ITERATIVE SCARCITY-GUIDED EXPLORATION: BOOTSTRAPPING GENERATIVE AUTO-BIDDING FROM NARROW SUPPORT

Qingmao Yao[1], **Guangzheng Hu**[2], **Yusen Huo**[3]*, **Zhilin Zhang**[2], **Chuan Yu**[2], **Jian Xu**[2], **Bo Zheng**[2], **Xiaotie Deng**[1]†
[1]Center on Frontiers of Computing Studies, School of Computer Science, Peking University
[2]Alimama Tech, Taobao & Tmall Group of Alibaba
[3]Independent Researcher

ABSTRACT

Auto-bidding is a core algorithmic component in large-scale auctions. In industrial cold-start settings, iterative training is commonly used to compensate for low-quality, narrowly supported historical data. Representative AI-Generated Bidding (AIGB) methods based on conditional diffusion planners have shown strong empirical performance. However, diffusion planners fail to sustain improvement in iterative training, becoming trapped in a suboptimal performance bottleneck. In this paper, we theoretically attribute this stagnation to the collapse of the effective signal-to-noise ratio (SNR): the sparse radial return signal is overwhelmed by a curvature-induced penalty, preventing extrapolation beyond the narrow support. To break this limit, we propose **Iterative Scarcity-Guided Exploration (ISGE)**, which alternates between a *Judger* that detects high-scarcity trajectories and an *Explorer* that performs guided generation toward high-scarcity and high-return regions. We theoretically show that ISGE elevates the effective SNR above its critical threshold and reshapes the optimization landscape, enabling the policy to extrapolate beyond the support. Extensive experiments on the industrial AuctionNet benchmark demonstrate that ISGE effectively bootstraps from narrow support and surpasses the performance of full-dataset baselines within three iterations.

1 INTRODUCTION

Online advertising has become indispensable for business growth, yet its sheer scale with trillions of daily impressions renders manual bidding impractical, making *auto-bidding* the industry standard Evans (2009). Fundamentally, auto-bidding is a real-time long-horizon sequential decision-making problem, where an agent must maximize returns (e.g., Gross Merchandise Value, GMV) under non-stationary market dynamics and strict budget constraints Zhang et al. (2014); He et al. (2021). Recently, advances in generative modeling have driven a paradigm shift from traditional optimization-based approaches to trajectory-generation frameworks capable of non-Markovian long-horizon planning. DiffBid Guo et al. (2024), the first AI-Generated Bidding (AIGB) method, has demonstrated superior performance in industrial deployments using a conditional *diffusion planner*.

However, the success of AIGB largely relies on the availability of high-quality expert demonstrations with broad data coverage. In many practical industrial scenarios, such as *cold-start phases* for new advertisers or emerging business sectors, this requirement is violated. Instead, models must learn from low-quality historical logs generated by conservative rule-based or manual strategies. Such data is inherently narrowly supported, lacking both expert demonstrations and diverse off-support trajectories needed for extrapolation. While online reinforcement learning (RL) may appear as a naive solution, online trial-and-error in advertising markets leads to severe financial risk, confining training to the offline setting. To mitigate the limitations of static offline training, *iterative training* is

*This work was completed while the author was at Alibaba.
†Corresponding Author. Email: xiaotie@pku.edu.cn.

widely adopted in practice, where a policy is repeatedly updated by generating new trajectories and retraining on the augmented dataset. Prior works on RL-based bidding Mou et al. (2022); Li et al. (2024) have explored iterative frameworks, but their exploration strategies rely on undirected noise, which are inefficient in the high-dimensional auction space. Conversely, current AIGB methods like DiffBid are predominantly static, leaving the potential of iterative generative exploration largely untapped. This raises a central question:

Can generative methods perform efficient exploration, bootstrapping a near-optimal policy solely from low-quality, narrowly supported data in just a few iterations?

In investigating this question, we observe that naively applying iterative self-training to diffusion planners reveals a distinct dual-phase pattern. Initially, diffusion planners demonstrate impressive extrapolation capability, benefiting from stochastic sampling and guidance mechanisms such as Classifier-Free Guidance (CFG). However, as training progresses, performance inevitably saturates at a suboptimal level, entering a bottleneck we term *Self-Locking*. We theoretically attribute this phenomenon to a fundamental geometric barrier in high-dimensional optimization: as the policy improves and enters out-of-distribution (OOD) regions, the effective signal-to-noise ratio (SNR) collapses because the weak radial component of the return gradient is insufficient to overcome the high-dimensional curvature penalty (i.e., entropic noise and tangential drift). This prevents sustained extrapolation beyond the current support. To overcome SNR collapse, we propose **Iterative Scarcity-Guided Exploration (ISGE)** by introducing scarcity as a geometric exploration signal that remains informative even in OOD regions. It alternates between a *Judger* that detects high-scarcity trajectories via reconstruction error and an *Explorer* that performs guided generation via CFG conditioned on both return and scarcity to seek high-scarcity, high-return "gift samples". We theoretically show that ISGE elevates the effective SNR above the critical locking threshold, reshapes the optimization landscape, and enables diffusion planners to extrapolate beyond narrow support. Extensive experiments on the industrial-scale AuctionNet benchmark demonstrate that ISGE effectively bridges the quality gap: even when restricted to low-quality, narrowly supported data (bottom 10th-30th return), ISGE breaks the self-locking limit and eventually surpasses the Full-Dataset Oracle *within three iterations*.

To summarize, our main contributions are:

- We theoretically attribute self-locking to the collapse of the effective SNR and prove that scarcity guidance elevates the effective SNR for OOD extrapolation.
- We propose ISGE, an efficient exploration framework that leverages scarcity to actively discover and aggregate "gift samples" beyond the training support.
- We demonstrate that ISGE bootstraps near-optimal policies from low-quality, narrowly supported data, achieving superior performance on industrial-scale benchmark AuctionNet.

2 PRELIMINARIES

Generative Auto-Bidding Auto-bidding in Real-Time Bidding (RTB) represents a canonical *long-horizon, sparse reward, non-Markovian sequential decision problem*. Over campaigns spanning thousands of auctions, the agent must maximize return by dynamically allocating a finite budget B while balancing ROI constraints. This creates a high-dimensional planning challenge: irreversible budget depletion, and the need to balance immediate opportunity against long-term constraint satisfaction. To tackle this, the representative AIGB framework DiffBid Guo et al. (2024) employs a conditional diffusion planner that directly models the distribution of feasible trajectories $p(\tau \mid y)$, where $\tau = (s_0, \dots, s_T)$ is a state trajectory and $y = (\text{Return}, \text{Budget}, \text{ROI Constraint}, \dots)$, then extracts actions via an inverse dynamics model $a_t = \phi(s_t, s_{t+1})$. Rather than solving a constrained optimization problem explicitly, the planner implicitly satisfies the objectives by learning to generate state trajectories τ that are consistent with the specified condition y. In this paper, we focus on the diffusion planner, which serves as the core component of this framework.

Conditional Diffusion Models and Classifier-Free Guidance Diffusion probabilistic models generate data by reversing a progressive noise-injection process. We focus on the score-based interpretation, where a parameterized network $\epsilon_\theta(x_t, t)$ approximates the score function $\nabla_{x_t} \log p(x_t)$ to denoise a latent variable $x_T \sim \mathcal{N}(0, I)$ into a clean trajectory τ. To steer generation towards high-value regions, we employ Classifier-Free Guidance (CFG) Ajay et al. (2022). Unlike classifier guidance which requires an auxiliary noisy regressor, CFG jointly trains a conditional model $\epsilon_\theta(x_t, t, y)$

and an unconditional counterpart $\epsilon_\theta(x_t, t, \emptyset)$ via random condition dropout. During sampling, the modified score estimate $\tilde{\epsilon}$ is given by: $\tilde{\epsilon}(x_t, t, y) = \epsilon_\theta(x_t, t, \emptyset) + w \cdot (\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t, \emptyset))$, where w is the guidance scale. This formulation is pivotal to our analysis: the term $(\epsilon_{cond} - \epsilon_{uncond})$ acts as an implicit gradient ascent step $\nabla_{x_t} \log p(y|x_t)$ on the return landscape. We will later theoretically show that in high-dimensional OOD regions, this gradient becomes unreliable and leads to SNR collapse.

3 Motivation and Theoretical Analysis

In this section, we first present a motivating example that empirically reveals the dual-phase pattern of iterative diffusion planner: the initial extrapolation capability, the subsequent self-locking bottleneck, and the efficacy of *scarcity-guided exploration* in breaking this limit. We then establish a theoretical framework to uncover the mechanisms behind these phenomena. All proofs are deferred to Appendix A.

A Motivating Example To investigate the boundaries of iterative diffusion planner, we design a high-dimensional (Horizon $= 96$) trajectory optimization task using real-world auto-bidding trajectories. The return of each trajectory is defined via an exponential transformation of its negative ℓ_1 distance to the optimal trajectory which is *not* in the training support. We simulate a cold-start scenario using a Narrowly Supported Dataset (max return ≤ 0.30) and compare the resulting planner against an Oracle trained on the Full Dataset (max return ≈ 0.90). Figure 1 illustrates the performance evolution, revealing several critical insights. When trained on the Narrowly Supported Dataset, the Vanilla self-training method achieves a remarkable leap in the first iteration (Avg: $0.34 \rightarrow 0.65$), demonstrating the power of generative extrapolation via CFG. However, this momentum is short-lived. The Vanilla framework hits a hard ceiling at Iteration 2, stagnating around 0.76. Crucially, since the successful Oracle employs the identical architecture and hyperparameters, this stagnation *cannot* be attributed to limited model capacity or insufficient denoising steps. Instead, it highlights a bottleneck in the structural information deficiency of the narrowly supported data, which vanilla self-training cannot overcome. We term this phenomenon "self-locking". In contrast, ISGE successfully breaks this lock and sustains improvement (Avg: $0.76 \rightarrow 0.90$), effectively matching the Oracle.

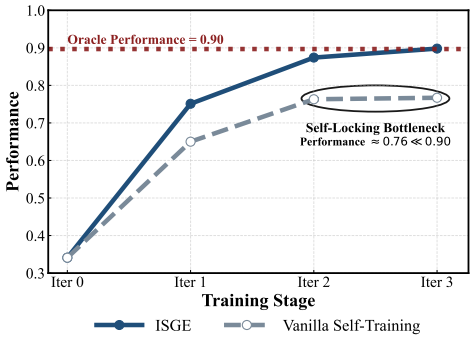


Figure 1: ISGE vs Vanilla Self-training

Theoretical Roadmap. These empirical findings raise three fundamental questions that we address in the subsequent analysis:

1. *What drives the initial generative extrapolation?*
2. *Why does self-locking occur in high-return regions?*
3. *How does ISGE theoretically guarantee improvement?*

3.1 What Drives the Initial Generative Extrapolation?

We theoretically attribute the extrapolation to the interaction between *intrinsic sampling stochasticity* and *implicit gradient guidance*. To delineate this mechanism, we first characterize the mathematical equivalent of Classifier-Free Guidance (CFG) in the score space.

Lemma 3.1 (CFG as Implicit Gradient)**.** *Let $\epsilon_\theta(x_t)$ denote the unconditional noise prediction. Leveraging the score matching property, the modified noise prediction $\tilde{\epsilon}_\theta(x_t)$ under CFG with scale w can be reformulated as:*

$$\tilde{\epsilon}_\theta(x_t) \approx \epsilon_\theta(x_t) - w\sqrt{1-\bar{\alpha}_t} \cdot \nabla_{x_t} \log p(y|x_t). \tag{1}$$

Lemma 3.1 reveals that CFG effectively injects a gradient term derived from the implicit discriminator $\log p(y|x)$ into the denoising process. By substituting this modified noise into the reverse diffusion step, we derive the physical dynamics of generation.

Proposition 3.2 (CFG as Biased Langevin Dynamics)**.** *By substituting the CFG-modified noise prediction $\tilde{\epsilon}_\theta$ into the reverse diffusion process, the sampling update at step t can be reformulated as a biased Langevin update:*

$$x_{t-1} \leftarrow \mu_\theta(x_t) + \underbrace{\lambda_t w \cdot \nabla_{x_t} \log p(y|x_t)}_{\textit{Guidance Drift}} + \underbrace{\sigma_t z}_{\textit{Noise}}, \tag{2}$$

where $\mu_\theta(x_t)$ is the standard mean reconstruction term, $z \sim \mathcal{N}(0, I)$ is the sampling stochasticity, and $\lambda_t = \frac{1-\alpha_t}{\sqrt{\alpha_t}}$ is a positive coefficient derived from the noise schedule.

Proposition 3.2 identifies the two forces driving extrapolation: The Gaussian noise (z) enables the trajectory to deviate from the exact training support. The guidance drift ($\nabla \log p(y|x)$) creates a momentum overshoot that biases the random walk towards high-value regions.

3.2 Why Does Self-Locking Occur?

As the iteration progresses and returns improve, the optimization problem shifts from traversing a broad slope to pinpointing a specific high-dimensional region. We characterize this as a collapse of the *Effective Signal-to-Noise Ratio (SNR$_{eff}$)* in exploration. To facilitate our analysis, we introduce a geometric abstraction of the optimization landscape in Section 3. We model the level set of trajectories $\mathbf{x} \in \mathbb{R}^D$ achieving return $R(\mathbf{x}) \approx r$ as a thin hyperspherical shell with radius $\epsilon(r)$ centered at the global optimum $\mathbf{x}^*$[1]. (Bold symbols denote vectors in a high-dimensional space) As $r \to R_{\max}$, the radius $\epsilon(r) \to 0$. We now analyze how this geometric contraction leads to signal collapse.

Theorem 3.3 (SNR Barrier in High-Dimensional Optimization)**.** *Consider a point $\mathbf{x}$ at distance $\epsilon = \|\mathbf{x}-\mathbf{x}^*\|$ from the optimum $\mathbf{x}^*$, and an update $\Delta\mathbf{x} = \mathbf{v}_{\text{det}} + \mathbf{v}_{\text{noise}}$, where $\mathbf{v}_{\text{det}}$ is a deterministic drift and $\mathbf{v}_{\text{noise}} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is isotropic noise. Let $\mathbf{n}^* = (\mathbf{x}^* - \mathbf{x})/\|\mathbf{x}^* - \mathbf{x}\|$ denote the radial unit vector pointing from the current point $\mathbf{x}$ toward the optimum $\mathbf{x}^*$. Let $v_{\text{rad}} = \langle \mathbf{v}_{\text{det}}, \mathbf{n}^* \rangle$ be the radial component of the drift, and $\|\mathbf{v}_{\tan}\|^2 = \|\mathbf{v}_{\text{det}} - v_{\text{rad}}\mathbf{n}^*\|^2 + D\sigma^2$ its total tangential energy. Then, the expected change in distance satisfies*

$$\mathbb{E}[\Delta\epsilon] \approx \underbrace{-v_{\text{rad}}}_{\textit{Signal}} + \underbrace{\frac{\|\mathbf{v}_{\tan}\|^2}{2\epsilon}}_{\textit{Penalty}}. \tag{3}$$

Consequently, distance reduction ($\mathbb{E}[\Delta\epsilon] < 0$) requires,

$$v_{\text{rad}} > \frac{\|\mathbf{v}_{\tan}\|^2}{2\epsilon} \iff \textit{SNR}_{\textit{eff}} = \frac{v_{\text{rad}}}{\|\mathbf{v}_{\tan}\|^2/(2\epsilon)} > 1. \tag{4}$$

This theorem reveals a fundamental trade-off: optimization succeeds only when the radial signal dominates the curvature-induced penalty (*SNR$_{eff}$* > 1). Since signal magnitude v_{rad} typically does not scale with the dimension D, the effective SNR scales as *SNR$_{eff}$* $\propto \epsilon/D$. This implies that fine-grained optimization (i.e., at small ϵ) becomes increasingly difficult in high-dimensional spaces (at large D). The following proposition shows how this signal degradation leads to self-locking of *diffusion planner*: even with optimal guidance scale, the planner stagnates when *SNR$_{eff}$* falls below unity.

