Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 28 May 2026
Revised 22 June 2026
Accepted 17 July 2026

Natural Resources for Human Health 2026, Vol. 6 (9s): 723–731

KEYWORDS:
Generative AI, Diffusion Models, Stochastic Optimal Control, Reinforcement Learning, Sequential Decision-Making, Policy Optimization, Markov Decision Process, Adaptive Control.

Generative AI and Stochastic Optimal Control and Reinforcement Learning Frameworks for Adaptive and Intelligent Decision-Making Systems

Ms. G. Brindha[1], Malyala Gayatri[2], Dr. Maganti Venkatesh[3], Amitava Podder[4], Dr. Arun Raj S.R[5], V. Purushothama Raju[6]

[1]Assistant Professor, Artificial Intelligence and Data Science, Dr. N. G. P. Institute of Technology, Coimbatore, Tamil Nadu, brindhasmg@gmail.com
[2]Senior Assistant professor, Department of CSE-AIML, Geethanjali college of engineering and technology, Medchal, Hyderabad, Telangana, mgayatri.cse@gcet.edu.in
[3]Associate Professor, Department of AIML, Aditya University, Surampalem, AP, India.
[4]Assistant Professor, Computer Science and Engineering -AI, Brainware University, Barasat, North 24 Parganas, West Bengal, amitavapodder24@gmail.com
[5]Assistant Professor, Department of Electronics and Communication Engineering, University B.D.T College of Engineering, Davanagere-577004, Visvesvaraya Technological University, Karnataka, Orchid id: 0009-0005-7604-2658
[6]Professor, Department of CSE,Shri Vishnu Engineering College for Women, Bhimavaram, A.P., India

ABSTRACT: Sequential decision-making under uncertainty has traditionally been formalized through two largely separate mathematical traditions: stochastic optimal control, which derives policies from explicit dynamics models and value functions, and reinforcement learning (RL), which learns policies directly from interaction with an environment. Generative artificial intelligence, and diffusion and flow-based models specifically, has recently emerged as a bridging technology that reframes both traditions around a shared computational substrate: treating decision-making, planning, and control-policy generation as instances of generative sequence modeling. This paper reviews the rapidly growing literature at the intersection of generative AI, stochastic optimal control, and reinforcement learning, synthesizing foundational work reformulating diffusion denoising as a sequential Markov decision process with control-theoretic treatments of stochastic optimal control via deep learning and RL-based approaches to nonlinear stochastic system stabilization. Particular attention is given to the specific mechanism connecting generative modeling to adaptive decision-making: the reviewed literature converges on the finding that framing multi-step denoising, trajectory generation, or policy sampling as a Markov decision process allows policy-gradient reinforcement learning to directly optimize non-differentiable, black-box, or human-preference-derived reward objectives that classical gradient-based diffusion training cannot address. Comparative tables map generative-RL frameworks to their algorithmic mechanism and application domain, cross-reference the specific decision-making tasks, robotic control, autonomous driving, text-to-image alignment, energy-system management, examined against the architectures and reported outcomes each study documents, and set the control-theoretic stability guarantees available for classical stochastic optimal control against the comparatively weaker guarantees available for generative policy architectures. The review concludes that while generative-AI-based decision-making frameworks have demonstrated substantial empirical gains in expressiveness and multi-modal action modeling relative to classical Gaussian policy classes, formal stability and safety guarantees comparable to those long established in stochastic optimal control theory remain the field's central unresolved research challenge.

I. INTRODUCTION

Sequential decision-making under uncertainty, choosing actions over time to optimize an expected outcome despite incomplete information and stochastic dynamics, has long been studied through two mathematically related but methodologically distinct research traditions. Reinforcement learning and stochastic optimization together form what a comprehensive unifying treatment has characterized as a single framework for sequential decisions, in which an agent must learn to make good decisions from data, typically by interacting with an uncertain environment, while stochastic optimization more broadly encompasses the mathematical machinery for optimizing decisions under uncertainty regardless of whether learning occurs online or offline [1]. Classical stochastic optimal control, by contrast, has historically approached the same underlying problem, choosing a control policy to optimize an expected cost or reward functional subject to stochastic system dynamics, through explicit dynamics models, value functions, and, in the deep-learning era, neural-network function approximation for solving the resulting high-dimensional partial differential equations [2].

