World Models for Autonomous Driving: An Initial Survey

Yanchen Guan*, Haicheng Liao*, Zhenning Li†, Jia Hu†, Runze Yuan, Guohui Zhang,
and Chengzhong Xu, *Fellow, IEEE*

Abstract—In the rapidly evolving landscape of autonomous driving, the capability to accurately predict future events and assess their implications is paramount for both safety and efficiency, critically aiding the decision-making process. World models have emerged as a transformative approach, enabling autonomous driving systems to synthesize and interpret vast amounts of sensor data, thereby predicting potential future scenarios and compensating for information gaps. This paper provides an initial review of the current state and prospective advancements of world models in autonomous driving, spanning their theoretical underpinnings, practical applications, and ongoing research efforts aimed at overcoming existing limitations. Highlighting the significant role of world models in advancing autonomous driving technologies, this survey aspires to serve as a foundational reference for the research community, facilitating swift access to and comprehension of this burgeoning field, and inspiring continued innovation and exploration.

Index Terms—World model, Autonomous driving, Foundational model, Model-based reinforcement learning.

I. INTRODUCTION

The quest to develop autonomous driving systems that seamlessly navigate the intricate tapestry of real-world scenarios remains a formidable frontier in contemporary technology. This challenge is not merely technical but also philosophical, probing the essence of cognition and perception that distinguishes human intelligence from artificial constructs. The crux of this challenge lies in imbuing machines with the kind of intuitive reasoning and 'common sense' that humans employ effortlessly. Current machine learning systems, despite their prowess, often stumble in pattern recognition tasks that humans resolve with ease, highlighting a significant gap in our quest for truly autonomous systems [1]. On the other hand, human decision-making is deeply rooted in sensory perceptions, constrained by both the memories of these perceptions and direct observations [2], [3]. Beyond mere perception, humans possess the uncanny ability to predict the outcomes of their actions, envision potential futures, and anticipate changes in

* Authors contributed equally; † Corresponding author.

Y. Guan, H. Liao, Z. Li, and C. Xu are with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau SAR, 999078, China. E-mail: zhenningli@um.edu.mo

J. Hu is with the College of Transportation Engineering, Tongji University, Shanghai, 200092, China. E-mail: hujia@tongji.edu.cn

R. Yuan is with the Department of Automation, Tsinghua University, Beijing, 100084, China.

G. Zhang is with the Department of Civil, Environmental and Construction Engineering, University of Hawaii at Manoa, Honolulu, HI 96848, United States.

sensory inputs—abilities that underpin our interaction with the world [4]. The endeavor to replicate such capabilities in machines is not just an engineering challenge but a step towards bridging the cognitive divide between human and machine intelligence.

Addressing this gap, world models have emerged as a critical solution, offering systems the ability to predict and adapt to dynamic environments by emulating human-like perception and decision-making processes. This evolution is essential in the face of real-world scenarios' complexity and unpredictability, where traditional AI approaches struggle to replicate the depth and variability of human cognitive processes. The necessity for world models is underscored by their potential to bridge the cognitive divide between human and machine intelligence, providing a pathway toward more sophisticated autonomous driving systems.

The journey of world models from conceptual frameworks in control theory during the 1970s to their current prominence in artificial intelligence research reflects a remarkable trajectory of technological evolution and interdisciplinary fusion. The initial formulations in control theory, as laid out by pioneers [5]–[7], were foundational, setting the stage for the integration of computational models in dynamic system management. The underlying logic of Model Predictive Control (MPC) in control theory aligns closely with that of world models. Furthermore, the proposal of the mental model systematically described how humans understand the surrounding world and the relationship between abstract concepts, providing theoretical support for world models to imitate humans' understanding of the relationship between the world and themselves [8]–[11]. These early efforts demonstrated the potential of applying mathematical models to predict and control complex systems, providing guidance for modeling human cognitive systems, which ultimately laid the foundation for the development of world models.

As the field progressed, the advent of neural networks introduced a paradigm shift, allowing for the modeling of dynamic systems with unparalleled depth and complexity [12]–[14]. This transition from static, linear models to dynamic, non-linear representations facilitated a deeper understanding of environmental interactions, laying the groundwork for the sophisticated world models we see today. The integration of recurrent neural networks (RNNs) was particularly transformative, marking a leap towards systems capable of temporal data processing, essential for predicting future states and enabling abstract reasoning [15], [16].

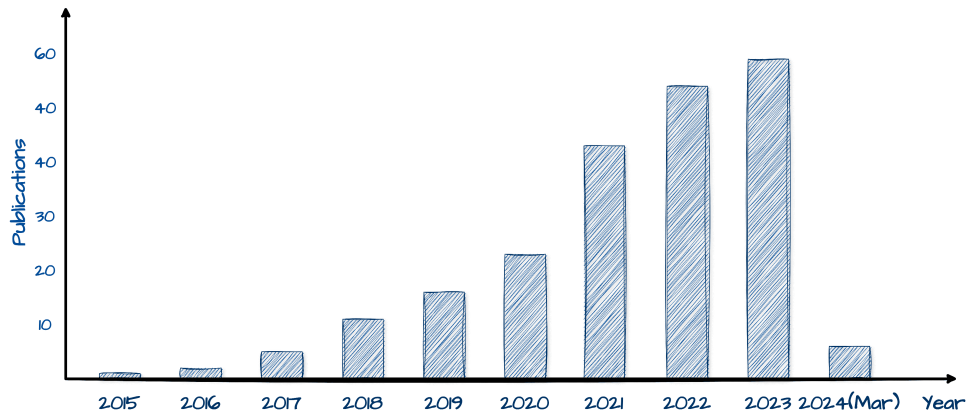In the early 1990s, using neural networks to learn models

Fig. 1: Number of Publications related to World Models since 2015. (Data sources: Web of Science Core Collection and Preprint Citation Index. Key words: "world model", "world models", "reinforcement learning".)

for control and applying predictive models in reinforcement learning for both learning and planning ignited pioneers' interest in the construction of learnable models, particularly in the contemplation of models' interactions with the world [17]–[20]. The Dyna architecture elucidated the interaction patterns between agents and the world, as well as the processes of reward and state updates in reinforcement learning [18]. Research using RNNs to develop internal models for inferring future problems offered a unified framework for the transition from learning to thinking within models [15], [21]–[25]. Following this, with a long exploration of learning dynamic models and employing these models to train policies, learnable models were back in the center of attention again [26], [27].

In Judea Pearl's hierarchy of causation, the base level is SEEING, which involves identifying associations between elements, a stage where most predictive models currently operate; The middle layer, DOING, refers to actions and interventions, exemplified by the exploration of reinforcement learning; The highest level, IMAGINING, focuses on counterfactual reasoning, understanding the causal relationships between things, enabling the transcendence from complete dependence on datasets [28]. This capability allows for dealing with situations beyond the training data through imagination. For learnable models, achieving counterfactual reasoning represents an arduous goal. While this ability is innate to humans, crossing this threshold poses a significant challenge for AI. Fortunately, the emergence of world models has introduced a turning point in this dilemma.

The formal unveiling of world models by Ha and Schmidhuber in 2018 was a defining moment that captured the collective aspiration of the AI research community to endow machines with a level of cognitive processing reminiscent of human consciousness [29], [30]. By harnessing the power of Mixture Density Networks (MDN) and RNNs, this work illuminated the path for unsupervised learning to extract and interpret the spatial and temporal patterns inherent in environmental data. The significance of this breakthrough cannot be overstated, it demonstrated that autonomous systems could achieve a nuanced understanding of their operational environments, predicting future scenarios with an accuracy that was previously unattainable.

In the realm of autonomous driving, the introduction of world models signifies a pivotal shift towards data-driven intelligence, where the capacity to predict and simulate future scenarios becomes a cornerstone for safety and efficiency. The challenge of data scarcity, particularly in specialized tasks such as BEV labeling, underscores the practical necessity of innovative solutions like world models [29], [31], [32]. By generating predictive scenarios from historical data, these models not only circumvent the limitations posed by data collection and labeling but also enhance the training of autonomous systems in simulated environments that can mirror, or even surpass, the complexity of real-world conditions. This approach heralds a new era where autonomous vehicles are equipped with predictive capabilities that reflect a form of intuition, enabling them to navigate and respond to their environment with an unprecedented level of sophistication.

This paper delves into the intricate tapestry of world models, exploring their foundational principles, methodological advancements, and practical applications within the realm of autonomous driving. It navigates through the challenges that beset this field, forecasting future research trajectories and contemplating the broader implications of integrating world models into autonomous systems. In doing so, this work aspires to not only chronicle the progress in this domain but also to inspire a deeper contemplation of the symbiosis between artificial intelligence and human cognition, heralding a new era in autonomous driving technology.

II. DEVELOPMENT OF WORLD MODELS

This section outlines the intricate architecture of world models, detailing their key components and significant applications across various research studies. These models are engineered to replicate the complex cognitive processes of the human brain, enabling autonomous systems to make decisions and understand their environment in a manner akin to human thinking.