Proposition 3.4 (Return Signal Collapse)**.** *Consider a trajectory $\mathbf{x}$ at distance $\epsilon = \|\mathbf{x} - \mathbf{x}^*\|$ from the OOD optimum $\mathbf{x}^*$ which lies outside the suboptimal training support, and the planner update $\Delta\mathbf{x}_t = \lambda_t w_R \mathbf{g}_R + \sigma_t \mathbf{z}$, where $\mathbf{g}_R = \nabla_{\mathbf{x}} \log p(y_R \mid \mathbf{x})$ and y_R is return condition. Assume the guidance scale $w_R > 0$ is chosen optimally. Then, the optimization stagnates ($\mathbb{E}[\Delta\epsilon] \geq 0$) when the radial signal falls below the threshold:*

$$\langle \mathbf{g}_R, \mathbf{n}^* \rangle \leq \frac{\sigma_t \sqrt{D}\, \|\mathbf{g}_{R,\tan}\|}{\epsilon}, \tag{5}$$

where $\mathbf{g}_{R,\tan} := \mathbf{g}_R - \langle \mathbf{g}_R, \mathbf{n}^* \rangle \mathbf{n}^*$. Consequently, there exists a minimal achievable radius $\epsilon_{\min} > 0$ such that further reduction of ϵ is statistically impossible under vanishing radial signal outside the suboptimal training support.

[1]This spherical model serves as a conservative proxy for general local optima. See Appendix B for justification in general smooth landscapes.

This threshold characterizes a fundamental limitation of return-guided diffusion in OOD regions. As the trajectory explores beyond the suboptimal training support, the return signal becomes increasingly scarce and misaligned, causing the radial component $\langle \mathbf{g}_R, \mathbf{n}^* \rangle$ to decay. Once this signal falls below the threshold required to overcome the curvature penalty induced by both tangential drift and isotropic noise, the optimization halts. That's why the diffusion planner with Gaussian noise cannot sustain improvement in OOD exploration.

3.3 How Does Scarcity-Guided Exploration Work?

In OOD regions, the return signal becomes increasingly sparse and misaligned, causing the diffusion planner to lose directional guidance, leading to self-locking in Proposition 3.4. To break this stagnation, we require a robust auxiliary signal that persists in the OOD vacuum and explicitly provides a non-vanishing outward impulse. Identifying scarcity as the geometric metric that naturally satisfies these requirements, we propose ISGE to introduce it as an auxiliary signal that remains geometrically coherent at and beyond the boundary of the training support.

Proposition 3.5 (Scarcity as Manifold Repulsion Field)**.** *Let the scarcity score $S(\mathbf{x})$ be defined as the reconstruction error of a denoising autoencoder (DAE) trained on data from manifold $\mathcal{M}$ with small isotropic corruption noise. The ISGE framework utilizes an implicit scarcity gradient $\mathbf{g}_S = \alpha(S) \cdot \nabla_x S(x), (\alpha(S) > 0)$ via CFG. Following the theory of DAE Alain & Bengio (2014), the gradient $\mathbf{g}_S$ satisfies:*

1. ***Normal Alignment:*** *$\mathbf{g}_S$ is approximately aligned with the outward-pointing unit normal vector $\mathbf{n}_\perp$ of $\mathcal{M}$.*
2. ***Non-Vanishing Magnitude:*** *There exists $\delta > 0$ such that $\|\mathbf{g}_S\| \geq \delta$ for all x in a neighborhood of boundary $\partial\mathcal{M}$ outside $\mathcal{M}$.*

Proposition 3.5 establishes that, unlike the return gradient which collapses in OOD space, the scarcity gradient $\mathbf{g_s}$ maintains a stable outward orientation and non-vanishing magnitude near the data manifold boundary. This geometric reliability makes it an ideal candidate to counteract the curvature-induced stagnation of diffusion dynamics.

Theorem 3.6 (Revival of Optimization via Scarcity)**.** *Consider the trajectory $\mathbf{x}$ in Proposition 3.4 satisfying the bottleneck condition $\mathbb{E}[\Delta\epsilon]_{base} = 0$. We apply the dual-guided Langevin update in ISGE: $\Delta\mathbf{x}_t = \lambda_t(w_R\mathbf{g}_R + w_S\mathbf{g}_S) + \sigma_t\mathbf{z}$. Under the geometry assumptions of the scarcity guidance (detailed in Appendix A.5), the optimization is revived ($\mathbb{E}[\Delta\epsilon]_{ISGE} < 0$) if the scarcity guidance scale satisfies:*

$$w_S \in (0, \frac{2\epsilon\langle \mathbf{g}_S, \mathbf{n}^* \rangle}{\lambda_t \|\mathbf{g}_{S,\tan}\|^2}) \tag{6}$$

Furthermore, by choosing the optimal guidance scale $w_S^* = \epsilon\langle \mathbf{g}_S, \mathbf{n}^* \rangle / \lambda_t \|\mathbf{g}_{S,\tan}\|^2$, the theoretical marginal effective SNR of the scarcity guidance is exactly 2, (i.e., $SNR_{\Delta eff}(w_S^*) = 2$), yielding a maximum optimization progress rate:

$$\max_{w_S} (-\mathbb{E}[\Delta\epsilon]_{ISGE}) = \frac{\epsilon}{2} \left(\frac{\langle \mathbf{g}_S, \mathbf{n}^* \rangle}{\|\mathbf{g}_{S,\tan}\|} \right)^2. \tag{7}$$

Theorem 3.6 quantifies this intuition: by injecting scarcity guidance with an appropriate scale w_S, ISGE provably breaks the self-locking barrier, enabling the discovery of high-return and high scarcity samples termed "gift samples" ($\mathcal{D}_{\text{gift}}$). However, relying on inference-time guidance alone is computationally expensive and geometrically transient. The ultimate goal is to consolidate these sporadic discoveries into the model's intrinsic landscape. The following theorem demonstrates how fine-tuning on $\mathcal{D}_{gift}$ fundamentally alters the optimization landscape, creating a permanent basin of attraction that guides future generation without explicit scarcity aid.

Theorem 3.7 (Return Gradient Restoration)**.** *Let $\mathcal{D}_{gift}$ be the set of high-return OOD trajectories discovered via ISGE under the effective SNR condition (Theorem 3.6). Consider the gradient update induced by training on the $\mathcal{D}_{gift}$. For a test sample $\mathbf{x}$ in the neighborhood of $\mathcal{D}_{gift}$, let $\Delta\mathbf{g}_R(\mathbf{x}) := \mathbf{g}_{R,new}(\mathbf{x}) - \mathbf{g}_{R,old}(\mathbf{x})$ denote the contribution of $\mathcal{D}_{gift}$ to the return gradient field. We show that this update positively aligns with the true optimum: $\mathbb{E}\big[\langle \Delta\mathbf{g}_R(\mathbf{x}), \mathbf{n}^* \rangle\big] > 0$.*

Theorem 3.7 provides the theoretical imperative for mixing $\mathcal{D}_{\text{gift}}$ into the training pipeline. Since the return gradient from the original distribution collapses at the OOD evaluation sample $\mathbf{x}$ ($\langle \mathbf{g}_{R,\text{old}}, \mathbf{n}^* \rangle \approx 0$, as per Proposition 3.4), the update $\Delta \mathbf{g}_R$ is not merely an enhancement but a necessary geometric correction. It supplies the missing radial component that re-orients the optimization trajectory, transforming the model's behavior from stagnation to sustained extrapolation towards the global optimum.

4 Iterative Scarcity-Guided Exploration

Building upon the theoretical insights in Section 3, we present Iterative Scarcity-Guided Exploration (ISGE), a practical framework designed to overcome self-locking in iterative training. As illustrated in Figure 2, ISGE operates as a "Generate-Judge-Explore-Update" cycle advancing from iteration k to $k+1$, driven by two core components: (1) a Scarcity Judger (green panel) that evaluates trajectory scarcity $S(\tau)$ via masked reconstruction error; and (2) a Dual-Condition Explorer (red panel) that leverages this signal to generate $\mathcal{D}_{\text{scarcity}}$, aimed at discovering "gift samples" conditioned on both high return y_R and high scarcity y_S. To close the loop, these novel exploration samples are integrated with conservative exploitation data ($\mathcal{D}_{gen}$) to robustly update the policy π_k into π_{k+1}.

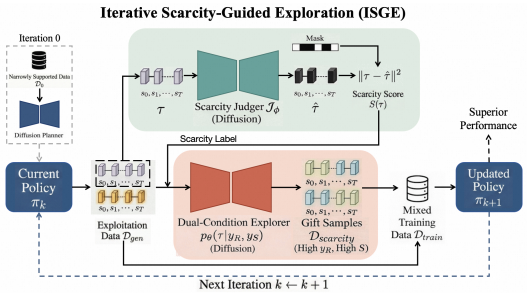


Figure 2: ISGE Framework

Evaluating Scarcity with Reconstruction Error To implement the scarcity guidance $\mathbf{g}_S$ in Proposition 3.5, we adopt a diffusion-based reconstruction error, leveraging its effectiveness in unsupervised anomaly detection Pang et al. (2021). We train an unconditional diffusion model, denoted as the *Judger* $\mathcal{J}_\phi$, to detect the scarcity of trajectories. For a generated trajectory τ, the scarcity score $S(\tau)$ is defined as the reconstruction residual: $S(\tau) = \|\tau - \hat{\tau}\|^2$, where $\hat{\tau}$ is the trajectory reconstructed by $\mathcal{J}_\phi$. Geometrically, a high $S(\tau)$ indicates that τ lies in the exterior of the manifold, thereby providing the necessary outward impulse to break self-locking. The distinction from recent exploration methods Wang et al. (2025); Ying et al. (2025) is discussed in the Related Work. To isolate strategic novelty from environmental randomness, we apply a masking operation $\text{Mask}(\cdot)$ that restricts $S(\tau)$ to strategic dimensions (e.g., remaining budget, historical bid prices), excluding environmental dimensions such as traffic fluctuations. This insight aligns with the feature space philosophy proposed in *Curiosity-driven Exploration* Pathak et al. (2017), which argues that an effective exploration signal should ignore uncontrollable environmental stochasticity.

Scarcity-Guided Exploration via Dual-Conditioning To implement the dual-guided Langevin dynamics derived in Theorem 3.6, we propose the *Dual-Condition Explorer*, a diffusion model $\epsilon_\theta(\tau_t, t \mid y_R, y_S)$ conditioned on both target return and scarcity. We employ disentangled Classifier-Free Guidance (CFG) to independently estimate the gradient fields. During sampling, we apply dual guidance to steer the generation:

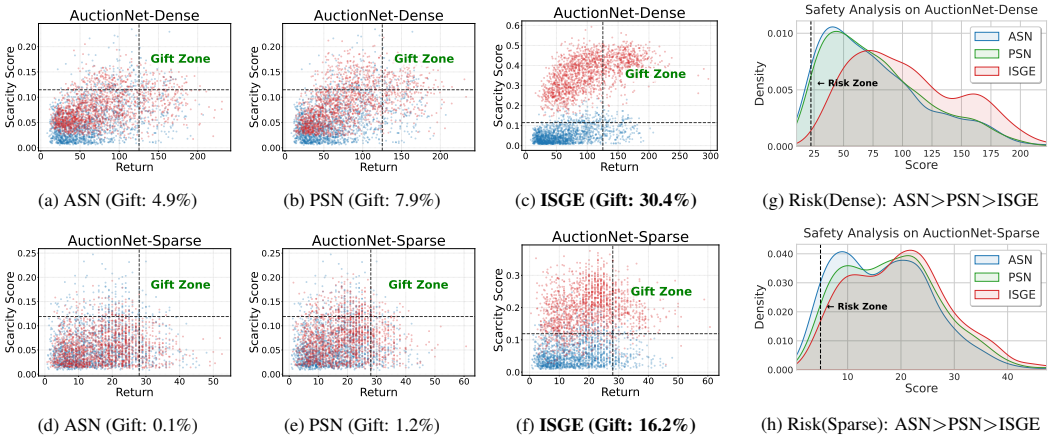
$$\tilde{\epsilon}_\theta(\tau_t) = \epsilon_{\text{uncond}} + w_R \underbrace{\left(\epsilon_\theta(\tau_t \mid y_R) - \epsilon_{\text{uncond}}\right)}_{\approx \mathbf{g}_R \text{ (Return Signal)}} + w_S \underbrace{\left(\epsilon_\theta(\tau_t \mid y_S) - \epsilon_{\text{uncond}}\right)}_{\approx \mathbf{g}_S \text{ (Scarcity Impulse)}}. \tag{8}$$

where w_R and w_S control the guidance strength. This formulation directly implements the theoretical update rule: the first term exploits the existing return landscape (which may be self-locked), while the second term injects the outward scarcity impulse (Proposition 3.5). By setting a high y_S, generation is steered along the manifold normal; together with a high y_R, trajectories are guided through the cone of improving directions to discover effective gift samples.

Iterative Training Framework The overall training process follows an iterative "Generate-Judge-Explore-Update" cycle. *(1) Exploitation Generation*: We sample a batch of trajectories $\mathcal{D}_{gen}$ from the current policy π_k conditioned on high returns. *(2) Scarcity Evaluation*: The Judger $\mathcal{J}_\phi$ is trained on $\mathcal{D}_{gen}$ to learn the current manifold density. It then evaluates scarcity scores $S(\tau)$ for

Table 1: **Main Results on AuctionNet.** Results on the AuctionNet-Dense and AuctionNet-Sparse are reported as mean ± std over 7 evaluation episodes. Metrics include Score, Reward, ER (CPA exceedance ratio), and WR (win rate). Improvements are reported relative to Iter 0 based on the Score. Results that outperform the oracle baseline are highlighted in blue, and the best results are highlighted in bold.

