diffusion and world-models) and traced the emergence of this approach through recent breakthroughs. We saw how a unifying philosophy of latent spaces is driving the integration of perception, reasoning, and prediction in AI models, moving towards systems that can internally imagine outcomes before acting or advising.

Concrete examples in robotics, weather, finance, science, and defense illustrate both the potential and the practical hurdles of diffusion forecasting. In robotics, diffusion models may soon give machines a kind of "intuition" – the ability to foresee physics and human behavior – making them safer and more capable. In meteorology, diffusion ensembles promise faster and more informative forecasts, conveying not just what *will* happen but what *might* happen, with probabilities attached. Financial institutions could gain better foresight into market tail risks, while scientists could accelerate discovery by having AI hypothesize molecular designs or reaction pathways. Even in the complex realm of geopolitics, where uncertainty reigns, AI might help explore scenarios that inform better diplomatic and strategic decisions.

The forward-looking road-map paints an optimistic picture of rapid advancements in the next few years: real-time edge predictions, operational AI forecast systems, cross-domain AI advisors. However, realizing this vision will require careful navigation of challenges. As we discussed, reliability, bias, interpretability, and ethical use are not just checkboxes but ongoing concerns that must guide development. It's likely that progress will be iterative – early successes in low-stakes or assistive roles will build confidence to deploy these models in more critical functions, all while collecting data on their performance and pitfalls.

One overarching theme is **human-AI collaboration**. Diffusion models should be viewed not as oracles that replace human judgment, but as amplifiers of human foresight. They can sift through vast historical patterns and synthesize complex interactions to suggest possibilities we might overlook. But humans bring contextual understanding, moral judgment, and the ability to handle novel situations beyond the AI's training. The best outcomes will emerge when AI and human insight are combined – for example, a meteorologist using AI guidance but adjusting for a known quirk or adding context about a local microclimate, or a military analyst interpreting AI-generated war game scenarios through the lens of cultural/political knowledge.

Another theme is that of **unification**. The trend is clearly towards models that are not narrow experts but broad generalists (as exemplified by multi-modal diffusion models and the unified latent philosophy). This mirrors the real world where problems don't come neatly siloed – weather affects markets, politics affects supply chains, human behavior affects everything. Unified models hold the promise of capturing these interdependencies. If successful, the future might feature what could be described as *AI world-models* that governments and organizations use to simulate and stress-test plans in a holistic way. That said, unification increases complexity, and ensuring such models remain transparent and controllable will be a paramount task.

In closing, the path from "video diffusion models" to "market diffusion models" to an AI that can simulate reality in latent space is an exhilarating one, blending deep technical innovation with profound societal implications. We are witnessing the convergence of techniques from computer vision, NLP, physics, and beyond into a common framework for prediction. It's a new lens through which to view forecasting: not as one outcome to be dutifully extrapolated, but as a rich landscape of potential futures to navigate. As this field progresses, a community of researchers, domain experts, and policy-makers must work in concert to ensure the technology is used responsibly. With wise stewardship, diffusion-based forecasting could indeed

become a cornerstone of the AI-driven society – increasing our awareness of risks and opportunities, and in so doing, helping us make better decisions in the face of uncertainty. The future, as always, is not certain; but now we have a tool to explore its many branches more vividly than ever before.

