FROM MASKS TO WORLDS: A HITCHHIKER'S GUIDE TO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

This is not a typical survey of world models, it is a guide for those who want to build worlds. We do not aim to catalog every paper that has ever mentioned a "world model". Instead, we follow one clear road: from early masked models that unified representation learning across modalities, to unified architectures that share a single paradigm, then to interactive generative models that close the action-perception loop, and finally to memory-augmented systems that sustain consistent worlds over time. We bypass noisy branches to focus on the core: the generative heart, the interactive loop, and the memory system. We show that this is the most promising path towards world models.

1 INTRODUCTION: THE NARROW ROAD TO WORLD MODELS

The term *world model* has been used to describe many different ideas: learned environment simulators for reinforcement learning (Ha & Schmidhuber, 2018; Hafner et al., 2019), agents that integrate learned models with planning (Schrittwieser et al., 2020), and large language models that simulate entire societies (Park et al., 2023). Yet despite hundreds of related works, there is no clear consensus on how to actually build a true world model. In this paper, we take a stance: the path is much narrower than it appears.

A true world model is not a monolithic entity, but a system synthesized from three core subsystems: a generative heart to produce coherent world states, an interactive loop to close the action-perception cycle in real time, and a persistent memory system to sustain coherence over long horizons. The history of the field can be understood as an evolutionary journey to first master these components in isolation, and now, to integrate them. Most works focuse on optimizing narrow tasks and drift away from the generative, interactive, and persistent nature required for a true world model.

To make this perspective concrete, we chart the historical evolution of world models as a sequence of five stages, shown in Figure 1. It begins with Stage I: Mask-based Models, which established a universal, token-based pretraining paradigm across modalities. This foundation enabled Stage II: Unified Models, where a single architecture learns to process and generate multiple modalities. The focus then shifts to closing the interactive loop in Stage III: Interactive Generative Models, transforming static generators into real-time simulators. To sustain these simulations over time, Stage IV: Memory and Consistency introduces mechanisms for durable and coherent state representation. Table 1 also summarizes representative models or methods across the four stages.

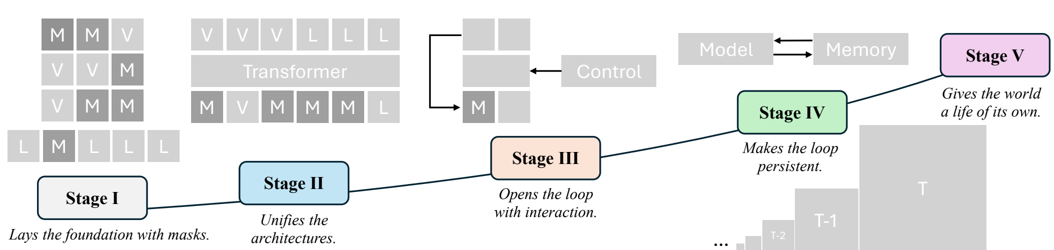

Figure 1: The evolution of world models across five stages.

Table 1: Representative models or methods along the narrow road to world models.

Stage I: Mask-based Models	
BERT (Devlin et al., 2019)	Bidirectional masked prediction for representation learning in language.
RoBERTa (Liu et al., 2019)	Dynamic masking and scale without next-sentence prediction strengthen BERT.
Gemini Diffusion (DeepMind, 2025)	Reported iterative denoising paradigm at commercial scale for generative language tasks.
BEiT (Bao et al., 2021)	Image patch masking for representation learning in vision.
MAE (He et al., 2022a)	High-ratio patch masking with lightweight decoder yields strong visual representations.
MaskGIT (Chang et al., 2022)	Non-autoregressive parallel masked tokens infilling for efficient image synthesis.
Meissonic (Bai et al., 2024)	Masked generative transformers achieving high fidelity text-to-image generation.
wav2vec 2.0 (Baevski et al., 2020)	Audio latent features masking for representation learning in speech.
Stage II: Unified Models	
EMU3 (Wang et al., 2024)	AR-based unified models with a single Transformer for text, image and video.
Chameleon (Chameleon Team, 2024)	AR-based unified models with a single Transformer for text and image.
VILA-U (Wu et al., 2024)	Language-prior AR-based unified models for text, image and video.
Janus-Pro (Chen et al., 2025)	Language-prior AR-based unified models for text and image.
MMaDA (Yang et al., 2025)	Language-prior mask-based (discrete-style denoising) unified models for text and image.
UniDiffuser (Bao et al., 2023)	Visual-prior diffusion-based unified models for text and image.
Muddit (Shi et al., 2025)	Visual-prior mask-based (discrete-style denoising) unified models for text and image.
UniDisc (Swerdlow et al., 2025)	Mask-based (discrete-style denoising) unified models.
Gemini (Comanici et al., 2025)	Google's multimodal model in a single system (but not in a single paradigm).
GPT-4o (Hurst et al., 2024)	OpenAI's multimodal model in a single system (but not in a single paradigm).
Stage III: Interactive Generative Models	
TextWorld (Côté et al., 2018)	Parser-based text game environments.
AI Dungeon (Latitude, 2024)	LLM-driven co-authored narrative with open-ended branching stories.
PVG (Menapace et al., 2021)	Stepwise playable video game conditioned on user action selection.
PE (Menapace et al., 2022)	3D playable environments conditioned on camera and multi-object control.
PGM (Menapace et al., 2024)	Promptable game model conditioned on semantic-level language control.
GameGAN (Kim et al., 2020)	GAN-based next frame generation conditioned on actions for 2D games.
Genie-1 (Bruce et al., 2024)	MaskGIT-based next frame generation conditioned on actions for 2D worlds.
Oasis (Decart et al., 2024)	Open-source Diffusion-based real-time generation conditioned on actions for 3D games.
GameNGen (Valevski et al., 2024)	Diffusion-based real-time next frame generation conditioned on actions for 3D games.
Genie-2 (Parker-Holder et al., 2024)	Diffusion-based generation conditioned on actions for 3D worlds initialized from images.
Genie-3 (Ball et al., 2025)	Real-time generation conditioned on actions and promptable world events for 3D worlds.
Mineworld (Guo et al., 2025)	Open-source MaskGIT-based generation conditioned on actions for 3D games.
Matrix-Game-2 (He et al., 2025)	Open-source diffusion-based real-time generation conditioned on actions for 3D games.
World Labs (World Labs, 2024)	Explorable 3D environments generation from a single image using geometry and depth.
Stage IV: Memory & Consistency	
RETRO (Borgeaud et al., 2022)	Improving LMs by conditioning on document chunks retrieved from a large corpus.
MemGPT (Packer et al., 2023)	OS-inspired virtual memory management framework for LLM workflows.
Transformer-XL (Dai et al., 2019)	Segment-level recurrence with relative positions for long-context sequence modeling.
Compressive Transformer (Rae et al., 2019)	Extends Transformer-XL by downsampling old states to retain long-range dependencies.
Mamba (Gu & Dao, 2023)	Selective state-space model with linear-time recurrence supporting near-infinite context.
FramePack (Zhang & Agrawala, 2025)	Packs long-frame histories into fixed context with inverted sampling to reduce drift.
MoC (Cai et al., 2025)	Learnable sparse attention routing that retrieves informative history chunks and anchors.
VMem (Li et al., 2025a)	Introduces surfel-indexed view memory using 3D surfels to enforce spatial coherence.