Dataset	Metrics	Baselines				Vanilla Self-Training			**ISGE (Ours)**		
		BC	DT	Oracle	Iter 0	Iter1	Iter2	Iter3	Iter1	Iter2	Iter3
Dense	Score ↑	56.0 ± 7.6	68.9 ± 7.4	102.0 ± 3.5	65.3 ± 3.2	74.9 ± 3.4	77.5 ± 3.1	77.8 ± 3.9	103.4 ± 5.3	104.1 ± 6.7	$\mathbf{108.4 \pm 8.2}$
	Reward ↑	68.7 ± 9.1	112.7 ± 7.1	208.3 ± 4.4	181.6 ± 2.8	192.5 ± 2.8	196.7 ± 4.0	198.4 ± 4.8	219.1 ± 4.6	220.3 ± 6.1	$\mathbf{223.1 \pm 7.0}$
	ER ↓	$48.0 \pm 19.4\%$	$\mathbf{43.4 \pm 4.0\%}$	$74.9 \pm 6.8\%$	$97.4 \pm 3.5\%$	$85.2 \pm 3.2\%$	$81.7 \pm 4.1\%$	$79.8 \pm 4.7\%$	$58.8 \pm 3.6\%$	$58.3 \pm 4.8\%$	$54.3 \pm 5.3\%$
	WR ↑	$1.6 \pm 0.2\%$	$3.3 \pm 0.2\%$	$5.8 \pm 0.2\%$	$5.4 \pm 0.2\%$	$5.6 \pm 0.1\%$	$5.7 \pm 0.2\%$	$5.7 \pm 0.2\%$	$6.1 \pm 0.1\%$	$6.1 \pm 0.1\%$	$\mathbf{6.2 \pm 0.1\%}$
	Improve ↑	–	–	56.2%	–	14.7%	18.7%	19.1%	58.3%	59.4%	**66.0%**
Sparse	Score ↑	0.5 ± 0.3	7.4 ± 0.3	20.5 ± 1.1	15.7 ± 1.0	16.9 ± 0.5	16.6 ± 0.6	16.0 ± 0.6	20.0 ± 0.9	20.6 ± 0.7	$\mathbf{21.1 \pm 1.0}$
	Reward ↑	0.6 ± 0.3	10.0 ± 0.5	27.6 ± 0.8	26.1 ± 0.9	27.2 ± 0.5	27.1 ± 0.5	26.6 ± 0.5	29.6 ± 0.6	30.1 ± 0.6	$\mathbf{30.3 \pm 0.8}$
	ER ↓	$452.9 \pm 102.2\%$	$32.4 \pm 5.5\%$	$\mathbf{10.8 \pm 2.9\%}$	$41.4 \pm 4.6\%$	$35.3 \pm 3.3\%$	$35.5 \pm 2.5\%$	$38.5 \pm 2.2\%$	$23.7 \pm 3.3\%$	$21.1 \pm 3.2\%$	$20.9 \pm 4.3\%$
	WR ↑	$0.1 \pm 0.1\%$	$2.8 \pm 0.3\%$	$7.4 \pm 0.1\%$	$7.8 \pm 0.2\%$	$8.1 \pm 0.1\%$	$8.2 \pm 0.1\%$	$8.2 \pm 0.1\%$	$8.5 \pm 0.1\%$	$8.5 \pm 0.1\%$	$\mathbf{8.6 \pm 0.1\%}$
	Improve ↑	–	–	30.6%	–	7.6%	5.7%	1.9%	27.4%	31.2%	**34.4%**



(a) ASN (Gift: 4.9%) (b) PSN (Gift: 7.9%) (c) **ISGE (Gift: 30.4%)** (g) Risk(Dense): ASN>PSN>ISGE

(d) ASN (Gift: 0.1%) (e) PSN (Gift: 1.2%) (f) **ISGE (Gift: 16.2%)** (h) Risk(Sparse): ASN>PSN>ISGE

Figure 3: **Exploration Efficiency and Safety Analysis. Left (a)-(f):** Scarcity–Return landscapes for Dense and Sparse environments. Red and blue points denote exploration and baseline (V1) trajectories, respectively. The *Gift Zone* (top-right) represents high-value strategies exceeding the top 10% thresholds of both return and scarcity computed on the baseline dataset. **Right (g)-(h):** Kernel density estimates of exploration score distributions for ASN (blue), PSN (green), and ISGE (red). The *Risk Zone* indicates performance falling below the 10% quantile of the baseline (V1) distribution. Risk proportion (Dense, Sparse): ASN (4.7%, 6.8%), PSN (3.6%, 4.1%), **ISGE (0.1%, 2.6%)**.

all samples in $\mathcal{D}_{gen}$. *(3) Guided Exploration*: The Dual-Condition Explorer is trained on the labeled dataset $\{(\tau, R(\tau), S(\tau))\}$. We then sample a new batch $\mathcal{D}_{scarcity}$ conditioned on both high return y_R and high scarcity y_S (in practice, chosen as the dataset maximum or a high quantile, e.g., the 99th percentile). *(4) Policy Update*: The next-iteration policy π_{k+1} is trained on a mixed dataset: $\mathcal{D}_{train} = \mathcal{D}_{gen} \cup \mathcal{D}_{scarcity}$. Crucially, this paradigm ensures *dynamic recalibration*: in each iteration, ISGE specifically supplements trajectories that are scarce relative to the *current* policy, thereby achieving efficient exploration without redundant interference from previous explorations.

5 Experiments

Experimental Setup We evaluate ISGE on AuctionNet Su et al. (2025), a large-scale industrial benchmark. To simulate the cold-start problem, we construct narrowly supported datasets retaining only trajectories in the return percentile range $[10, 30]$, creating two settings: Dense (strict constraints) and Sparse (sparse rewards). We compare against BC, Decision Transformer (DT), and Vanilla Self-Training (Iter 0-3). The Oracle is a diffusion planner trained on the Full Dataset. Performance is measured by Score (primary metric), Total Reward (total conversions), CPA Exceedance Ratio (ER), and Win Rate (WR). See Appendix C for full definitions and implementation details.

5.1 Main Results: Bootstrapping from Narrow Support

Table 1 highlights a stark contrast in performance evolution. Traditional imitation baselines (BC, DT) struggle significantly, particularly in the *Sparse* setting where reward signals are rare (e.g., BC Score 0.5). Even the stronger diffusion baseline, Vanilla Self-Training, succumbs to self-locking:

it saturates far below the Oracle in *Dense* (77.5 vs. 102.0) and regresses in *Sparse* (Iter $1 \rightarrow 3$: $16.9 \rightarrow 16.0$), validating our Proposition 3.4 that the return signal is drowned out by noise. In contrast, ISGE effectively breaks this bottleneck. It not only outperforms baselines but surpasses the Full-Dataset Oracle in both settings. Notably, in *Dense*, ISGE achieves a Score of 108.4 (vs. Oracle 102.0) with significantly better constraint satisfaction (ER 54.3% vs. 74.9%), demonstrating Pareto-optimal improvement. By Iteration 3, ISGE attains relative gains of **+66.0%** (Dense) and **+34.4%** (Sparse). This consistent monotonic growth confirms that dynamic scarcity injection continuously repairs the optimization landscape, enabling bootstrapping from severely suboptimal initialization.

5.2 ISGE Achieves Efficient and Safe Exploration

To validate that our gains stem from scarcity guidance rather than simple randomness, we compare ISGE against two standard undirected exploration baselines in Auto-bidding: Action Space Noise (ASN) Mou et al. (2022) and Parameter Space Noise (PSN) Li et al. (2024) (see Appendix C for details). Using Iteration 1 Vanilla (V1) as the baseline, Table 2 shows that ASN and PSN yield only marginal improvements ($\sim 9\%$ in Dense, $\sim 8\%$ in Sparse). This stagnation supports our hypothesis that in high-dimensional manifolds, isotropic noise is statistically orthogonal to sparse directions of improvement. In contrast, ISGE achieves a transformative leap, improving the Score by **38.0%** in Dense and **19.0%** in Sparse settings. Crucially, ISGE dominates in both reward maximization and constraint satisfaction, successfully navigating the trade-off between aggressive exploration and strict CPA limits.

Table 2: **Performance of Directed vs. Undirected Exploration.** Metrics are reported for Iteration 1 relative to Vanilla Self-Training (V1). ISGE significantly outperforms undirected baselines (ASN/PSN) across all metrics.

Dataset	Metrics	Vanilla (V1)	ASN	PSN	**ISGE**
Dense	Score ↑	74.9 ± 3.4	81.7 ± 4.0	81.8 ± 3.8	$\mathbf{103.4 \pm 5.3}$
	Reward ↑	192.5 ± 2.8	200.8 ± 3.9	201.0 ± 3.9	$\mathbf{219.1 \pm 4.6}$
	ER ↓	$85.2 \pm 3.2\%$	$77.7 \pm 4.0\%$	$77.4 \pm 4.1\%$	$\mathbf{58.8 \pm 3.6}\%$
	WR ↑	$5.6 \pm 0.1\%$	$5.8 \pm 0.1\%$	$5.8 \pm 0.1\%$	$\mathbf{6.1 \pm 0.1}\%$
	Improve ↑	–	9.1%	9.2%	**38.0%**
Sparse	Score ↑	16.8 ± 0.9	18.0 ± 0.5	18.2 ± 0.6	$\mathbf{20.0 \pm 0.9}$
	Reward ↑	27.1 ± 0.7	28.1 ± 0.5	28.2 ± 0.5	$\mathbf{29.6 \pm 0.6}$
	ER ↓	$36.0 \pm 3.8\%$	$30.9 \pm 2.9\%$	$30.4 \pm 3.0\%$	$\mathbf{23.7 \pm 3.3}\%$
	WR ↑	$8.1 \pm 0.1\%$	$8.3 \pm 0.1\%$	$8.2 \pm 0.1\%$	$\mathbf{8.5 \pm 0.1}\%$
	Improve ↑	–	7.1%	8.3%	**19.0%**

Mechanism Analysis: Efficiency and Safety. We further investigate this success by visualizing the *Scarcity-Return landscape* of the explored trajectories in Figure 3(a)-(f). We define the "Gift Zone" (top-right quadrant) as the target region containing high-scarcity, high-return trajectories. As observed in the scatter plots, undirected baselines (ASN/PSN) are inefficient, capturing negligible samples in the Gift Zone (e.g., $\leq 1.2\%$ in Sparse), proving that random perturbations fail to stumble upon high-value strategies in data vacuums. Conversely, ISGE's dual-condition explorer functions as a directed guide, efficiently achieving capture rates of **30.4%** (Dense) and **16.2%** (Sparse). Furthermore, regarding *safety* (downside risk), Figure 3(g)-(h) reveals that undirected methods exhibit heavy left-tail distributions falling into the "Risk Zone" (up to 6.8% risk). In contrast, ISGE explicitly truncates this downside risk ($\leq 0.1\%$ in Dense, $\leq 2.6\%$ in Sparse) via return conditioning. This demonstrates that ISGE robustly expands the performance upper bound without compromising the reliable lower bound.

5.3 Ablation Studies

In this section, we analyze key components of ISGE on AuctionNet-Dense (Iteration 1). We focus on the necessity of Strategic Masking and hyperparameter analysis.

Effect of Strategic Masking. ISGE calculates scarcity solely on controllable policy states, masking out uncontrollable environmental dimensions. Removing this mask degrades performance and significantly reduces the gift sample capture rate from 30.4% to 19.1% (Table 3a). This decline stems from a fundamental distraction: granular analysis (Table 3b) reveals that environmental features like *Traffic Value* exhibit negligible scarcity growth because the agent cannot "invent" market conditions. Consequently, including these immutable dimensions introduces futile noise that distracts the explorer from controllable strategic features (e.g., *Budget Left*), thereby diluting the gradient signal of scarcity for valid exploration.

Table 3(a): **Masking vs. No Masking.**

Metrics	No Masking	**Masking**
Score	94.7	**103.4**
Gift %	19.1%	**30.4%**

Table 3(b): **Feature Scarcity Growth.**

	State Feature	**Pre**	**Post**
Strat.	Budget Left	0.056	0.105
	History Bid	0.025	0.486
Env.	Traffic Volume	0.052	0.062
	Traffic Value	0.229	0.235

Hyperparameter Sensitivity & Mechanism Validation. Finally, we analyze the impact of guidance scales and data mixture ratios (Table 4) to validate the core design principles of ISGE.

(a) The Necessity of Proper Scarcity Guidance. Table 4(a) reveals that weak scarcity guidance ($w_S = 1$) fails to effectively capture gift samples (19.7%), dropping performance to 90.2. This confirms that scarcity identification alone is insufficient; a sufficiently large w_S is required to mechanically "push" the trajectory away from the high-density manifold against the strong attraction of the diffusion prior. This empirically validates our theoretical analysis (Theorem 3.6) that proper guidance scale w_S is necessary for scarcity-guided exploration.

(b) The Necessity of Data Mixing. Table 4(b) demonstrates that while *Explore-Only* (96.7) significantly outperforms *Base-Only* (74.9), it still lags behind the mixed policy (103.4). This highlights the complementary nature of the datasets: exploration data fixes policy "blind spots" (high-value/sparse regions) but sacrifices global coverage due to its targeted bias. Consequently, mixing is essential to combine the evolutionary signal of exploration with the structural stability of the base distribution.

Table 4: **Hyperparameters Analysis.** We study the impact of (a) Guidance Scales balance and (b) Data Mixture Ratios. The default settings are highlighted in gray.

(a) CFG Guidance Balance ($w_R : w_S$)

Setting (w_R, w_S)	Ratio	Exploration Data Quality			Final Policy
		Avg Score	Avg Scarcity	**Gift %**	**Score**
Reward-Heavy $(3, 1)$	3:1	88.6	0.14	19.7%	90.2 ± 3.8
Balanced $(3, 3)$	1:1	102.9	0.36	30.4%	103.4 ± 5.3
Scarcity-Heavy $(3, 5)$	3:5	**105.4**	**0.59**	**32.2%**	$\mathbf{105.5 \pm 6.1}$

(b) Data Mixture Ratio (Explore : Base)

Mixture Ratio	**Score**
Base-Only $(0 : 1)$	74.9 ± 3.4
Conservative $(1 : 2)$	103.3 ± 6.9
Balanced (5 : 4)	$\mathbf{103.4 \pm 5.3}$
Aggressive $(2 : 1)$	101.6 ± 5.3
Explore-Only $(1 : 0)$	96.7 ± 6.0

6 Conclusion and Related Work

In this paper, we addressed the critical industrial challenge of bootstrapping bidding policies from low-quality narrowly supported historical data. We identified the "self-locking" phenomenon in iterative training, revealing that the geometric collapse of the return gradient within the training manifold acts as the fundamental bottleneck preventing OOD extrapolation. To break this stagnation, we introduced ISGE, a novel framework that leverages scarcity as an auxiliary impulse to revive directional guidance in the gradient vacuum. Our theoretical analysis proves that scarcity-driven updates successfully restore the missing radial component of the return signal. We implemented this theory via a robust iterative training protocol that aggregates discovered "gift samples". Extensive experiments on AuctionNet benchmarks demonstrate that ISGE significantly outperforms baselines, validating its capability to achieve sustained performance growth beyond the low-quality and narrowly supported dataset. Related work is discussed in the following paragraphs.

Auto-bidding and Generative Auto-bidding Traditionally, auto-bidding has been formulated as a constrained optimization problem, typically addressed through PID controllers Zhang et al. (2016); Yang et al. (2019) or primal-dual frameworks Balseiro et al. (2021); Aggarwal et al. (2024). To handle dynamic market environments, Reinforcement Learning (RL) methods were subsequently introduced to maximize cumulative returns Cai et al. (2017); Wu et al. (2018); He et al. (2021); Wen et al. (2022); Ou et al. (2023). However, they struggle with ineffective credit assignment in non-Markovian long-horizon scenarios. Recently, the field has undergone a paradigm shift toward AIGB, which reformulates auto-bidding as a generative sequence modeling task. Pioneered by DiffBid Guo et al. (2024), this paradigm utilizes conditional diffusion models to capture the joint distribution of bidding trajectories Li et al. (2025a), while subsequent Transformer-based approaches like GAVE Gao et al. (2025), GAS Li et al. (2025b) and GRAD Lei et al. (2025) employ autoregressive planning with return-to-go conditioning. To mitigate the limitations of static offline datasets, recent research has focused on offline improvement mechanism Peng et al. (2025); Mou et al. (2025); Jiang et al. (2025). However, these methods typically rely on high-quality expert demonstrations or broad data coverage. In contrast, our work targets the iterative cold-start setting, focusing on bootstrapping policies purely from low-quality, narrowly supported data without expert priors.