A. Architectural Foundations of World Models

The architecture of a world model is designed to mimic the coherent thinking and decision-making processes of the human
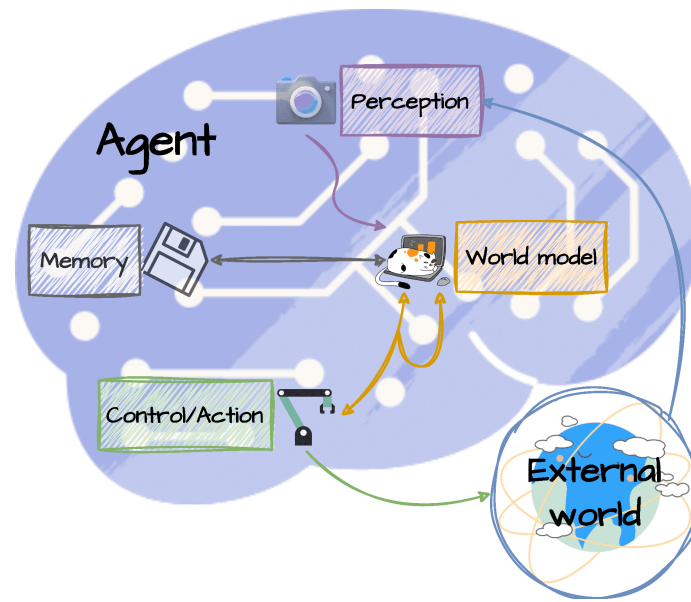
Fig. 2: Diagram of an Agent's World Model Framework.

brain, integrating several critical components:

a) Perception Module: This foundational element acts as the system's sensory input, analogous to human senses. It employs advanced sensors and encoder modules, such as Variational Autoencoder (VAE) [33]–[35], Masked Autoencoder (MAE) [36], [37], and Discrete Autoencoder (DAE) [31], [38], to process and compress environmental inputs (images, videos, text, control commands) into a more manageable format. The effectiveness of this module is crucial for the accurate perception of complex, dynamic environments, facilitating a detailed understanding that informs the model's subsequent predictions and decisions.

b) Memory Module: Serving a role similar to the human hippocampus, the memory module is pivotal for recording and managing past, present, and predicted world states along with their associated costs or rewards [1]. It supports both short-term and long-term memory functions by replaying recent experiences, a process that enhances learning and adaptation by incorporating past insights into future decisions [39]. This module's ability to synthesize and retain crucial information is essential for developing a nuanced understanding of environmental dynamics over time.

c) Control/Action Module: This component is directly responsible for interacting with the environment through actions. It evaluates the current state and predictions provided by the world model to determine the optimal sequence of actions aimed at achieving specific goals, such as minimizing costs or maximizing rewards. The sophistication of this module lies in its ability to integrate sensory data, memory, and predictive insights to make informed, strategic decisions that navigate the complexities of real-world scenarios. This module distinguishes the decision-making process from the intricate world model module and trains it independently using a minimal parameter set. This design allows for the application of more unconventional training methods, such as evolution strategies [40], to address challenging reinforcement learning tasks where credit assignment poses significant difficulties.

d) World Model Module: As the heart of the architecture, the world model module performs two primary functions: estimating any missing information about the current world state and predicting future states of the environment. This dual capability allows the system to generate a comprehensive, predictive model of its surroundings, accounting for uncertainties and dynamic changes. By simulating potential future scenarios, this module enables the system to prepare and adjust its strategies proactively, mirroring the predictive and adaptive thought processes found in human cognition.

Together, these components form a robust framework that empowers world models to simulate cognitive processes and decision-making akin to humans. By integrating these modules, world models achieve a comprehensive and predictive understanding of their environment, which is crucial for the development of autonomous systems capable of navigating and interacting with the complexity of the real world with unprecedented sophistication.

In high-dimensional sensory input scenarios, world models utilize latent dynamical models to abstractly represent observed information, enabling compact forward predictions within a latent state space. These latent states, being significantly more space-efficient than direct predictions of high-dimensional data, facilitate the execution of numerous parallel predictions [41], thanks to advancements in deep learning and latent variable models. Take, for instance, the ambiguity of a car's direction at an intersection, a scenario emblematic of the inherent unpredictability of real-world dynamics. Latent variables serve as a powerful tool in representing these uncertain outcomes, setting the stage for world models to envision a spectrum of future possibilities grounded in the present state. The crux of this endeavor lies in harmonizing the deterministic aspects of prediction with the intrinsic uncertainty of real-world phenomena, a balancing act central to the efficacy of world models.
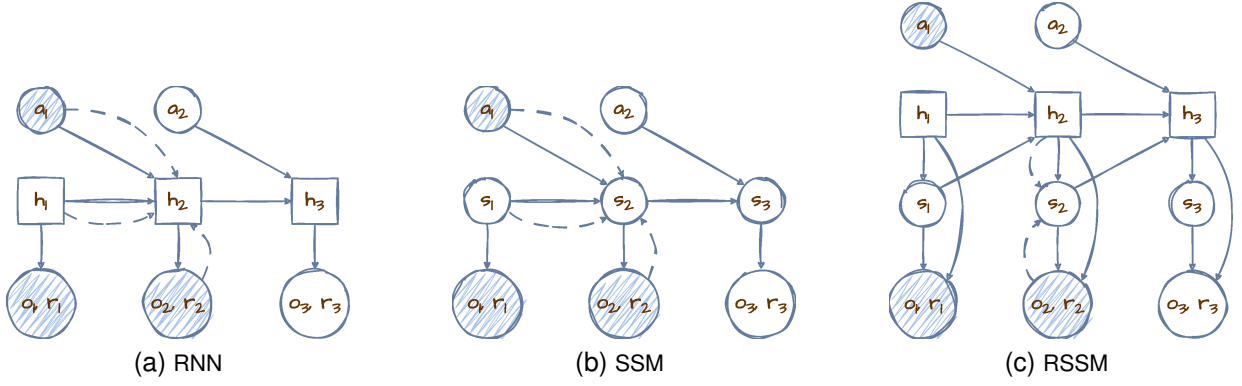
Fig. 3: Comparative Schematic of RNN, SSM, and RSSM Architectures in Latent Dynamics Modeling.

To navigate this challenge, a variety of strategies have been proposed, from introducing uncertainty via a temperature variable [29] to adopting structured frameworks like the Recurrent State Space Model (RSSM) [41]–[45] and the Joint Embedding Predictive Architecture (JEPA) [1], [36], [46], [47]. These methodologies strive to fine-tune the balance between precision and flexibility in prediction. Moreover, leveraging Top-k sampling and transitioning from CNN-based models to transformer architectures [31], [48], [49], such as the Transformer State Space Model (TSSM) or Spatial Temporal Patchwise Transformer (STPT), have shown promise in enhancing model performance by better approximating the complexity and uncertainty of the real world. These solutions strive to align the outputs of world models more closely with the probable developments of the real world. This alignment is crucial because the real world, compared to gaming environments, has a much wider range of influencing factors and a greater degree of randomness in future outcomes. Over-reliance on the highest probability predictions may lead to repetitive cycles in long-term forecasting. Conversely, excessive randomness in predictions can lead to absurd futures that significantly diverge from reality.

In particular, the core structures most commonly employed in world model research are RSSM and JEPA:

1) Recurrent State Space Model (RSSM) [42] stands as a pivotal model within the Dreamer series of world models, designed to facilitate forward prediction purely within the latent space. This innovative structure enables the model to predict through the latent state space, where both stochastic and deterministic paths within the transition model play critical roles in successful planning.

Fig. 3 illustrates a schematic of the latent dynamics models across three time steps. Initially observing two time steps, these models then forecast the third. Here, stochastic variables (circles) and deterministic variables (squares) interact within the models' architecture—solid lines illustrate the generative processes, while dashed lines represent inference pathways. The initial deterministic inference approach in Fig. 3a reveals its limitation in capturing diverse potential futures due to its fixed nature. Conversely, an entirely stochastic approach in Fig. 3b presents challenges in information retention across time steps, given its inherent unpredictability.

RSSM's innovation lies in its strategic decomposition of states into stochastic and deterministic components in Fig. 3c, effectively harnessing the predictive stability of deterministic elements and the adaptive potential of stochastic elements. This hybrid structure ensures robust learning and prediction capabilities, accommodating the unpredictability of real-world dynamics while preserving information continuity. By marrying the strengths of RNNs with the flexibility of State Space Models (SSM) [50], RSSM establishes a comprehensive framework for world models, enhancing their ability to predict future states with a balance of precision and adaptability.

We denote the sequence of observations and actions as $(\mathbf{x}_0, \boldsymbol{a}_1, \mathbf{x}_1, \boldsymbol{a}_2, \mathbf{x}_2, \ldots, \boldsymbol{a}_T, \mathbf{x}_T)$. Namely, the agent takes action $\boldsymbol{a}_{t+1}$ after observing $\mathbf{x}_t$, and receives the next observation $\mathbf{x}_{t+1}$. We omit the reward for simplicity. RSSM models the observations and state transitions through the following generative process:

$$p\left(\mathbf{x}_{0:T} \mid \boldsymbol{a}_{1:T}\right) = \int \prod_{t=0}^{T} p\left(\mathbf{x}_t \mid \mathbf{z}_{\leq t}, \boldsymbol{a}_{\leq t}\right) p\left(\mathbf{z}_t \mid \mathbf{z}_{<t}, \boldsymbol{a}_{\leq t}\right) \mathrm{d}\mathbf{z}_{0:T}$$
(1)

where $\mathbf{z}_{0:T}$ are the stochastic latent states. The approximate posterior is defined as:

$$q\left(\mathbf{z}_{0:T} \mid \mathbf{x}_{0:T}, \boldsymbol{a}_{1:T}\right) = \prod_{t=0}^{T} q\left(\mathbf{z}_t \mid \mathbf{z}_{<t}, \boldsymbol{a}_{\leq t}, \mathbf{x}_t\right)$$
(2)

The conditioning on previous states $\mathbf{z}_{<t}$ and actions $\boldsymbol{a}_{\leq t}$ appears multiple times. RSSM uses a shared GRU to compress $\mathbf{z}_{<t}$ and $\boldsymbol{a}_{\leq t}$ into a deterministic encoding $\boldsymbol{h}_t$:

$$\boldsymbol{h}_t = \mathrm{GRU}\left(\boldsymbol{h}_{t-1}, \mathrm{MLP}\left(\mathrm{concat}\left[\mathbf{z}_{t-1}, \boldsymbol{a}_t\right]\right)\right)$$
(3)

