REGRET-AWARE ADAPTIVE SCHEDULING FOR RESOURCE-CONSTRAINED AUTO-BIDDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Frequent decision updates are crucial in dynamic environments, yet strict computational resource constraints severely restrict execution frequency. Existing methods for managing this trade-off either fail to align with horizon-wide system utility, rely on inflexible schedules, or incur profound search complexity through coupled action-timing exploration. To address these limitations, we propose **AIGB-ReAL** (**Re**gret-Aware **A**daptive Schedu**L**ing), a novel framework for resource-constrained temporal scheduling. Conceptually, an unconstrained system can optimize actions across the horizon at every time step, whereas a limited compute budget requires merging adjacent steps to share the same action. We introduce *Merge Regret* to quantify the resulting reduction in optimal cumulative value after global re-optimization and estimate it using a *Decision Landscape*. For auto-bidding, this landscape is represented by volume-price response curves, guiding low-regret merging and responsive triggering with reduced reliance on joint exploration. Extensive experiments in simulated and real-world advertising systems demonstrate the superiority of our approach, achieving near-oracle performance with significantly fewer updates.

1 INTRODUCTION

In Real-Time Bidding (RTB), automated algorithms dynamically adjust bids to maximize advertising value under the budget (Cai et al., 2017). While frequent updates enable timely adaptation to market changes (Zhang et al., 2016), industrial platforms operate under strict computational constraints (Zhu et al., 2017; Mou et al., 2022), primarily Queries-Per-Second (QPS) limits. Practical systems therefore often update parameters at fixed intervals (Wu et al., 2018; Li et al., 2024). Despite their widespread use, these fixed schedules waste updates during stable periods and delay responses to sudden shifts. This necessitates a more intelligent paradigm that adaptively balances market responsiveness with strict resource constraints.

Similar trade-offs extend (Wu et al., 2019; Lin et al., 2023) to domains like autonomous systems, financial trading and energy management (Postoyan et al., 2015; Paden et al., 2016; Dolk et al., 2017; Pan et al., 2023). To address this, a prevalent mechanism is *Event-Triggered Control (ETC)*, which triggers policy updates based on predefined state deviation thresholds (Tabuada, 2007; Heemels et al., 2012; Dimarogonas et al., 2012; Hertneck et al., 2026). While supporting responsive control, it is inherently limited to a *short horizon*, as the magnitude of a localized state deviation does not directly measure the long-term benefit of an update.(Demirel et al., 2017; Solowjow & Trimpe, 2020).

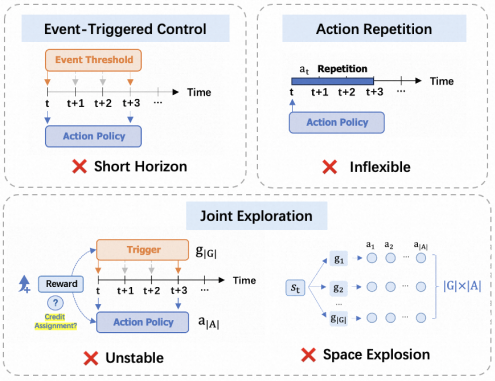

Figure 1: Limitations of existing paradigms.

Alternatively, temporal abstraction in reinforcement learning reduces decision frequency using *Action Repetition*, which selects how long actions persist (Srinivas et al., 2016; 2017; Sharma et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2017; Metelli et al., 2020; Biedenkapp & Lindauer, 2021; Seohong Park & Kim, 2021). However, because the duration is determined in advance, these methods cannot revise the action using observations received during its execution, rendering them *inflexible* in responding to abrupt environmental changes. (Yu & Zhang, 2021). To retain responsiveness, some approaches adaptively consider *Joint Exploration* (Neunert et al., 2020; Chen et al., 2021a). One development in this direction is LICRA (Mguni et al., 2023), which jointly trains an acting policy with a binary trigger that determines whether to execute the new action. Despite these advances, jointly learning actions and update timing remains challenging. This joint learning paradigm inherently results in ambiguous *credit assignment* (Mesnard et al., 2021) and noisy optimization signals, leading to *unstable* policy convergence. Moreover, this joint formulation suffers from a *space* explosion as the combined search space grows rapidly over long horizons, while sparse triggers provide limited samples for action learning.

These limitations (as directly shown in Figure 1) necessitate a decoupled scheduling paradigm that explicitly accounts for global horizon. Our core insight is that, without QPS constraints, optimizing bids at every timestep yields the highest achievable value, whereas a limited update budget requires merging adjacent steps to share the same action, potentially reducing this optimal value. We introduce *Merge Regret* to quantify the global loss in optimal cumulative value caused by merging adjacent intervals. To estimate this global metric into a reactive decision boundary, we construct a *Decision Landscape*, evaluating structural changes in achievable outcomes. For auto-bidding, the *Decision Landscape* is represented by volume-price response curves that relate bidding multipliers to expected advertising value and spending.

Building on this theoretical foundation, we propose the **AIGB-ReAL** (**Re**gret-Aware **A**daptive Schedu**L**ing) framework. The overall methodology operates through three cohesive stages as shown in Figure 2. First, a trigger mechanism guided by merge regret is employed to derive an optimal partition, establishing a globally optimized schedule. Second, once this optimal partition is fixed, the action policy is optimized via reinforcement learning to determine the best action strategies within the partitioned intervals. By decoupling the timing of updates from action selection, this stage ensures stable policy training. Finally, during online execution, the framework continuously evaluates the merge regret in real time, enabling flexible adaptation to dynamic market conditions.

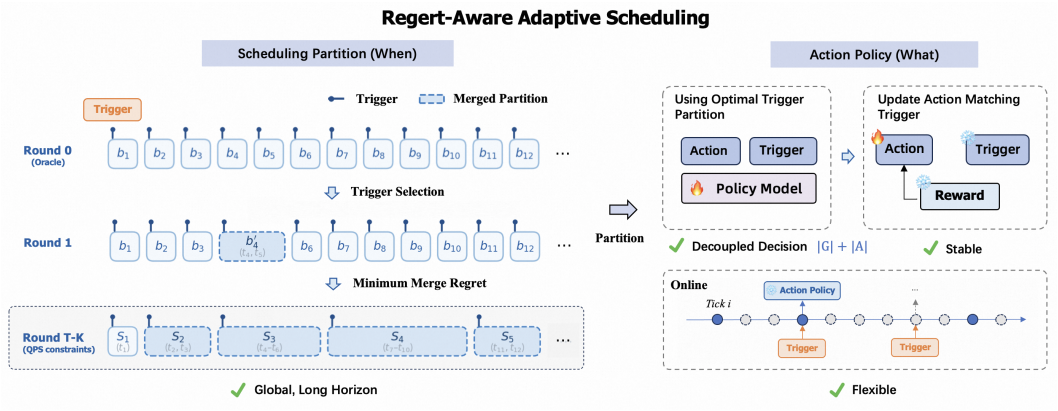

Figure 2: The overall framework of AIGB-ReAL. This architecture decouples the problem into two components. The left part illustrates the **Scheduling Partition (When)** module, which iteratively merges adjacent intervals based on minimum merge regret to satisfy QPS constraints over a global horizon. The right panel depicts the **Action Policy (What)** module, showing the decoupled decision-making design ($|G| + |A|$) that ensures stable RL updates and flexible online execution.

Our contributions are summarized as follows: (i) To the best of our knowledge, this is the first work to formulate auto-bidding under a limited update budget (e.g., QPS limits) as a resource-constrained problem, evaluating update opportunities through their horizon-wide economic value. (ii) We introduce *Merge Regret*, a decision-centric metric that quantifies the performance gap to the unconstrained optimal, and link it to the *Decision Landscape*. This provides a theoretical foundation for regret-aware scheduling, fundamentally decoupling update timing from policy optimization. (iii) Extensive offline experiments on AuctionNet and real-world online A/B tests demonstrate that

AIGB-ReAL substantially outperforms other baselines, achieving near-oracle performance with significantly fewer decision updates.

2 PRELIMINARIES

In Real-Time Bidding (RTB) platforms, an auto-bidding agent interacts with an auction environment over a finite horizon T. At each time step t, the agent applies a continuous bidding multiplier $b_t \in \mathcal{B}$ to impression-level value estimates, yielding an expected Gross Merchandise Volume (GMV) $G_t(b_t)$ at an expected cost $C_t(b_t)$ over that tick (Wu et al., 2018).