Decision Making and Exploration with Diffusion Models Diffusion models have revolutionized decision-making, either by treating planning as trajectory generation Janner et al. (2022); Ajay et al. (2022); Dong et al. (2024) or representing policies Wang et al. (2022); Chen et al. (2024); Ni et al.

(2023). Regarding exploration, methods like PGR Wang et al. (2025) and ExDM Ying et al. (2025) rely on *transition-level* ($s_t, a_t \to s_{t+1}$) novelty. However, auto-bidding is a strictly non-Markovian problem governed by global constraints: the budget spent at step t limits the feasible actions at step T, thereby transition-level metrics are insufficient for capturing the long-horizon strategic structure. In contrast, ISGE adopts *trajectory-level* exploration via reconstruction error, aligning with the sequential nature of diffusion planners. Moreover, unlike AdaptDiffuser Liang et al. (2023) which passively filters plans within existing support, ISGE actively injects scarcity signals via CFG during sampling. This mechanism enables the explorer to extrapolate beyond the current support, purposefully discovering "gift samples" in high-scarcity regions.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 62572010), supported by Alibaba Group through Alibaba Innovative Research Program, and supported by Peking University Alimama Joint Laboratory of AI Innovation. We thank all anonymous reviewers for their helpful feedback.

REFERENCES

Gagan Aggarwal, Ashwinkumar Badanidiyuru, Santiago R Balseiro, Kshipra Bhawalkar, Yuan Deng, Zhe Feng, Gagan Goel, Christopher Liaw, Haihao Lu, Mohammad Mahdian, et al. Auto-bidding and auctions in online advertising: A survey. *ACM SIGecom Exchanges*, 22(1):159–183, 2024.

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022.

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *J. Mach. Learn. Res.*, 15(1):3563–3593, January 2014. ISSN 1532-4435.

Santiago R Balseiro, Yuan Deng, Jieming Mao, Vahab S Mirrokni, and Song Zuo. The landscape of auto-bidding auctions: Value versus utility maximization. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 132–133, 2021.

Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the tenth ACM international conference on web search and data mining*, pp. 661–670, 2017.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021. URL https://arxiv.org/abs/2106.01345.

Tianyu Chen, Zhendong Wang, and Mingyuan Zhou. Diffusion policies creating a trust region for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 37:50098–50125, 2024.

Zibin Dong, Jianye Hao, Yifu Yuan, Fei Ni, Yitian Wang, Pengyi Li, and Yan Zheng. Diffuserlite: Towards real-time diffusion planning. *Advances in Neural Information Processing Systems*, 37: 122556–122583, 2024.

David S Evans. The online advertising industry: Economics, evolution, and privacy. *Journal of economic perspectives*, 23(3):37–60, 2009.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration, 2019. URL https://arxiv.org/abs/1812.02900.

Jingtong Gao, Yewen Li, Shuai Mao, Peng Jiang, Nan Jiang, Yejing Wang, Qingpeng Cai, Fei Pan, Peng Jiang, Kun Gai, Bo An, and Xiangyu Zhao. Generative auto-bidding with value-guided explorations. SIGIR '25, pp. 244–254, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3729987. URL https://doi.org/10.1145/3726302.3729987.

Jiayan Guo, Yusen Huo, Zhilin Zhang, Tianyu Wang, Chuan Yu, Jian Xu, Bo Zheng, and Yan Zhang. Generative auto-bidding via conditional diffusion modeling. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 5038–5049, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671526. URL https://doi.org/10.1145/3637528.3671526.

Yue He, Xiujun Chen, Di Wu, Junwei Pan, Qing Tan, Chuan Yu, Jian Xu, and Xiaoqiang Zhu. A unified solution to constrained bidding in online display advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 2993–3001, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467199. URL https://doi.org/10.1145/3447548.3467199.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Hao Jiang, Yongxiang Tang, Yanxiang Zeng, Pengjia Yuan, Yanhua Cheng, Teng Sha, Xialong Liu, and Peng Jiang. Optimal return-to-go guided decision transformer for auto-bidding in advertisement. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 1033–1037, 2025.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Yu Lei, Jiayang Zhao, Yilei Zhao, Zhaoqi Zhang, Linyou Cai, Qianlong Xie, and Xingxing Wang. Generative large-scale pre-trained models for automated ad bidding optimization, 2025. URL https://arxiv.org/abs/2508.02002.

Haoming Li, Yusen Huo, Shuai Dou, Zhenzhe Zheng, Zhilin Zhang, Chuan Yu, Jian Xu, and Fan Wu. Trajectory-wise iterative reinforcement learning framework for auto-bidding. In *Proceedings of the ACM Web Conference 2024*, pp. 4193–4203, 2024.

Yewen Li, Jingtong Gao, Nan Jiang, Shuai Mao, Ruyi An, Fei Pan, Xiangyu Zhao, Bo An, Qingpeng Cai, and Peng Jiang. Generative auto-bidding in large-scale competitive auctions via diffusion completer-aligner, 2025a. URL https://arxiv.org/abs/2509.03348.

Yewen Li, Shuai Mao, Jingtong Gao, Nan Jiang, Yunjian Xu, Qingpeng Cai, Fei Pan, Peng Jiang, and Bo An. Gas: Generative auto-bidding with post-training search. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 315–324, 2025b.

Zhixuan Liang, Yao Mu, Mingyu Ding, Fei Ni, Masayoshi Tomizuka, and Ping Luo. Adaptdiffuser: Diffusion models as adaptive self-evolving planners. In *International Conference on Machine Learning*, 2023.

Zhiyu Mou, Yusen Huo, Rongquan Bai, Mingzhou Xie, Chuan Yu, Jian Xu, and Bo Zheng. Sustainable online reinforcement learning for auto-bidding. *Advances in Neural Information Processing Systems*, 35:2651–2663, 2022.

Zhiyu Mou, Yiqin Lv, Miao Xu, Qi Wang, Yixiu Mao, Qichen Ye, Chao Li, Rongquan Bai, Chuan Yu, Jian Xu, and Bo Zheng. Enhancing generative auto-bidding with offline reward evaluation and policy search, 2025. URL https://arxiv.org/abs/2509.15927.

Fei Ni, Jianye Hao, Yao Mu, Yifu Yuan, Yan Zheng, Bin Wang, and Zhixuan Liang. Metadiffuser: Diffusion model as conditional planner for offline meta-rl, 2023. URL https://arxiv.org/abs/2305.19923.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1–2): 1–179, March 2018. ISSN 1935-8261. doi: 10.1561/2300000053. URL http://dx.doi.org/10.1561/2300000053.

Weitong Ou, Bo Chen, Xinyi Dai, Weinan Zhang, Weiwen Liu, Ruiming Tang, and Yong Yu. A survey on bid optimization in real-time bidding display advertising. *ACM Trans. Knowl. Discov. Data*, 18(3), December 2023. ISSN 1556-4681. doi: 10.1145/3628603. URL https://doi.org/10.1145/3628603.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Yunshan Peng, Wenzheng Shu, Jiahao Sun, Yanxiang Zeng, Jinan Pang, Wentao Bai, Yunke Bai, Xialong Liu, and Peng Jiang. Expert-guided diffusion planner for auto-bidding. CIKM '25, pp. 5963–5970, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Kefan Su, Yusen Huo, Zhilin Zhang, Shuai Dou, Chuan Yu, Jian Xu, Zongqing Lu, and Bo Zheng. Auctionnet: a novel benchmark for decision-making in large-scale games. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Renhao Wang, Kevin Frans, Pieter Abbeel, Sergey Levine, and Alexei A Efros. Prioritized generative replay. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5IkDAfabuo.

Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*, 2022.

Chao Wen, Miao Xu, Zhilin Zhang, Zhenzhe Zheng, Yuhui Wang, Xiangyu Liu, Yu Rong, Dong Xie, Xiaoyang Tan, Chuan Yu, et al. A cooperative-competitive multi-agent framework for auto-bidding in online advertising. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 1129–1139, 2022.

Di Wu, Xiujun Chen, Xun Yang, Hao Wang, Qing Tan, Xiaoxun Zhang, Jian Xu, and Kun Gai. Budget constrained bidding by model-free reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1443–1451, 2018.

Xun Yang, Yasong Li, Hao Wang, Di Wu, Qing Tan, Jian Xu, and Kun Gai. Bid optimization by multivariable control in display advertising. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1966–1974, 2019.

Chengyang Ying, Huayu Chen, Xinning Zhou, Zhongkai Hao, Hang Su, and Jun Zhu. Exploratory diffusion model for unsupervised reinforcement learning. *arXiv preprint arXiv:2502.07279*, 2025.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34:28954–28967, 2021.

Weinan Zhang, Shuai Yuan, and Jun Wang. Optimal real-time bidding for display advertising. KDD '14, pp. 1077–1086, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569. doi: 10.1145/2623330.2623633. URL https://doi.org/10.1145/2623330.2623633.

Weinan Zhang, Yifei Rong, Jun Wang, Tianchi Zhu, and Xiaofan Wang. Feedback control of real-time display advertising. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp. 407–416, 2016.

A MISSING PROOF IN SECTION 3

A.1 PROOF OF LEMMA 3.1

Lemma 3.1 (Restated). *Let $\epsilon_\theta(x_t)$ denote the unconditional noise prediction. Leveraging the score matching property $\epsilon_\theta(x_t) \approx -\sqrt{1-\bar{\alpha}_t}\,\nabla_{x_t}\log p_t(x_t)$, the modified noise prediction $\tilde{\epsilon}_\theta(x_t)$ under CFG with scale w can be reformulated as:*

$$\tilde{\epsilon}_\theta(x_t) \approx \epsilon_\theta(x_t) - w\sqrt{1-\bar{\alpha}_t}\cdot\nabla_{x_t}\log p(y \mid x_t). \tag{9}$$

Proof. Recall the definition of Classifier-Free Guidance (CFG). The modified noise prediction is a linear interpolation between the unconditional prediction $\epsilon_\theta(x_t)$ and the conditional prediction $\epsilon_\theta(x_t, y)$:

$$\tilde{\epsilon}_\theta(x_t) = \epsilon_\theta(x_t) + w\cdot\big(\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t)\big). \tag{10}$$

According to the denoising score matching objective Song et al. (2021), the optimal noise prediction network approximates the score of the data distribution (scaled by the noise level):

$$\epsilon_\theta(x_t) \approx -\sqrt{1-\bar{\alpha}_t}\,\nabla_{x_t}\log p_t(x_t), \tag{11}$$

$$\epsilon_\theta(x_t, y) \approx -\sqrt{1-\bar{\alpha}_t}\,\nabla_{x_t}\log p_t(x_t \mid y). \tag{12}$$

We apply Bayes' rule to the score function:

$$\nabla_{x_t}\log p_t(x_t \mid y) = \nabla_{x_t}\log p_t(x_t) + \nabla_{x_t}\log p(y \mid x_t). \tag{13}$$

Now, we substitute Eqs. equation 11 and equation 12 into the CFG term $(\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t))$ from Eq. equation 10:

$$\begin{aligned}
\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t) &\approx -\sqrt{1-\bar{\alpha}_t}\,(\nabla_{x_t}\log p_t(x_t \mid y) - \nabla_{x_t}\log p_t(x_t)) \\
&= -\sqrt{1-\bar{\alpha}_t}\,\big(\big(\nabla_{x_t}\log p_t(x_t) + \nabla_{x_t}\log p(y \mid x_t)\big) - \nabla_{x_t}\log p_t(x_t)\big) \\
&= -\sqrt{1-\bar{\alpha}_t}\cdot\nabla_{x_t}\log p(y \mid x_t).
\end{aligned} \tag{14}$$

Finally, substituting this back into Eq. equation 10, we obtain the effective noise prediction:

$$\tilde{\epsilon}_\theta(x_t) \approx \epsilon_\theta(x_t) - w\sqrt{1-\bar{\alpha}_t}\cdot\nabla_{x_t}\log p(y \mid x_t). \tag{15}$$

□

A.2 PROOF OF PROPOSITION 3.2

Proposition 3.2 (Restated). *By substituting $\tilde{\epsilon}_\theta(x_t)$ into the standard DDPM update rule, the sampling process at step t is reformulated as a biased Langevin update:*

$$x_{t-1} \leftarrow \mu_\theta(x_t) + \lambda_t w\cdot\nabla_{x_t}\log p(y \mid x_t) + \sigma_t z, \tag{16}$$

where $\lambda_t = \frac{1-\alpha_t}{\sqrt{\alpha_t}}$ is a positive coefficient derived from the noise schedule.

Proof. In the standard DDPM sampling process Ho et al. (2020), the transition from x_t to x_{t-1} is given by:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\tilde{\epsilon}_\theta(x_t)\right) + \sigma_t z, \tag{17}$$

where $z \sim \mathcal{N}(0, I)$ and $\tilde{\epsilon}_\theta(x_t)$ is the predicted noise used for denoising.

Substituting the result from Lemma 3.1 into Eq. equation 17:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\left[\epsilon_\theta(x_t) - w\sqrt{1-\bar{\alpha}_t}\,\nabla_{x_t}\log p(y \mid x_t)\right]\right) + \sigma_t z. \tag{18}$$

We expand the terms inside the parentheses. Let $\mu_\theta(x_t)$ denote the standard reconstruction mean toward the unconditional manifold:

$$\mu_\theta(x_t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t) \right). \tag{19}$$

Now, we focus on the additional term introduced by the guidance gradient. Note that the coefficient $\sqrt{1-\bar{\alpha}_t}$ in the numerator (from Lemma 3.1) cancels perfectly with the $\sqrt{1-\bar{\alpha}_t}$ in the denominator of the DDPM variance scaling term:

$$\begin{aligned} \text{Drift Term} &= \frac{1}{\sqrt{\alpha_t}} \left(-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \right) \cdot \left(-w\sqrt{1-\bar{\alpha}_t}\, \nabla_{x_t} \log p(y \mid x_t) \right) \\ &= \frac{1}{\sqrt{\alpha_t}} \cdot (1-\alpha_t) \cdot w \cdot \nabla_{x_t} \log p(y \mid x_t) \\ &= \frac{1-\alpha_t}{\sqrt{\alpha_t}}\, w \cdot \nabla_{x_t} \log p(y \mid x_t). \end{aligned} \tag{20}$$

Defining $\lambda_t = \frac{1-\alpha_t}{\sqrt{\alpha_t}}$, the update rule simplifies to:

$$x_{t-1} = \mu_\theta(x_t) + \lambda_t w \nabla_{x_t} \log p(y \mid x_t) + \sigma_t z. \tag{21}$$

Since $\alpha_t \in (0,1)$, the coefficient λ_t is strictly positive. This confirms that CFG manifests as an additive gradient ascent term (drift) superimposed on the standard reverse diffusion process. □

A.3 Proof of Geometric Barrier and Dynamics Failure (Section 3.3)

In this section, we provide the proofs for Theorem 3.3 and Proposition 3.4. We first establish the general geometric condition for distance reduction in high-dimensional space (Theorem 3.3) and then apply this condition to the specific Langevin dynamics of diffusion planners to demonstrate the self-locking mechanism (Proposition 3.4).