This is then used to compute the sufficient statistics of the prior, likelihood, and posterior:

$$\begin{aligned}
p\left(\mathbf{z}_t \mid \mathbf{z}_{<t}, \boldsymbol{a}_{\leq t}\right) &= \mathrm{MLP}\left(\boldsymbol{h}_t\right), \\
p\left(\mathbf{x}_t \mid \mathbf{z}_{\leq t}, \boldsymbol{a}_{\leq t}\right) &= \mathcal{N}\left(\hat{\mathbf{x}}_t, \mathbf{1}\right), \\
\hat{\mathbf{x}}_t &= \mathrm{Decoder}\left(\mathrm{concat}\left[\boldsymbol{h}_t, \mathbf{z}_t\right]\right), \\
q\left(\mathbf{z}_t \mid \mathbf{z}_{<t}, \boldsymbol{a}_{\leq t}, \mathbf{x}_t\right) &= \mathrm{MLP}\left(\mathrm{concat}\left[\boldsymbol{h}_t, \boldsymbol{e}_t\right]\right), \\
\boldsymbol{e}_t &= \mathrm{Encoder}\left(\mathbf{x}_t\right).
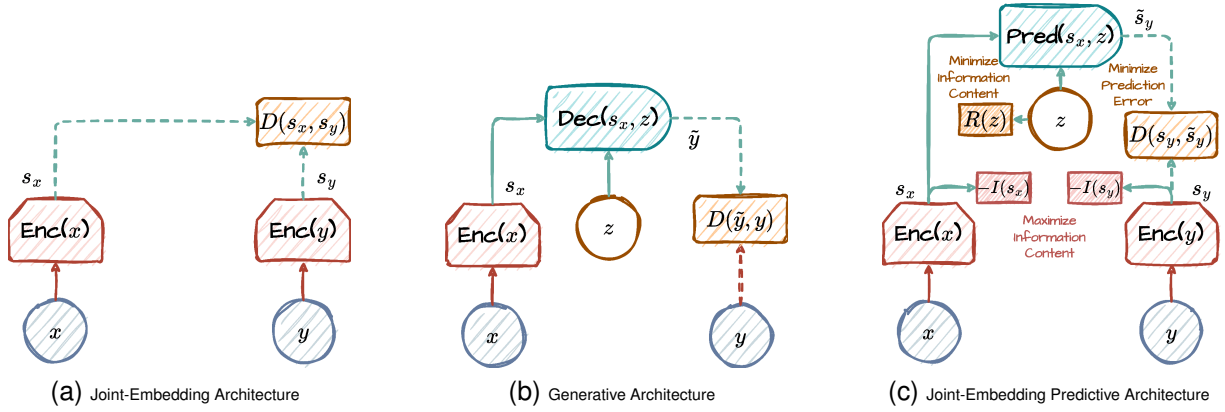\end{aligned}$$
(4)

Fig. 4: Comparative Schematic of Joint-Embedding Architecture, Generative Architecture, and Joint-Embedding Predictive Architecture.

The training objective is to maximize the evidence lower bound (ELBO):

$$\log p\left(\mathbf{x}_{0:T} \mid \boldsymbol{a}_{1:T}\right) \geq \mathbb{E}_q\left[\sum_{t=0}^{T} \log p\left(\mathbf{x}_t \mid \mathbf{z}_{\leq t}, \boldsymbol{a}_{\leq t}\right) \right.$$
$$\left. - \mathcal{L}_{\text{KL}}\left(q\left(\mathbf{z}_t \mid \mathbf{z}_{<t}, \boldsymbol{a}_{\leq t}, \mathbf{x}_t\right), p\left(\mathbf{z}_t \mid \mathbf{z}_{<t}, \boldsymbol{a}_{\leq t}\right)\right)\right] \quad (5)$$

2) Joint-Embedding Predictive Architecture (JEPA) [1] marks a paradigm shift in predictive modeling by focusing on representation space rather than direct, detailed predictions. As shown in Fig. 4, by abstracting input ($\mathbf{x}$) and target ($\mathbf{y}$) through dual encoders into representations ($\mathbf{s}_x$ and $\mathbf{s}_y$), and leveraging a latent variable ($\mathbf{z}$) for prediction, JEPA achieves a significant leap in efficiency and accuracy. This model excels in filtering out noise and irrelevancies, concentrating on the essence of the predictive task. The strategic use of the latent variable ($\mathbf{z}$) to manage uncertainties further refines the model's focus, enabling it to predict abstract outcomes with heightened precision. By prioritizing relevant features and embracing the inherent uncertainties of predictive tasks, JEPA not only streamlines the prediction process but also ensures outcomes are both relevant and reliable, setting a new standard for predictive modeling in complex environments.

JEPA is underpinned by a multifaceted mathematical model that synthesizes concepts from statistics, machine learning, and optimization. At the heart of JEPA lies the energy function $\mathbf{E}_w\left(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta\right)$, which captures the prediction error within the model. Here, $\mathbf{x}$ and $\mathbf{y}$ represent the input and target data, respectively, while $\mathbf{z}$ stands for the latent variables, and θ denotes the model parameters. Mathematically, the energy function $\mathbf{E}_w$ is defined as:

$$\mathbf{E}_w\left(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta\right) = \|\mathbf{s}_y - \text{Pred}\left(\mathbf{s}_x, \mathbf{z} : \phi\right)\|_2^2 + \lambda\|\mathbf{z}\|_2^2 \quad (6)$$

Therefore, the core of JEPA's predictive capability is encapsulated in the following key expressions: 1) The squared L2 norm $\|\mathbf{s}_y - \text{Pred}\left(\mathbf{s}_x, \mathbf{z} : \phi\right)\|_2^2$, measures the Euclidean distance between the target representation $\mathbf{s}_y$ and the predicted representation, highlighting the model's prediction error. The predictive function Pred maps the input representation $\mathbf{s}_x$ and

latent variables $\mathbf{z}$ to the target space, parameterized by ϕ; and 2) The regularization term $\lambda\|\mathbf{z}\|_2^2$ penalizes the model's complexity to prevent overfitting by imposing a cost on the magnitude of the latent variables $\mathbf{z}$, with λ being a nonnegative scalar controlling the regularization strength.

Following this, the optimization process aims to minimize $\mathbf{E}_w$ through finding θ, ϕ, and $\mathbf{z}$. This can be expressed as a complex Lagrangian optimization problem with constraints on the data distribution:

$$\text{L}\left(\theta, \phi, \mathbf{z}; \mathbf{x}, \mathbf{y}, \alpha\right) = \mathbf{E}_w\left(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta\right) - \alpha\left(h\left(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta, \phi\right) - \mathbf{c}\right) \quad (7)$$

where $\text{L}\left(\theta, \phi, \mathbf{z}; \mathbf{x}, \mathbf{y}, \alpha\right)$ represents the Lagrange equation commonly used in constrained optimization problems, α is a Lagrange multiplier enforcing the constraint $h\left(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta, \phi\right) = \mathbf{c}$, $h\left(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta, \phi\right)$ represents the constraint function of the optimization problem parameterized by θ and ϕ, and $\mathbf{c}$ is a constant as the target value of function h.

The training of JEPA involves higher-order optimization methods, considering the second-order derivatives to ensure convergence in complex landscapes:

$$\theta_{\mathbf{t}+1} = \theta_{\mathbf{t}} - \eta\nabla_\theta^2 \text{L}\left(\theta_{\mathbf{t}}, \phi_{\mathbf{t}}, \mathbf{z}_{\mathbf{t}}; \mathbf{x}, \mathbf{y}, \alpha_{\mathbf{t}}\right) \quad (8)$$

where, $\theta_{\mathbf{t}+1}$ represents the updated parameter vector at time $(\mathbf{t}+1)$. η is the learning rate that determines the step size of the update. $\eta\nabla_\theta^2 \text{L}\left(\theta_{\mathbf{t}}, \phi_{\mathbf{t}}, \mathbf{z}_{\mathbf{t}}; \mathbf{x}, \mathbf{y}, \alpha_{\mathbf{t}}\right)$ represents the Hessian matrix of second-order partial derivatives of the Lagrangian L with respect to the parameters θ at time $\mathbf{t}$.

To accommodate the complexity of the optimization landscape, JEPA may utilize higher-order optimization methods that consider second-order derivatives. Furthermore, given the high-dimensional nature of $\mathbf{z}$ and the possibility of multimodal distributions, JEPA might employ a variational approximation to the intractable posterior $p\left(\mathbf{z} \mid \mathbf{x}, \mathbf{y}; \theta\right)$, leading to a variational lower bound:

$$\log p\left(\mathbf{y} \mid \mathbf{x}; \theta, \phi\right) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x};\psi)}\left[\log p\left(\mathbf{y} \mid \mathbf{x}, z; \theta, \phi\right)\right] - \text{KL}\left[q\left(\mathbf{z} \mid \mathbf{x}; \psi\right)\|p\left(\mathbf{z} \mid \mathbf{x}; \theta\right)\right] \quad (9)$$

where $\log p\left(\mathbf{y} \mid \mathbf{x}; \theta, \phi\right)$ is the log-likelihood of the data $\mathbf{y}$ given $\mathbf{x}$ under the model parameters θ and ϕ, $\mathbb{E}_{q(\mathbf{z}|\mathbf{x};\psi)}[\cdot]$

is expectation with respect to the distribution $q\left(\mathbf{z} \mid \mathbf{x}; \psi\right)$, $\mathrm{KL}\left[q\left(\mathbf{z} \mid \mathbf{x}; \psi\right) \| p\left(\mathbf{z} \mid \mathbf{x}; \theta\right)\right]$ is the Kullback-Leibler divergence between the variational distribution $q\left(\mathbf{z} \mid \mathbf{x}; \psi\right)$ and the prior distribution $p\left(\mathbf{z} \mid \mathbf{x}; \theta\right)$.

This inequality, Equation (9), is further used to maximize ELBO to approximate the true posterior distribution.

B. Broad Spectrum Applications

As shown in TABLE I, with the evolution of world models, they have been applied across multiple domains, showcasing unparalleled performance in diverse environments, particularly in gaming, where their capabilities are prominently displayed. In the competitive landscape of the Atari 100k leaderboard, world models dominate, with four out of the top five positions held by these innovative architectures [31], [32], [49], [53], [68]. Among these, EfficientZero stands out by significantly enhancing sampling efficiency in image-based reinforcement learning. By leveraging the foundational principles of MuZero and incorporating lightweight improvements, EfficientZero achieved human-comparable gaming proficiency within a mere two hours of training, securing the top position on the leaderboard [52]. In the Minecraft game, DreamerV3 marks a milestone as the inaugural model to autonomously mine diamonds, a feat accomplished without leveraging human-generated data or predefined learning curricula. This achievement is attributed to its novel use of Symlog Predictions, facilitating the model's adaptability across diverse environmental scales by employing static symlog transformations [44], [80]. Conversely, HarmonyDream introduces a dynamic approach to loss scaling within world model learning, optimizing multi-task learning efficiency through an intricate balance of scale, dimensionality, and training dynamics [68]. The synergistic integration of DreamerV3's symlog transformation with HarmonyDream's dynamic loss adjustment holds the potential to further elevate world models' performance and versatility.

The Image-based Joint-Embedding Predictive Architecture (I-JEPA) [47] illustrates an approach for learning highly semantic image representations without relying on hand-crafted data augmentations. I-JEPA predicts missing target information using abstract representations, effectively eliminating unnecessary pixel-level details. This enables the model to learn more semantic features, achieving more accurate analysis and completion of incomplete images through self-supervised learning of the world's abstract representations. Beyond images, this architecture demonstrates high scalability with the Audio-based Joint-Embedding Predictive Architecture (A-JEPA) [46], setting new state-of-the-art performance on multiple audio and speech classification tasks, outperforming models that rely on externally supervised pre-training.