This progression culminates in Stage V: True World Models. This stage is not defined by adding a new component, but by the synthesis of the preceding stages into an autonomous whole. At this threshold, models begin to exhibit the defining properties of persistence, agency, and emergence, moving from engines of prediction to living worlds. By analyzing each stage's key innovations and unsolved challenges, this paper offers a clear and opinionated roadmap from today's components to tomorrow's living worlds.

2 WHAT IS A WORLD MODEL?

2.1 HISTORICAL AND CONTEMPORARY PERSPECTIVES

The concept of a world model originated in reinforcement learning, where Ha and Schmidhuber (Ha & Schmidhuber, 2018) first proposed learning a latent dynamics simulator for agent planning. This control-oriented view was advanced by systems like Dreamer (Hafner et al., 2019), which learned policies purely through latent imagination, and MuZero (Schrittwieser et al., 2020), which integrated tree-based planning with a learned, abstract model. In parallel, the rise of large-scale generative modeling broadened this definition. With generative agents (Park et al., 2023) and large multimodal systems (Reed et al., 2022), the concept evolved from a predictive simulator for an agent to a rich, generative system that could be an entire interactive world. This has led to the contemporary view of a "world simulator", a term that now informally encompasses three major paradigms: explicit 3D scene generators (World Labs, 2024), passive video generators that go beyond pixels to approximate physical dynamics (Brooks et al., 2024), and interactive games and environments for agents, whether text-based (Niesz & Holland, 1984) or video-based, as exemplified by the Genie series (Bruce et al., 2024; Parker-Holder et al., 2024; Ball et al., 2025).

2.2 THE ANATOMY OF A TRUE WORLD MODEL

To bring clarity to these diverse threads, we define a true world model by the three essential subsystems it must integrate, which in turn enable the core properties that define each stage of our evolutionary roadmap. Figure 2 presents the high-level architecture of a true world model, showing how the generative, interactive, and memory subsystems integrate.

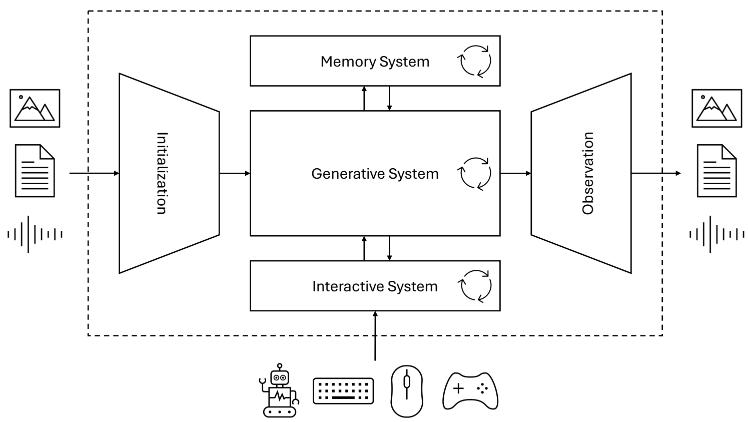

Figure 2: The architecture of a true world model.

The Generative Heart ($\mathcal{G}$). The foundation of a world model is its generative heart: a learned model of the world's dynamics and appearance, formally described by the generative process p_θ. It must be able to predict future states, observations, and the task-relevant outcomes.

$$\mathcal{G} = \Big(\underbrace{p_\theta(z_{t+1} \mid z_t, a_t)}_{\text{Dynamics}}, \underbrace{p_\theta(o_t \mid z_t)}_{\text{Observation}}, \underbrace{p_\theta(r_t \mid z_t, a_t)}_{\text{Reward}}, \underbrace{p_\theta(\gamma_t \mid z_t, a_t)}_{\text{Discount/Termination}} \Big)$$

This subsystem, which models state transitions, observations, rewards, and terminations, is the foundation for the property of **Generation**.