A.3.1 Proof of Theorem 3.3 (Geometric Barrier)

Theorem 3.3 (Restated). *Consider a point $\mathbf{x}$ at distance $\epsilon = \|\mathbf{x} - \mathbf{x}^*\|$ from the optimum $\mathbf{x}^*$, and an update $\Delta\mathbf{x} = \mathbf{v}_{\text{det}} + \mathbf{v}_{\text{noise}}$, where $\mathbf{v}_{\text{det}}$ is a deterministic drift and $\mathbf{v}_{\text{noise}} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is isotropic noise. Let $v_{\text{rad}} = \langle \mathbf{v}_{\text{det}}, \mathbf{n}^* \rangle$ be the radial component of the drift, and $\|\mathbf{v}_{\tan}\|^2 = \|\mathbf{v}_{\text{det}} - v_{\text{rad}}\mathbf{n}^*\|^2 + D\sigma^2$ its total tangential energy. Then, the expected change in distance satisfies*

$$\mathbb{E}[\Delta\epsilon] \approx \underbrace{-v_{\text{rad}}}_{\text{Signal}} + \underbrace{\frac{\|\mathbf{v}_{\tan}\|^2}{2\epsilon}}_{\text{Penalty}}. \tag{22}$$

Proof. Without loss of generality, place the optimum at the origin ($\mathbf{x}^* = \mathbf{0}$), so that $\epsilon = \|\mathbf{x}\|$ and the unit inward radial direction is $\mathbf{n}^* = -\mathbf{x}/\epsilon$. Decompose the update as:

$$\Delta\mathbf{x} = \mathbf{v}_{\text{det}} + \mathbf{v}_{\text{noise}} = v_{\text{rad}}\mathbf{n}^* + \mathbf{v}_{\text{det,tan}} + \mathbf{v}_{\text{noise}},$$

where $\mathbf{v}_{\text{det,tan}} = \mathbf{v}_{\text{det}} - v_{\text{rad}}\mathbf{n}^*$ is the deterministic tangential component.

The new state is $\mathbf{x}' = \mathbf{x} + \Delta\mathbf{x}$, and its squared norm is:

$$\begin{aligned} \|\mathbf{x}'\|^2 &= \|\mathbf{x} + v_{\text{rad}}\mathbf{n}^* + \mathbf{v}_{\text{det,tan}} + \mathbf{v}_{\text{noise}}\|^2 \\ &= \|\underbrace{\mathbf{x} + v_{\text{rad}}\mathbf{n}^*}_{=(\epsilon - v_{\text{rad}})\mathbf{n}^*}\|^2 + \|\mathbf{v}_{\text{det,tan}} + \mathbf{v}_{\text{noise}}\|^2 \quad \text{(orthogonality)} \\ &= (\epsilon - v_{\text{rad}})^2 + \|\mathbf{v}_{\text{det,tan}} + \mathbf{v}_{\text{noise}}\|^2. \end{aligned} \tag{23}$$

Expanding and taking expectation:

$$\begin{aligned} \mathbb{E}[\|\mathbf{x}'\|^2] &= \epsilon^2 - 2\epsilon v_{\text{rad}} + v_{\text{rad}}^2 + \mathbb{E}[\|\mathbf{v}_{\text{det,tan}} + \mathbf{v}_{\text{noise}}\|^2] \\ &= \epsilon^2 - 2\epsilon v_{\text{rad}} + v_{\text{rad}}^2 + \|\mathbf{v}_{\text{det,tan}}\|^2 + \mathbb{E}[\|\mathbf{v}_{\text{noise}}\|^2], \end{aligned} \tag{24}$$

since $\mathbb{E}[\langle \mathbf{v}_{\text{det,tan}}, \mathbf{v}_{\text{noise}} \rangle] = 0$.

Under the high-dimensional assumption ($D \gg 1$) and small step size ($\|\Delta \mathbf{x}\| \ll \epsilon$), we apply the Taylor expansion $\|\mathbf{x}'\| = \sqrt{\mathbb{E}[\|\mathbf{x}'\|^2]} \approx \epsilon + \frac{\mathbb{E}[\|\mathbf{x}'\|^2] - \epsilon^2}{2\epsilon}$ to obtain the expected distance change:

$$\begin{aligned} \mathbb{E}[\Delta\epsilon] &= \mathbb{E}[\|\mathbf{x}'\|] - \epsilon \\ &\approx \frac{-2\epsilon v_{\text{rad}} + v_{\text{rad}}^2 + \|\mathbf{v}_{\text{det,tan}}\|^2 + D\sigma^2}{2\epsilon}. \end{aligned} \tag{25}$$

Neglecting the second-order radial term $v_{\text{rad}}^2/(2\epsilon)$ (valid for small steps), we arrive at:

$$\mathbb{E}[\Delta\epsilon] \approx -v_{\text{rad}} + \frac{\|\mathbf{v}_{\text{det,tan}}\|^2 + D\sigma^2}{2\epsilon} = -v_{\text{rad}} + \frac{\|\mathbf{v}_{\text{tan}}\|^2}{2\epsilon}, \tag{26}$$

where we define the total tangential energy as $\|\mathbf{v}_{\text{tan}}\|^2 := \|\mathbf{v}_{\text{det,tan}}\|^2 + D\sigma^2$.

Consequently, distance reduction ($\mathbb{E}[\Delta\epsilon] < 0$) requires:

$$v_{\text{rad}} > \frac{\|\mathbf{v}_{\text{tan}}\|^2}{2\epsilon} \quad \Longleftrightarrow \quad \text{SNR}_{\text{eff}} := \frac{v_{\text{rad}}}{\|\mathbf{v}_{\text{tan}}\|^2/(2\epsilon)} > 1. \tag{27}$$

This completes the proof. □

2D Geometric Intuition Figure 4 provides a geometric illustration of Theorem 3.3. Consider two points $\mathbf{x}_1$ and $\mathbf{x}_2$ located at distances ϵ_1 and ϵ_2 from the optimum, respectively, with $\epsilon_2 \ll \epsilon_1$. The update direction at any point is decomposed into a deterministic drift component $\mathbf{v}_{\text{det}}$ and an isotropic noise component $\mathbf{v}_{\text{noise}}$. In the outer region (Large Sphere Shell, radius ϵ_1), the combined effect of drift and noise produces a gradient $\Delta\mathbf{x}_1 = \mathbf{v}_{\text{det}} + \mathbf{v}_{\text{noise}}$ that reduces the distance to the optimum ($\Delta\epsilon_1 < 0$), resulting in convergent behavior. In contrast, in the inner region (Small Sphere Shell, radius ϵ_2), the same decomposition produces $\Delta\mathbf{x}_2 = \mathbf{v}_{\text{det}} + \mathbf{v}_{\text{noise}}$ that increases the distance from the optimum ($\Delta\epsilon_2 > 0$), leading to divergent behavior.

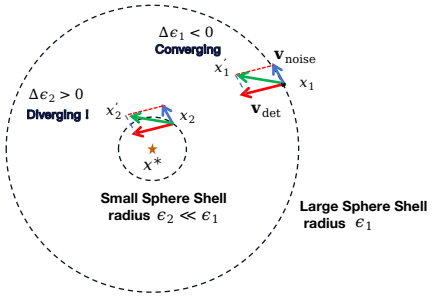


Figure 4: 2D Geometric illustration of Theorem 3.3

A.3.2 Proof of Proposition 3.4

Proposition 3.4 (Restated). *Assuming the guidance scale w_R is chosen optimally to maximize convergence speed, the diffusion planner optimization stagnates ($\mathbb{E}[\Delta\epsilon] \geq 0$) when the radial signal intensity drops below the critical threshold:*

$$\langle \mathbf{g}_R, \mathbf{n}^* \rangle \leq \frac{\sigma_t \sqrt{D}\, \|\mathbf{g}_{R,\tan}\|}{\epsilon}. \tag{28}$$

Proof. The proof proceeds by analyzing the expected change in the distance to the optimum, $\mathbb{E}[\Delta\epsilon]$, under the update rule $\Delta x_t = \lambda_t w_R \mathbf{g}_R + \sigma_t \mathbf{z}$. We derive the optimal guidance scale w_R^* and show that even under this optimal condition, distance reduction is impossible if the stated inequality holds.

Step 1: Decomposition of Distance Dynamics Based on the geometric expansion for high-dimensional distances (Theorem 3.3), the expected change in distance is approximated by the competition between radial drift and tangential dispersion:

$$\mathbb{E}[\Delta\epsilon] \approx -\mathbb{E}[v_{\text{rad}}] + \frac{\mathbb{E}[\|\mathbf{v}_{\tan}\|^2]}{2\epsilon}. \tag{29}$$

We define the "Convergence Velocity" $V(w_R) := -\mathbb{E}[\Delta\epsilon]$. Stagnation occurs when $V(w_R) \leq 0$.

We decompose the update vector Δx_t into radial and tangential components:

- **Radial Component (v_{rad}):** The projection of the update onto $\mathbf{n}^*$:

  $$\mathbb{E}[v_{\text{rad}}] = \mathbb{E}[\langle \lambda_t w_R \mathbf{g}_R + \sigma_t \mathbf{z}, \mathbf{n}^* \rangle] = \lambda_t w_R \langle \mathbf{g}_R, \mathbf{n}^* \rangle. \tag{30}$$

  Let $g_{\text{rad}} := \langle \mathbf{g}_R, \mathbf{n}^* \rangle$.

- **Tangential Component ($\mathbf{v}_{\tan}$):** The projection onto the tangent space orthogonal to $\mathbf{n}^*$:

  $$\mathbf{v}_{\tan} = \lambda_t w_R \mathbf{g}_{R,\tan} + \sigma_t \mathbf{z}_{\tan}. \tag{31}$$

  The expected squared norm (tangential energy) is:

  $$\mathbb{E}[\|\mathbf{v}_{\tan}\|^2] = \lambda_t^2 w_R^2 \|\mathbf{g}_{R,\tan}\|^2 + D\sigma_t^2. \tag{32}$$

  (Note: The cross-term vanishes because $\mathbb{E}[\mathbf{z}_{\tan}] = 0$, and $\mathbb{E}[\|\sigma_t \mathbf{z}_{\tan}\|^2] \approx D\sigma_t^2$ for $D \gg 1$.)

Substituting these into the velocity function:

$$V(w_R) = \lambda_t w_R g_{\text{rad}} - \frac{\lambda_t^2 w_R^2 \|\mathbf{g}_{R,\tan}\|^2 + D\sigma_t^2}{2\epsilon}. \tag{33}$$

Step 2: Optimization of Guidance Scale w_R To find the theoretical limit of convergence, we solve for the optimal guidance scale w_R^* that maximizes $V(w_R)$. We compute the derivative with respect to w_R:

$$\frac{dV}{dw_R} = \lambda_t g_{\text{rad}} - \frac{2\lambda_t^2 w_R \|\mathbf{g}_{R,\tan}\|^2}{2\epsilon} = 0. \tag{34}$$

Solving for w_R, we obtain the optimal scale:

$$w_R^* = \frac{\epsilon g_{\text{rad}}}{\lambda_t \|\mathbf{g}_{R,\tan}\|^2}. \tag{35}$$

This derivation confirms that simply increasing w_R indefinitely is detrimental, as the quadratic penalty term ($\propto w_R^2$) eventually dominates the linear signal term ($\propto w_R$).

Step 3: Deriving the Stagnation Condition We substitute the optimal scale w_R^* back into the velocity function to find the maximum possible convergence speed $V_{\max}$:

$$V(w_R^*) = \lambda_t \left(\frac{\epsilon g_{\text{rad}}}{\lambda_t \|\mathbf{g}_{R,\tan}\|^2} \right) g_{\text{rad}} - \frac{\lambda_t^2 \left(\frac{\epsilon g_{\text{rad}}}{\lambda_t \|\mathbf{g}_{R,\tan}\|^2} \right)^2 \|\mathbf{g}_{R,\tan}\|^2 + D\sigma_t^2}{2\epsilon} \tag{36}$$

$$= \frac{\epsilon g_{\text{rad}}^2}{\|\mathbf{g}_{R,\tan}\|^2} - \frac{\frac{\epsilon^2 g_{\text{rad}}^2}{\|\mathbf{g}_{R,\tan}\|^2} + D\sigma_t^2}{2\epsilon} \tag{37}$$

$$= \frac{\epsilon g_{\text{rad}}^2}{\|\mathbf{g}_{R,\tan}\|^2} - \left(\frac{\epsilon g_{\text{rad}}^2}{2\|\mathbf{g}_{R,\tan}\|^2} + \frac{D\sigma_t^2}{2\epsilon} \right) \tag{38}$$

$$= \frac{\epsilon g_{\text{rad}}^2}{2\|\mathbf{g}_{R,\tan}\|^2} - \frac{D\sigma_t^2}{2\epsilon}. \tag{39}$$

Optimization stagnates when even this maximum velocity is non-positive ($V(w_R^*) \leq 0$):

$$\frac{\epsilon g_{\text{rad}}^2}{2\|\mathbf{g}_{R,\tan}\|^2} \leq \frac{D\sigma_t^2}{2\epsilon}. \tag{40}$$

Multiplying by 2ϵ and rearranging:

$$\epsilon^2 g_{\text{rad}}^2 \leq D\sigma_t^2 \|\mathbf{g}_{R,\tan}\|^2. \tag{41}$$

Taking the square root (assuming all terms are positive):

$$\epsilon g_{\text{rad}} \leq \sigma_t \sqrt{D}\, \|\mathbf{g}_{R,\tan}\|. \tag{42}$$

Finally, solving for the radial signal $g_{\text{rad}} = \langle \mathbf{g}_R, \mathbf{n}^* \rangle$:

$$\langle \mathbf{g}_R, \mathbf{n}^* \rangle \leq \frac{\sigma_t \sqrt{D}\, \|\mathbf{g}_{R,\tan}\|}{\epsilon}. \tag{43}$$

This inequality demonstrates that when the distance ϵ is small or the misalignment $\|\mathbf{g}_{R,\tan}\|$ is large, the required radial signal to maintain convergence exceeds the capability of the reward model, enforcing a minimal achievable radius $\epsilon_{\min}$. □

A.4 Analysis for Proposition 3.5

We now formalize the geometric behavior of the scarcity gradient. Our analysis proceeds in two steps: first, we establish the geometric properties of the explicit scarcity score gradient $\nabla_{\mathbf{x}} S(\mathbf{x})$ under the framework of Denoising Autoencoders (DAEs); second, we prove that the implicit CFG gradient $\mathbf{g}_S$ used in ISGE aligns with this explicit gradient via the chain rule.

A.4.1 Geometry of Explicit Scarcity Gradient

Our analysis builds upon the foundational result of Alain & Bengio (2014), which connects reconstruction error to the underlying data manifold.