Generative artificial intelligence has recently introduced a third, increasingly influential perspective that partially unifies these two traditions. Foundational research reformulating the diffusion denoising process explicitly as a sequential decision-making problem demonstrated that posing denoising as a multi-step decision-making problem enables a class of policy gradient algorithms, denoising diffusion policy optimization (DDPO), that use exact likelihoods at each denoising step in place of the approximate likelihoods induced by a full denoising process, making it possible to optimize a diffusion model for downstream tasks using only a black-box reward function [3]. This reformulation is significant precisely because it treats a generative model's internal sampling process, the sequence of denoising steps that produces a final generated sample, as itself a Markov decision process amenable to standard reinforcement-learning policy-gradient methods, thereby importing RL's capacity to optimize non-differentiable, human-preference-derived, or otherwise black-box objectives directly into generative model training [3].

This generative-RL synthesis has direct relevance to adaptive and intelligent decision-making systems because it addresses a specific limitation of classical RL policy architectures. Research on generative models for robotic decision-making observed that many RL studies increasingly investigate leveraging the powerful expressivity of generative models, particularly flow-based and diffusion-based models, to effectively master the capability of modeling multi-modal action distributions, a capability that significantly overcomes the inherent limitations of traditional Gaussian-based policies, which can only represent unimodal action distributions [4]. This architectural advantage, the ability to represent complex, multi-modal action or trajectory distributions rather than being restricted to unimodal Gaussian policies, explains why generative models have rapidly displaced classical policy parameterizations across robotics, autonomous driving, and other continuous-control decision-making domains.

This paper reviews the generative-AI-for-reinforcement-learning literature alongside classical and RL-augmented stochastic optimal control research within a single comparative framework, examining how diffusion- and flow-based generative architectures are being integrated into sequential decision-making systems, what algorithmic mechanisms enable this integration, and what formal guarantees, if any, these hybrid frameworks currently provide relative to the stability guarantees long established in classical control theory.

Research Objectives

- To review the foundational reformulation of diffusion-model denoising as a sequential Markov decision process amenable to reinforcement-learning policy optimization.
- To synthesize the stochastic-optimal-control literature addressing deep-learning approximation of high-dimensional control problems and RL-based stabilization of nonlinear stochastic systems.
- To examine the application-specific generative-RL frameworks developed for robotic control, autonomous driving, and adaptive planning under uncertainty.

- To evaluate the formal stability and safety guarantees available for generative-model-based decision-making relative to classical stochastic-optimal-control guarantees.
- To identify future research priorities for adaptive, intelligent decision-making systems combining generative AI, stochastic optimal control, and reinforcement learning.

II. LITERATURE REVIEW

2.1 Foundational Reformulation: Diffusion Denoising as Sequential Decision-Making

The conceptual foundation connecting generative AI to reinforcement learning rests on a specific mathematical reframing of the diffusion sampling process. The original denoising diffusion policy optimization (DDPO) framework addressed a problem inherent to diffusion probabilistic models used as the de facto standard for generative modeling in continuous domains: training these models to satisfy downstream objectives directly, rather than merely matching a training data distribution, is challenging because exact likelihood computation with diffusion models is intractable, making it difficult to apply many conventional reinforcement-learning algorithms directly [3]. DDPO's solution was to frame each individual denoising step, rather than the full generation trajectory, as an action within a Markov decision process, using the exact, tractable likelihood available at each single denoising step as a substitute for the intractable full-trajectory likelihood [3]. Empirically, this reformulation allowed DDPO to adapt text-to-image diffusion models to objectives difficult to express via prompting, such as image compressibility, objectives derived directly from human feedback, such as aesthetic quality, and prompt-image alignment feedback from a vision-language model without requiring additional human annotation [3].

Subsequent research has extended and refined this MDP-based reformulation for policy learning specifically, as opposed to purely generative image synthesis. Diffusion Policy Policy Optimization (DPPO) introduced a comprehensive algorithmic framework, including best practices, for fine-tuning diffusion-based robot-control policies using policy-gradient methods from reinforcement learning, directly addressing a belief in prior literature that policy-gradient methods were inefficient for training diffusion-based policies in continuous control tasks [5]. The DPPO framework's explicit interest in how such fine-tuning fits together with other decision-making tools, including model-based planning and video-prediction-aided decision-making, signals the field's broader ambition to integrate generative-model-based policy learning with the classical model-based planning apparatus long central to control theory [5].