Standard auto-bidding maximizes cumulative GMV under a total budget B, theoretically allowing per-tick multiplier updates across the entire horizon T. However, policy recomputation consumes limited platform serving capacity, subject to constraints such as Queries Per Second (QPS) limits. Motivated by this resource limitation, we study an advertiser-level *update budget* of at most K decisions ($K \ll T$).

To incorporate this update constraint, we introduce a binary adaptive trigger $g_t \in \{0, 1\}$. The agent recomputes the bidding multiplier if $g_t = 1$, and retains the previous value otherwise. The initial decision counts toward the budget, so $g_1 = 1$. The auto-bidding problem is formulated as follows:

$$\begin{aligned} \max_{\pi_b, \boldsymbol{\pi}_g} \quad & \mathbb{E}_{\pi_b, \boldsymbol{\pi}_g}\left[\sum_{t=1}^{T} G_t(b_t)\right] \quad \text{s.t.} \quad \mathbb{E}_{\pi_b, \boldsymbol{\pi}_g}\left[\sum_{t=1}^{T} C_t(b_t)\right] \le B \\ & (1 - \boldsymbol{g}_t)(b_t - b_{t-1}) = 0 \ (t \ge 2), \quad \boldsymbol{g}_1 = 1, \quad \sum_{t=1}^{T} \boldsymbol{g}_t \le K \end{aligned} \tag{1}$$

Here, π_g determines whether to update and π_b selects the multiplier when an update occurs. The first line of Eq. (1) extends the classic auto-bidding formulation to account for the expected returns and costs under the update trigger $\boldsymbol{\pi}_g$, while the second line explicitly defines our proposed constraints to manage the limited update capacity.

3 THEORETICAL ANALYSIS

Overview. Our key idea is to view Eq. (1) as a temporal compression problem that minimizes the regret induced by removing decision boundaries. Specifically, allowing more updates enlarges the feasible policy space and can improve the optimal achievable bidding performance (Section 3.1). Theoretically, the best achievable performance is attained at the highest control frequency, with bid updates allowed at every time step. QPS constraints restrict this flexibility and can therefore incur performance loss. To quantify this loss, we introduce Merge Regret as the reduction in the globally optimized objective when adjacent intervals are required to share a bidding decision under the same spending budget (Section 3.2). This definition captures how a local scheduling decision can affect budget allocation across the entire horizon. Evaluating every candidate merge exactly, however, requires repeated global optimization. We therefore characterize the decision landscape using the GMV and cost response curves. Their marginal trade-off is captured by marginal efficiency, defined as the ratio of marginal GMV to marginal cost (Section 3.3). We further show that, under suitable regularity conditions, small discrepancies between adjacent intervals' marginal efficiency curves yield a small upper bound on the loss of sharing a bidding decision (Section 3.4). This result motivates using marginal efficiency discrepancies as a proxy for evaluating Merge Regret.

3.1 MONOTONICITY OF CONTROL FREQUENCY

Let $\mathcal{J}^*(K)$ denote the optimal cumulative objective achieved by a bidding strategy restricted to at most K adaptive triggers (i.e., the update budget) over the horizon T. Since increasing K enlarges the feasible policy space, the optimal achievable performance is non-decreasing with respect to the trigger allowance, as formalized in Lemma 1.

Lemma 1 (Monotonicity of Control Frequency)**.** *For any two decision budgets K_1 and K_2 such that $1 \le K_1 \le K_2 \le T$, the optimal achievable performance monotonically increases:*

$$\mathcal{J}^*(K_1) \le \mathcal{J}^*(K_2).$$

Proof sketch. Any bidding-triggering policy that is feasible under budget K_1 is also feasible under budget K_2, because K_2 allows at least as many updates. Therefore, the feasible set under K_1 is contained in the feasible set under K_2. Details are provided in Appendix B.1.

3.2 Temporal Segmentation and Merge Regret

Lemma 1 establishes that the highest optimal bidding performance is attainable when updates are allowed at every time step. However, strict QPS constraints physically limit the allowable number of bid changes to a maximum budget K. We therefore view frequency reduction as progressively merging adjacent intervals, starting from the finest partition $\mathcal{P}_{\text{dense}} = \{\{1\}, \ldots, \{T\}\}$. Each merge removes one decision boundary and reduces the number of intervals by one. Thus, $T-K$ successive merges produce a final partition with K intervals, satisfying the decision budget.

For any partition $\mathcal{P} = \{I_1, \ldots, I_M\}$, where $1 \le M \le T$, all time steps within interval I_m share a single bidding multiplier u_m. For any interval I, define the aggregate GMV and cost responses as $G_I(b) = \sum_{t \in I} G_t(b), C_I(b) = \sum_{t \in I} C_t(b)$. The optimal global value associated with a partition $\mathcal{P}$ is:

$$\mathcal{J}(\mathcal{P}) = \max_{\substack{u_1,\ldots,u_M \in \mathcal{B} \\ \sum_{m=1}^{M} C_{I_m}(u_m) \le B}} \sum_{m=1}^{M} G_{I_m}(u_m). \tag{2}$$

We assume the planning problems considered in Eq. (2) are feasible and attain their maxima. Restricting adjacent intervals to share a multiplier reduces the feasible set and can therefore incur a loss in optimized value. To mathematically quantify this loss, we introduce the concept of *Merge Regret*.

Definition 1 (Merge Regret)**.** *For adjacent intervals $I, J \in \mathcal{P}$, let $\mathcal{P}' = (\mathcal{P} \setminus \{I, J\}) \cup \{I \cup J\}$ denote the partition where I and J are constrained to share a bid. The Merge Regret is defined as the performance discrepancy:*

$$\Delta L(I, J \mid \mathcal{P}) = \mathcal{J}(\mathcal{P}) - \mathcal{J}(\mathcal{P}') \ge 0. \tag{3}$$

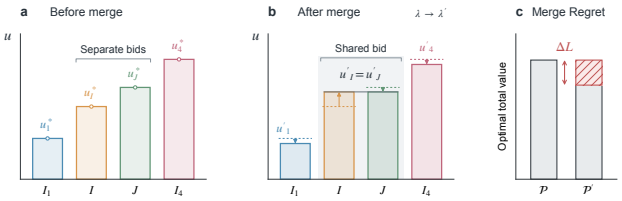

Figure 3: Global Merge Regret under a shared action.

Global objective. The total compression loss of a partition is

$$E(\mathcal{P}) = \mathcal{J}(\mathcal{P}_{\text{dense}}) - \mathcal{J}(\mathcal{P}). \tag{4}$$

Maximizing $\mathcal{J}(\mathcal{P})$ over partitions with $|\mathcal{P}| \le K$ is equivalent to minimizing $E(\mathcal{P})$.

3.3 Characterizing the Decision Landscape

Exact Merge Regret requires solving the budget constrained problems before and after each candidate merge. We therefore seek a response based surrogate for evaluating update opportunities. For any interval I, its marginal efficiency is defined as:

$$\rho_I(b) = \frac{G'_I(b)}{C'_I(b)}, \tag{5}$$

For a singleton interval $I = \{t\}$, this definition coincides with the marginal efficiency $\rho_t(b)$ at time step t.

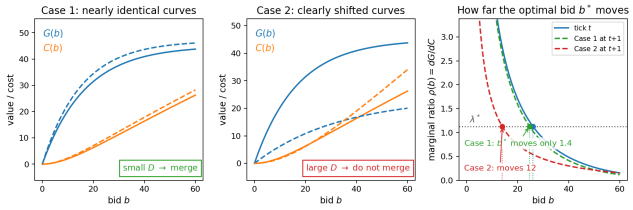

Figure 4: Examples of response landscape. When adjacent GMV and Cost curves exhibit high similarity (Case 1), merging their corresponding timesteps causes minimal perturbation to the optimal bid alignment. Conversely, distinct shifts in these curves (Case 2) result in severe bid displacement and profound utility loss.

At an interior optimum satisfying the first order conditions with a common budget multiplier λ^*, each interval satisfies $\rho_I(u_I^*) = \lambda^*$. Similar efficiency curves near an optimal multiplier can therefore indicate that adjacent intervals can share a decision with limited loss. Figure 4 illustrates this geometric relationship.