The Interactive Loop ($\mathcal{F}, \mathcal{C}$). To be more than a passive movie generator, the model must support a closed interactive loop. For partially observable worlds, it requires an *inference filter* (q_ϕ) for the agent to interpret observations in real-time, and a *policy* (π_η) for it to act upon its understanding of the world, often paired with a value function (v_ω) to evaluate trajectories.

$$\mathcal{F}: \quad \underbrace{q_\phi(z_t \mid h_{t-1}, o_t)}_{\text{State Inference}}, \qquad \mathcal{C} = \Big(\underbrace{\pi_\eta(a_t \mid z_t, h_t)}_{\text{Policy}}, \underbrace{v_\omega(z_t, h_t)}_{\text{Value}} \Big)$$

This loop is what enables true **Interaction** and **Real-time Adaptation**.

The Memory System ($\mathcal{M}$). Finally, to ensure coherence over time, the model needs a memory system that allows past events to inform the future. This is formally captured by a recurrent state, h_t, which is updated based on past memory, the current inferred state, and the last action.

$$\mathcal{M}: \quad \underbrace{h_t = f_\psi(h_{t-1}, z_t, a_{t-1})}_{\text{Memory Update}}$$

This component is the basis for the property of **Memory**.

A detailed formalism of each component is provided in Appendix A. This definition clarifies why a system like a Unified Model (Stage II) is a precursor, not a true world model. While it may possess a powerful generative heart, it typically lacks the dedicated interactive loop and explicit memory system required to sustain a persistent, agent-inhabited world.

3 STAGE I: MASK-BASED MODELS ACROSS MODALITIES

The first stage in the evolution toward world models is the era of *mask-based modeling*, where a system learns by reconstructing missing or corrupted parts of its input. This paradigm, which can be summarized as mask, infill, and generalize, has proven to be strikingly universal across modalities. It provides a unified way of tokenizing, representing, and pretraining large models, establishing the foundation for all subsequent stages.

3.1 LANGUAGE MODALITY

Masked language modeling (MLM) has played a foundational role in modern natural language processing. BERT (Devlin et al., 2019) introduced bidirectional context prediction, where 15% of tokens in each input are randomly replaced with a [MASK] symbol and predicted from surrounding context. SpanBERT (Joshi et al., 2020) refined this approach by masking contiguous spans rather than isolated tokens, improving extraction and reasoning tasks. Sequence-to-sequence variants such as MASS (Song et al., 2019), T5 (Raffel et al., 2020), and BART (Lewis et al., 2019) reformulated MLM as a denoising autoencoding objective. ELECTRA (Clark et al., 2020) improved sample efficiency by replacing the MLM objective with a discriminative replacement-detection task.

Beyond fixed-ratio masking, a line of non-autoregressive work introduces dynamic masking and unmasking through iterative refinement. RoBERTa (Liu et al., 2019) demonstrated that simply optimizing BERT's training recipe with more data and dynamic masking yielded significant gains. Mask-Predict (Ghazvininejad et al., 2019) introduced iterative refinement, re-masking low-confidence tokens over several passes. This concept culminated in discrete diffusion models (Li et al., 2022; He et al., 2022b; Gong et al., 2022; Ou et al., 2024; Sahoo et al., 2024; Shi et al., 2024), which replace fixed masking with a time-indexed noise schedule and train the model to iteratively denoise. As demonstrated by industrial systems like Mercury (Inception Labs et al., 2025) and Gemini Diffusion (DeepMind, 2025), this dynamic denoising paradigm has matured to rival or exceed autoregressive baselines in both quality and inference speed, solidifying the power of masking as a core generative principle (Yu et al., 2025c; Li et al., 2025b).

3.2 VISION MODALITY

The masked image modeling (MIM) paradigm extended this principle to perception. Early works established two main branches. For representation learning, BEiT (Bao et al., 2021) and especially MAE (He et al., 2022a) created direct visual analogues to BERT, reconstructing masked tokens or patches to learn powerful features. This spurred a family of related works exploring different reconstruction targets and self-distillation techniques (Xie et al., 2022; Zhou et al., 2021; Wei et al., 2022).

For generative modeling, MaskGIT (Chang et al., 2022) and MUSE (Chang et al., 2023) pioneered the use of masked infilling for high-quality parallel image synthesis. This generative trajectory has recently culminated in models like Meissonic (Bai et al., 2024), which demonstrates that masked generative transformers can achieve fidelity rivaling large diffusion models while offering superior efficiency and control.

This mask-reconstruct-generalize principle scaled effectively to video. VideoMAE (Tong et al., 2022) and MaskFeat(Wei et al., 2022) showed that high-ratio tube masking was a data-efficient method for learning spatiotemporal representations, confirming that masking could capture not just static scenes but also their dynamics.

3.3 Other Modalities

The universality of the masking paradigm was confirmed by its rapid adoption in other fields. In audio, models like wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), and Audio-MAE (Huang et al., 2022) applied masked prediction to latent speech representations. In 3D domains, Point-BERT (Yu et al., 2022) and Point-MAE (Pang et al., 2023) adapted masking to point clouds. The principle was even extended to structured data with models like GraphMAE (Hou et al., 2022). These successes reinforced masking as a cross-domain general approach to self-supervised learning.