2.2 Stochastic Optimal Control via Deep Learning

Parallel to the generative-RL research reviewed in Section 2.1, a distinct but structurally related literature has applied deep learning directly to classical stochastic optimal control problems, particularly those whose dimensionality exceeds what traditional numerical PDE-solution methods can handle. Foundational work on deep-learning approximation for stochastic control problems established neural-network-based algorithms for approximating solutions to the Hamilton-Jacobi-Bellman partial differential equations that classical stochastic optimal control theory uses to characterize optimal value functions and policies, extending techniques for solving high-dimensional partial differential equations using deep learning to the control setting specifically [6]. Building on this foundation, an efficient on-policy deep-learning framework for stochastic optimal control was developed specifically to improve sample efficiency and training stability relative to earlier deep-learning-based stochastic control approaches, situating itself explicitly within the classical stochastic-control tradition traced back to Fleming and Rishel's foundational treatment of deterministic and stochastic optimal control [7].

This deep-learning-for-stochastic-control literature is important for the present review because it represents a methodologically distinct, though objective-equivalent, path toward the same class of adaptive decision-making problems the generative-RL literature addresses: both traditions seek tractable computational methods for high-dimensional, stochastic sequential decision problems, but the deep-learning-stochastic-control tradition retains an explicit connection to the underlying PDE and value-function formalism of classical control theory, while the generative-RL tradition reviewed in Section 2.1 instead reformulates the problem directly as black-box policy optimization over a learned generative sampling process.

2.3 Reinforcement Learning for Stability-Guaranteed Nonlinear Stochastic Control

A further strand of research has sought to directly import classical control-theoretic stability guarantees into reinforcement-learning-based controllers for stochastic nonlinear systems, addressing the stability-guarantee gap that purely generative or purely value-based RL approaches otherwise leave open. A 2025 study developing reinforcement learning for optimal control of stochastic nonlinear systems constructed a stochastic control Lyapunov function (SCLF) candidate using neural networks as an approximator to the value function, then developed a control policy based on this SCLF specifically to ensure the stability in probability of the resulting stochastic nonlinear system [8]. The associated RL algorithm iteratively updates both the value function and the control policy, driving them toward their optimal values while preserving the Lyapunov-based stability guarantee throughout training, with the framework further extended to remain feasible under sample-and-hold implementation of control actions, a practically important consideration for real-world digital control systems that cannot apply continuously varying control signals [8]. The study's demonstrated application to two chemical reactor examples illustrates the framework's applicability to genuinely safety-critical industrial process-control settings, where the stability guarantees classical control theory provides cannot simply be abandoned in favor of the raw empirical performance gains that dominate much of the generative-RL literature reviewed in Section 2.1 [8].

2.4 Generative Models Within the Reinforcement Learning Training Loop

Beyond directly reformulating generative sampling as policy optimization, a broader body of research has examined how generative models can address specific structural challenges within the conventional RL training loop itself. A comprehensive review of the duality between generative AI and reinforcement learning in robotics organized this literature around three specific RL training-loop challenges generative models address: reward-signal generation in the presence of sparse rewards and under goal-specification constraints, state-representation learning to improve sample efficiency, and planning and exploration mechanisms that support generalization to new tasks [4]. This three-part organizing framework is analytically valuable because it demonstrates that generative AI's contribution to reinforcement learning is not limited to the policy-parameterization role examined in Sections 2.1–2.2, but extends to auxiliary components of the RL pipeline, reward modeling and state representation specifically, where generative models' distributional modeling capacity offers distinct advantages over conventional discriminative approaches.

A complementary treatment of generative AI for deep reinforcement learning characterized deep reinforcement learning itself as a transformative approach that incorporates the principles of deep learning and reinforcement learning to offer promising solutions to complex decision-making problems across various domains, framing generative AI's integration into this existing DRL paradigm as an extension of, rather than a replacement for, established deep-RL methodology [9]. Complex, high-dimensional decision-making trajectories are also increasingly modeled as sequence-generation problems directly: research on adaptive planning with generative models under uncertainty traced this development to work reimagining reinforcement learning as a sequence modeling problem, employing a Transformer architecture to model distributions over state trajectories, a conceptual shift away from conventional reinforcement-learning techniques that primarily concentrate on estimating value functions or policy gradients directly [10]. The same review specifically highlighted the Diffuser planning framework, which differs from traditional model-based planning by predicting entire trajectories at once rather than incrementally, a design choice that enhances scalability specifically for long-horizon planning problems where classical step-by-step value-based planning methods struggle with compounding prediction error [10].