In robotic manipulation, such as Fetch [81], DeepMind Control Suite [82], and Meta-world [83], the Latent Explorer Achiever (LEXA) [45] outperforms previous unsupervised methods in 40 robot operation and movement tasks by training an explorer and an achiever simultaneously through imagination. Additionally, in these tasks, $\mathrm{L}^3\mathrm{P}$ [84] devises a novel algorithm to learn latent landmarks scattered across the goal space, achieving dominant performance in both

learning speed and test-time generalization in three robotic manipulation environments. Google's team has innovatively applied the concept of world models to robot navigation tasks, utilizing them to acquire information about the surrounding environment and enable the intelligent agent to predict the consequences of its actions in specific contexts. Pathdreamer's implementation in robot navigation leverages world models for enhanced environmental awareness and predictive planning, achieving a notable improvement in navigation success rates through its innovative use of 3D point clouds for environment representation [54]. Furthermore, SafeDreamer integrates a Lagrangian-based approach into the Dreamer framework for safety reinforcement learning, demonstrating the feasibility of high-performance, low-cost safety applications [67].

The rapid training capabilities of world models, exemplified by DayDreamer's real-world robot learning efficiency, contrast starkly with traditional methods, highlighting the transformative potential of these models in accelerating learning processes and enhancing performance [63], [85].

Virtual scene and video generation emerge as pivotal applications, with SORA and Genie leading advancements in this space. SORA's capability to produce coherent, high-definition videos from diverse prompts represents a significant step towards simulating complex world dynamics. Despite its challenges in physical interaction simulation, SORA's consistent 3D spatial representation underscores its potential as a foundational world model [86]. Genie's interactive environment generation, though not as advanced in video quality as SORA, introduces a novel dimension of user-driven world manipulation, offering a glimpse into future applications of world models in creating immersive, controllable virtual realities [77].

In summary, this comprehensive review in this section underscores the exceptional versatility and advancing frontiers of world models, illustrating their foundational role in driving innovation across gaming, robotics, virtual environment generation, and beyond. The convergence of these models' capabilities with dynamic adaptation and multi-domain generalization heralds a new era of AI, where world models not only serve as tools for specific tasks but also as platforms for broader exploration, learning, and discovery.

III. WORLD MODELS IN AUTONOMOUS DRIVING

This section delves into the transformative application of world models within the autonomous driving sector, underscoring their pivotal contributions to environmental comprehension, dynamic prediction, and the elucidation of physical principles governing motion. As an emergent frontier in the application of world models, the autonomous driving domain presents unique challenges and opportunities for leveraging these advanced computational frameworks. Despite the burgeoning interest, the integration of world models into autonomous driving predominantly revolves around scenario generation, and planning and control mechanisms, areas ripe for exploration and innovation.

TABLE I: Summary of Recent World Model Applications.

| Year	World Model	Core Structure	Contribution/Key Idea
2018	World models [30]	MDN, RNN	Pioneered the first world model for unsupervised learning of spatial and temporal representation of environments.
2019	PlaNet [42]	RSSM	Introduced a latent dynamics model featuring both deterministic and stochastic transition components for latent prediction.
	DreamerV1 [41]	RSSM	Efficient policy learning through the propagation of analytic value gradients for optimized decision-making processes.
2020	DreamerV2 [43]	RSSM	Utilized discrete multi-class vectors instead of Gaussian distribution for more precise state representation predictions.
	AWML [51]	γ-Progress	Developed a world model that employs curiosity-driven exploration.
	MuZero [52]	MCTS	Integrated a tree-based search with a learned model for improved strategic planning and decision-making.
2021	EfficientZero [53]	MCTS	Crafted a sample efficient model-based visual RL algorithm built on MuZero.
	Pathdreamer [54]	RedNet, PatchGAN	Engineered a world model for Indoor Vision-and-Language Navigation.
2022	MILE [55]	Probabilistic Generative Model	Adopted a model-based imitation learning approach for joint policy learning in autonomous driving.
	SEM2 [56]	RSSM	Implemented filters to extract key features and balance data distribution, optimizing model performance and accuracy.
	TRANSDREAMER [48]	Transformer State-Space Model	Introduced the first transformer-based stochastic world model.
	IRIS [31]	DAE, Autoregressive Transformer	Constructed a world model combining a discrete autoencoder with an autoregressive transformer.
	DreamingV2 [57]	RSSM	Introduced a novel approach by integrating both the discrete representation and the reconstruction-free objective.
	DreamerPro [58]	RSSM, Convolutional Encoder	Enhanced resistance to visual interference, building upon the Dreamer model's foundation.
	MaxEnt Dreamer [59]	RSSM, Maximum Entropy	Implemented maximum entropy reinforcement learning to significantly enhance the world model's exploration capabilities.
2023	DriveDreamer [60]	ActionFormer, Diffusion	Enabled controllable generation of driving scene videos equipped with simplified interactive features.
	Adriver-I [61]	MLLM, VDM	Developed a world model that leverages image-text pairs, combined with large language models.
	GAIA-1 [62]	Autoregressive Transformer	Crafted a system to generate realistic driving scenes using video, text, and action inputs, providing fine-grained control over the generated environment.
	I-JEPA [47]	JEPA	Implemented a non-generative approach for self-supervised learning of images.
	A-JEPA [46]	JEPA	Extended the I-JEPA self-supervised learning approach to the audio spectrum.
	MC-JEPA [36]	JEPA	Employed self-supervised optical flow estimation from videos to jointly learn motion features.
	DreamerV3 [44]	RSSM	Adopted symlog predictions to seamlessly adapt across all ranges using a tailored set of hyperparameters.
	Daydreamer [63]	RSSM	Utilized robots to deploy the Dreamer model in real-world settings, enabling direct learning without simulators.
	Drive-WM [64]	VAE, Diffusion	Introduced the first multi-view, end-to-end autonomous driving world model.
	TrafficBots [65]	CVAE, Transformer	Developed a world model for multimodal motion prediction and end-to-end driving.
	Uniworld [66]	Transformer	Crafted a world model utilizing multi-view, label-free image-LiDAR pairs to model the surrounding scene.
	Safe DreamerV3 [67]	RSSM	Introduced a world model specifically designed for safe reinforcement learning.
	HarmonyDream [68]	RSSM	Incorporated a comprehensive approach by considering the loss of per-dimension scales, dimensions, and training dynamics together.
	Dr.G [69]	RSSM, DCL, RSID	Employed a zero-shot MBRL framework that resists visual interference.
	STORM [49]	VAE, Transformer	Combined transformers with variational autoencoders to introduce random noise into model-based reinforcement learning.
	TWM [70]	Transformer-XL	Utilized the Transformer-XL architecture to learn long-term dependencies while staying computationally efficient.
	DiffDreamer [71]	Conditional Diffusion Model	Developed a world model for scene extrapolation.
	MUVO [72]	MLP, GRU	Utilized world model to predict 3D occupancy and camera and lidar observations.
	SWIM [73]	RSSM	Trained a world model using videos of human hands to enable robots to interact with the environment.
	OccWorld [74]	S-T Generative Transformer	Generalized the world model to 3D Occupancy space.
2024	WorldDreamer [75]	S-T Patchwise Transformer	Facilitated learning of dynamic visual signal using attention.
	S4WM [76]	Parallelizable SSMs	Modified world model backbones for improving long-term memory.
	Genie [77]	ST-Transformer	Trained in an unsupervised manner on unlabeled Internet videos to generate interactive environments.
	V-JEPA [78]	JEPA	Extent I-JEPA to videos for self-supervised video pretraining.
	Think2Drive [79]	RSSM	Efficiently addressed CALAR v2 scenarios with world models.
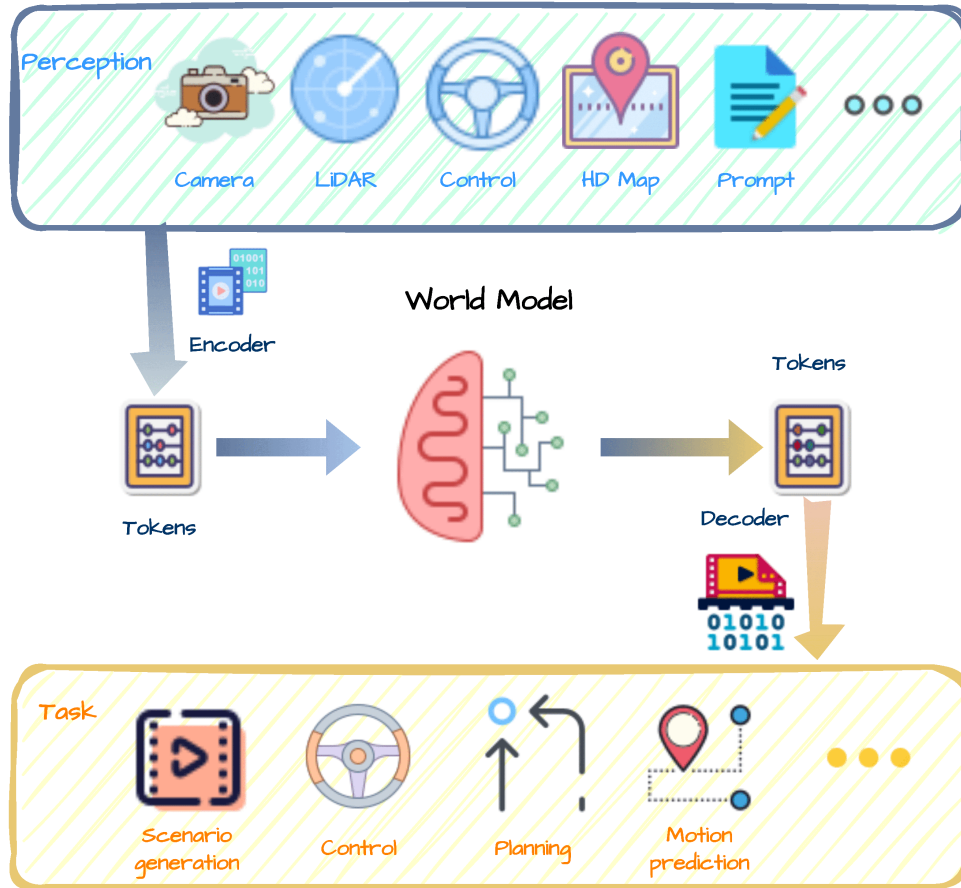
Fig. 5: World Models in Autonomous Driving Pipelines.