In summary, Stage I established the principle of masking as a universal foundation for representation learning. While this unified the pretraining paradigm, the models themselves remained specialized architectures. The inability of these separate models to form a holistic worldview motivated Stage II: the pursuit of a single, unified architecture.

4 Stage II: Unified Models

Stage I established a universal paradigm for representation learning, but the models themselves remained specialists locked within their own modalities. Stage II takes the crucial next step: unifying the models themselves. We define a unified model as a system that processes and generates across different modalities with the shared backbone and the same paradigm. By collapsing modality-specific pipelines, these models simplify scaling, enable powerful cross-modal transfer, and represent the first decisive synthesis on the path toward a true world model.

4.1 Representative Works

Leading unified modeling efforts span several trajectories, distinguished by their foundational paradigm. We exclude simple glue models that stitch different paradigms for different modalities, such as using autoregression for text and diffusion for image, as well as models limited to text generation without extending to image generation or other modalities.

Extending Language Model Pre-training: Language-Prior Modeling. The dominant trajectory has been to extend the paradigm of autoregressive large language models (LLMs) (Radford et al., 2019; Brown et al., 2020). This began by connecting pre-trained vision encoders to frozen LLMs, as pioneered by BLIP-2 (Li et al., 2023) and popularized by LLaVA (Liu et al., 2023b; 2024), which was built upon LLaMA (Touvron et al., 2023). This approach was pushed further into grounded multimodal reasoning by Kosmos-2 (Peng et al., 2023) and embodied reasoning by PaLM-E (Driess et al., 2023). More recently, systems like the EMU family (Sun et al., 2024; Wang et al., 2024), Chameleon (Chameleon Team, 2024), VILA-U (Wu et al., 2024), and Janus-Pro (Chen et al., 2025) have advanced towards true end-to-end unified generation, creating both text and images within shared token space and unified autoregressive paradigm. In parallel, A notable offshoot of this trend is rooted in mask-based language modeling. LLaDA (Nie et al., 2025) abandons the autoregressive framework and models text through a masked diffusion process with a single Transformer. Its multimodal extension, MMaDA (Yang et al., 2025), introduces a unified discrete diffusion architecture for text and image, a mixed chain-of-thought fine-tuning strategy, and a policy-gradient RL algorithm (UniGRPO) to unify reasoning and generation across modalities within a single model.

Extending Vision Model Pre-training: Visual-Prior Modeling. A parallel effort started from vision-centric foundations, primarily along two paths. The first path built upon latent diffusion models, the foundation laid by Stable Diffusion (Rombach et al., 2022) was later generalized to a unified, joint diffusion process over text and images in models like UniDiffuser (Bao et al., 2023).

The second path built upon the masked image modeling (MIM) paradigm, with models like Muddit (Shi et al., 2025) extending Meissonic (Bai et al., 2024) into a unified discrete diffusion system that produces both images and captions within shared architecture and paradigm. Besides, UniDisc (Swerdlow et al., 2025) trained a unified discrete-diffusion model from scratch for both language and vision modalities.

Industrial-Scale Unified Systems. At production scale, Gemini (Comanici et al., 2025) and GPT-4o (Hurst et al., 2024) unify language and vision modalities in a single model, although not in a single paradigm. These demonstrate that unified modeling has transcended research to become a foundational industrial paradigm.

4.2 Benefits and Gaps

The primary benefit of Stage II is the reduction of fragmentation, leading to powerful cross-modal transfer and emergent capabilities. This paradigm now underpins productized multimodal interaction at scale, as demonstrated by industrial systems like Gemini (Comanici et al., 2025) and GPT-4o (Hurst et al., 2024). However, despite the impressive progress of language-prior unified models in interactive dialogue, visual-prior unified models for text-to-image and text-to-video remain limited to single-shot synthesis or stepwise editing. They lack the capacity for continuous, real-time closed-loop interaction. Thus while Stage II unified architectures, the creation of truly dynamic and interactive worlds remains an open challenge and motivates Stage III.

5 Stage III: Interactive Generative Models

Here, models are no longer static predictors or one-shot generators, but participants in a closed action-perception loop, sustaining interaction through low-latency response and action-conditioned evolution. We define interactive generative models as systems whose outputs are conditioned on streamed inputs or user actions, supported by internal state. We explore this evolution across three distinct domains: language-based, video-based and scene-based.

5.1 Language-based Worlds: Interaction as Narrative

Classic interactive fiction (IF) (Niesz & Holland, 1984; Montfort, 2011; Ammanabrolu et al., 2020) established the paradigm of text-driven worlds where players interact through textual descriptions and actions. These took several forms: parser-based games where the player types text commands character by character, choice-based games where the player selects from a set of predefined action options, hypertext-based games where the player clicks on links embedded in the narrative. Choice-based visual novels, such as Memories Off (KID, 2004), exemplify emotionally branching narratives in which player decisions directly affect relationships and endings. These static worlds naturally evolved into benchmarks for artificial intelligence. A significant line of research, supported by platforms like TextWorld (Côté et al., 2018) and Jericho (Hausknecht et al., 2020), was dedicated to training agents that could master them. In these settings, the world was a fixed puzzle to be solved, and the locus of intelligence was the agent who navigated a static world, not the world itself.

A fundamental shift occurred when large language models (LLMs) (Hurst et al., 2024; Comanici et al., 2025) themselves became the world engine. AI Dungeon (Latitude, 2024) pioneered this transition, dynamically generating new narrative branches in response to free-form user prompts. Players could explore unbounded story spaces limited only by imagination and the model's generative capacity. This marked the transition from solving pre-authored worlds to co-creating open-ended ones, envisioning a future where visual novels such as Memories Off (KID, 2004) could be interactively generated, offering unique storylines and relationships for each player.