3.4 Theoretical Bounds for Merge Regret

We next bound the globally re-optimized Merge Regret using adjacent intervals' efficiency curves.

Theorem 1 (Global Pairwise Merge-Loss Bound)**.** *Fix a partition $\mathcal{P}$ and an optimal solution, and let $x = u_I^*$ and $y = u_J^*$ be the interior multipliers of two adjacent intervals. Suppose there is a common $\lambda^* \geq 0$ such that*

$$G_I'(x) = \lambda^* C_I'(x), \qquad G_J'(y) = \lambda^* C_J'(y). \tag{6}$$

On an interval $\mathcal{D} \subseteq \mathcal{B}$ containing x and y, assume that, for each $A \in \{I, J\}$, the responses are twice continuously differentiable and $C_A'(b) > 0$. Let $M_A \geq 0$ be a finite uniform bound on the absolute curvature of the Lagrangian-adjusted response $V_A(b; \lambda^\star) = G_A(b) - \lambda^\star C_A(b)$. Assume that, for all $b \in \mathcal{D}$,

$$|V_A''(b; \lambda^\star)| = |G_A''(b) - \lambda^\star C_A''(b)| \leq M_A, \qquad |\rho_J'(b)| \geq L > 0. \tag{7}$$

Then, with $\overline{M} = M_I + M_J$,

$$0 \leq \Delta L(I, J \mid \mathcal{P}) \leq \frac{\overline{M}}{2L^2} \left|\rho_I(x) - \rho_J(x)\right|^2. \tag{8}$$

Proof sketch. Because the cost responses are increasing, there exists a common multiplier z between x and y that preserves their combined spending: $C_I(z) + C_J(z) = C_I(x) + C_J(y)$. Replacing only these two multipliers by z gives a feasible candidate after merging. Global re-optimization can only improve upon this candidate. Writing $V_A(b; \lambda^*) = G_A(b) - \lambda^* C_A(b)$, spending preservation and the curvature bounds yield a candidate loss of at most $\overline{M}|x - y|^2/2$. The common first-order conditions and the mean value theorem give $|x - y| \leq |\rho_I(x) - \rho_J(x)|/L$. Combining these inequalities proves the result; details are provided in Appendix B.2.

The preceding theorem bounds merge loss through the true marginal-efficiency discrepancy at an optimal multiplier. Since this multiplier and the true curves are generally unknown in practice, we instead evaluate estimated efficiency curves at the b_{off} observed in the offline data. The following corollary connects this observable discrepancy to the merge-loss bound.

Corollary 1 (Behavior-Anchored Bound with Estimation Error)**.** *Under Theorem 1, let $b_{\text{off}} \in \mathcal{D}$ be an observed reference multiplier. Assume $\Delta\rho(b) = \rho_I(b) - \rho_J(b)$ is K_L-Lipschitz on $\mathcal{D}$ and $|x - b_{\text{off}}| \leq \epsilon$. Suppose the estimated efficiency curves satisfy $|\widehat{\rho}_A(b_{\text{off}}) - \rho_A(b_{\text{off}})| \leq \eta_A$ for $A \in \{I, J\}$. Then $\Delta L(I, J \mid \mathcal{P}) \leq \frac{\overline{M}}{2L^2}\left(|\widehat{\rho}_I(b_{\text{off}}) - \widehat{\rho}_J(b_{\text{off}})| + K_L\epsilon + \eta_I + \eta_J\right)^2$.*

The corollary follows from the triangle inequality and Lipschitz continuity (in Appendix B.3).

Greedy merging and accumulated regret. We construct a compressed schedule through greedy merging of adjacent intervals. Starting from $\mathcal{P}_{\text{dense}}$, we score each adjacent pair using the merge-loss bound in Corollary 1 and merge the pair with the smallest score. The aggregated response curves and candidate scores are updated after each merge. After $T - K$ merges, the resulting partition $\mathcal{P}_g$ contains K intervals. Let $\mathcal{P}^{(r)}$ denote the partition after r merges, with $\mathcal{P}^{(0)} = \mathcal{P}_{\text{dense}}$ and $\mathcal{P}^{(T-K)} = \mathcal{P}_g$. The actual global Merge Regret at step r is

$$\Delta L_r = \mathcal{J}(\mathcal{P}^{(r-1)}) - \mathcal{J}(\mathcal{P}^{(r)}),$$

where both objective values are globally re-optimized under the same spending budget. Let U_r denote the score used to select this merge. Under the Corollary 1, $\Delta L_r \leq U_r$. Consequently, the actual total performance loss relative to the dense schedule satisfies

$$E(\mathcal{P}_g) = \mathcal{J}(\mathcal{P}_{\text{dense}}) - \mathcal{J}(\mathcal{P}_g) = \sum_{r=1}^{T-K} \Delta L_r \leq \sum_{r=1}^{T-K} U_r. \tag{9}$$

The intermediate objective values cancel when the per-merge losses are summed. Thus, the accumulated scores bound the actual global performance loss of the resulting schedule.

4 ALGORITHM DESIGN

4.1 REPRESENTATION OF MARKET DYNAMICS

We represent market dynamics through the GMV and cost response functions, $G_t(b)$ and $C_t(b)$. In our implementation, we adopt the following exponential parameterization $G_t(b) = \alpha_t(1 - e^{-\beta_t b}), C_t(b) = \gamma_t b(1 - e^{-\beta_t b})$, where $\theta_t = (\alpha_t, \beta_t, \gamma_t) > 0$ dictates the market regime at time t.

This specific parameterization guarantees physical and economic validity through its explicit analytical derivatives. For the expected GMV curve, its first and second derivatives with respect to the bid b are given by: $G'_t(b) = \alpha_t\beta_t e^{-\beta_t b} > 0, G''_t(b) = -\alpha_t\beta_t^2 e^{-\beta_t b} < 0$, These properties mathematically ensure that the GMV monotonically increases with higher bids, while exhibiting diminishing marginal returns and eventually saturating at the market capacity limit α_t.

Similarly, as the bid increases, the expenditure initially grows at a moderate pace and subsequently transitions toward a linear asymptotic growth regime, scaling as $C_t(b) \sim \gamma_t b$ when $b \to \infty$. Once a bid reaches a sufficiently high threshold to secure nearly all available traffic impressions, the winning probability converges to one. Consequently, any further elevation in the bid fails to acquire additional volume and strictly inflates the clearing price, leading to the aggressive budget depletion observed in highly competitive environments.

By leveraging this infinitely differentiable formulation, the marginal efficiency $\rho_t(b) = \frac{G'_t(b)}{C'_t(b)}$ is analytically derived as:

$$\rho_t(b) = \frac{G'_t(b)}{C'_t(b)} = \frac{\alpha_t\beta_t}{\gamma_t(e^{\beta_t b} - 1 + \beta_t b)}, b > 0, \tag{10}$$

Because $\rho'_t(b) < 0$, the function strictly guarantees a monotonically decreasing marginal efficiency space across all positive bids. Defining the market landscape through these smooth analytic curves establishes a mathematically rigorous foundation for calculating the continuous structural divergence required by our optimal scheduling bounds. We verify that this exponential formulation satisfies all necessary structural assumptions for our theoretical bounds in Appendix C.

With these functional forms specified, the GMV and cost response curves at each tick are determined by the time-varying parameters $\theta_t = (\alpha_t, \beta_t, \gamma_t)$. We therefore learn these parameters from traffic observations to construct a curve-based representation of market dynamics.

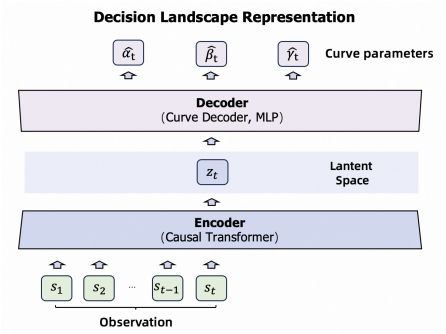

Figure 5: Architecture of the Decision Landscape representation.

Specifically, at time t, a causal Transformer encoder processes the observation history $s_{1:t}$ and extracts a latent state z_t (Chen et al., 2021b).
To ensure z_t captures the volume price properties, we introduce a Curve Decoder (parameterized by a Multi Layer Perceptron). The decoder projects the latent state z_t into the parameter space:

$$\hat{\theta}_t = (\hat{\alpha}_t, \hat{\beta}_t, \hat{\gamma}_t) = \text{Decoder}(z_t). \tag{11}$$

The representation network is optimized via a reconstruction loss that forces the estimated parameters to align with the a posteriori ground truth curves observed from offline logs.