2.5 Advanced Policy-Optimization Algorithms for Diffusion-Based Policies

Building on the DDPO and DPPO foundations reviewed in Section 2.1, recent research has developed increasingly sophisticated policy-optimization algorithms specifically addressing the training instability and computational cost that naive reinforcement-learning fine-tuning of diffusion policies otherwise incurs. Decision Flow Policy Optimization applied flow-based generative models specifically, rather than diffusion models, to reinforcement learning, enabling the effective modeling of complex multi-modal action distributions and achieving superior robotic control in continuous action spaces, while noting that previous methods usually adopt generative models purely as behavior models fitting state-conditioned action distributions from datasets, with policy optimization conducted separately through additional value-based sample weighting or gradient-based updates [11]. A study on Forward KL Regularized Preference Optimization for aligning diffusion policies observed that conventional reinforcement-learning algorithms for sequential decision-making problems typically adopt Gaussian or deterministic policy classes, which, although successful on various challenging tasks, can be limited in learning multi-modal policies and lead to sub-optimal behaviors in complex environments, directly motivating diffusion-based policy classes as a solution to this expressiveness limitation [12].

A recently proposed Dichotomous Diffusion Policy Optimization (DIPOLE) framework addressed a specific training-stability trade-off recurring across this literature: existing methods either suffer from unstable training due to directly maximizing value objectives, or face computational issues due to relying on crude Gaussian likelihood approximations that require a large number of sufficiently small denoising steps [13]. DIPOLE's proposed solution revisited the KL-regularized objective in reinforcement learning, which offers a desirable weighted-regression objective for diffusion-policy extraction but often struggles to balance greediness (aggressive value maximization) and stability (bounded deviation from the reference policy), illustrating that the field's current research frontier centers specifically on managing this greediness-stability trade-off within diffusion-based policy optimization [13].

2.6 Domain-Specific Applications: Autonomous Driving and Robotic Motion Planning

Autonomous driving and robotic motion planning have emerged as particularly active application domains for generative-AI-based decision-making frameworks, precisely because these domains require both the multi-modal trajectory expressiveness diffusion models provide and the capacity to optimize genuinely non-differentiable safety and effectiveness objectives. Research on non-differentiable reward optimization for diffusion-based autonomous motion planning observed that critical objectives for planning related to safety and effectiveness are non-differentiable, and cannot be explicitly optimized with conventional training schemes, motivating a reinforcement-learning approach built directly upon the DDPO framework that integrates non-differentiable reward signals into diffusion-based planners by modeling the diffusion generation process itself as a Markov decision process [14]. This direct extension of DDPO's core MDP reformulation from image generation to safety-critical motion planning illustrates the framework's methodological generality across substantially different downstream domains.

The Gen-Drive framework extended this generation-then-evaluation paradigm specifically to autonomous driving, proposing a scene-evaluator reward model trained with pairwise preference data collected through vision-language-model assistance, thereby reducing human labeling workload, combined with a reinforcement-learning fine-tuning framework to improve the generation quality of the underlying diffusion model for planning tasks specifically [15]. Evaluated on the nuPlan closed-loop planning dataset, the study found that this generation-then-evaluation strategy outperforms other learning-based approaches, and specifically that the learned reward model, when used for either evaluation or RL fine-tuning, leads to better planning performance compared to relying on human-designed reward functions [15]. This finding, that learned reward models can outperform hand-designed rewards even in a safety-critical planning domain, is significant because it suggests the generative-RL paradigm's advantage extends beyond policy expressiveness to reward specification itself, a task classical control theory has traditionally required domain experts to design manually.