Assumption A.1 (Scarcity as Squared Distance Proxy)**.** Let the training data lie on or near a smooth d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$. Let $S(\mathbf{x})$ be the expected squared reconstruction error of a DAE trained with small isotropic Gaussian corruption noise of variance η^2. Then, in the limit $\eta \to 0$, there exists a tubular neighborhood $\mathcal{N}_r(\mathcal{M}) = \{\mathbf{x} \in \mathbb{R}^D : \text{dist}(\mathbf{x}, \mathcal{M}) < r\}$ for some $r > 0$, such that

$$S(\mathbf{x}) = c \cdot \text{dist}(\mathbf{x}, \mathcal{M})^2 + o(\text{dist}(\mathbf{x}, \mathcal{M})^2),$$

where $c > 0$ is a constant independent of $\mathbf{x}$, and $\text{dist}(\mathbf{x}, \mathcal{M}) = \inf_{\mathbf{y} \in \mathcal{M}} \|\mathbf{x} - \mathbf{y}\|$.

Lemma A.2 (Geometric Properties of ∇S)**.** *In the tubular neighborhood $\mathcal{N}_r(\mathcal{M})$, the explicit scarcity gradient $\mathbf{v}_S(\mathbf{x}) := \nabla_{\mathbf{x}} S(\mathbf{x})$ satisfies:*

1. ***Normal Alignment:*** *For any $\mathbf{x} \in \mathcal{N}_r(\mathcal{M}) \setminus \mathcal{M}$, let $\mathbf{y} = \Pi_{\mathcal{M}}(\mathbf{x})$ be the unique projection of $\mathbf{x}$ onto $\mathcal{M}$. Then $\mathbf{v}_S(\mathbf{x})$ lies in the normal space of $\mathcal{M}$ at $\mathbf{y}$, i.e., $\mathbf{v}_S(\mathbf{x}) \in \mathcal{N}_{\mathbf{y}}\mathcal{M}$.*
2. ***Non-Vanishing Magnitude:*** *There exists $\delta > 0$ such that for all $\mathbf{x}$ with $\text{dist}(\mathbf{x}, \mathcal{M}) \geq \epsilon_0 > 0$ (for some fixed $\epsilon_0 < r$), we have $\|\mathbf{v}_S(\mathbf{x})\| \geq \delta$.*

Proof. By Assumption A.1, we have $S(\mathbf{x}) = c \cdot \text{dist}(\mathbf{x}, \mathcal{M})^2 + o(\text{dist}(\mathbf{x}, \mathcal{M})^2)$. It is a classical result in differential geometry Federer (1959) that the squared distance function to a smooth manifold is differentiable in its tubular neighborhood, and its gradient is given by

$$\nabla_{\mathbf{x}} \text{dist}(\mathbf{x}, \mathcal{M})^2 = 2(\mathbf{x} - \Pi_{\mathcal{M}}(\mathbf{x})).$$

The vector $(\mathbf{x} - \Pi_{\mathcal{M}}(\mathbf{x}))$ is orthogonal to the tangent space $T_{\Pi_{\mathcal{M}}(\mathbf{x})}\mathcal{M}$ by definition of the projection, and thus lies in the normal space $\mathcal{N}_{\Pi_{\mathcal{M}}(\mathbf{x})}\mathcal{M}$. Therefore,

$$\mathbf{v}_S(\mathbf{x}) = \nabla_{\mathbf{x}} S(\mathbf{x}) = 2c(\mathbf{x} - \Pi_{\mathcal{M}}(\mathbf{x})) + o(\|\mathbf{x} - \Pi_{\mathcal{M}}(\mathbf{x})\|),$$

which establishes the orthogonality property. For the magnitude, note that $\|\mathbf{x} - \Pi_{\mathcal{M}}(\mathbf{x})\| = \text{dist}(\mathbf{x}, \mathcal{M}) \geq \epsilon_0 > 0$. Hence,

$$\|\mathbf{v}_S(\mathbf{x})\| = \|2c(\mathbf{x} - \Pi_{\mathcal{M}}(\mathbf{x})) + o(\epsilon_0)\| \geq 2c\epsilon_0 - o(\epsilon_0).$$

Choosing ϵ_0 sufficiently small such that the higher-order term is dominated by the leading term, we obtain $\|\mathbf{v}_S(\mathbf{x})\| \geq \delta$ for some $\delta > 0$. □

A.4.2 ALIGNMENT OF IMPLICIT CFG GRADIENT

We now bridge the theoretical result above to the actual implementation in ISGE. The ISGE framework utilizes an implicit gradient via Classifier-Free Guidance (CFG):

$$\mathbf{g}_S(\mathbf{x}) := \nabla_{\mathbf{x}} \log p(y_S|\mathbf{x}),$$

where y_S represents the condition of "high scarcity".

Justification of the Chain Rule. Crucially, during the construction of the exploration dataset, the scarcity label y_S for any trajectory $\mathbf{x}$ is derived *exclusively* from the scarcity metric $S(\mathbf{x})$ (i.e., the reconstruction error). This establishes a deterministic dependency structure: $\mathbf{x} \to S(\mathbf{x}) \to y_S$. By the Markov property of this labeling process, y_S is conditionally independent of $\mathbf{x}$ given $S(\mathbf{x})$. Therefore, the likelihood satisfies the exact relation:

$$p(y_S|\mathbf{x}) = p(y_S|S(\mathbf{x})).$$

Consequently, the gradient calculation follows the chain rule strictly without approximation:

$$\mathbf{g}_S(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(y_S|S(\mathbf{x})) = \underbrace{\frac{\partial \log p(y_S|S)}{\partial S}}_{\alpha(S)} \cdot \underbrace{\nabla_{\mathbf{x}} S(\mathbf{x})}_{\mathbf{v}_S(\mathbf{x})}. \tag{44}$$

Directional Consistency. The scalar coefficient $\alpha(S) = \frac{\partial \log p(y_S|S)}{\partial S}$ determines the alignment. In our formulation, y_S is set to a high quantile (e.g., the 99th percentile) of the scarcity distribution. This implies that the likelihood $p(y_S|S)$ is a monotonically increasing function of S within the region of interest (as higher S indicates the sample is closer to the target condition y_S). Consequently, the derivative is strictly positive:

$$\alpha(S) > 0.$$

Thus, $\mathbf{g}_S(\mathbf{x})$ is a positive scaling of $\nabla_{\mathbf{x}} S(\mathbf{x})$. By Lemma A.2, the implicit CFG gradient $\mathbf{g}_S(\mathbf{x})$ inherits the geometric properties of **Orthogonality** (aligning with the manifold normal $\mathbf{n}_\perp$) and **Non-Vanishing Magnitude** (scaled by $\alpha(S)$). This justifies the use of $\mathbf{g}_S$ as a valid repulsion field in Theorem 3.6.

A.5 PROOF OF THEOREM 3.6

Assumption A.3 (Geometry of Scarcity Guidance)**.** To analyze the mechanics of Scarcity Guidance in high-dimensional OOD regions, we make the following geometric assumptions based on the theory of regularized autoencoders and high-dimensional statistics:

1. **Orthogonality (Low Tangential Drift):** The scarcity gradient $\mathbf{g}_S$ approximates the geodesic normal of the data manifold. Its projection onto the tangent space of the exploration sphere is negligible compared to the radial component: $\|\mathbf{g}_{S,\tan}\| \ll \langle \mathbf{g}_S, \mathbf{n}^* \rangle$.

2. **Frontier Alignment (Selection-Induced Optimality):** For the self-locked trajectories $\mathbf{x}$ residing on the high-return frontier $\mathcal{F} \subset \partial\mathcal{M}$, we assume the scarcity gradient aligns positively with the global optimum: $\langle \mathbf{g}_S(\mathbf{x}), \mathbf{n}^* \rangle > 0$.

   Justification: In the general case of an arbitrary return landscape, self-locking implies the exhaustion of improving directions within the manifold's tangent space (i.e., $\nabla_{\tan} R(\mathbf{x}) \approx 0$). Consequently, the true gradient of improvement $\mathbf{n}^*$ is dominated by its normal component. Since the training data represents a suboptimal subset, the direction of higher return at this boundary naturally points towards the unexplored exterior (outward polarity) rather than the interior. Thus, the scarcity gradient $\mathbf{g}_S$, which drives outward traversal, shares the same outward polarity with $\mathbf{n}^*$ with high probability. This geometric correlation ensures that scarcity-guided exploration reliably uncovers the cone of improving directions.

3. **Tangential Uncorrelation:** The tangential component of the return gradient, $\mathbf{g}_{R,\tan}$, arising from epistemic uncertainty, is statistically uncorrelated with the tangential component of the scarcity gradient, $\mathbf{g}_{S,\tan}$. Thus, their expected inner product vanishes: $\mathbb{E}[\langle \mathbf{g}_{R,\tan}, \mathbf{g}_{S,\tan} \rangle] \approx 0$.

Proof. We analyze the optimization dynamics using an incremental approach. Let the "Net Convergence Velocity" be $V := -\mathbb{E}[\Delta\epsilon]$. Stagnation of the baseline implies $V_{\text{base}} \approx 0$. The new velocity under dual guidance is $V_{\text{ISGE}}(w_S) = V_{\text{base}} + \Delta V(w_S)$. We aim to prove $\Delta V(w_S) > 0$ and find its maximum.

Step 1: Incremental Velocity Formulation The scarcity term $\lambda_t w_S \mathbf{g}_S$ contributes a radial gain and increases the tangential curvature penalty.

1. **Radial Gain:**
$$\Delta v_{\text{rad}} = \langle \lambda_t w_S \mathbf{g}_S, \mathbf{n}^* \rangle = \lambda_t w_S \mathcal{V}_S, \tag{45}$$
where $\mathcal{V}_S = \langle \mathbf{g}_S, \mathbf{n}^* \rangle > 0$ by Assumption A.3.2.

2. **Penalty Increase:** The tangential velocity becomes $\mathbf{v}_{\text{tan}}^{\text{new}} = \mathbf{v}_{\text{tan}}^{\text{base}} + \lambda_t w_S \mathbf{g}_{S,\tan}$. The change in expected tangential energy is:
$$\Delta\mathbb{E}[\|\mathbf{v}_{\text{tan}}\|^2] = \lambda_t^2 w_S^2 \|\mathbf{g}_{S,\tan}\|^2 + 2\lambda_t w_S \, \mathbb{E}[\langle \mathbf{v}_{\text{tan}}^{\text{base}}, \mathbf{g}_{S,\tan} \rangle]. \tag{46}$$
The cross-term decomposes as:
$$\mathbb{E}[\langle \mathbf{v}_{\text{tan}}^{\text{base}}, \mathbf{g}_{S,\tan} \rangle] = \lambda_t w_R^* \, \mathbb{E}[\langle \mathbf{g}_{R,\tan}, \mathbf{g}_{S,\tan} \rangle] + \sigma_t \, \mathbb{E}[\langle \mathbf{z}_{\tan}, \mathbf{g}_{S,\tan} \rangle].$$
The second term vanishes because $\mathbb{E}[\mathbf{z}_{\tan}] = 0$. The first term vanishes by Assumption A.3.3. Thus,
$$\Delta\text{Penalty} = \frac{\Delta\mathbb{E}[\|\mathbf{v}_{\text{tan}}\|^2]}{2\epsilon} = \frac{\lambda_t^2 w_S^2 \|\mathbf{g}_{S,\tan}\|^2}{2\epsilon}. \tag{47}$$

Combining these, the net velocity increment is a quadratic function of w_S:
$$\Delta V(w_S) = \lambda_t w_S \mathcal{V}_S - \frac{\lambda_t^2 w_S^2 \|\mathbf{g}_{S,\tan}\|^2}{2\epsilon}. \tag{48}$$

Step 2: Feasible Interval and Optimization Optimization is revived when $\Delta V(w_S) > 0$:
$$\lambda_t w_S \mathcal{V}_S > \frac{\lambda_t^2 w_S^2 \|\mathbf{g}_{S,\tan}\|^2}{2\epsilon} \implies w_S < \frac{2\epsilon \mathcal{V}_S}{\lambda_t \|\mathbf{g}_{S,\tan}\|^2}. \tag{49}$$

This defines the feasible interval $(0, w_{\max})$. To find the optimal scale, we set $\frac{d\Delta V}{dw_S} = 0$:
$$\lambda_t \mathcal{V}_S - \frac{\lambda_t^2 w_S \|\mathbf{g}_{S,\tan}\|^2}{\epsilon} = 0 \implies w_S^* = \frac{\epsilon \mathcal{V}_S}{\lambda_t \|\mathbf{g}_{S,\tan}\|^2}. \tag{50}$$

Step 3: Marginal Effective SNR Analysis We define the marginal effective SNR as the ratio of radial gain to penalty increase at w_S^*:
$$\text{SNR}_{\Delta\text{eff}}(w_S^*) = \frac{\Delta v_{\text{rad}}(w_S^*)}{\Delta\text{Penalty}(w_S^*)}. \tag{51}$$

Substituting w_S^*:

- Numerator: $\lambda_t \left(\frac{\epsilon \mathcal{V}_S}{\lambda_t \|\mathbf{g}_{S,\tan}\|^2} \right) \mathcal{V}_S = \frac{\epsilon \mathcal{V}_S^2}{\|\mathbf{g}_{S,\tan}\|^2}$,
- Denominator: $\frac{\lambda_t^2}{2\epsilon} \left(\frac{\epsilon \mathcal{V}_S}{\lambda_t \|\mathbf{g}_{S,\tan}\|^2} \right)^2 \|\mathbf{g}_{S,\tan}\|^2 = \frac{\epsilon \mathcal{V}_S^2}{2\|\mathbf{g}_{S,\tan}\|^2}$.

The ratio is exactly:
$$\text{SNR}_{\Delta\text{eff}}(w_S^*) = \frac{\epsilon \mathcal{V}_S^2 / \|\mathbf{g}_{S,\tan}\|^2}{\epsilon \mathcal{V}_S^2 / (2\|\mathbf{g}_{S,\tan}\|^2)} = 2. \tag{52}$$

This constant ratio reflects the intrinsic geometry of the scarcity field and confirms that the incremental update is always beneficial when $\mathcal{V}_S > 0$.

Step 4: Maximum Progress Rate Finally, substituting w_S^* into $\Delta V(w_S)$ yields the maximum progress:

$$\Delta V(w_S^*) = \frac{\epsilon \mathcal{V}_S^2}{\|\mathbf{g}_{S,\tan}\|^2} - \frac{\epsilon \mathcal{V}_S^2}{2\|\mathbf{g}_{S,\tan}\|^2} = \frac{\epsilon}{2}\left(\frac{\mathcal{V}_S}{\|\mathbf{g}_{S,\tan}\|}\right)^2. \tag{53}$$

Under Assumption A.3.1 (Orthogonality), the ratio $\mathcal{V}_S/\|\mathbf{g}_{S,\tan}\|$ is large, ensuring significant optimization progress. Since $\Delta V(w_S^*) > 0$ and $V_{\text{base}} \approx 0$, we conclude $V_{\text{ISGE}} > 0$, i.e., $\mathbb{E}[\Delta\epsilon]_{\text{ISGE}} < 0$, which proves the theorem. □

A.6 Proof of Theorem 3.7

Problem Setup. We analyze the functional gradient update at a test point $\mathbf{x}$ in the OOD region after fine-tuning on the gift sample set $\mathcal{D}_{\text{gift}} = \{\mathbf{x}_i\}_{i=1}^M$. Here, each sample $\mathbf{x}_i$ is generated starting from $\mathbf{x}$ (or its neighborhood) via the ISGE dynamics described in Theorem 3.6. The fine-tuning loss is $\mathcal{L}(\theta) = \frac{1}{2}\sum_{i=1}^M (f_\theta(\mathbf{x}_i) - y_{\text{gift}})^2$, where y_{gift} is the target high return. We assume $y_{\text{gift}} > f_{\text{old}}(\mathbf{x}_i)$, implying a positive prediction error $\delta_i := y_{\text{gift}} - f_{\text{old}}(\mathbf{x}_i) > 0$.