5.2 Video-based and Scene-Based Worlds: Interaction as Experience

Interactive generation in video and spatial domains has progressed from offline frame prediction to real-time, controllable simulation. Early work on world models (Ha & Schmidhuber, 2018) used latent rollouts to "dream" trajectories for policy training, demonstrating the potential of closed-loop simulation. GameGAN (Kim et al., 2020) advanced this idea into a neural game engine, rendering successive frames from user input while implicitly learning game rules from observation. User control evolved from stepwise action selection in Playable Video Generation (PVG) (Menapace et al., 2021), through 3D scenes with camera and multi-object control in Playable Environments (PE) (Menapace et al., 2022), to natural language prompts in Promptable Game Models (PGM) (Menapace et al., 2024), which enabled semantic-level direction of play.

Building on these conceptual foundations, a decisive trajectory emerged with the Genie series. Genie-1 (Bruce et al., 2024) learned latent action interfaces from Internet-scale videos to create controllable 2D environments. Genie-2 (Parker-Holder et al., 2024) extended this capability to larger, quasi-3D spaces, initialized from a single image and playable via standard controls. Genie-3 (Ball et al., 2025) scaled further, producing real-time text-to-world experiences at 720p and 24 fps with minutes of coherent play, a marked shift from passive video generation to active interaction.

Community and industrial efforts soon followed. Systems such as Oasis (Decart et al., 2024), GameNGen (Valevski et al., 2024), Mineworld (Guo et al., 2025), and Matrix-Game (He et al., 2025) demonstrated *real-time* open environments with emergent physics and streaming diffusion. For a comprehensive overview, see the survey by Yu et al. (2025b).

Beyond frame synthesis, scene-based approaches emerged. World Labs (World Labs, 2024) proposed large world models that generate explorable 3D environments from a single image, enabling interactive navigation through generated geometry and depth rather than sequential video.

Taken together, these advances trace a trajectory from offline video generators to real-time, action-conditioned world simulators. They ultimately transform generative models into engines of interactive human experiences.

5.3 Challenges

Despite the leap to real-time interaction, sustaining long-horizon consistency remains unsolved. Two paradigms illustrate the tension: explicit scene generators like NeRFs and Gaussian Splatting (*e.g.*, World Labs) offer stable 3D navigation environments but depend on explicit spatial modeling; implicit frame-by-frame generators offer flexibility but are brittle, prone to losing context and hallucinating objects, especially over extended play. The Genie series highlights this tradeoff: from Genie-1's short 16-frame memory (Bruce et al., 2024), to Genie-2's object permanence (Parker-Holder et al., 2024), to Genie-3's few minutes of coherence (Ball et al., 2025), progress is clear yet far from persistence. At the object level, implicit video models rely on KV caches or control signals to maintain identity, while explicit 3D approaches embed spatial location directly but still struggle with dynamic elements, as explored in 4D Gaussian Splatting. These challenges reveal a deeper gap: the reactive action–perception loop enables interaction, but without dedicated memory and state management, it cannot sustain persistent worlds, which is the central theme of Stage IV.

6 Stage IV: Memory and Consistency

A world model that acts without memory is reactive yet forgetful. This stage aims to endow models with mechanisms that sustain coherent state across long horizons. The central question emerges: can world models not only generate but also sustain coherent histories, preserve identities, and resist drift? We organize this section around three questions: where to anchor memory, how to extend its span and capacity, and how to govern it to preserve consistency.

6.1 Externalized Memory

Retrieval augments parametric models with non-parametric, often editable, knowledge stores. Early explorations such as Neural Turing Machines (Graves et al., 2014), Differentiable Neural Computers (Graves et al., 2016), and End-to-End Memory Networks (MemN2N) (Sukhbaatar et al., 2015) first explored learnable read–write memory slots. While conceptually groundbreaking, their complexity gave way to more pragmatic, decoupled designs. kNN-LM (Khandelwal et al., 2019), REALM (Guu et al., 2020), and RAG (Lewis et al., 2020) showed that conditioning on retrieved passages could dramatically expand effective context while keeping knowledge traceable and updatable. DPR (Karpukhin et al., 2020) and RETRO (Borgeaud et al., 2022) scaled this approach to dense retrievers and trillion-token databases, rivaling far larger dense LMs while providing traceable and updatable evidence.

Beyond simple retrieval, research has sought to make memory more scalable and dynamic. Product Key Memory (PKM) (Lample et al., 2019) supported massive lookup capacity through factorized keys; MemGPT (Packer et al., 2023) reframed LLMs as operating systems with explicit virtual memory management; LONGMEM (Wang et al., 2023) extends KV caches beyond 65k tokens through decoupled readers; and From RAG to Memory (Gutiérrez et al., 2025) extended retrieval into continual learning, enabling dynamic knowledge updates without retraining. These systems collectively signal a shift from retrieval as a tool to memory as a co-evolving substrate.

6.2 Extending Capacity and Span

Parallel efforts seek to build persistence directly into the architecture, moving beyond fixed-length attention windows. Within Transformers, Universal Transformer (Dehghani et al., 2018) introduced depth-wise recurrence; Transformer-XL (Dai et al., 2019) propagated segment states across windows; the Compressive Transformer (Rae et al., 2019) down-sampled older activations. Subsequent designs such as the Memorizing Transformer (Wu et al., 2022) and Recurrent Memory Transformer (RMT) (Bulatov et al., 2022) attached associative key–value stores or persistent memory tokens, reaching million-token horizons in practice; Infini-attention (Munkhdalai et al., 2024) added a compressive long-term path for unbounded streaming. In parallel, Perceiver-AR (Hawthorne et al., 2022) introduced a latent cross-attention bottleneck, compressing long inputs into a compact representation and enabling autoregression over 100k tokens across text, images, and music. Together, this line of work represents a reformist trajectory that extends attention through recurrence and compression.