2.7 Generative AI Decision-Making Beyond Robotics: Health and Complex Service Systems

The reviewed generative-RL and stochastic-control literature has concentrated heavily on robotics, autonomous driving, and continuous-control domains, but generative AI's role in adaptive decision-making extends to complex organizational and service-system contexts where the relevant "control" problem concerns human institutional decisions rather than robotic actuation. A rapid review of generative AI decision-making attributes in complex health services found that generative AI incorporates programming rules that are normative but also has a descriptive component based on its programmers' subjective preferences and any discrepancies in the underlying training data, and specifically observed that generative AI generates both truth and falsehood, supports both ethical and unethical decisions, and is neither fully transparent nor fully accountable [16]. This finding is directly relevant to the present review's stability-and-guarantee theme, extended from a control-theoretic to an institutional-decision-making register: just as the reviewed robotic and control-theoretic literature identifies formal stability guarantees as generative-RL's central unresolved gap (Section 2.3), this health-services review identifies transparency and accountability as the analogous unresolved gap for generative-AI-supported decisions in complex service systems such as health policy and health regulation [16].

2.8 Research Gap

While the generative-RL literature reviewed in Sections 2.1, 2.4, and 2.5 has established substantial empirical evidence for diffusion- and flow-based policies' superior expressiveness relative to classical Gaussian policy classes, and the stochastic-optimal-control literature reviewed in Sections 2.2 and 2.3 has established formal, Lyapunov-based stability guarantees for RL-augmented nonlinear stochastic control, comparatively few studies bridge these two bodies of work directly: no work reviewed here establishes a formal stability guarantee, analogous to the stochastic control Lyapunov function approach [8], for a diffusion-or flow-based generative policy of the DDPO, DPPO, or DIPOLE type. This asymmetry, rigorous stability theorems for value-function-based nonlinear stochastic control on the one hand, purely empirical validation for the generative policy architectures that increasingly dominate high-expressiveness robotic and planning applications on the other, constitutes the central research gap this review identifies.

III. METHODOLOGY

This study adopts a qualitative, descriptive, and comparative research design synthesizing peer-reviewed and archival literature on generative AI, stochastic optimal control, and reinforcement learning. Reviewed works were compared along four axes: methodological tradition (classical/deep-learning stochastic optimal control, generative-model-based policy optimization, or hybrid RL-training-loop integration); generative architecture examined (diffusion model, flow-based model, or transformer sequence model); application domain (robotics, autonomous driving, text-to-image/creative generation, energy systems, or health/institutional decision-making); and guarantee type (formal stability theorem, monotone-improvement proof, or empirical benchmark validation only).

IV. Results and Discussion

4.1 Generative-RL Frameworks Mapped to Mechanism and Application Domain

Table 1 maps the principal generative-AI-and-reinforcement-learning frameworks reviewed onto their core algorithmic mechanism and primary application domain.

Table 1. Generative-RL Frameworks Mapped to Mechanism and Application Domain

Framework	Core Mechanism	Application Domain	Key Reference
DDPO	Denoising step as MDP action; policy-gradient fine-tuning	Text-to-image generation, aesthetic/alignment objectives	Black et al., DDPO [3]
DPPO	Policy-gradient fine-tuning of diffusion-based robot policies	Continuous control, robot learning	Diffusion Policy Policy Optimization [5]
Decision Flow Policy Optimization	Flow-based (Jacobian-free) generative policy modeling	Robotic control, multi-modal action distributions	Decision Flow Policy Optimization [11]
DIPOLE	KL-regularized dichotomous policy improvement	Stable diffusion-policy fine-tuning	Dichotomous Diffusion Policy Optimization [13]
Diffuser / sequence-modeling planners	Transformer/diffusion trajectory-level prediction	Long-horizon planning	Adaptive planning under uncertainty review [10]
Gen-Drive	Generation-then-evaluation with VLM-assisted reward model	Autonomous driving (nuPlan)	Gen-Drive [15]
Non-differentiable reward diffusion planner	DDPO-based MDP reformulation for safety/effectiveness rewards	Autonomous motion planning	Non-differentiable reward optimization [14]

4.2 Stability and Guarantee Type Mapped to Methodological Tradition

Table 2 consolidates the guarantee type each methodological tradition provides, directly illustrating the gap identified in Section 2.8.