We adopt standard NTK assumptions for the implicit return function f_θ:

1. **NTK Regime:** The network evolves linearly with a fixed kernel $\Theta(\mathbf{x}, \mathbf{x}') = \kappa(\langle \mathbf{x}, \mathbf{x}' \rangle)$.
2. **Normalized Inputs:** $\|\mathbf{x}\| = \|\mathbf{x}_i\| = 1$.
3. **Kernel Monotonicity:** $\kappa'(z) > 0$ for $z \in [-1, 1]$ (e.g., as in ReLU NTK).

Derivation of Restored Gradient. Under gradient descent with step size η, the functional change at the test point $\mathbf{x}$ is:

$$\Delta f(\mathbf{x}) = \eta \sum_{i=1}^M \delta_i \, \Theta(\mathbf{x}, \mathbf{x}_i). \tag{54}$$

Taking the spatial gradient with respect to $\mathbf{x}$ yields the restored return signal:

$$\Delta \mathbf{g}_R(\mathbf{x}) = \nabla_{\mathbf{x}} \Delta f(\mathbf{x}) = \eta \sum_{i=1}^M \delta_i \, \kappa'(\cos\phi_i) \, \mathbf{v}_i, \tag{55}$$

where $\phi_i = \angle(\mathbf{x}, \mathbf{x}_i)$ is the angle, and $\mathbf{v}_i = \mathbf{x}_i - \langle \mathbf{x}, \mathbf{x}_i \rangle \mathbf{x}$ represents the projection of $\mathbf{x}_i$ onto the tangent space $T_{\mathbf{x}}\mathbb{S}^{D-1}$.

Structure of Gift Samples from ISGE. According to the ISGE update rule (Theorem 3.6), each sample $\mathbf{x}_i$ is generated by:

$$\mathbf{x}_i = \mathbf{x} + \lambda w_S \mathbf{g}_S(\mathbf{x}) + \sigma \mathbf{z}_i + o(1), \tag{56}$$

where $\mathbf{g}_S(\mathbf{x})$ is the scarcity guidance direction (aligned with the manifold normal per Proposition 3.5), and $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I})$ is isotropic Gaussian noise. Projecting this update onto the tangent space $T_{\mathbf{x}}\mathbb{S}^{D-1}$ gives:

$$\mathbf{v}_i = \underbrace{\lambda w_S \mathbf{g}_{S,\text{proj}}}_{\mathbf{d}} + \underbrace{\sigma \mathbf{z}_{i,\text{proj}}}_{\mathbf{n}_i} + o(1), \tag{57}$$

where $\mathbf{g}_{S,\text{proj}} = \mathbf{g}_S - \langle \mathbf{g}_S, \mathbf{x} \rangle \mathbf{x}$ is the effective tangential scarcity component, $\mathbf{d}$ is the deterministic drift vector, and $\mathbf{n}_i$ is the projected noise vector. In the following proof, we first show $\mathbb{E}\big[\langle \Delta \mathbf{g}_R(\mathbf{x}), \mathbf{g}_S(\mathbf{x}) \rangle\big] > 0$, then we further show $\mathbb{E}[\langle \Delta \mathbf{g}_R, \mathbf{n}^* \rangle] > 0$.

Expected Alignment Analysis. We analyze the expected alignment between the restored return gradient and the scarcity direction:

$$\mathcal{I} := \mathbb{E}\big[\langle \Delta \mathbf{g}_R(\mathbf{x}), \mathbf{g}_S(\mathbf{x}) \rangle\big] = \eta \sum_{i=1}^M \mathbb{E}_{\mathbf{z}_i}\left[c_i(\mathbf{z}_i) \, \langle \mathbf{d} + \mathbf{n}_i, \mathbf{g}_S \rangle\right],$$

where the scalar weight $c_i(\mathbf{z}_i) := \delta_i(\mathbf{z}_i) \, \kappa'(\cos\phi_i(\mathbf{z}_i))$ captures the sensitivity of the update. Note that $c_i > 0$ due to the positive error δ_i and kernel monotonicity.

We perform a first-order Taylor expansion of $c_i(\mathbf{z}_i)$ around the noise-free center ($\mathbf{z}_i = 0$):

$$c_i(\mathbf{z}_i) \approx \bar{c} + (\nabla_{\mathbf{z}} c)_0^\top \mathbf{z}_i, \tag{58}$$

where $\bar{c} = c_i(0) > 0$ is the base coefficient determined by the deterministic scarcity drift.

Substituting this into $\mathcal{I}$ and expanding the product terms, we obtain:

$$\mathcal{I} = \eta M \Big(\underbrace{\bar{c}\,\langle \mathbf{d}, \mathbf{g}_S \rangle}_{\text{Term 1: Main Signal}} + \underbrace{\bar{c}\,\mathbb{E}[\langle \mathbf{n}_i, \mathbf{g}_S \rangle]}_{\text{Term 2: Zero Mean}} + \underbrace{\mathbb{E}\big[((\nabla c)_0^\top \mathbf{z}_i)\,\langle \mathbf{n}_i, \mathbf{g}_S \rangle\big]}_{\text{Term 3: Covariance Drag}} \Big). \tag{59}$$

We analyze each term individually:

1. **Term 1 (Main Signal):** Substituting $\mathbf{d} = \lambda w_S \mathbf{g}_{S,\text{proj}}$, we have:
$$\langle \mathbf{d}, \mathbf{g}_S \rangle = \lambda w_S \langle \mathbf{g}_{S,\text{proj}}, \mathbf{g}_S \rangle = \lambda w_S \|\mathbf{g}_{S,\text{proj}}\|^2.$$
By Proposition 3.5, $\|\mathbf{g}_S\| > 0$ and is orthogonal to the manifold, ensuring a non-vanishing projection. Thus, Term 1 is strictly positive and scales linearly with the guidance strength w_S.
2. **Term 2 (Zero Mean):** Since the noise $\mathbf{n}_i = \sigma \mathbf{z}_{i,\text{proj}}$ is zero-mean isotropic Gaussian, its projection onto any fixed vector vanishes: $\mathbb{E}[\langle \mathbf{n}_i, \mathbf{g}_S \rangle] = 0$.
3. **Term 3 (Covariance Drag):** This term captures the correlation between the supervision weight c_i and the noise direction. Geometrically, moving further in the direction of the gradient $\mathbf{g}_S$ brings the sample $\mathbf{x}_i$ closer to the target distribution (higher return), thereby reducing the prediction error δ_i and the weight c_i. This implies a negative correlation between $\mathbf{z}_i$ (along $\mathbf{g}_S$) and $c_i(\mathbf{z}_i)$. Consequently, Term 3 acts as a drag force (non-positive). Crucially, its magnitude scales with the noise variance:
$$|\text{Term 3}| = O(\mathbb{E}[\|\mathbf{n}_i\|^2]) = O(\sigma^2).$$

Applying the SNR Condition. Theorem 3.6 establishes that ISGE operates in the high Effective SNR regime, where the deterministic radial signal dominates the noise. In terms of energy, this implies:

$$\|\mathbf{d}\|^2 = (\lambda w_S)^2 \|\mathbf{g}_{S,\text{proj}}\|^2 \gg \mathbb{E}[\|\mathbf{n}_i\|^2] \propto \sigma^2 D.$$

Since Term 1 scales with $\|\mathbf{d}\|$ (linear in w_S) while Term 3 scales with σ^2 (independent of w_S), the high-SNR condition ensures that the main signal dominates the second-order covariance drag:

$$\text{Term 1} \gg |\text{Term 3}|.$$

Therefore, the total expected alignment is strictly positive:

$$\mathcal{I} > 0. \tag{60}$$

Vector Decomposition Analysis. While the scalar alignment $\mathcal{I} > 0$ proves that the update has a positive component along $\mathbf{g}_S$, establishing the connection to the optimal direction $\mathbf{n}^*$ requires analyzing the vector structure of the expected update $\mathbb{E}[\Delta \mathbf{g}_R]$.

Recall the update equation:

$$\Delta \mathbf{g}_R = \eta \sum_{i=1}^{M} c_i(\mathbf{z}_i)(\mathbf{d} + \mathbf{n}_i).$$

Taking the expectation over the isotropic noise $\mathbf{z}_i$, we decompose the result into a signal vector and a noise-induced drift vector:

$$\mathbb{E}[\Delta \mathbf{g}_R] = \eta M \Big(\underbrace{\mathbb{E}[c_i]\mathbf{d}}_{\text{Signal Vector}} + \underbrace{\mathbb{E}[c_i(\mathbf{z}_i)\mathbf{n}_i]}_{\text{Correlation Vector}} \Big). \tag{61}$$

1. **Signal Vector:** Since $\mathbf{d} = \lambda w_S \mathbf{g}_{S,\text{proj}}$, this term is strictly aligned with the scarcity gradient $\mathbf{g}_S$. Since $c_i > 0$, this vector points in the direction of $\mathbf{g}_S$.

2. **Correlation Vector:** Using the first-order Taylor expansion $c_i(\mathbf{z}_i) \approx \bar{c} + \mathbf{g}_c^\top \mathbf{z}_i$, the zero-mean term vanishes ($\bar{c}\mathbb{E}[\mathbf{n}_i] = 0$). The remaining term is the covariance $\mathbb{E}[(\mathbf{g}_c^\top \mathbf{z}_i)\mathbf{n}_i]$. Geometrically, the coefficient c_i (reflecting prediction error) decreases most rapidly when moving along the gradient of improvement. Since the dominant signal direction is $\mathbf{g}_S$, the sensitivity vector $\mathbf{g}_c$ is primarily anti-parallel to $\mathbf{g}_S$. Thus, this correlation vector acts as a "drag" force opposing $\mathbf{d}$, effectively reducing the magnitude but not altering the primary direction. Any orthogonal components in $\mathbf{g}_c$ arise from random local curvature and result in a residual vector $\mathbf{v}_\perp$.

Combining these terms, the expected gradient update takes the form:

$$\mathbb{E}[\Delta \mathbf{g}_R] = k_{\text{eff}} \cdot \mathbf{g}_S + \mathbf{v}_\perp, \tag{62}$$

where $k_{\text{eff}} > 0$ is the effective scalar gain (guaranteed positive by the high-SNR condition in Theorem 3.6), and $\mathbf{v}_\perp$ represents tangential noise residuals.

Projection onto the Optimal Direction $\mathbf{n}^*$. Finally, we evaluate the projection of this restored gradient onto the true optimal direction $\mathbf{n}^*$:

$$\langle \mathbb{E}[\Delta \mathbf{g}_R], \mathbf{n}^* \rangle = k_{\text{eff}} \langle \mathbf{g}_S, \mathbf{n}^* \rangle + \langle \mathbf{v}_\perp, \mathbf{n}^* \rangle.$$

1. The first term is strictly positive due to Assumption A.3 (Point 2: Outward Optimality), which posits $\langle \mathbf{g}_S, \mathbf{n}^* \rangle > 0$.
2. The second term vanishes due to Assumption A.3 (Point 3: Tangential Uncorrelation), which states that random tangential noise in high dimensions is statistically uncorrelated with the fixed optimal direction $\mathbf{n}^*$ (i.e., $\mathbb{E}[\langle \mathbf{v}_\perp, \mathbf{n}^* \rangle] \approx 0$).

Conclusion. Therefore, we conclude that $\mathbb{E}[\langle \Delta \mathbf{g}_R, \mathbf{n}^* \rangle] > 0$. The fine-tuning process explicitly restores the radial component of the return gradient, enabling the model to effectively extrapolate towards the global optimum. □

B Generalization to Quadratic Landscapes

In the main text, we derived the geometric barrier under a spherical reward landscape $R(x) \approx R(x^*) - \frac{1}{2}\|x - x^*\|^2$ and isotropic noise $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. Here, we justify its applicability to general smooth landscapes near a strict local optimum.

1. Local Quadratic Model. Let x^* be a strict local maximum of the reward function $R(x)$. By Taylor expansion,

$$R(x) \approx R(x^*) - \frac{1}{2}(x - x^*)^\top \mathbf{A}(x - x^*), \tag{63}$$

where $\mathbf{A} = -\nabla^2 R(x^*) \succ 0$ is positive definite. The level sets $\{x \mid R(x) = r\}$ are hyper-ellipsoids defined by $(x - x^*)^\top \mathbf{A}(x - x^*) = 2(R(x^*) - r)$.

The Langevin dynamics for reward maximization are:

$$dx_t = \nabla R(x_t)\, dt + \sigma\, dW_t \approx -\mathbf{A}(x_t - x^*)\, dt + \sigma\, dW_t, \tag{64}$$

with isotropic noise $\sigma\, dW_t$, consistent with standard diffusion samplers.

2. Tangential Noise Energy is Dimension-Dependent, Not Curvature-Dependent. The key quantity in Theorem 3.3 is the expected tangential noise energy $\mathbb{E}[\|\mathbf{v}_{\text{noise,tan}}\|^2]$, where $\mathbf{v}_{\text{noise,tan}}$ is the projection of the noise vector onto the tangent space of the level set at x.

For any smooth level set of codimension 1, the tangent space is $(D-1)$-dimensional. Under isotropic noise $\mathbf{v}_{\text{noise}} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, the expected squared norm of its tangential projection is:

$$\mathbb{E}[\|\mathbf{v}_{\text{noise,tan}}\|^2] = \sigma^2 \cdot \text{Tr}(\mathbf{P}_{\text{tan}}) = \sigma^2 (D - 1), \tag{65}$$

where $\mathbf{P}_{\text{tan}}$ is the orthogonal projector onto the tangent space. Crucially, this result depends only on the dimensionality of the tangent space, not on the eigenvalues of $\mathbf{A}$.

Thus, the curvature penalty term $\frac{\|\mathbf{v}_{\text{tan}}\|^2}{2\epsilon}$ in Theorem 3.3 remains valid for general quadratic basins under isotropic noise.

3. Why the Spherical Model is Conservative. In an ill-conditioned landscape (e.g., $\mathbf{A}$ has small eigenvalues), the radial signal $v_{\text{rad}} = \langle \nabla R(x), \mathbf{n}^* \rangle$ becomes weaker in flat directions, while the tangential noise energy remains $\sigma^2(D-1)$. This makes the effective SNR even smaller than in the spherical case with the same average curvature.

Therefore, the spherical analysis—where all directions have equal curvature—provides a *best-case* scenario for optimization. Real-world anisotropic landscapes are strictly harder, making our spherical-derived SNR threshold a *conservative lower bound* on the difficulty of high-dimensional exploration.

Conclusion. The hyperspherical model in Theorem 3.3 isolates the universal effect of dimensionality: under isotropic noise, the entropic barrier scales as $O(D)$ regardless of landscape conditioning. This justifies its use as a canonical model for analyzing self-locking in high-dimensional optimization.