A. Driving Scenario Generation

The acquisition of data in autonomous driving encounters substantial hurdles, including high costs associated with data collection and annotation, legal constraints, and safety considerations. World models, through self-supervised learning paradigms, offer a promising solution by enabling the extraction of valuable insights from vast quantities of unlabeled data, thereby enhancing model performance cost-effectively. The application of world models in driving scenario generation is particularly noteworthy, as it facilitates the creation of varied and realistic driving environments. This capability significantly enriches training datasets, equipping autonomous systems with the robustness to navigate rare and intricate driving scenarios [87].

GAIA-1 [62] represents a novel autonomous generative AI model capable of creating realistic driving videos using video, text, and action inputs. Trained on extensive real-world driving data from British cities via Wayve, GAIA-1 learns and understands some real-world rules and key concepts in driving scenarios, including different types of vehicles, pedestrians, buildings, and infrastructure. It can predict and generate subsequent driving scenarios based on a few seconds of video input. Notably, the generated future driving scenarios are not closely tied to the prompt video but are based on GAIA-1's understanding of world rules. Employing an autoregressive transformer network at its core, GAIA-1 predicts upcoming image tokens conditioned on the input image, text, and action tokens, and then decodes these predictions back to pixel space. GAIA-1 can predict multiple potential futures and generate diverse videos or specific driving scenarios based on prompts (e.g., changing weather, scenes, traffic participants, vehicle actions), even including actions and scenes beyond its training set (e.g., forced entry onto sidewalks). This demonstrates its ability to understand and infer driving concepts not present in its training set, which also proves its capabilities of counterfactual reasoning. In the real world, such driving behaviors are hard to acquire data for due to their riskiness. Driving scenario generation allows for simulated testing, enriching data composition, enhancing system capabilities in complex scenarios, and better evaluating existing driving models. Moreover, GAIA-1 generates coherent actions and effectively captures the perspective influences of 3D geometric structures, showcasing its understanding of contextual information and physical rules. Combined with its demonstrated ability for counterfactual
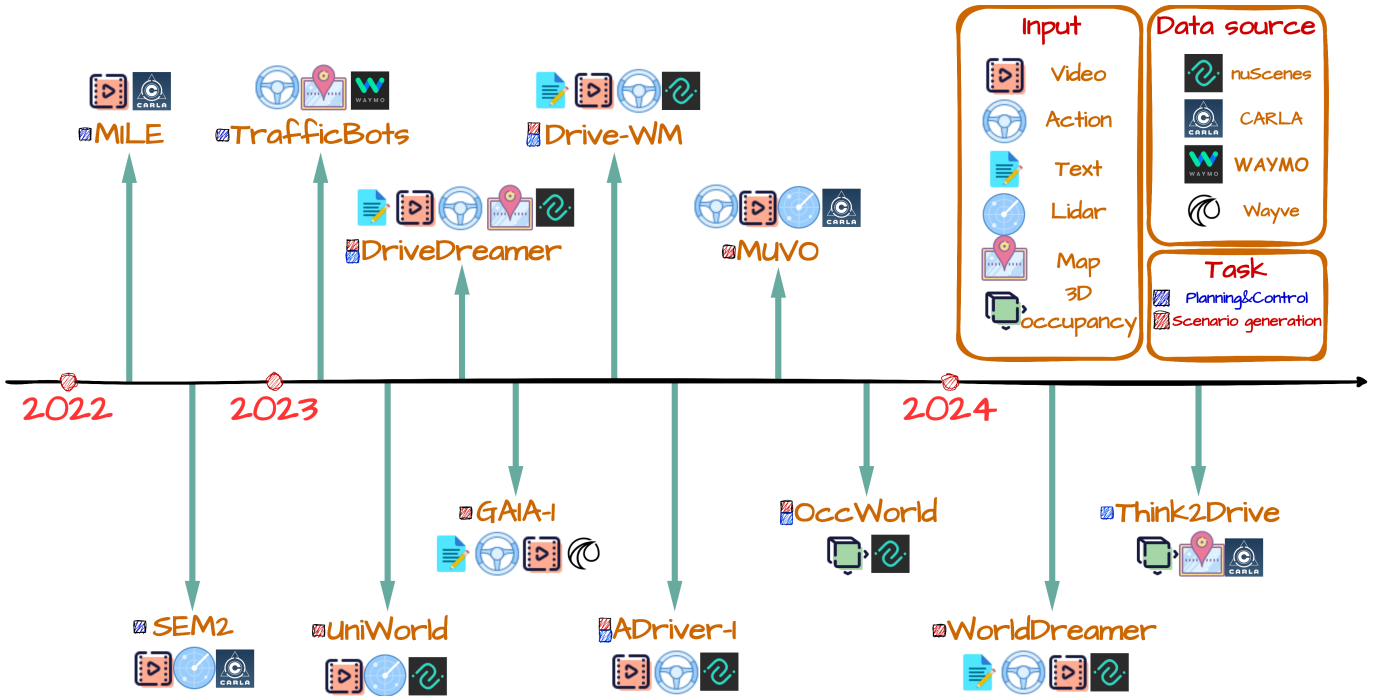
Fig. 6: Chronological Overview of World Models in Autonomous Driving.

reasoning, it can be said that GAIA-1 marks a high level of achievement in the world models of autonomous driving, both in the understanding of abstract concepts and in causal reasoning.

DriveDreamer [60], also dedicated to driving scenario generation, differs from GAIA-1 as it is trained on the nuScenes dataset [88]. Its model inputs include more elements like HD maps and 3D boxes, allowing for more precise control over driving scenario generation and deeper understanding, thus improving video generation quality. Additionally, Drive-Dreamer can generate future driving actions and corresponding predictive scenarios, aiding in decision-making.

ADriver-I employs current video frames and historical vision-action pairs as inputs for a multimodal large language model (MLLM) [89], [90] and a video latent diffusion model (VDM) [91]. The MLLM outputs control signals in an autoregressive manner, which serve as prompts for VDM to predict subsequent video outputs. Through continuous prediction cycles, ADriver-I achieves infinite driving in the predictive world. In ADricer-I, the combination of world model and MLLM significantly improves the interpretability of prediction and decision-making, and also indicates the feasibility of combining the world model as a fundamental model with other models.

Drawing inspiration from the success of large language models, WorldDreamer [75] approaches world modeling as an unsupervised visual sequence modeling challenge. It utilizes the STPT to concentrate attention on local patches within spatiotemporal windows. This focus facilitates dynamic learning of visual signals and accelerates the convergence of the training process. Although World Dreamer is a general-purpose video generation model, it has demonstrated excep-

tional performance in generating autonomous driving videos.

Beyond visual information, driving scenarios also include a plethora of critical physical data. MUVO [72] leverages the world model framework for the prediction and generation of driving scenes and integrates both LIDAR point clouds and visual inputs to predict videos, point clouds, and 3D occupancy grids of future driving scenes. Such a comprehensive approach significantly enhances the quality of predictions and generated outcomes. Particularly, the outcome of 3D occupancy grids can be directly applied to downstream tasks. Taking it a step further, OccWorld [74] and Think2Drive [79] directly utilize the 3D occupancy information as system inputs to predict the evolution of the surrounding environment and plan the actions of the autonomous vehicles. It's evident that as research progresses, world model studies for scenario generation in the autonomous driving domain are gradually evolving towards a multi-modal approach. And world models have demonstrated a versatile capability in processing multi-modal information.

B. Planning and Control

Beyond scenario generation, world models are instrumental in learning within driving contexts, evaluating potential futures, and refining planning and control strategies. For instance, Model-based Imitation LEarning (MILE) [55] adopts a model-based imitation learning approach to jointly learn the dynamics model and driving behavior in CARLA from offline datasets. MILE employs a 'generalized inference algorithm' for rational and visualizable imagination and prediction of future driving environments, using imagination to compensate for missing perceptual information. This ability enables planning future actions, allowing autonomous vehicles to operate without high-definition maps. In inexperienced test

scenarios within the CARLA simulator, MILE significantly outperformed the state-of-the-art models, improving the Driving Score from 46 to 61 (compared to an expert data score of 88). MILE is characterized by long temporal and highly diversified future predictions. Using a decoder on the predicted future states, MILE demonstrates stable driving across various scenarios.

SEM2 [56], building upon RSSM, introduces the semantic masked world model to enhance the sampling efficiency and robustness of end-to-end autonomous driving. The authors contend that world models' latent states contain too much task-irrelevant information, adversely affecting sampling efficiency and system robustness. Moreover, due to imbalanced training data, world models struggle to handle unexpected situations. To address these issues, a signature filter is introduced to extract key task features, reconstructing semantic masks using the filtered features. For data imbalance, a sampler is used to balance the data distribution. In each batch of training, samples from various scenarios are evenly added, to achieve an even and balanced distribution of training samples, which is conducive to generalization and solving corner cases. After being trained and tested in CARLA, the performance of SEM2 showed substantial improvement over DreamerV2.

Considering that most autonomous vehicles typically have multiple cameras, multi-view modeling is also a crucial aspect of world models. Drive-WM [64] is the first multi-view world model designed to enhance the safety of end-to-end autonomous driving planning. Drive-WM, through multi-view and temporal modeling, jointly generates frames for multiple views, and then predicts intermediate views from adjacent ones, significantly improving consistency across multiple views. Additionally, Drive-WM introduces a simple unified conditional interface, flexibly applying images, actions, text, and other conditions, simplifying the conditional generation process. Trained and validated on the nuScenes dataset [88] with six views, Drive-WM selects the best trajectory by sampling predicted candidate trajectories and using an image-based reward function. Furthermore, consistent with GAIA-1, Drive-WM's ability to navigate in non-drivable areas showcases the world model's understanding and potential in handling out-of-domain cases. Besides, drawing inspiration from the seminal work of Alberto Elfes [92], UniWorld [66] introduces an innovative approach by utilizing multi-frame point cloud fusion as the ground truth for generating 4D occupancy labels. This method accounts for the temporal-spatial correlations present in images from multi-camera systems. By leveraging unlabeled image-lidar pairs, UniWorld undergoes pre-training for world models, significantly enhancing the understanding of environmental dynamics. When tested on the nuScenes dataset, UniWorld demonstrates notable improvements in IoU for tasks such as motion prediction and semantic scene completion compared to methods that rely on monocular pre-training.