The loss function is defined as:

$$\mathcal{L}_{\text{rec}} = w_G \cdot \text{MSE}(\hat{G}_t(b), G_t(b)) + w_C \cdot \text{MSE}(\hat{C}_t(b), C_t(b)), \tag{12}$$

where w_G and w_C are balancing weights. Besides, our framework can be generalized to other parametric families, such as the Hill Formulation (Appendix E).

4.2 ONLINE ADAPTIVE TRIGGERING

The execution pipeline (Algorithm 1), integrates the representations with the algorithmic scheduling. For every incoming micro step, the online agent calculates the localized geometric distance d_t

between the current curver dynamics $\hat{\theta}_t$ and the historical anchor $\hat{\theta}_\tau$. If this functional divergence exceeds the optimal threshold τ^* extracted from the offline greedy partition, the system activates the trigger, issuing a newly optimized bid and resetting the temporal anchor. Otherwise, the agent suppresses the action. This execution mechanism avoids unnecessary computational overhead and system volatility.

5 Experiment Results

5.1 Experimental Setup

Dataset and Environment. We conduct our evaluation using the **AuctionNet** simulation environment, a large-scale, industry-standard benchmark (Su et al., 2024). The environment incorporates 48 diverse auto-bidding advertisers. We evaluate the target advertiser (Index 0, Category 0) initialized with a daily budget of 2,900 and a Cost-Per-Action (CPA) constraint of 100. At any active trigger point, the agent dictates a bid multiplier α, such that the real-time bid is calculated as $bid = \alpha \cdot \text{pValue}$. To account for traffic stochasticity, all evaluations are averaged over 32 distinct episodes initialized with different page-view (PV) generator random seeds.

Baselines. We benchmark our proposed method against several baselines: (i) **Random**, which selects the K trigger points randomly across the entire horizon; (ii) **Uniform (TTC)**, a standard strategy (Wu et al., 2018) that executes updates at equal time intervals; (iii) **WaitWatch (ETC)**, a rule-based Event-Triggered Control strategy that determines when to adjust based on accumulated conversions; (iv) **Interval-RL (DAR)** (Biedenkapp & Lindauer, 2021), an action-repetition approach based on Dynamic Action Repetition that samples a discrete hold-length (i.e., a time threshold) to maintain the current action before the next update; and (v) **Safe Action Repetition (SAR)** (Seohong Park & Kim, 2021), a condition-driven baseline that outputs a safe region radius d at the trigger point state; (vi) **JointExploration (LICRA)** (Mguni et al., 2023), which jointly explores actions and triggers.

To establish performance ceilings, we report two upper bounds: **Oracle (Practical UB)**, the empirical practical upper bound achieved when the agent triggers bid updates continuously without any strict frequency or QPS constraints; and **Oracle (Theoretical UB)**, the global optimum that acts as an offline oracle, sorting all traffic by cost-effectiveness in descending order and greedily acquiring the highest-value impressions until the budget is exhausted (He et al., 2021).

Evaluation Metrics. Our primary performance indicator is the **Score** defined in the AuctionNet benchmark. This metric reflects the effective volume of acquisitions while severely penalizing any violations of the CPA (Cost-Per-Action) constraint, formulated as:

$$\text{Score} = \text{Conversions} \times \min\left(1, \frac{\text{CPA}_{\text{target}}}{\text{CPA}_{\text{realized}}}\right)^2 \tag{13}$$

In addition to the overall Score, we report the **Conversions** and the total **Cost**.

5.2 Overall Evaluation

Table 1 summarizes the performance across varying temporal control granularities ($T \in \{48, 96, 144, 288\}$) under a strict decision budget of $K = 10$. The results clearly demonstrate the superiority of the AIGB-ReAL framework in consistently balancing objective maximization with severe decision-frequency constraints.

The Importance of the Decision Landscape. Traditional rule-based methods, such as fixed-frequency pacing (Uniform) and conversion-based heuristics (WaitWatch), consistently underperform across all granularities. For instance, at the finest granularity ($T = 288$), Uniform and WaitWatch achieve scores of only 27.17 and 25.28, respectively. Heuristic methods naively assume that resource consumption correlates higher with market volatility. However, raw resource depletion fails to capture structural shifts in the auction's action-value landscape. *By dynamically tracking the geometric divergence of the volume-price curves (ρ), our regret-aware scheduling explicitly minimizes Merge Regret, allocating trigger budgets only when the underlying market geometry fundamentally shifts.* This underscores the critical need to identify a decision landscape that is intrinsically linked to the fundamental decision-making process, rather than relying on simplistic observational proxies.

Table 1: Main comparison across control granularities (T). Score, converted volume (Conv), and Cost are reported as mean ± s.e. over 32 episodes. Best score among practical methods is in **bold**.

$T = 48$				$T = 96$			
Method	Score	Conv	Cost	Method	Score	Conv	Cost
Random	22.97 ± 2.46	23.68 ± 2.37	1889.05 ± 177.03	Random	20.87 ± 1.68	25.34 ± 1.14	2750.82 ± 45.13
Uniform (TTC)	22.27 ± 2.61	23.00 ± 2.52	1810.51 ± 184.36	Uniform (TTC)	21.77 ± 1.65	25.86 ± 1.03	2789.50 ± 33.53
WaitWatch (ETC)	18.91 ± 2.42	19.53 ± 2.36	1427.71 ± 162.77	WaitWatch (ETC)	22.32 ± 1.65	24.22 ± 1.41	2189.94 ± 107.33
Interval-RL (DAR)	21.98 ± 2.58	22.78 ± 2.48	1831.08 ± 184.93	Interval-RL (DAR)	18.94 ± 1.79	24.44 ± 1.13	2797.30 ± 49.46
Safe Action Rep (SAR)	24.63 ± 2.04	25.25 ± 1.98	2106.66 ± 168.05	Safe Action Rep (SAR)	23.86 ± 1.72	26.91 ± 1.17	2716.35 ± 73.58
JointExploration (LICRA)	9.50 ± 1.31	18.78 ± 0.89	2899.95 ± 0.01	JointExploration (LICRA)	7.57 ± 0.91	17.53 ± 0.78	2899.95 ± 0.00
Ours (AIGB-ReAL)	$\mathbf{26.02 \pm 2.16}$	27.00 ± 2.06	2165.95 ± 154.21	Ours (AIGB-ReAL)	$\mathbf{22.88 \pm 1.75}$	26.38 ± 1.23	2687.28 ± 77.90
Oracle (Practical UB)	27.74 ± 1.89	29.34 ± 1.59	2568.62 ± 87.22	Oracle (Practical UB)	26.75 ± 1.74	28.34 ± 1.48	2608.86 ± 107.54
Oracle (Theoretical UB)	28.32 ± 1.62	30.79 ± 1.04	2898.72 ± 0.13	Oracle (Theoretical UB)	29.45 ± 1.73	31.71 ± 1.19	2898.92 ± 0.12
$T = 144$				$T = 288$			
Method	Score	Conv	Cost	Method	Score	Conv	Cost
Random	23.29 ± 1.92	27.24 ± 1.30	2790.61 ± 37.56	Random	26.09 ± 1.59	28.92 ± 1.17	2759.12 ± 45.44
Uniform (TTC)	24.44 ± 1.90	27.93 ± 1.26	2814.59 ± 36.75	Uniform (TTC)	27.17 ± 1.74	29.72 ± 1.29	2779.92 ± 46.17
WaitWatch (ETC)	25.50 ± 2.04	27.00 ± 1.73	2327.31 ± 109.45	WaitWatch (ETC)	25.28 ± 1.84	26.31 ± 1.69	2066.56 ± 125.83
Interval-RL (DAR)	23.45 ± 2.01	27.16 ± 1.30	2808.04 ± 44.24	Interval-RL (DAR)	25.08 ± 2.11	28.22 ± 1.60	2607.86 ± 84.70
Safe Action Rep (SAR)	24.54 ± 2.13	28.22 ± 1.48	2791.36 ± 70.88	Safe Action Rep (SAR)	25.52 ± 1.95	28.66 ± 1.38	2790.54 ± 56.94
JointExploration (LICRA)	8.09 ± 1.11	17.91 ± 0.78	2899.96 ± 0.00	JointExploration (LICRA)	7.34 ± 1.15	17.25 ± 0.86	2899.96 ± 0.01
Ours (AIGB-ReAL)	$\mathbf{25.62 \pm 2.11}$	28.75 ± 1.49	2727.75 ± 54.82	Ours (AIGB-ReAL)	$\mathbf{28.69 \pm 1.90}$	30.16 ± 1.58	2632.22 ± 85.87
Oracle (Practical UB)	27.14 ± 2.15	28.94 ± 1.77	2539.76 ± 114.36	Oracle (Practical UB)	28.95 ± 1.90	29.84 ± 1.76	2491.44 ± 121.77
Oracle (Theoretical UB)	29.82 ± 1.75	32.07 ± 1.22	2899.10 ± 0.10	Oracle (Theoretical UB)	30.27 ± 1.73	32.37 ± 1.24	2899.04 ± 0.11