Table 2. Guarantee Type Mapped to Methodological Tradition

Tradition	Guarantee Type	Formal Mechanism	Key Reference
Classical/deep-learning stochastic optimal control	Value-function/PDE-based optimality	Hamilton-Jacobi-Bellman approximation via deep learning	Han & E [6]; On-policy deep learning framework [7]
RL for nonlinear stochastic control	Formal stability-in-probability guarantee	Stochastic Control Lyapunov Function (SCLF)	Zhu, RL for stochastic nonlinear systems [8]
DDPO / diffusion policy optimization	Empirical benchmark validation only	Policy-gradient on exact per-step likelihoods	Black et al. [3]; DPPO [5]
Flow-based / dichotomous policy optimization	Empirical benchmark + training-stability heuristics	KL-regularized weighted regression	Decision Flow [11]; DIPOLE [13]
Generation-then-evaluation (autonomous driving)	Empirical closed-loop planning validation	VLM-assisted preference-based reward model	Gen-Drive [15]

4.3 Structural Advantages of Generative Policies Over Classical Policy Classes

Table 3 summarizes the specific structural advantages the reviewed literature attributes to generative policy architectures relative to classical Gaussian or deterministic policy classes.

Table 3. Generative Policy Advantages Over Classical Policy Classes

Limitation of Classical Policies	Generative-Model Solution	Reported Benefit	Key Reference
Gaussian policies represent only unimodal action distributions	Diffusion/flow-based multi-modal action modeling	Superior robotic control in continuous action spaces	Duality of GenAI and RL review [4]; Decision Flow [11]
Non-differentiable objectives (safety, aesthetics, alignment) cannot be optimized via standard gradients	Denoising-as-MDP reformulation enables policy-gradient RL	Optimization of black-box, human-feedback-derived rewards	DDPO [3]; Non-differentiable reward planner [14]
Step-by-step value-based planning suffers compounding prediction error over long horizons	Whole-trajectory generative prediction (Diffuser-style)	Enhanced scalability for long-horizon planning	Adaptive planning review [10]
Hand-designed reward functions require domain expertise and may be sub-optimal	Learned, preference-based scene-evaluator reward models	Outperforms human-designed rewards in closed-loop tests	Gen-Drive [15]
Sparse rewards and costly state representations limit RL sample efficiency	Generative models for reward signal / state representation	Improved sample efficiency, generalization to new tasks	Duality of GenAI and RL review [4]

The consolidated evidence across Tables 1–3 indicates that generative AI's integration into reinforcement learning and stochastic optimal control has produced substantial, empirically well-documented gains in policy expressiveness and the capacity to optimize non-differentiable objectives, but that these gains have so far been achieved almost entirely without the formal stability guarantees classical stochastic optimal control theory has long provided for value-function-based nonlinear control [8]. This asymmetry is particularly consequential for safety-critical adaptive decision-making systems, autonomous driving, industrial process control, and health-service decision support, where the reviewed literature's own governance-oriented findings caution that generative AI is neither transparent nor accountable by default [16], underscoring why the Lyapunov-based stability approach developed for RL-augmented nonlinear stochastic control [8] represents a methodologically important, if currently narrow, counterexample to this general pattern.

V. FUTURE RESEARCH DIRECTIONS

Three priorities emerge from the reviewed literature. First, extending the stochastic control Lyapunov function methodology developed for value-function-based nonlinear stochastic control [8] to diffusion- and flow-based generative policies represents the field's most consequential open mathematical problem; a plausible first step would involve characterizing the effective value function implicitly induced by a DDPO- or DPPO-style denoising policy and asking whether a Lyapunov certificate can be constructed over that induced value function. Second, the generation-then-evaluation paradigm validated for autonomous driving [15] should be extended to other safety-critical domains, such as the chemical-process-control applications already validated for Lyapunov-based RL [8], to directly test whether learned, preference-based reward models can match or exceed hand-designed rewards in domains where the stability-guarantee literature is comparatively more mature. Third, given the health-services literature's finding that generative AI decision-making lacks default transparency and accountability [16], future research should develop interpretability and auditability tools specifically for the MDP-reformulated denoising process that underlies DDPO and its successors, since the per-step action structure DDPO already exposes for policy-gradient training [3] may offer a natural, underexploited entry point for step-level auditability that end-to-end generative models otherwise lack.