A more radical line argues that persistence requires abandoning quadratic attention entirely. Structured state-space and linear-time models such as S4 (Lu et al., 2023), Mamba (Gu & Dao, 2023), and RetNet (Sun et al., 2023) replace attention with recurrent state updates that achieve linear complexity and thereby, in principle, support infinite context. Precursors such as Linear Transformers (Katharopoulos et al., 2020), together with more recent variants such as Hyena (Poli et al., 2023), pointed in this direction with kernels and long-range convolutions. Together, this line of work represents a revolutionary trajectory that abandons attention in favor of continuous dynamical systems.

Scaling strategies and engineering refinements extend these capacities further. LongNet (Ding et al., 2023) employs dilated attention for billion-token contexts; Ring Attention (Liu et al., 2023a) distributes computation across devices for million-token horizons; LSSVWM (Po et al., 2025) adapts state-space updates for long causal video generation. Practical techniques such as ALiBi (Press et al., 2021), LongLoRA (Chen et al., 2023), and StreamingLLM (Zeng et al., 2024) retrofit long-context ability into existing models. Together, this line of work represents a pragmatic trajectory that extends persistence through scaling strategies and engineering refinements.

Ultimately, these three trajectories, reformist, revolutionary, and pragmatic, converge on the same goal: to achieve genuine continuity, creating models that can read a book, watch a film, or play for hours without losing the thread.

6.3 Regulating Memory for Consistency

Persistence without discipline degenerates into drift. The nature of this challenge depends critically on the underlying world representation, which has largely followed two paradigms: implicit 2D video frames and explicit 3D scenes.

In implicit, autoregressive video models, the primary challenge is preventing two entangled failures: forgetting, where early content fades, and drifting, where errors compound. Efforts to mitigate one often aggravate the other (Zhang & Agrawala, 2025). The Genie series highlights this progression: Genie-1 (Bruce et al., 2024) suffers from short memory and drifts after only a few frames; Genie-2 (Parker-Holder et al., 2024) introduces object permanence and sustains coherence for about a minute; Genie-3 (Ball et al., 2025) reaches emergent multi-minute consistency. This underscores a broader challenge: autoregressively generating an environment is fundamentally harder than producing a pre-rendered video, since small inaccuracies accumulate over time. To tackle this, FramePack (Zhang & Agrawala, 2025) uses keyframe anchoring and context compression; Self-Forcing (Huang et al., 2025) and CausVid (Yin et al., 2025) impose stronger causal constraints; Context-as-Memory (Yu et al., 2025a) retrieves overlapping past frames to stabilize long video rollouts, and Mixture of Contexts (MoC) (Cai et al., 2025) learns sparse routing policies that focus attention on salient history.

Conversely, explicit 3D representations that built upon generative assets from models like Trellis (Xiang et al., 2025), or TripoSG (Li et al., 2025c), inherently provide strong spatial consistency. Here, the challenge shifts to representing dynamic changes and long-term object states. Methods like WorldMem (Xiao et al., 2025), geometry-grounded spatial memory (Wu et al., 2025) and surfel-indexed view memory (VMem) (Li et al., 2025a) leverage this explicit 3D structure to maintain a coherent world state over time, including dynamic representations that capture evolving geometry and supporting revisitations across long horizons. Beyond perceptual consistency, maintaining logical and factual coherence in reasoning remains crucial, addressed by techniques that learn to critique their own outputs (Asai et al., 2024).

The overarching lesson is that longer context alone is insufficient. Consistency emerges from explicit policies over memory: what to write, what to retrieve, how to update, and when to forget.

6.4 Summary

Stage IV reframes generation as stateful computation. Externalized memory makes knowledge editable. Architectural persistence makes it durable. Consistency policies make it reliable. At production scale, multimodal systems such as Gemini (Comanici et al., 2025) and Claude (Anthropic, 2024) extend these ideas, sustaining million-token contexts across text, audio, and video and coupling long horizons with reasoning for agentic workflows.

A deeper question remains. Are elaborate memory systems fundamental solutions, or are they sophisticated workarounds for the current constraints of hardware and data? The existence of models with massive, brute-force context windows suggests that some memory problems might simply dissolve with sufficient scale, much like how larger models unlocked emergent abilities. Similarly, consistency failures may also stem from limited data diversity or flaws in the data itself, such as contradictory or erroneous text and videos that are only a few seconds long. The answer will determine whether persistence in world models emerges as a natural property of scale, or as the product of carefully engineered memory discipline. When we ask if world models can dream consistently, the answer we seek is not just an engineering target, but a deeper understanding of the interplay between architecture, scale, and data.

7 Stage V: Towards True World Models

The preceding stages constructed the necessary components: a universal generative paradigm (I), a unified architecture (II), a real-time interactive loop (III), and a persistent memory system (IV). Stage V is not the addition of another component, but the synthesis of these parts into a cohesive, autonomous whole. A true world model is not merely a sophisticated simulator controlled by a user; it is a self-sustaining computational ecosystem. Its defining properties are not just programmed but emergent. We show that this synthesis gives rise to three defining properties: Persistence, Agency, and Emergence. More details can be found in Appendix B.