Robustness Against Learned RL Baselines. Unlike standard reinforcement learning baselines, AIGB-ReAL maintains stability across varying horizon lengths. Jointly trained RL approaches (**LICRA**) suffer from severe action-space non-stationarity, collapsing to scores below 10.0 across all settings. Specifically, *as the decision granularity becomes finer, this instability is further amplified*, leading to a noticeable performance degradation from 9.50 down to 7.34. Furthermore, **Interval-RL** fluctuates significantly, dropping as low as 18.94 at T=96. These failures highlight the *fundamental limitations of standard step-wise temporal difference (TD) learning paradigms in highly volatile, multi-agent auction environments.* AIGB-ReAL leverages an inherently decoupled learning paradigm. By structurally separating the bidding and updating decision processes, the architecture effectively maintains robust performance across all evaluation horizons.

Gap to the Full-Adjustment Upper Bound. Finally, we evaluate the performance gap between our sparse-triggering framework and the unconstrained upper bounds. Remarkably, AIGB-ReAL achieves a score of 28.69 at $T = 288$, exhibiting only a marginal gap of 0.26 compared to this full-adjustment benchmark. It also maintains a reasonably small gap to the offline theoretical Oracle (30.27). These results demonstrate AIGB-ReAL's capability to effectively identify critical decision moments within the dynamic auction. By executing a limited number of updates during shifts in market dynamics, the agent avoids the volatility introduced by continuous adjustments. This indicates that *capturing appropriate timing for sparse interventions is highly effective, allowing the model to approach the performance ceiling while maintaining high decision efficiency.*

5.3 Ablation Study

To validate the components of our trigger signal representation, we conduct an ablation study across different control granularities (T). As shown in Table 2, we compare the full **AIGB-ReAL** model against two degraded variants: **w/o latent** z (which bypasses representation learning to use raw states directly) and **w/o curve** ρ (which uses the learned latent representation but omits the curve-property signal).

The results demonstrate that both components are essential. The full AIGB-ReAL consistently outperforms the **w/o curve** ρ variant across all granularities, notably increasing the score from 26.58 to 28.69 at $T = 288$. This confirms that explicitly modeling curve properties provides critical, forward-looking geometric guidance for optimal stopping. Furthermore, *the latent representation becomes increasingly vital at finer granularities*. While relying on the raw state (**w/o latent** z) remains competitive at $T = 48$, its performance degrades significantly at longer horizons ($T \geq 96$). Ultimately, *the synergy between the noise-filtering latent state z and the curve-based decision landscape characterized by ρ* enables AIGB-ReAL to achieve the highest performance.

Table 2: Ablation on the trigger signal representation. **AIGB-ReAL** uses the full curve-property signal (ρ). **w/o curve** ρ relies only on the learned latent (z), and **w/o latent** z falls back to the raw state. Score, converted volume (Conv), and Cost are reported as mean $\pm$ s.e. over 32 episodes. Best score among practical methods is in **bold**.

$T = 48$				$T = 96$			
Method	Score	Conv	Cost	Method	Score	Conv	Cost
w/o curve ρ	23.75 ± 2.07	24.84 ± 1.98	2095.28 ± 168.36	w/o curve ρ	20.35 ± 1.73	24.63 ± 1.07	2763.46 ± 60.73
w/o latent z	24.64 ± 2.06	25.41 ± 1.96	2099.24 ± 154.83	w/o latent z	19.82 ± 1.66	24.03 ± 1.13	2616.15 ± 91.92
Ours (AIGB-ReAL)	$\mathbf{26.02 \pm 2.16}$	27.00 ± 2.06	2165.95 ± 154.21	Ours (AIGB-ReAL)	$\mathbf{22.88 \pm 1.75}$	26.38 ± 1.23	2687.28 ± 77.90
$T = 144$				$T = 288$			
Method	Score	Conv	Cost	Method	Score	Conv	Cost
w/o curve ρ	22.63 ± 2.24	26.69 ± 1.55	2760.21 ± 56.14	w/o curve ρ	26.58 ± 1.98	29.38 ± 1.46	2731.57 ± 58.19
w/o latent z	22.21 ± 2.01	26.38 ± 1.36	2822.33 ± 50.28	w/o latent z	23.11 ± 2.17	27.16 ± 1.41	2695.33 ± 84.22
Ours (AIGB-ReAL)	$\mathbf{25.62 \pm 2.11}$	28.75 ± 1.49	2727.75 ± 54.82	Ours (AIGB-ReAL)	$\mathbf{28.69 \pm 1.90}$	30.16 ± 1.58	2632.22 ± 85.87

6 Online A/B Test

To evaluate the real-world effectiveness of our proposed framework, we conducted a large-scale online A/B test involving approximately 600,000 advertisers on a leading e-commerce platform. The experiment ran for ten days, from September 3 to September 12, 2026, and compared the low-frequency baseline (15 min) with our high-frequency (1 min) AIGB-ReAL framework. As shown in Table 3, AIGB-ReAL drives business growth while satisfying advertiser constraints.

Table 3: Online A/B test results comparing the low-frequency baseline (15 min) with our high-frequency AIGB-ReAL framework.

Method	Cost	GMV	Orders (7d)	TargetValue	ROI Attain.
Low-freq (15min)	139,028,893.27	683,427,568.11	11,572,766	121,955,280.97	71.33%
High-freq (AIGB-ReAL)	143,048,317.54	699,206,051.90	11,955,862	125,552,867.28	72.56%
Improvement (Δ)	$+\mathbf{2.89}\%$	$+\mathbf{2.31}\%$	$+\mathbf{3.31}\%$	$+\mathbf{2.95}\%$	$+\mathbf{1.23}$ pt

Note: Orders (7d) denotes the number of transactions attributed within a seven-day window. TargetValue denotes the aggregate advertising spend corresponding to the realized GMV under advertisers' specified target ROI levels and is computed as $\sum_p \mathrm{GMV}_p / \mathrm{TargetROI}_p$ over all plans p.

7 Conclusion

We propose the AIGB-ReAL, a scheduling architecture designed to break the fundamental trade-off between the performance gains of high-frequency updates and strict infrastructure query (QPS) limits in real-time auto-bidding. While the majority of current systems rely on rigid fixed-frequency updates, adaptive paradigms proposed to address these challenges still exhibit distinct inherent limitations. Specifically, Event-Triggered Control is restricted to a short horizon due to its reliance on local state deviations, Action Repetition remains fundamentally inflexible, and Joint Exploration frequently suffers from severe instability and space explosion.

To address these limitations, we adopt a global perspective across the entire decision horizon. We mathematically ground this temporal allocation by introducing Merge Regret to explicitly quantify the global loss in optimal cumulative value caused by merging adjacent action intervals. To translate this global metric into a reactive decision boundary, we construct a Decision Landscape that evaluates structural changes in achievable outcomes. By implementing this regret-aware mechanism, the architecture effectively decouples the temporal triggering module from the downstream action optimization process. This provides theoretical and empirical advantages over existing methods. First, by evaluating global Merge Regret, the framework overcomes the short-sightedness of Event-Triggered Control heuristics. Second, by continuously evaluating the triggering condition at every step, it breaks the rigidity of Action Repetition. Furthermore, this decoupled architecture bypasses the action-space explosion and instability in Joint Exploration. While AIGB-ReAL achieves superior performance in complex auto-bidding environments, this paradigm provides a highly transferable foundation for broader resource-constraints decision domains.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, improving the clarity and presentation of the manuscript, and providing suggestions during code development and debugging. The core research problem, methodology, theoretical analysis, experimental design, and interpretation of the results were developed and validated by the authors. All AI-assisted text, code, and research suggestions were carefully reviewed and revised by the authors.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our work from both methodological and experimental perspectives. The complete formulation and algorithmic procedure of the proposed method are described in the main paper, including the model architecture, training objectives, and inference procedure. For theoretical results, all assumptions are explicitly stated and complete derivations and proofs are included in the Appendix.