TrafficBots [65], also an end-to-end autonomous driving model, places a greater emphasis on predicting the actions of individual agents within a scene. Conditioning on each agent's destination, TrafficBots employs a Conditional Variational Autoencoder (CVAE) [93] to learn distinct personalities for each agent, thereby facilitating action prediction from a BEV perspective. Compared to alternative approaches, TrafficBots offers faster operation speeds and can scale to accommodate more agents. Although its performance may not yet rival state-of-the-art open-loop strategies, TrafficBots showcases the potential of closed-loop strategies for action prediction.

C. Conclusion

Fig. 6 represents a chronological overview of existing world models in the field of autonomous driving, including inputs, tasks, and training datasets. Since employing world models in the autonomous driving sector remains a nascent topic, the operational tasks and input-output mechanisms of different world models vary considerably. In the domain of scenario generation, this not only encompasses the generation of predicted scenario videos but also includes subdivisions such as scenario information completion and 3D occupancy prediction, among others. In the control domain, it involves autonomous driving based on input from sensors, vehicle control based on prompt words, and more. Additionally, it can be integrated with scenario generation, outputting predicted scenarios corresponding to control info, thereby offering a pathway to enhance the interpretability of autonomous driving systems.

For those limitations, comparing the performance of different world models faces numerous challenges, including variations in tasks, validation datasets, and the criteria used for measuring performance. For world models in the context of scenario video generation using the nuScenes dataset, they employ FID (Fréchet Inception Distance) and FVD (Fréchet Video Distance) as metrics for assessing video quality. These metrics are critical for understanding how closely the generated scenes resemble real-world scenarios captured in the dataset, with lower scores indicating higher similarity and, therefore, better model performance. According to the comparison result of FID and FVD, we can find that Adriver-I has better video quality than DriveWM and DriveDreamer.

For other world models, although a direct horizontal comparison might not be feasible due to the diversity of tasks and performance metrics, they have achieved state-of-the-art results in their tasks, compared with the traditional methods.

IV. CHALLENGES AND FUTURE PERSPECTIVES

The advancement of world models in the realm of autonomous driving presents a frontier of innovation with the potential to redefine vehicular mobility. However, this promising landscape is not without its challenges. Addressing these hurdles and exploring future perspectives necessitates a deep dive into both the technical complexities and the broader societal implications.

A. Technical and Computational Challenges

a) Long-Term Scalable Memory Integration: The quest to imbue world models with long-term, scalable memories that mirror the intricacies of human cognitive processes remains a formidable challenge in the realm of autonomous driving.

Contemporary models grapple with issues such as vanishing gradients [94] and catastrophic forgetting [95], which severely limit their long-term memory capabilities. Transformer architectures, despite their advancements in facilitating access to historical data through self-attention mechanisms, encounter obstacles in scalability and speed when processing lengthy sequences. Innovative approaches, exemplified by studies like TRANSDREAMER [48] and S4WM [76], explore alternative neural architectures aiming to surmount these barriers. Notably, S4WM has demonstrated superior performance in maintaining high-quality generation over sequences up to 500 steps, markedly surpassing traditional architectures. Yet, the decline in performance observed beyond 1000 steps accentuates the existing disparity between the capacities of artificial and biological memory systems.

To bridge this gap, future research endeavors may pivot towards a multi-pronged strategy encompassing the augmentation of network capacities, the integration of sophisticated external memory modules, and the exploration of iterative learning strategies. These efforts aim not only to extend the temporal reach of memory in world models but also to enhance their ability to navigate the complex decision-making processes inherent in autonomous driving. By fostering a deeper synergy between computational efficiency and memory scalability, these advancements could significantly propel the capabilities of autonomous vehicles, enabling them to adapt and respond to the ever-changing dynamics of real-world driving environments with unprecedented precision and reliability.

b) Simulation-to-Real-World Generalization: The disparity between simulated training environments and the multifaceted nature of real-world conditions presents a critical bottleneck in the evolution of autonomous driving technologies. Current simulation platforms, while advanced, fall short of perfectly mirroring the unpredictability and variability of real-world scenarios. This discordance, manifesting in discrepancies in physical properties, sensor noise, and the occurrence of unforeseen events, critically undermines the applicability of world models trained solely in simulated environments [96].

Although this challenge is still inevitable due to technological limitations, the endeavor to develop world models capable of seamless generalization from simulation to real-world driving scenarios is paramount. This requires not only the refinement of simulation technologies to capture the subtleties and unpredictability of real-world environments more accurately but also the development of models that are inherently robust to the variances between simulated and real-world data. Moreover, the integration of advanced sensory fusion techniques and the exploration of novel learning paradigms, such as meta-learning and reinforcement learning from diverse data sources, could further empower world models to adapt dynamically to the complexities of real-world driving. These advancements are pivotal for the realization of truly autonomous driving systems capable of navigating the myriad challenges posed by real-world environments with agility, accuracy, and safety.

c) Theory and Hardware Breakthroughs: World models currently excel in generative tasks more than in pure predictive tasks, such as motion prediction. This is partly because these models still fall short of perfectly mimicking the evolution of the real world, including the balance between determinism and randomness. Additionally, after processing through sensors and encoders, the information entering the latent space loses a significant amount of detail compared to the real world. Although it's possible to ignore some details and focus on the main factors likely to impact the future through attention mechanisms or special model structures—thus reducing data volume and avoiding irrelevant detail interference—neglecting these nuances inevitably leads to performance losses, creating a bottleneck for the model's predictive capacity. Addressing and restoring sufficient detail, on the other hand, poses a significant challenge to memory and computational power. Despite the encouraging achievements of world models, these newly developed models still require further refinement in theory and structure to serve as foundational AI models. Consequently, transitioning world models from simulators to real-world deployment is expected to take a considerable amount of time.

In summary, while continuously improving the theory of world models, advancements in hardware are also necessary. This includes lossless collection and processing of multimodal information, along with enhancements in the computational capabilities of hardware facilities.

B. Ethical and Safety Challenges

a) Decision-Making Accountability: The imperative of ensuring accountability within the autonomous decision-making frameworks of vehicles stands as a paramount ethical concern, necessitating the development of systems characterized by an unparalleled level of transparency. The complexity inherent in the algorithms guiding autonomous vehicles necessitates a mechanism that not only facilitates decision-making in critical and routine scenarios but also enables these systems to articulate the rationale underpinning their decisions. This transparency is vital for building and maintaining trust among end-users, regulatory bodies, and the broader public.

To achieve this, there is a pressing need for the integration of explainable AI (XAI) principles directly into the development of world models [97]. XAI aims to make AI decisions more interpretable to humans, providing clear, understandable explanations for the actions taken by autonomous vehicles. This involves not merely an exposition of the decision-making process but a comprehensive delineation of the ethical, logical, and practical considerations influencing these decisions. Implementing XAI within autonomous driving systems necessitates a multidisciplinary approach, drawing on expertise from AI development, ethics, legal standards, and user experience design [98]. In this respect, the regulatory guidance that has emerged regarding explainable automated decision-making under the General Data Protection Regulation (GDPR) should be of assistance.

b) Privacy and Data Integrity: The reliance on autonomous driving technologies on extensive datasets for operation and continuous improvement brings to the forefront significant concerns regarding privacy and data security. For companies developing autonomous driving systems, the vehicle-related data collected, including information about passengers,

origins and destinations, travel times, and routes, constitutes both a legally obtained and highly valuable commercial asset [99]. The safeguarding of personal information against unauthorized access and breaches is a critical priority, requiring a robust framework for the ethical handling and protection of data.

Addressing these concerns involves a multifaceted strategy that extends beyond compliance with existing privacy regulations such as GDPR in Europe or California Consumer Privacy Act (CCPA) in US [100]. It entails the establishment of stringent data governance policies that dictate the collection, processing, storage, and sharing of data. These policies should be designed to minimize data exposure and ensure data minimization principles, whereby only the data necessary for specific legitimate purposes is processed. Moreover, the deployment of advanced cybersecurity measures is critical to protect data integrity and confidentiality. This includes the utilization of encryption technologies, secure data storage solutions, and regular security audits to identify and mitigate potential vulnerabilities. Additionally, fostering transparency with users about how their data is collected, used, and protected is fundamental [101]. This can be achieved through clear, accessible privacy policies and mechanisms that empower users with control over their personal information, including options for data access, correction, and deletion [102].

c) Responsibility and Criterion: As world models support or take over driving tasks in autonomous driving systems, human responsibility is not lessened or eliminated but rather redistributed among the network of individuals and organizations involved in their creation, deployment, and usage. This shift demands different requirements from the participants, calling for new research and policies to govern this transformation in demand.

Obligation-wise, policymakers should encourage research activities in ethics design, aiming to establish it as a solid academic field, akin to medical ethics [103]. Moreover, there is a need to create suitable educational environments to promote citizen education on the obligations of various stakeholders. Ethically, policymakers, manufacturers, and deployers should set up mechanisms to reward those who actively take on responsibilities within organizations or professional associations focused on ethical design and deployment, thus driving a shift in culture [104].

Given the complexity of the socio-technical systems autonomous driving systems belong to, traditional, moral, and legal standards for attributing responsibility to individual human agents may not easily apply to behaviors emerging from interactions between humans and intelligent systems. Policymakers, in collaboration with researchers, manufacturers, and deployers, should establish clear and fair legal rules for assigning liability in case of incidents. This could involve developing new insurance models. Such rules must find a balance between avoiding culpability gaps and fulfilling the demands of corrective justice [105].

C. Future Perspectives

a) Bridging Human Intuition and AI Precision: One groundbreaking perspective is the evolution of world models towards facilitating a cognitive co-piloting framework within autonomous vehicles. Unlike traditional autonomous systems, which rely solely on pre-defined algorithms and sensor inputs for decision-making, cognitive co-piloting aims to blend the nuanced, intuitive decision-making capabilities of human drivers with the precision and reliability of AI. By leveraging advanced world models, vehicles can gain an unprecedented level of environmental awareness and predictive capability, mirroring human cognitive processes such as anticipation, intuition, and the ability to navigate complex socio-technical environments.

This integration enables autonomous vehicles to not only react to the immediate physical world but also to understand and adapt to the social and psychological dimensions of driving—interpreting gestures, predicting human behaviors, and making decisions that reflect a deeper understanding of human norms and expectations. For instance, a world model equipped with cognitive co-piloting capabilities could accurately predict pedestrian movements in urban settings, navigate social driving conventions at four-way stops, or adapt driving styles in response to passenger comfort and feedback.

b) Harmonizing Vehicles with the Urban Ecosystem: Another visionary perspective involves the role of world models in transforming autonomous vehicles into agents of ecological engineering, harmonizing with urban ecosystems through adaptive, responsive behaviors that contribute to environmental sustainability. World models, with their deep understanding of complex systems and dynamics, can enable autonomous vehicles to optimize routes and driving patterns not just for efficiency and safety, but also for environmental impact, such as minimizing emissions, reducing congestion, and promoting energy conservation.