VI. CONCLUSION

This review has examined the intersection of generative AI, stochastic optimal control, and reinforcement learning for adaptive and intelligent decision-making systems, synthesizing the foundational reformulation of diffusion-model denoising as a sequential Markov decision process with classical and RL-augmented stochastic-optimal-control research and domain-specific generative-RL applications in robotics and autonomous driving. The reviewed evidence establishes that framing generative sampling processes as sequential decision problems allows policy-gradient reinforcement learning to directly optimize non-differentiable, black-box, and human-preference-derived objectives that classical gradient-based training of generative models cannot address, and that the resulting diffusion- and flow-based policy architectures offer substantial expressiveness advantages, particularly multi-modal action-distribution modeling, over classical Gaussian policy classes. However, this expressiveness advantage has been achieved largely without the formal stability and safety guarantees long established in classical stochastic optimal control theory, where Lyapunov-function-based approaches already provide rigorous stability-in-probability certificates for RL-augmented nonlinear stochastic control. Bridging this guarantee gap, extending formal stability theory to generative policy architectures, represents the central unresolved challenge for deploying generative-AI-based decision-making frameworks in the safety-critical adaptive control settings where such guarantees matter most.

REFERENCES

[1] W. B. Powell, *Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions*. Hoboken, NJ: Wiley, 2022.
[2] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, vol. 1. New York: Springer Science & Business Media, 2012.
[3] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, "Training Diffusion Models with Reinforcement Learning (DDPO)," *arXiv preprint arXiv:2305.13301*, 2023.
[4] "The Duality of Generative AI and Reinforcement Learning in Robotics: A Review," *arXiv preprint arXiv:2410.16411*, 2024.
A. Ren et al., "Diffusion Policy Policy Optimization (DPPO)," *arXiv preprint arXiv:2409.00588*, 2024.
[5] J. Han and W. E, "Deep learning approximation for stochastic control problems," *arXiv preprint arXiv:1611.07422*, 2016.
[6] "An Efficient On-Policy Deep Learning Framework for Stochastic Optimal Control," *arXiv preprint arXiv:2410.05163*, 2024.
[7] Zhu et al., "Reinforcement learning for optimal control of stochastic nonlinear systems," *AIChE Journal*, vol. 71, no. 7, e18840, 2025.
[8] "Generative AI for Deep Reinforcement Learning: Framework, Analysis, and Use Cases," *arXiv preprint arXiv:2405.20568*, 2024.
[9] "Adaptive Planning with Generative Models under Uncertainty," *arXiv preprint arXiv:2408.01510*, 2024.

[10] J. Hu, S. Huang, S. Guo, Z. Liu, L. Shen, L. Sun, H. Chen, Y. Chang, and D. Tao, "Decision Flow Policy Optimization," *arXiv preprint arXiv:2505.20350*, 2025.
[11] "Forward KL Regularized Preference Optimization for Aligning Diffusion Policies," *arXiv preprint arXiv:2409.05622*, 2024.
[12] "Dichotomous Diffusion Policy Optimization (DIPOLE)," *arXiv preprint arXiv:2601.00898*, 2026.
[13] "Non-differentiable Reward Optimization for Diffusion-based Autonomous Motion Planning," *arXiv preprint arXiv:2507.12977*, 2025.
[14] "Gen-Drive: Enhancing Diffusion Generative Driving Policies with Reward Modeling and Reinforcement Learning Fine-tuning," *arXiv preprint arXiv:2410.05582*, 2024.
[15] N. Doreswamy and L. Horstmanshof, "Generative AI Decision-Making Attributes in Complex Health Services: A Rapid Review," *Cureus*, vol. 17, no. 1, e78257, 2025.
[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA: MIT Press, 2018.
[17] S. Meyn, *Control Systems and Reinforcement Learning*. Cambridge, U.K.: Cambridge University Press, 2022.
[18] J. Han, A. Jentzen, and W. E, "Solving high-dimensional partial differential equations using deep learning," *Proceedings of the National Academy of Sciences*, vol. 115, no. 34, pp. 8505–8510, 2018.
[19] J. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis (Diffuser)," in *Proc. Int. Conf. Machine Learning (ICML)*, 2022.
[20] L. Chen et al., "Decision transformer: Reinforcement learning via sequence modeling," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
[21] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[22] Ajay, Y. Du, A. Gupta, J. B. Tenenbaum, T. S. Jaakkola, and P. Agrawal, "Is conditional generative modeling all you need for decision making?," arXiv preprint arXiv:2211.15657, 2022.
[23] Chi et al., "Diffusion policy: Visuomotor policy learning via action diffusion," in Proc. Robotics: Science and Systems (RSS), 2023.
[24] Silver, S. Singh, D. Precup, and R. S. Sutton, "Reward is enough," Artificial Intelligence, vol. 299, 2021.