REFERENCES

Thomas Avé, Matthias Hutsebaut-Buysse, and Kevin Mets. Temporal distillation: Compressing a policy in space and time. *Machine Learning*, 114(11), 2025.

Raghu Rajan Frank Hutter Biedenkapp, Andre and Marius Lindauer. Temporl: Learning when to act. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 914–924. PMLR, 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.

Chen Chen, Hongyao Tang, Jianye Hao, Wulong Liu, and Zhaopeng Meng. Addressing action oscillations through learning policy inertia. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7020–7027, 2021a.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097, 2021b.

Batuhan Demirel, Vijay Gupta, Daniel E. Quevedo, and Karl Henrik Johansson. On the trade-off between communication and control cost in event-triggered dead-beat control. *IEEE Transactions on Automatic Control*, 62(6):2973–2980, 2017.

Dimos V. Dimarogonas, Emilio Frazzoli, and Karl H. Johansson. Distributed event-triggered control for multi-agent systems. *IEEE Transactions on Automatic Control*, 57(5):1291–1297, 2012.

Victor S. Dolk, Jeroen Ploeg, and W. P. M. H. Heemels. Event-triggered control for string-stable vehicle platooning. *IEEE Transactions on Intelligent Transportation Systems*, 18(12):3486–3500, 2017.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Jean Harb, Pierre-Luc Bacon, Martin Klissarov, and Doina Precup. When waiting is not an option: Learning options with a deliberation cost. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3165–3172, 2018.

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 and Data Mining*, pp. 2993–3001, 2021.

W. P. M. H. Heemels, Karl Henrik Johansson, and Paulo Tabuada. An introduction to event-triggered and self-triggered control. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 3270–3285, 2012.

Michael Hertneck, David Meister, and Frank Allgöwer. Current trends and future directions in event-based control. *European Journal of Control*, 87:101387, 2026.

Niklas Karlsson. Feedback control in programmatic advertising: The frontier of optimization in real-time bidding. *IEEE Control Systems Magazine*, 40(5):40–77, 2020.

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*, WWW '24, pp. 4193–4203, 2024.

Yiheng Lin, James A. Preiss, Emile Anand, Yingying Li, Yisong Yue, and Adam Wierman. Online adaptive policy selection in time-varying systems: No-regret via contractive perturbations. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53508–53521, 2023.

Thomas Mesnard, Theophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S. Stepleton, Nicolas Heess, Arthur Guez, Eric Moulines, Marcus Hutter, Lars Buesing, and Remi Munos. Counterfactual credit assignment in model-free reinforcement learning. In *International Conference on Machine Learning*, pp. 7654–7664, 2021.

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. Control frequency adaptation via action persistence in batch reinforcement learning. volume 119, pp. 6862–6873. PMLR, 2020.

David Henry Mguni, Aivar Sootla, Juliusz Krzysztof Ziomek, Oliver Slumbers, Zipeng Dai, Kun Shao, and Jun Wang. Timing is everything: Learning to act selectively with costly actions and budgetary constraints. In *International Conference on Learning Representations*, 2023.

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

Michael Neunert, Abbas Abdolmaleki, Markus Wulfmeier, Thomas Lampe, Tobias Springenberg, Roland Hafner, Francesco Romano, Jonas Buchli, Nicolas Heess, and Martin Riedmiller. Continuous-discrete reinforcement learning for hybrid control in robotics. In *Proceedings of the Conference on Robot Learning*, volume 100, pp. 735–751. PMLR, 2020.

Brian Paden, Michal Cáp, Sze Zheng Yong, Dmitry Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, 1(1):33–55, 2016.

Zhicheng Pan, Yihang Wang, Yingying Zhang, Sean Bin Yang, Yunyao Cheng, Peng Chen, Chenjuan Guo, Qingsong Wen, Xiduo Tian, Yunliang Dou, Zhiqiang Zhou, Chengcheng Yang, Aoying Zhou, and Bin Yang. Magicscaler: Uncertainty-aware, predictive autoscaling. *Proceedings of the VLDB Endowment*, 16(12):3808–3821, 2023.

Romain Postoyan, Marcos Cesar Bragagnolo, Ernest Galbrun, Jamal Daafouz, Dragan Nesic, and Eugênio B. Castelan. Event-triggered tracking control of unicycle mobile robots. *Automatica*, 52: 302–308, 2015.

Jaekyeom Kim Seohong Park and Gunhee Kim. Time discretization-invariant safe action repetition for policy gradient methods. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Sahil Sharma, Aravind Lakshminarayanan, and Balaraman Ravindran. Learning to repeat: Fine grained action repetition for deep reinforcement learning. In *International Conference on Learning Representations*, 2017.

Friedrich Solowjow and Sebastian Trimpe. Event-triggered learning. *Automatica*, 117:109009, 2020.

A. Srinivas, Sahil Sharma, and Balaraman Ravindran. Dynamic action repetition for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2133–2139, 2017.

Aravind Srinivas, Sahil Sharma, and Balaraman Ravindran. Dynamic frame skip deep q network. In *IJCAI 2016 Workshop on Deep Reinforcement Learning: Frontiers and Challenges*, 2016.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 94428–94452, 2024.

Paulo Tabuada. Event-triggered real-time scheduling of stabilizing control tasks. *IEEE Transactions on Automatic Control*, 52(9):1680–1685, 2007.

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 (CIKM)*, pp. 1443–1451, 2018.

Qingyun Wu, Huazheng Wang, Yanen Li, and Hongning Wang. Dynamic ensemble of contextual bandits to satisfy users' changing interests. In *The World Wide Web Conference*, WWW '19, pp. 2080–2090, 2019.

Wei Xu Yu, Haonan and Haichao Zhang. Taac: Temporally abstract actor-critic for continuous control. In *Advances in Neural Information Processing Systems*, pp. 29021–29033, 2021.

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.

Han Zhu, Junqi Jin, Chang Tan, Fei Pan, Yifan Zeng, Han Li, and Kun Gai. Optimized cost per click in Taobao display advertising. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

CONTENTS

1 Introduction **1**

2 Preliminaries **3**

3 Theoretical Analysis **3**

3.1 Monotonicity of Control Frequency 3

3.2 Temporal Segmentation and Merge Regret 4

3.3 Characterizing the Decision Landscape 4

3.4 Theoretical Bounds for Merge Regret 5

4 Algorithm Design **6**

4.1 Representation of Market Dynamics 6

4.2 Online Adaptive Triggering 6

5 Experiment Results **7**

5.1 Experimental Setup 7

5.2 Overall Evaluation 7

5.3 Ablation Study 8

6 Online A/B Test **9**

7 Conclusion **9**

A Related Work **15**

A.1 Budget-Constrained Auto-Bidding under Low Fixed-Frequency 15

A.2 Event-Triggered Control and Local Short-Sightedness 15

A.3 Action Repetition and Structural Inflexibility 15

A.4 Joint Exploration and Instability 15

B Proofs of Theoretical Results **16**

B.1 Proof of Lemma 1 16

B.2 Proof of Theorem 1 16

B.3 Proof of Corollary 1 17

C Verification of Structural Assumptions for the Exponential Bidding Model **17**

D Empirical Validation of the Parametric Landscape **18**

E Generalization and Structural Verification: The Generalized Hill Formulation **19**

F Instability of Joint RL Baselines **20**

G Algorithm of AIGB-ReAL **20**

H Online Experiment: Bidding Policy via BidDreamer **21**

H.1 World Model Representation Learning 21

H.2 Policy Optimization via Latent Imagination 22

A Related Work

A.1 Budget-Constrained Auto-Bidding under Low Fixed-Frequency

Budget-constrained bidding is fundamentally a sequential optimization problem where current bids directly restrict future spending capabilities (Cai et al., 2017; Wu et al., 2018). To navigate dynamic market conditions, reinforcement learning is widely employed to continuously adapt overarching bidding parameters (Zhang et al., 2016; Karlsson, 2020). In practical industrial systems, this parameter optimization is typically decoupled from the per-auction bid computation to meet strict latency requirements, relying heavily on rigid, fixed-frequency updates (e.g., 15-minute (Wu et al., 2018) or 30-minute (Li et al., 2024) intervals). While recent advancements have focused extensively on improving the action policy within these fixed intervals (Mou et al., 2022; Li et al., 2024), the temporal inefficiency of the update schedule itself remains largely unexplored. Theoretically, high-frequency updates yield superior performance by instantly capturing market volatility. However, real-world bidding infrastructures are strictly bounded by finite query per second (QPS) constraints. Our focus explicitly addresses this trade-off. Rather than settling for suboptimal fixed intervals, we adaptively allocate a strictly limited QPS update budget across the decision horizon.