A global survey [106] reveals that over 60% of respondents believe that with the advancement in urban system automation, the pollution generated by transportation systems and the likelihood of vehicle collisions will decrease. Furthermore, more than 70% of the respondents anticipate improvements in traffic noise. Meanwhile, those who believe that roadworks and transport system management will see enhancements exceed 80%. This data reflects a widespread optimism towards the potential environmental and safety benefits of increasingly automated urban transport systems. World models represent a crucial direction for development that can facilitate a higher degree of automation in vehicles and traffic systems, showcasing their significance in advancing urban infrastructure towards safer and more sustainable futures.

V. CONCLUSION

In conclusion, this survey has delved into the transformative potential of world models in the autonomous driving landscape, highlighting their pivotal role in advancing vehicle autonomy through enhanced prediction, simulation, and decision-making capabilities. Despite the significant progress

documented, challenges such as long-term memory integration, simulation-to-real-world generalization, and ethical considerations underscore the complexity of deploying these models in real-world applications. Addressing these challenges necessitates a multidisciplinary approach that combines advancements in AI research with ethical frameworks and innovative computational solutions. Looking ahead, the evolution of world models promises to not only enhance autonomous driving technologies but also to redefine our interaction with automated systems, underscoring the need for continued research and collaboration across fields. As we stand on the cusp of this technological frontier, it is imperative that we navigate the ethical implications and societal impacts with diligence and foresight, ensuring that the development of autonomous driving technologies remains aligned with broader societal values and safety standards.

ACKNOWLEDGMENT

REFERENCES

[1] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27," *Open Review*, vol. 62, no. 1, 2022.

[2] L. Chang and D. Y. Tsao, "The code for facial identity in the primate brain," *Cell*, vol. 169, no. 6, pp. 1013–1028, 2017.

[3] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, "Invariant visual representation by single neurons in the human brain," *Nature*, vol. 435, no. 7045, pp. 1102–1107, 2005.

[4] G. B. Keller, T. Bonhoeffer, and M. Hübener, "Sensorimotor mismatch signals in primary visual cortex of the behaving mouse," *Neuron*, vol. 74, no. 5, pp. 809–815, 2012.

[5] J. Rault, A. Richalet, J. Testud, and J. Papon, "Model predictive heuristic control: Application to industrial processes," *Automatica*, vol. 14, no. 5, pp. 413–428, 1978.

[6] A. E. Bryson, *Applied optimal control: optimization, estimation and control*. Routledge, 2018.

[7] A. E. Bryson, "Optimal control-1950 to 1985," *IEEE Control Systems Magazine*, vol. 16, no. 3, pp. 26–33, 1996.

[8] J. W. Forrester, "Counterintuitive behavior of social systems," *Theory and decision*, vol. 2, no. 2, pp. 109–140, 1971.

[9] P. N. Johnson-Laird, "The history of mental models," in *Psychology of reasoning*, Psychology Press, 2004, pp. 189–222.

[10] N. A. Jones, H. Ross, T. Lynam, P. Perez, and A. Leitch, "Mental models: An interdisciplinary synthesis of theory and methods," *Ecology and society*, vol. 16, no. 1, 2011.

[11] P. N. Johnson-Laird, *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, 1983.

[12] S. N. Kumpati, P. Kannan, *et al.*, "Identification and control of dynamical systems using neural networks," *IEEE Transactions on neural networks*, vol. 1, no. 1, pp. 4–27, 1990.

[13] A. U. Levin and K. S. Narendra, "Control of nonlinear dynamical systems using neural networks: Controllability and stabilization," *IEEE Transactions on neural networks*, vol. 4, no. 2, pp. 192–206, 1993.

[14] A. Draeger, S. Engell, and H. Ranke, "Model predictive control using neural networks," *IEEE Control Systems Magazine*, vol. 15, no. 5, pp. 61–66, 1995.

[15] J. Schmidhuber, "An on-line algorithm for dynamic reinforcement learning and planning in reactive environments," in *1990 IJCNN international joint conference on neural networks*, IEEE, 1990, pp. 253–258.

[16] J. Schmidhuber, "On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models," *arXiv preprint arXiv:1511.09249*, 2015.

[17] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Transactions on neural networks*, vol. 1, no. 1, pp. 4–27, 1990.

[18] R. S. Sutton, "Dyna, an integrated architecture for learning, planning, and reacting," *ACM Sigart Bulletin*, vol. 2, no. 4, pp. 160–163, 1991.

[19] K. J. Hunt, D. Sbarbaro, R. Żbikowski, and P. J. Gawthrop, "Neural networks for control systems—a survey," *Automatica*, vol. 28, no. 6, pp. 1083–1112, 1992.

[20] M. I. Jordan and D. E. Rumelhart, "Forward models: Supervised learning with a distal teacher," in *Backpropagation*, Psychology Press, 2013, pp. 189–236.

[21] J. Schmidhuber, *Making the world differentiable: on using self supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments*. Inst. für Informatik, 1990, vol. 126.

[22] S. Chiappa, S. Racaniere, D. Wierstra, and S. Mohamed, "Recurrent environment simulators," *arXiv preprint arXiv:1704.02254*, 2017.

[23] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in atari games," *Advances in neural information processing systems*, vol. 28, 2015.

[24] D. Silver, H. Hasselt, M. Hessel, *et al.*, "The predictron: End-to-end learning and planning," in *International Conference on Machine Learning*, PMLR, 2017, pp. 3191–3199.

[25] N. Watters, D. Zoran, T. Weber, P. Battaglia, R. Pascanu, and A. Tacchetti, "Visual interaction networks: Learning a physics simulator from video," *Advances in neural information processing systems*, vol. 30, 2017.

[26] M. Deisenroth and C. E. Rasmussen, "Pilco: A model-based and data-efficient approach to policy search," in

Proceedings of the 28th International Conference on machine learning (ICML-11), 2011, pp. 465–472.

[27] T. M. Moerland, J. Broekens, A. Plaat, C. M. Jonker, *et al.*, "Model-based reinforcement learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 16, no. 1, pp. 1–118, 2023.

[28] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic books, 2018.

[29] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[30] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," *Advances in neural information processing systems*, vol. 31, 2018.

[31] V. Micheli, E. Alonso, and F. Fleuret, "Transformers are sample efficient world models," *arXiv preprint arXiv:2209.00588*, 2022.

[32] L. Kaiser, M. Babaeizadeh, P. Milos, *et al.*, "Model-based reinforcement learning for atari," *arXiv preprint arXiv:1903.00374*, 2019.

[33] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[34] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[35] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*, PMLR, 2014, pp. 1278–1286.

[36] A. Bardes, J. Ponce, and Y. LeCun, "Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features," *arXiv preprint arXiv:2307.12698*, 2023.

[37] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[38] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[39] D. J. Foster, "Replay comes of age," *Annual review of neuroscience*, vol. 40, pp. 581–602, 2017.

[40] I. Rechenberg, "Evolutionsstrategien," in *Simulationsmethoden in der Medizin und Biologie: Workshop, Hannover, 29. Sept.–1. Okt. 1977*, Springer, 1978, pp. 83–114.

[41] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.

[42] D. Hafner, T. Lillicrap, I. Fischer, *et al.*, "Learning latent dynamics for planning from pixels," in *International conference on machine learning*, PMLR, 2019, pp. 2555–2565.

[43] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering atari with discrete world models," *arXiv preprint arXiv:2010.02193*, 2020.

[44] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.

[45] R. Mendonca, O. Rybkin, K. Daniilidis, D. Hafner, and D. Pathak, "Discovering and achieving goals via world models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 379–24 391, 2021.

[46] Z. Fei, M. Fan, and J. Huang, "A-jepa: Joint-embedding predictive architecture can listen," *arXiv preprint arXiv:2311.15830*, 2023.

[47] M. Assran, Q. Duval, I. Misra, *et al.*, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 619–15 629.

[48] C. Chen, Y.-F. Wu, J. Yoon, and S. Ahn, "Transdreamer: Reinforcement learning with transformer world models," *arXiv preprint arXiv:2202.09481*, 2022.

[49] W. Zhang, G. Wang, J. Sun, Y. Yuan, and G. Huang, "Storm: Efficient stochastic transformer based world models for reinforcement learning," *arXiv preprint arXiv:2310.09615*, 2023.

[50] J. D. Hamilton, "State-space models," *Handbook of econometrics*, vol. 4, pp. 3039–3080, 1994.

[51] K. Kim, M. Sano, J. De Freitas, N. Haber, and D. Yamins, "Active world model learning with progress curiosity," in *International conference on machine learning*, PMLR, 2020, pp. 5306–5315.

[52] J. Schrittwieser, I. Antonoglou, T. Hubert, *et al.*, "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.

[53] W. Ye, S. Liu, T. Kurutach, P. Abbeel, and Y. Gao, "Mastering atari games with limited data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 476–25 488, 2021.

[54] J. Y. Koh, H. Lee, Y. Yang, J. Baldridge, and P. Anderson, "Pathdreamer: A world model for indoor navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 738–14 748.

[55] A. Hu, G. Corrado, N. Griffiths, *et al.*, "Model-based imitation learning for urban driving," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 703–20 716, 2022.

[56] Z. Gao, Y. Mu, R. Shen, *et al.*, "Enhance sample efficiency and robustness of end-to-end urban autonomous driving via semantic masked world model," *arXiv preprint arXiv:2210.04017*, 2022.

[57] M. Okada and T. Taniguchi, "Dreamingv2: Reinforcement learning with discrete world models without reconstruction," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2022, pp. 985–991.

[58] F. Deng, I. Jang, and S. Ahn, "Dreamerpro: Reconstruction-free model-based reinforcement learning with prototypical representations," in *Interna-

tional Conference on Machine Learning, PMLR, 2022, pp. 4956–4975.

[59] H. Ma, W. Xue, R. Ying, and P. Liu, "Maxent dreamer: Maximum entropy reinforcement learning with world model," in *2022 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2022, pp. 1–9.

[60] X. Wang, Z. Zhu, G. Huang, X. Chen, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," *arXiv preprint arXiv:2309.09777*, 2023.