8 Conclusion: Building Living Worlds

This paper has charted a narrow road: a logical progression from the universal paradigm of masking to the threshold of a new reality. We have argued that this path is defined by the sequential mastery of three fundamental capabilities: unified generation, real-time interaction, and persistent memory. These are not ends in themselves, but the necessary foundations for worlds that can truly be called living worlds that persist with their own history, that are inhabited by goal-directed agents, and that give rise to unforeseen emergence.

The pursuit of isolated benchmarks for static tasks is a detour. The true frontier lies in embracing the architectural and theoretical commitments required to build these self-sustaining computational ecosystems. Therefore, the great choice ahead is whether we build worlds as mere tools for entertainment and escapism, or as scientific instruments for comprehending our complexity. The narrow road we have charted leads to this horizon: a future where we forge not just better models, but new mirrors in which to see ourselves.

REFERENCES

Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark Riedl. Bringing stories alive: Generating interactive fiction worlds. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 16, pp. 3–9, 2020.

Anthropic. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. *arXiv preprint arXiv:2410.08261*, 2024.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models, 2025. URL https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pp. 1692–1717. PMLR, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091, 2022.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. Textworld: A learning environment for text-based games. In *Workshop on Computer Games*, pp. 41–75. Springer, 2018.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Etched Decart, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. *URL: https://oasis-model. github. io*, 2024.

DeepMind. Gemini diffusion, 2025. URL https://deepmind.google/models/gemini-diffusion/. Accessed: 2025-08-19.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pp. 8469–8488. PMLR, 2023.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538 (7626):471–476, 2016.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*, 2025.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7903–7910, 2020.

Curtis Hawthorne, Andrew Jaegle, Cătălina Cangea, Sebastian Borgeaud, Charlie Nash, Mateusz Malinowski, Sander Dieleman, Oriol Vinyals, Matthew Botvinick, Ian Simon, et al. General-purpose, long-context autoregressive modeling with perceiver ar. In *International Conference on Machine Learning*, pp. 8535–8558. PMLR, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022a.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Cyrus Wu, Wei Li, Xuchen Song, Yang Liu, Eric Li, and Yahui Zhou. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*, 2022b.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 594–604, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Inception Labs, Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, et al. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*, 2025.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77, 2020.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pp. 6769–6781, 2020.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019.

KID. Memories off: Sorekara. Visual novel video game, first released for PlayStation, 2004. Publisher: KID.

Seung Wook Kim, Yuhao Zhou, Jonah Philion, Antonio Torralba, and Sanja Fidler. Learning to simulate dynamic environments with gamegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1231–1240, 2020.

Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Large memory layers with product keys. *Advances in Neural Information Processing Systems*, 32, 2019.

Latitude. Ai dungeon. https://play.aidungeon.com/, 2024.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:204960716.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. *arXiv preprint arXiv:2506.18903*, 2025a.

Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. A survey on diffusion language models. *arXiv preprint arXiv:2508.10875*, 2025b.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35: 4328–4343, 2022.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025c.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Chris Lu, Yannick Schroecker, Albert Gu, Emilio Parisotto, Jakob Foerster, Satinder Singh, and Feryal Behbahani. Structured state space models for in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36:47016–47031, 2023.

Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci. Playable video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10061–10070, 2021.

Willi Menapace, Stéphane Lathuilière, Aliaksandr Siarohin, Christian Theobalt, Sergey Tulyakov, Vladislav Golyanik, and Elisa Ricci. Playable environments: Video manipulation in space and time. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 3584–3593, 2022.

Willi Menapace, Aliaksandr Siarohin, Stéphane Lathuilière, Panos Achlioptas, Vladislav Golyanik, Sergey Tulyakov, and Elisa Ricci. Promptable game models: Text-guided game simulation via masked diffusion models. *ACM Transactions on Graphics*, 43(2):1–16, 2024.

Nick Montfort. Toward a theory of interactive fiction. *IF theory reader*, 25, 2011.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 101, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Anthony J Niesz and Norman N Holland. Interactive fiction. *Critical Inquiry*, 11(1):110–129, 1984.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*, 2024.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Yatian Pang, Eng Hock Francis Tay, Li Yuan, and Zhenghua Chen. Masked autoencoders for 3d point cloud self-supervised learning. *World Scientific Annual Review of Artificial Intelligence*, 1: 2440001, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model, 2024. URL https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. *arXiv preprint arXiv:2505.20171*, 2025.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pp. 28043–28078. PMLR, 2023.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167, 2024.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, et al. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 2025.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543, 2023.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14668–14678, 2022.

World Labs. World labs, 2024. URL https://www.worldlabs.ai/. Accessed: 2025-08-19.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *arXiv preprint arXiv:2203.08913*, 2022.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21469–21480, 2025.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9653–9663, 2022.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22963–22974, 2025.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025a.

Jiwen Yu, Yiran Qin, Haoxuan Che, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Hao Chen, and Xihui Liu. A survey of interactive generative video. *arXiv preprint arXiv:2504.21853*, 2025b.

Runpeng Yu, Qi Li, and Xinchao Wang. Discrete diffusion in large language and multimodal models: A survey. *arXiv preprint arXiv:2506.13759*, 2025c.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19313–19322, 2022.

Zihao Zeng, Bokai Lin, Tianqi Hou, Hao Zhang, and Zhijie Deng. In-context kv-cache eviction for llms via attention-gate. *arXiv preprint arXiv:2410.12876*, 2024.