A.2 Event-Triggered Control and Local Short-Sightedness

Event-triggered control (ETC) dynamically schedules updates by reacting to specific monitored conditions, classically utilizing local state errors or Lyapunov conditions to reduce sampling frequency (Tabuada, 2007; Hertneck et al., 2026). While this mechanism effectively filters out continuous updates, applying standard ETC to budget-constrained auto-bidding exposes a fundamental limitation: it is inherently myopic. Standard ETC heuristics treat immediate, raw state deviations as a direct proxy for economic benefit. However, in sequential auction environments, a significant local state variation does not necessarily translate to a loss in overall campaign value. Our approach radically departs from this short-sighted paradigm by adopting a macroscopic, global perspective. Instead of reacting to local state deviations, we introduce Merge Regret as our primary scheduling criterion. Merge Regret explicitly evaluates the horizon-wide reduction in total campaign value when adjacent intervals share a bidding multiplier, fully accounting for the horizon-wide re-optimization of the remaining budget. This ensures that a temporal update is triggered strictly by its structural impact on global economic outcomes, rather than local, short-term fluctuations.

A.3 Action Repetition and Structural Inflexibility

Action repetition paradigms attempt to streamline temporal decisions by forcing the agent to commit to a prolonged action duration, with approaches ranging from dynamic duration learning (Srinivas et al., 2016; 2017; Biedenkapp & Lindauer, 2021; Yu & Zhang, 2021) to policy factorization and persistence evaluation (Sharma et al., 2017; Metelli et al., 2020). However, this advance duration commitment is fundamentally inflexible, severely delaying the system's ability to react to sudden market disturbances or traffic shifts. While more recent variants such as SAR (Seohong Park & Kim, 2021), and related adaptations (Neunert et al., 2020; Chen et al., 2021a) introduce observation-dependent termination to mitigate this rigidity, they fail to resolve the underlying computational inefficiency. Our approach fundamentally resolves this inefficiency. Rather than rigidly committing to future repetitions or constantly computing redundant actions, we leverage global response-curve estimates to inform when bidding multipliers should be recomputed. This continuous, lightweight evaluation ensures the agent remains highly flexible against market volatility while bypassing redundant downstream optimization overhead.

A.4 Joint Exploration and Instability

Joint exploration paradigms attempt to learn temporal triggers concurrently with action policies. These methods typically frame timing as a selective execution or deliberation process, utilizing approaches such as explicit switching costs (Harb et al., 2018), policy combination (Mguni et al., 2023), or temporal distillation (Avé et al., 2025). While these frameworks can accommodate finite execution budgets, they inherently entangle the timing decision with the downstream bid optimization. This coupled architecture triggers a multiplicative expansion of the state-action space,

Table 4: Comparison of action-triggering paradigms. Our AIGB-ReAL framework specifically addresses the four key limitations of existing methods illustrated in Figure 1: *Short Horizon*, *Inflexible* action repetition, *Unstable* credit assignment, and *Space Explosion*.

Method	**Global Horizon** (Long-term value)	**Flexible** (Online re-check)	**Credits Assignment** (Stable updates)	**Decoupled Search** (Avoids $\|G\|\times\|A\|$)
Periodic bidding	×	×	–	✓
Classical ETC	×	✓	–	✓
Interval-RL	✓	×	×	×
SAR	✓	✓	×	×
LICRA	✓	✓	×	×
Ours (AIGB-ReAL)	✓	✓	✓	✓

✓: present; ×: absent; –: not explicitly formulated.
Global Horizon: Overcomes the "Short Horizon" defect of basic ETC by optimizing long-term cumulative return.
Flexible: Avoids the "Inflexible" action repetition by allowing dynamic online evaluations.
Credits Assignment: Solves the "Unstable" reward allocation problem typical in joint RL exploration.
Decoupled Search: Prevents the "Space Explosion" of joint exploration by decoupling decisions into $|G| + |A|$.

inevitably leading to severe instability and ambiguous credit assignment during training. Specifically, the agent struggles to isolate whether a suboptimal outcome stems from poor update timing or an ineffective bidding multiplier. AIGB-ReAL fundamentally bypasses this joint exploration bottleneck. Instead of forcing the agent to simultaneously learn when and how to bid, we structurally decouple the temporal trigger. Guided by the globally re-optimized value difference of Merge Regret and a response-curve surrogate, our architecture evaluates the structural necessity of an update independently.

B Proofs of Theoretical Results

B.1 Proof of Lemma 1

Proof. Let Π_K denote the set of consecutive partitions of $\{1, \dots, T\}$ with at most K intervals. For $K_1 \le K_2$, we have $\Pi_{K_1} \subseteq \Pi_{K_2}$. All partitions are evaluated under the same response curves, multiplier domain, and spending budget. Consequently,

$$\mathcal{J}^*(K_1) = \max_{\mathcal{P}\in\Pi_{K_1}} \mathcal{J}(\mathcal{P}) \le \max_{\mathcal{P}\in\Pi_{K_2}} \mathcal{J}(\mathcal{P}) = \mathcal{J}^*(K_2). \tag{14}$$

□

B.2 Proof of Theorem 1

Proof. Fix an optimal multiplier vector for $\mathcal{P}$. Let $x = u_I^*$ and $y = u_J^*$, and write $\mathcal{P}'$ for the partition after merging I and J. The feasible set for $\mathcal{P}'$ is contained in that for $\mathcal{P}$, so $\Delta L(I, J \mid \mathcal{P}) \ge 0$.

A spending-preserving feasible candidate. Set $a = \min\{x, y\}$, $c = \max\{x, y\}$, and $S = C_I(x) + C_J(y)$. Since both cost functions are continuous and strictly increasing on $\mathcal{D}$,

$$C_I(a) + C_J(a) \le S \le C_I(c) + C_J(c). \tag{15}$$

The intermediate value theorem therefore gives a unique $z \in [a, c]$ satisfying

$$C_I(z) + C_J(z) = C_I(x) + C_J(y). \tag{16}$$

When $x = y$, take $z = x$. Replace the multipliers of I and J by z, leaving all other multipliers at their values in the selected optimum. This is feasible for $\mathcal{P}'$ and preserves total spending, whether or not the budget is active.

Let $\widetilde{F}$ denote this candidate's total advertising value. Since $\mathcal{J}(\mathcal{P}')$ re-optimizes all multipliers, $\mathcal{J}(\mathcal{P}') \ge \widetilde{F}$. Hence,

$$\begin{aligned} \Delta L(I, J \mid \mathcal{P}) &\le \mathcal{J}(\mathcal{P}) - \widetilde{F} \\ &= G_I(x) + G_J(y) - G_I(z) - G_J(z). \end{aligned} \tag{17}$$

This construction does not constrain the actual re-optimized solution to keep any multiplier or shadow price fixed.

Bounding the candidate loss. Define $V_A(b) = G_A(b) - \lambda^* C_A(b)$ for $A \in \{I, J\}$. By spending preservation, the right-hand side of Eq. (17) equals

$$V_I(x) - V_I(z) + V_J(y) - V_J(z). \tag{18}$$

The assumed first-order conditions give $V_I'(x) = V_J'(y) = 0$. Taylor's theorem and $|V_A''| \le M_A$ imply

$$\begin{aligned} V_I(x) - V_I(z) &\le \frac{M_I}{2}(x - z)^2, \\ V_J(y) - V_J(z) &\le \frac{M_J}{2}(y - z)^2. \end{aligned} \tag{19}$$

Because z lies between x and y, both $|x - z|$ and $|y - z|$ are at most $|x - y|$. Therefore,

$$\Delta L(I, J \mid \mathcal{P}) \le \frac{M_I + M_J}{2}|x - y|^2. \tag{20}$$

No third-order remainder estimate or strong-duality argument is required for this step.