[61] F. Jia, W. Mao, Y. Liu, *et al.*, "Adriver-i: A general world model for autonomous driving," *arXiv preprint arXiv:2311.13549*, 2023.

[62] A. Hu, L. Russell, H. Yeo, *et al.*, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[63] P. Wu, A. Escontrela, D. Hafner, P. Abbeel, and K. Goldberg, "Daydreamer: World models for physical robot learning," in *Conference on Robot Learning*, PMLR, 2023, pp. 2226–2240.

[64] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," *arXiv preprint arXiv:2311.17918*, 2023.

[65] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "Trafficbots: Towards world models for autonomous driving simulation and motion prediction," *arXiv preprint arXiv:2303.04116*, 2023.

[66] C. Min, D. Zhao, L. Xiao, Y. Nie, and B. Dai, "Uniworld: Autonomous driving pre-training via world models," *arXiv preprint arXiv:2308.07234*, 2023.

[67] W. Huang, J. Ji, B. Zhang, C. Xia, and Y. Yang, "Safe dreamerv3: Safe reinforcement learning with world models," *arXiv preprint arXiv:2307.07176*, 2023.

[68] H. Ma, J. Wu, N. Feng, J. Wang, and M. Long, "Harmony world models: Boosting sample efficiency for model-based reinforcement learning," *arXiv preprint arXiv:2310.00344*, 2023.

[69] J. Ha, K. Kim, and Y. Kim, "Dream to generalize: Zero-shot model-based reinforcement learning for unseen visual distractions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 7802–7810.

[70] J. Robine, M. Höftmann, T. Uelwer, and S. Harmeling, "Transformer-based world models are happy with 100k interactions," *arXiv preprint arXiv:2303.07109*, 2023.

[71] S. Cai, E. R. Chan, S. Peng, *et al.*, "Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2139–2150.

[72] D. Bogdoll, Y. Yang, and J. M. Zöllner, "Muvo: A multimodal generative world model for autonomous driving with geometric representations," *arXiv preprint arXiv:2311.11762*, 2023.

[73] R. Mendonca, S. Bahl, and D. Pathak, "Structured world models from human videos," *arXiv preprint arXiv:2308.10901*, 2023.

[74] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3d occupancy world model for autonomous driving," *arXiv preprint arXiv:2311.16038*, 2023.

[75] X. Wang, Z. Zhu, G. Huang, B. Wang, X. Chen, and J. Lu, "Worlddreamer: Towards general world models for video generation via predicting masked tokens," *arXiv preprint arXiv:2401.09985*, 2024.

[76] F. Deng, J. Park, and S. Ahn, "Facing off world model backbones: Rnns, transformers, and s4," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[77] J. Bruce, M. Dennis, A. Edwards, *et al.*, *Genie: Generative interactive environments*, 2024. arXiv: 2402. 15391 [cs.LG].

[78] A. Bardes, Q. Garrido, J. Ponce, *et al.*, "V-jepa: Latent video prediction for visual representation learning," 2023.

[79] Q. Li, X. Jia, S. Wang, and J. Yan, "Think2drive: Efficient reinforcement learning by thinking in latent world model for quasi-realistic autonomous driving (in carla-v2)," *arXiv preprint arXiv:2402.16720*, 2024.

[80] J. B. W. Webber, "A bi-symmetric log transformation for wide-range data," *Measurement Science and Technology*, vol. 24, no. 2, p. 027 001, 2012.

[81] M. Plappert, M. Andrychowicz, A. Ray, *et al.*, "Multigoal reinforcement learning: Challenging robotics environments and request for research," *arXiv preprint arXiv:1802.09464*, 2018.

[82] Y. Tassa, Y. Doron, A. Muldal, *et al.*, "Deepmind control suite," *arXiv preprint arXiv:1801.00690*, 2018.

[83] T. Yu, D. Quillen, Z. He, *et al.*, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on robot learning*, PMLR, 2020, pp. 1094–1100.

[84] L. Zhang, G. Yang, and B. C. Stadie, "World model as a graph: Learning latent landmarks for planning," in *International Conference on Machine Learning*, PMLR, 2021, pp. 12 611–12 620.

[85] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, PMLR, 2018, pp. 1861–1870.

[86] T. Brooks, B. Peebles, C. Homes, *et al.*, "Video generation models as world simulators," 2024. [Online]. Available: https : / / openai . com / research / video - generation-models-as-world-simulators.

[87] F. Tian, Z. Li, F.-Y. Wang, and L. Li, "Parallel learning-based steering control for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 379–389, 2022.

[88] H. Caesar, V. Bankiti, A. H. Lang, *et al.*, "Nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[89] W.-L. Chiang, Z. Li, Z. Lin, *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt

quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[90] H. Touvron, L. Martin, K. Stone, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[91] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[92] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, 1989.

[93] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[94] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, Pmlr, 2013, pp. 1310–1318.

[95] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[96] A. Stocco, B. Pulfer, and P. Tonella, "Mind the gap! a study on the transferability of virtual vs physical-world testing of autonomous driving systems," *IEEE Transactions on Software Engineering*, 2022.

[97] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, and A. Taly, "Explainable ai in industry," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 3203–3204.

[98] A. Ignatiev, "Towards trustable explainable ai.," in *IJCAI*, 2020, pp. 5154–5158.

[99] L. Collingwood, "Privacy implications and liability issues of autonomous vehicles," *Information & Communications Technology Law*, vol. 26, no. 1, pp. 32–45, 2017.

[100] G. D. P. Regulation, "General data protection regulation (gdpr)," *Intersoft Consulting, Accessed in October*, vol. 24, no. 1, 2018.

[101] M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner, *Autonomous driving: technical, legal and social aspects*. Springer Nature, 2016.

[102] J.-F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles," *Science*, vol. 352, no. 6293, pp. 1573–1576, 2016.

[103] C. Véliz, "Three things digital ethics can learn from medical ethics," *Nature Electronics*, vol. 2, no. 8, pp. 316–318, 2019.

[104] L. Schiebinger, *Gendered innovations in science and engineering*. Stanford University Press, 2008.

[105] J.-F. Bonnefon, D. Černy, J. Danaher, *et al.*, "Ethics of connected and automated vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility," 2020.

[106] M. A. Goddard, Z. G. Davies, S. Guenat, *et al.*, "A global horizon scan of the future impacts of robotics and autonomous systems on urban ecosystems," *Nature ecology & evolution*, vol. 5, no. 2, pp. 219–230, 2021.

Yanchen Guan is currently a PhD student at the State Key Laboratory of Internet of Things for Smart City and the Department of Civil Engineering at the University of Macau. He holds a master's degree in Mobility Engineering from Politecnico di Milano (2023) and a bachelor's degree in Mechatronics Engineering from Harbin Institute of Technology (2019). His research primarily focuses on autonomous driving, intelligent transportation systems, mechanical structures, and data analysis.

Haicheng Liao (Student Member, IEEE) received the B.S. degree in software engineering from the University of Electronic Science and Technology of China (UESTC) in 2022. He is currently pursuing the Ph.D. degree at the State Key Laboratory of Internet of Things for Smart City and the Department of Computer and Information Science, University of Macau. His research interests include connected autonomous vehicles and the application of deep reinforcement learning to autonomous driving.

Zhenning Li (Member, IEEE) received his Ph.D. in Civil Engineering from the University of Hawaii at Manoa in 2019. Currently, he holds the position of Assistant Professor at the State Key Laboratory of Internet of Things for Smart City, as well as the Departments of Civil and Environmental Engineering and Computer and Information Science at the University of Macau, Macau SAR. He has published over 50 papers in top journals and conferences. His main areas of research focus on the intersection of connected autonomous vehicles and Big Data applications in urban transportation systems. He has been honored with several awards, including the TRB Best Young Researcher award and the CICTP Best Paper Award, amongst others.
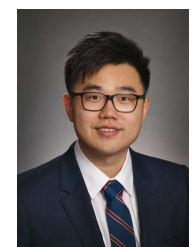
Jia Hu (Member, IEEE) is currently working as a Zhongte Distinguished Chair of Cooperative Automation with the College of Transportation Engineering, Tongji University. Before joining Tongji University, he was a Research Associate with the Federal Highway Administration (FHWA), USA. He is an Editorial Board Member of the Journal of Intelligent Transportation Systems and the International Journal of Transportation Science and Technology. He is a member of TRB (a Division of the National Academies) Vehicle Highway Automation Committee, the Freeway Operations Committee, the Simulation subcommittee of the Traffic Signal Systems Committee, and the Advanced Technologies Committee of the ASCE Transportation and Development Institute. He is the Chair of the Vehicle Automation and Connectivity Committee of the World Transport Convention. He is an Associate Editor of the American Society of Civil Engineers Journal of Transportation Engineering and IEEE OPEN JOURNAL OF INTELLIGENT TRANSPORTATION SYSTEMS.

Runze Yuan received a bachelor's degree from Tsinghua University, Beijing, in 2017. He received the PhD degree from the University of Hawaii at Manoa, Honolulu, in 2023. He is currently a postdoc in the Department of Automation at Tsinghua University, Beijing. His main research areas include microscopic traffic flow modeling, artificial intelligence in transportation, traffic flow detection and control, and intelligent transportation systems.

Guohui Zhang is currently a Professor with the Department of Civil and Environmental Engineering, University of Hawai'i at Manoa. His primary research interests include transportation system resilience, transportation systems modeling, planning, and operation, traffic sensing and sensor data analytics, artificial intelligence in transportation, connected and autonomous vehicle systems, and transportation safety and security. He serves as a member of the American Society of Civil Engineers Connected and Autonomous Vehicle Impact Committee and the Transportation Research Board Information Systems and Technology Committee and as a panelist for multiple National Collaborative Highway Research Program projects.

Chengzhong Xu (Fellow, IEEE) received the Ph.D. degree from The University of Hong Kong, in 1993. He is currently the chair professor of computer science and the dean with the Faculty of Science and Technology, University of Macau. Prior to this, he was with the faculty at Wayne State University, USA, and the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. He has published more than 400 papers and more than 100 patents. His research interests include cloud computing and data-driven intelligent applications. He was the Best Paper awardee or the Nominee of ICPP2005, HPCA2013, HPDC2013, Cluster2015, GPC2018, UIC2018, and AIMS2019. He also won the Best Paper award of SoCC2021. He was the Chair of the IEEE Technical Committee on Distributed Processing from 2015 to 2019.