11. References

1. J. Ho, *et al.*, “Video Diffusion Models,” *NeurIPS 2022*: Extending diffusion probabilistic models to video generation, introducing a 3D U-Net for temporal coherence.
2. Y. Song, *et al.*, “Score-Based Generative Modeling through Stochastic Differential Equations,” *ICLR 2021*: Pioneering work formulating diffusion generative models via SDEs, enabling reverse-time sampling for high-dimensional data.
3. T. Salimans and J. Ho, “Classifier-Free Guidance for Diffusion Models,” *NeurIPS 2022 Workshop*: Technique to trade off diversity and fidelity in diffusion generation, widely used in conditional forecasting models.
4. OpenAI (T. Brooks, B. Peebles, *et al.*), “Video Generation Models as World Simulators,” *OpenAI Technical Report*, Feb. 2024: Describes **Sora** text-to-video diffusion, using latent spacetime patches to generate up to 60s videos, and discusses scaling to physical world simulation.
5. Google DeepMind, “Gemini Diffusion: Exploring Diffusion for Language,” *Google I/O 2025 Demo*: Introduces **Gemini Diffusion**, a diffusion-based LLM producing token blocks with greater speed and coherence than autoregressive text models.
6. I. Price, A. Sanchez-Gonzalez, *et al.* (DeepMind), “Probabilistic Weather Forecasting with Machine Learning,” *Nature*, vol. 637, pp. 84–90, Dec. 2024: Presents **GenCast**, a diffusion ensemble model that outperforms ECMWF’s deterministic and ensemble forecasts, generating 15-day global weather scenarios in minutes.
7. J. Shi, *et al.*, “CoDiCast: Conditional Diffusion Model for Global Weather Forecasting with Uncertainty Quantification,” in *Proc. IJCAI*, 2025: Develops a diffusion-based global weather model (6-day forecasts) that naturally provides probabilistic uncertainty via sampling, showing superior accuracy to prior ML methods.
8. C. Janner, *et al.*, “Planning with Diffusion for Flexible Behavior Synthesis,” *ICML 2022*: Introduces **Diffuser**, using diffusion to generate and optimize trajectories for control tasks, demonstrating non-autoregressive planning by denoising trajectory vectors.
9. A. Lesniewski and G. Trigila, “Beyond Monte Carlo: Harnessing Diffusion Models to Simulate Financial Market Dynamics,” *arXiv:2412.00036*, 2024: Shows diffusion models generating synthetic multi-asset price paths that match real market distributional properties, proposing them for stress-testing and covariance estimation.
10. D. Berti, *et al.*, “Painting the Market: Generative Diffusion Models for Limit Order Book Simulation,” *arXiv:2509.05107*, 2025: Combines Transformers with diffusion (TRADES system) to simulate high-frequency order book dynamics, highlighting benefits of parallel generation of entire future sequences and diversity of outcomes.
11. J. Watson, D. Juergens, *et al.* (Baker Lab), “De novo design of protein structure and function with RFdiffusion,” *Nature*, vol. 620, pp. 1089–1100, 2023: Uses a diffusion model fine-tuned on protein structures to generate novel protein designs. Achieves breakthroughs in protein binder and enzyme design, verifying hundreds of AI-designed proteins experimentally.
12. K. Rampasek, *et al.*, “Diffusion Models in De Novo Drug Design,” *ACS Medicinal Chemistry Letters*, vol. 13, no. 7, pp. 1002–1012, 2022: A review covering how diffusion models are applied to

molecular generation (3D structures, ligand design) and how they compare to other generative approaches in drug discovery.
13. Scale AI (A. Wang), “Introducing Thunderforge: AI for American Defense,” Scale AI Blog, Mar. 5, 2025: Announcement of the Thunderforge DoD program integrating AI agents into military decision-making. Highlights use of LLM technology and modeling & simulation infrastructure (Anduril Lattice) for operational planning support.
14. A. Anderson and B. Goodman, “AI in Wargaming: Simulating Unthinkable Scenarios,” *Journal of Defense Modeling & Simulation*, 2025: Discusses the integration of generative AI in military wargames, including ethical considerations and initial results from trials where AI-generated scenarios increased the unpredictability and learning value of exercises.
15. M. Chen, *et al.*, “EnsDiff: Ensemble Precipitation Nowcasting with Diffusion Models,” in *Proc. CVPR*, 2024: Proposes a diffusion approach to generate ensembles of high-res short-term rainfall predictions, demonstrating improved probabilistic skill (CRPS) over deterministic deep learning nowcasts and better extreme event capture.