C Experiments Details

C.1 Setup

Benchmark and Dataset We evaluate our method on **AuctionNet** Su et al. (2025), a large-scale industrial auto-bidding benchmark derived from real-world advertising campaigns. AuctionNet provides a flexible and researcher-friendly experimental environment for large-scale auto-bidding studies. To enable efficient iterative experimentation, we adopt a lightweight configuration of the environment. Specifically, we reduce the total traffic volume from 500,000 to 50,000 per evaluation while scaling the mean traffic value by a factor of ten, thereby preserving the overall order of magnitude of total conversions.

The environment consists of 48 competing agents with varying budgets and CPAs, instantiated from a diverse set of bidding strategies, including PID, OnlineLP, BC, BCQ Fujimoto et al. (2019), TD3+BC Fujimoto & Gu (2021), IQL Kostrikov et al. (2021), CQL Kumar et al. (2020), MBRL-MOPO Yu et al. (2020), and MBRL-COMBO Yu et al. (2021). These strategies span a wide spectrum from rule-based methods to reinforcement learning approaches, covering both conservative and aggressive bidding behaviors. This diversity is designed to faithfully emulate the complexity and heterogeneity of real-world online advertising markets. Since the environment is operated under a lightweight configuration, we retrain all built-in agents to ensure that their policies remain sufficiently competitive under the modified traffic dynamics. In each evaluation episode, our method sequentially interacts with the environment by assuming the role of each of the 48 agents. The final performance of an episode is reported as the average score across all agents.

To faithfully simulate the cold-start dilemma discussed in Section 1, we construct a *narrowly supported dataset* by rigorously filtering the full offline logs, retaining only trajectories in the return percentile range $[10, 30]$. This creates a challenging learning environment where the data is both low-quality (far below optimality) and narrowly supported (lacking both high-performing demonstrations and diverse failure cases), forcing the model to extrapolate into the unknown. We establish two distinct experimental settings based on traffic characteristics: (1) **AuctionNet-Dense**, characterized by high traffic value ($p^d_{avg} \approx 0.05$) under strict safety constraints; and (2) **AuctionNet-Sparse**, representing a harder exploration landscape with extremely low traffic value ($p^s_{avg} \approx 0.005$) and sparse reward signals. For each iteration, we conduct 63 evaluation episodes and log the complete trajectory data generated by all agents. This evaluation procedure requires approximately 3 hours using four GPUs in parallel, whereas running the same protocol under the default AuctionNet configuration would require over 20 hours each iteration. Detailed parameters are shown in Table 5.

In detail, the state includes the following information:

- time_left: The remaining time steps left in the current advertising period.
- budget_left: The remaining budget that the advertiser has available to spend in the current advertising period.
- historical_bid_mean: The average values of bids made by the advertiser over past time steps.

Table 5: The Parameters of AuctionNet-Dense and AuctionNet-Sparse.

Params	AuctionNet-Dense	AuctionNet-Sparse
Impressions (each episode)	50000	50000
Time steps in a trajectory	48	48
Advertiser Number	48	48
State dimension	16	16
Action dimension	1	1
CPA range	[6, 13]	[60, 130]
Mean Traffic Value	0.05	0.005
Trajectories (full)	5376	5040
Max Score (full)	276.6	63.0
Avg Score (full)	86.1	13.5
Trajectories (cold start)	1092	1318
Max Score (cold start)	111.1	21.0
Avg Score (cold start)	55.7	5.1
Trajectories for Iter 1-3	3024	3024

- last_three_bid_mean: The average values of bids over the last three time steps.
- historical_LeastWinningCost_mean: The average of the least cost required to win an impression over previous time steps.
- historical_pValues_mean: The average of historical p-values over past time steps.
- historical_conversion_mean: The average number of conversions (e.g., sales, clicks, etc.) the advertiser achieved in previous time steps.
- historical_xi_mean: The average winning status of advertisers in impression opportunities, where 1 represents winning and 0 represents not winning.
- last_three_LeastWinningCost_mean: The average of the least winning costs over the last three time steps.
- last_three_pValues_mean: The average of conversion probability of advertising exposure to users over the last three time steps.
- last_three_conversion_mean: The average number of conversions over the last three time steps.
- last_three_xi_mean: The average winning status of advertisers over the last three time steps.
- current_pValues_mean: The mean of p-values at the current time step.
- current_pv_num: The number of impressions served at the current time step.
- last_three_pv_num_total: The total number of impressions served over the last three time steps.
- historical_pv_num_total: The total number of impressions served over past time steps.

Baselines. We compare ISGE against a comprehensive suite of methods trained strictly on the narrowly supported data. To contextualize performance against established paradigms, we first evaluate Behavior Cloning (BC) Osa et al. (2018), the classical baseline for imitation learning, and Decision Transformer (DT) Chen et al. (2021), the representative autoregressive model for generative decision-making. Building on the diffusion paradigm, we employ Diffusion-planner (Iter 0), a diffusion planner Guo et al. (2024) equipped with a simulation-based inverse controller (see C.2), serving as a strong offline baseline. Second, we evaluate Vanilla Self-Training, the standard iterative framework utilizing CFG but lacking directed scarcity exploration. Finally, we include an Oracle, a diffusion planner trained on the Full Dataset, which represents the theoretical ceiling of the architecture with given data.

Metrics We first give the problem statement of auto-bidding He et al. (2021). Consider a scenario where there are H impression opportunities arriving sequentially, each labeled by an index i. An advertiser wins an impression if its bid b_i surpasses those of other advertisers and incurs a cost c_i. The goal is to maximize the total value of the impressions won, represented by $\sum_i o_i v_i$, where v_i denotes the impression value and o_i is a binary variable indicating whether the advertiser wins

impression i. Additionally, it is essential to consider the budget and various KPI constraints, such as limiting the unit cost of specific advertising events like CPC and CPA. Therefore, the auto-bidding problem could be formulated as:

$$\begin{aligned} \text{maximize} \quad & \sum\nolimits_i o_i v_i \\ \text{s.t.} \quad & \sum\nolimits_i o_i c_i \le B \\ & \frac{\sum_i c_{ij} o_i}{\sum_i p_{ij} o_i} \le C_j, \quad \forall j \\ & o_i \in \{0,1\}, \quad \forall i \end{aligned} \tag{66}$$

where B is the budget, C_j is the upper bound of the j-th constraint provided by the advertiser. p_{ij} can be any performance indicator, such as conversions, and c_{ij} is the cost of constraint j.

We report four key metrics averaged among 48 advertisers over 7 evaluation episodes (total 336 bidding trajectories):

- **WR** (Win Rate): The fraction of page views (PVs) in which the advertiser wins the auction, $\text{WR} = \frac{1}{H}\sum_{i=1}^{H} o_i$, where H is the total number of participated impressions.
- **Value**: The cumulative conversion value (GMV) $\sum_i o_i v_i$;
- **ER** (Exceeding ratio of the CPA constraints): The exceedance ratio of Cost-Per-Action (CPA) constraints $\text{ER} = \frac{1}{J}\sum_j C_j^{real}/C_j = \frac{1}{J}\sum_j (\sum_i c_{ij} o_i / \sum_i p_{ij} o_i)/C_j$, where C_j is the CPA limit.
- **Score**: A composite metric balancing revenue and constraints, $\text{Score} = (\sum_i o_i v_i) \times min\{penalty_j\}_{j=1\sim J}$. where $\text{penalty}_j = \min\{(\frac{C_j}{\sum_i c_{ij} o_i / \sum_i p_{ij} o_i})^\beta, 1\}, \beta = 2$.

C.2 Simulation-based Inverse Control

In our experiments, we employ a *Simulation-based Inverse Control* mechanism for diffusion planner. This design choice allows us to isolate the trajectory planning capability of the diffusion model from the execution errors typically introduced by heuristic controllers. The simulation-based inverse controller solves for the precise action required to realize the planner's state transitions.

Formally, let $\tau_{plan} = \{\hat{s}_t, \hat{s}_{t+1}, \ldots, \hat{s}_{t+H}\}$ denote the trajectory generated by the diffusion planner at time step t. The planner dictates that the remaining budget should transition from the current state b_t to $\hat{b}_{t+1}$ in the next step. Consequently, the target expenditure for the current step is derived as $c^*_{target} = \max(0, b_t - \hat{b}_{t+1})$. The objective of the inverse control is to find a bidding scalar α_t such that the actual cost incurred in the auction environment, $\mathcal{C}(\alpha_t, \mathbf{B}_{-i})$, approximates c^*_{target} as closely as possible, where $\mathbf{B}_{-i}$ represents the bids of competitors. We leverage the reproducibility of the simulation environment. During the evaluation of this baseline, we grant the controller oracle access to a "frozen" copy of the current environment state, including the exact bids of all competing agents. Leveraging the monotonicity of the cost function with respect to the bid scaling factor α, we employ a **Simulation-based Binary Search** to numerically invert the cost function. This mechanism ensures that the realized trajectory aligns almost perfectly with the planned trajectory, barring stochastic deviations in impression exposure (e.g., Bernoulli trials). By effectively eliminating control latency and execution variance, this setup enables a rigorous evaluation of the *planning quality* itself, validating that performance improvements are attributable to the diffusion planner rather than the low-level controller.

Remark on Real-World Feasibility. We acknowledge that the oracle access used in the simulation-based binary search represents an idealized control setting. However, this design is intentional and serves to rigorously decouple the evaluation of our generative planning framework from the noise of low-level execution. In practical production environments where precise environment inversion is unavailable, this module is modular and can be seamlessly substituted with standard engineering approximations without altering the core diffusion planning mechanism:

- **Learned Inverse Dynamics:** A separate lightweight network (Inverse Dynamics Model Ajay et al. (2022)) can be trained to predict the action a_t required to achieve the state transition $s_t \to s_{t+1}$.

- **Feedback Control (PID):** Rule-based PID controllers, which are ubiquitous in industrial auto-bidding systems, can dynamically adjust the bidding scalar α_t to track the target expenditure c^*_{target} with bounded error.

While replacing the simulation-based controller with these approximate methods may introduce minor execution noise, the *relative performance advantage* provided by the superior strategic guidance of the diffusion planner remains robust. The planner determines *where* to go, which is the dominant factor in performance, while the controller merely handles *how* to get there. Table 6 shows the performance difference between simulation-based inverse controller and standard inverse dynamic controller in AuctionNet-Dense. Remarkably, equipped with the standard controller, ISGE (Iter 1) still outperforms the Full Data Oracle.

Table 6: Comparison of Scores with Different Controllers in AuctionNet-Dense.

Method	**Controller Type**	
	Simulation-based	**Standard Inverse**
Oracle (Full Dataset)	102.0	95.1
Vanilla (Iter 1)	74.9	69.1
ISGE (Iter 1, Ours)	103.4	97.4

C.3 Hyperparameters

We list the hyperparameter details in Table 7 for reproduction.

Table 7: Hyperparameter Settings for DFUSER and DFUSER$_{\text{SC}}$ (Explorer)

Category	**Parameter**	**Value**
Diffusion Process	Diffusion Steps (T)	20
	Horizon (H)	48
	Beta Schedule	Cosine
	Guidance Weight (w)	3.0
	Condition Dropout	0.25 (0.30 for DFUSER$_{\text{SC}}$)
	Replanning Prob.	0.5
	Prediction Target	ϵ
Network Architecture	Model Type	Temporal U-Net
	Base Channel Dim	64
	Channel Multipliers	(1, 2)
	Kernel Size	5
	Activation	Mish
	InvDyn Hidden Dim	256
Optimization	Optimizer	Adam
	Learning Rate	1×10^{-4}
	Discount Factor (γ)	1.0

D Implementation Details of Undirected Exploration Baselines

To evaluate the effectiveness of our proposed method, we compare it against two standard exploration strategies: Action Space Noise (ASN) Mou et al. (2022) and Parameter Space Noise (PSN) Li et al. (2024). We introduce these noises during the inference phase to simulate different levels of exploration behavior. The implementation details and hyperparameter tuning processes are described below.

D.1 Action Space Noise (ASN)

ASN is a straightforward exploration strategy commonly used in reinforcement learning, where perturbation is directly applied to the agent's output action. In our automated bidding scenario, the

action is represented by the bidding coefficient α. At each decision step t, after the model calculates the baseline bidding coefficient α_t, we add a noise term ϵ_{asn} sampled from a Gaussian distribution. Since bids must be non-negative, we apply a rectification operation:

$$\alpha'_t = \max(0, \alpha_t + \epsilon_{\text{asn}}), \quad \text{where } \epsilon_{\text{asn}} \sim \mathcal{N}(0, \sigma^2_{\text{asn}}) \tag{67}$$

This corresponds to the logic where the final calculated bid coefficient is perturbed before interacting with the auction environment.

Since ASN operates directly on the action value, its scale has a direct physical interpretation (i.e., modifying the bid multiplier). Consequently, a relatively larger noise scale is required to induce significant exploration. We performed a grid search for σ_{asn} over the set $\{0.01, 0.05, 0.1, 0.5, 1.0\}$. We observed that smaller scales had negligible impact on the bidding outcome, while larger scales led to unstable performance. We selected $\sigma_{\text{asn}} = 0.1$ as the optimal configuration for both AuctionNet-Dense and AuctionNet-Sparse.

D.2 Parameter Space Noise (PSN)

PSN induces state-dependent exploration by perturbing the weights of the neural network during the planning phase. This approach allows the agent to "imagine" diverse trajectories consistent with the model's internal uncertainty.

We introduce noise into the Diffusion U-Net during the **inference-time planning** process. Before generating a trajectory, we inject Gaussian noise into the learnable parameters θ of the Temporal U-Net:

$$\tilde{\theta} = \theta + \epsilon_{\text{psn}}, \quad \text{where } \epsilon_{\text{psn}} \sim \mathcal{N}(0, \sigma^2_{\text{psn}}) \tag{68}$$

Crucially, our implementation utilizes a *transient injection* mechanism: the noise is applied solely for the current planning step's reverse diffusion loop and is immediately removed (weights are restored) afterwards. This ensures that the model parameters remain stable for subsequent evaluations, isolating the exploration effect to the current decision. The noise is applied only to the trajectory generation model (U-Net), leaving the Inverse Dynamics model untouched to preserve action execution precision.

Deep neural networks are highly sensitive to weight perturbations, as small changes are amplified through multiple layers of non-linear transformations. Therefore, the scale for PSN must be significantly smaller than that of ASN. We tuned σ_{psn} over the set $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}\}$. Our experiments showed that $\sigma_{\text{psn}} > 10^{-2}$ caused trajectory collapse (e.g., generating invalid or negative budgets), rendering the planning infeasible. We selected $\sigma_{\text{psn}} = 1 \times 10^{-3}$ for AuctionNet-Dense and $\sigma_{\text{psn}} = 5 \times 10^{-4}$ for AuctionNet-Sparse to balance exploration diversity with trajectory validity.

Table 8: Summary of Baseline Implementation and Hyperparameters

Method	**Injection Target**	**Noise Type**	**Search Candidates** (σ)	**Selected**
ASN	Final Action (α)	Gaussian	$\{0.01, 0.05, \mathbf{0.1}, 0.5, 1.0\}$	0.1
PSN	U-Net Weights (θ)	Gaussian	{1e-4, **5e-4**, **1e-3**, 5e-3, 1e-2}	1e-3 (Dense), 5e-4 (Sparse)