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

APPENDIX

A A FORMALIZATION OF THE THREE SUBSYSTEMS

This appendix provides a detailed breakdown of the components formalized in Section 2. We consider a standard partially observable Markov decision process (POMDP) formulation where at each timestep t, an agent takes an action a_t, receives an observation o_t, and a reward r_t. The world terminates based on γ_t. The model maintains a latent belief state z_t and a deterministic memory state h_t.

The Generative Heart ($\mathcal{G}$). This subsystem models the world's underlying generative process and comprises three components:

- **Dynamics Model** $p_\theta(z_{t+1} \mid z_t, a_t)$: Predicts the next latent state given the current state and an action. This is the core of the model's ability to "dream" futures.
- **Observation Model** $p_\theta(o_t \mid z_t)$: Maps a latent state back to a sensory observation (*e.g.*, a video frame), grounding the latent space in perceptible reality.
- **Outcome Model** $p_\theta(r_t, \gamma_t \mid z_t, a_t)$: Predicts task-relevant outcomes like rewards and termination signals from the latent state.

The Interactive Loop ($\mathcal{F}, \mathcal{C}$). This subsystem enables a closed-loop exchange between an agent and the world model. It consists of:

- **Inference Model (Filter)** $q_\phi(z_t \mid h_{t-1}, o_t)$: Infers the current latent belief state z_t from the new observation o_t and past memory h_{t-1}.
- **Control Model (Policy & Value)** $\pi_\eta(a_t \mid z_t, h_t), v_\omega(z_t, h_t)$: The policy selects the next action based on the current belief and memory, while the value function estimates future outcomes, guiding the policy.

The Memory System ($\mathcal{M}$). This subsystem ensures long-horizon coherence. It has one core component:

- **Memory Update Model** $h_t = f_\psi(h_{t-1}, z_t, a_{t-1})$: Updates the deterministic memory state based on the previous memory, the inferred state, and the last action, creating a persistent representation of history.

This component-based formalization provides a unified lens through which to view the historical evolution of the field, from early control-oriented models that focused on specific components (*e.g.*, Ha & Schmidhuber (2018)) to modern generative systems that aim to integrate them all. It forms the analytical foundation for the five-stage roadmap presented in this paper.

B STAGE V: TOWARDS TRUE WORLD MODELS

B.1 THE THRESHOLD: PERSISTENCE, AGENCY, AND EMERGENCE

A true world model ceases to be a program one runs, but a world one enters. Its defining properties are:

- **Persistence**: The world's state and history exist independently of any single user session, accumulating consequence over time. It has a past that can be revisited and a future that unfolds continuously. This is the ultimate fulfillment of the Memory System ($\mathcal{M}$), transforming the property of Memory into an enduring reality.
- **Agency**: The world is inhabited by multiple, goal-directed agents (human or AI) that interact within a shared context. This property is enabled by the Interactive Loop ($\mathcal{F}, \mathcal{C}$), elevating the properties of Interaction and Adaptation into a multi-agent society.

- **Emergence**: The world's macro-level dynamics arise from the micro-level interactions of its agents and underlying rules, rather than being explicitly scripted. The model becomes a crucible for discovering unforeseen social structures, behaviors, and causal chains. This is the critical synthesis that occurs only when the Generative Heart ($\mathcal{G}$), Interactive Loop ($\mathcal{F}, \mathcal{C}$), and Memory System ($\mathcal{M}$) operate in unison over time.

B.2 The Frontier: Three Defining Challenges

The path to this threshold is defined by three fundamental, unsolved research problems. These are not merely technical hurdles, but grand challenges that constitute the frontier of the field.

The Coherence Problem (Evaluation). For conventional models, fidelity is measured against external ground truth. A true world model, however, writes its own history. The challenge is to evaluate the "truth" of a self-generating reality: to formalize and measure its internal logical, causal, and narrative coherence, and to define what it means for such a world to be consistent.

The Compression Problem (Scaling). An ever-growing history risks computational collapse. The challenge is to learn causally sufficient state abstractions that preserve consequence while discarding noise, approaching the information-theoretic bounds of predictive representation. Yet even with abstraction, long-horizon dynamics may be computationally irreducible, forcing us to treat world models not only as engineered systems but as objects of scientific observation.

The Alignment Problem (Safety). An autonomous, persistent world model is a technology with profound societal implications. The alignment challenge for a true world model operates on two distinct levels. At its base, the model itself can be viewed as a single environment whose generating process must align with human values. However, the complexity arises when this model becomes the substrate for a multi-agent society. The alignment problem then becomes squared: it requires aligning not only the world's underlying laws (the substrate), but also the emergent, unpredictable dynamics of the agents interacting within it. This is the harder challenge, distinguishing a true world model from a mere single environment simulator.

B.3 The Horizon: From Simulator to Scientific Instrument

The journey detailed in this paper, from masks to worlds, has been about forging a new kind of technology. Yet, the ultimate promise of a true world model lies beyond its function as a simulator for entertainment or training.

Once a world model crosses the threshold of persistence, agency, and emergence, it transforms from a technological artifact into a new kind of scientific instrument. It becomes a computational crucible for running experiments on complex adaptive systems such as economies, cultures, and cognitive ecosystems that are impossible to conduct in reality.

The quest for true world models, therefore, is not merely an engineering endeavor. It is a pursuit of the ultimate tool for understanding complexity itself. The narrow road leads here: to a future where we build worlds not to escape our own, but to comprehend it.

C The Use of Large Language Models

During the preparation of this paper, large language models were used only for language polishing and minor editing. All research ideas, methods, and experimental results were carried out entirely by the human authors.