Relating multiplier displacement to efficiency. Since $C_I'(x), C_J'(y) > 0$, the first-order conditions imply

$$\rho_I(x) = \rho_J(y) = \lambda^*. \tag{21}$$

For $x \neq y$, the mean value theorem gives a ξ between x and y such that

$$\begin{aligned} |\rho_I(x) - \rho_J(x)| &= |\rho_J(y) - \rho_J(x)| \\ &= |\rho_J'(\xi)|\,|y - x| \\ &\ge L|y - x|. \end{aligned} \tag{22}$$

Combining this with Eq. (20) proves

$$\Delta L(I, J \mid \mathcal{P}) \le \frac{M_I + M_J}{2L^2}|\rho_I(x) - \rho_J(x)|^2. \tag{23}$$

If $x = y$, the original optimum is already feasible after merging, so the loss is zero and the same inequality holds. □

B.3 Proof of Corollary 1

Proof. Define $\widehat{\Delta\rho}(b) = \widehat{\rho}_I(b) - \widehat{\rho}_J(b)$. The estimation-error assumptions imply

$$|\Delta\rho(b_{\text{off}}) - \widehat{\Delta\rho}(b_{\text{off}})| \le \eta_I + \eta_J. \tag{24}$$

Using the triangle inequality, Lipschitz continuity, and the coverage assumption gives

$$\begin{aligned} |\Delta\rho(x)| &\le |\Delta\rho(b_{\text{off}})| + |\Delta\rho(x) - \Delta\rho(b_{\text{off}})| \\ &\le |\widehat{\Delta\rho}(b_{\text{off}})| + \eta_I + \eta_J + K_L\epsilon. \end{aligned} \tag{25}$$

□

C Verification of Structural Assumptions for the Exponential Bidding Model

In our empirical evaluations, the volume price responses are modeled using the standard exponential parametric functions:

$$G(b) = \alpha(1 - e^{-\beta b}), \qquad C(b) = \gamma b(1 - e^{-\beta b}), \tag{26}$$

where $\alpha, \beta, \gamma > 0$. We now rigorously verify that this formulation satisfies the structural assumptions required by Theorem 1 on a compact domain $[b_{\min}, b_{\max}]$ where $b_{\min} > 0$.

Cost Monotonicity. The first derivative of the cost function is:

$$C'(b) = \gamma \left(1 - e^{-\beta b} + b\beta e^{-\beta b}\right). \tag{27}$$

For any $b > 0$, we have $1 - e^{-\beta b} > 0$ and $b\beta e^{-\beta b} > 0$. Therefore, $C'(b) > 0$ strictly holds, ensuring the cost structure is continuous and strictly increasing, which guarantees the unique existence of the spending preserving feasible candidate.

Bounded Curvature. The second derivatives of the responses are:

$$\begin{aligned} G''(b) &= -\alpha\beta^2 e^{-\beta b}, \\ C''(b) &= \gamma\beta(2 - b\beta)e^{-\beta b}. \end{aligned} \tag{28}$$

Since these are smooth exponential functions, their absolute values are bounded by finite constants on the compact domain $[b_{\min}, b_{\max}]$. This bounds the second derivative of the Lagrange function, satisfying the condition $|V''(b)| \leq M$.

Strictly Bounded Marginal Efficiency Slope. The most critical assumption for the Merge Regret bound is the existence of a non zero lower bound L for the absolute slope of the marginal efficiency curve $|\rho'(b)|$. The marginal efficiency is defined as $\rho(b) = \frac{G'(b)}{C'(b)}$. Applying the quotient rule, the numerator of $\rho'(b)$ is given by $N(b) = G''(b)C'(b) - G'(b)C''(b)$.

Substituting the derivatives yields:

$$\begin{aligned} N(b) = &\left(-\alpha\beta^2 e^{-\beta b}\right) \cdot \gamma \left(1 - e^{-\beta b} + b\beta e^{-\beta b}\right) \\ &- \left(\alpha\beta e^{-\beta b}\right) \cdot \gamma\beta(2 - b\beta)e^{-\beta b}. \end{aligned} \tag{29}$$

Factoring out the common term $-\alpha\gamma\beta^2 e^{-\beta b}$, we obtain:

$$\begin{aligned} N(b) &= -\alpha\gamma\beta^2 e^{-\beta b} \left[\left(1 - e^{-\beta b} + b\beta e^{-\beta b}\right) + e^{-\beta b}(2 - b\beta)\right] \\ &= -\alpha\gamma\beta^2 e^{-\beta b} \left(1 - e^{-\beta b} + b\beta e^{-\beta b} + 2e^{-\beta b} - b\beta e^{-\beta b}\right). \end{aligned} \tag{30}$$

The terms containing $b\beta e^{-\beta b}$ exactly cancel out, reducing the expression to a strictly negative elegant form:

$$N(b) = -\alpha\gamma\beta^2 e^{-\beta b} \left(1 + e^{-\beta b}\right) < 0. \tag{31}$$

Because the denominator $[C'(b)]^2$ is bounded and bounded away from zero on $[b_{\min}, b_{\max}]$, and the numerator $N(b)$ is strictly negative, the derivative $\rho'(b)$ never vanishes. Consequently, there exists a strictly positive constant $L > 0$ such that $|\rho'(b)| \geq L$, completely fulfilling the requirements of Theorem 1.

D Empirical Validation of the Parametric Landscape

A core structural assumption of AIGB-ReAL is the parametric representation of the Decision Landscape. To rigorously justify whether our chosen functional forms for the GMV and Cost curves accurately capture real world traffic dynamics, we conduct an empirical goodness of fit analysis on the offline logged data from the AuctionNet environment. As visually demonstrated in Figure 6, the physical constraints of diminishing marginal returns are highly prevalent in the large-scale RTB environment.

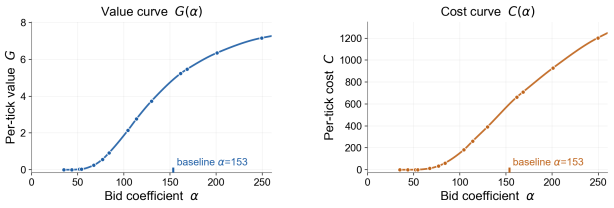

Figure 6: Empirical validation of the parametric assumption. The discrete points represent the resulting instantaneous GMV (left) and Cost (right) evaluations for each tested multiplier, which are sequentially connected to visualize the localized decision landscape.

To isolate the immediate economic impact without the compounding effects of subsequent state transitions, these curves are generated by evaluating a single specific timestep (tick). At this exact moment, the policies of all other agents are held constant while only the bidding multiplier of the focal agent is systematically varied

Table 5: Generalization of the curve-property trigger to the Hill curve form. **AIGB-ReAL Hill** replaces the default exponential curve formulation with a generalized Hill function. Score, converted volume (Conv), and Cost are reported as mean ± s.e. over 32 episodes. Best score per granularity is in **bold**.

$T = 48$				$T = 96$			
Method	Score	Conv	Cost	Method	Score	Conv	Cost
Random	22.97 ± 2.46	23.68 ± 2.37	1889.05 ± 177.03	Random	20.87 ± 1.68	25.34 ± 1.14	2750.82 ± 45.13
Uniform (TTC)	22.27 ± 2.61	23.00 ± 2.52	1810.51 ± 184.36	Uniform (TTC)	21.77 ± 1.65	25.86 ± 1.03	2789.50 ± 33.53
WaitWatch (ETC)	18.91 ± 2.42	19.53 ± 2.36	1427.71 ± 162.77	WaitWatch (ETC)	22.32 ± 1.65	24.22 ± 1.41	2189.94 ± 107.33
Interval-RL (DAR)	21.98 ± 2.58	22.78 ± 2.48	1831.08 ± 184.93	Interval-RL (DAR)	18.94 ± 1.79	24.44 ± 1.13	2797.30 ± 49.46
Safe Action Rep (SAR)	24.63 ± 2.04	25.25 ± 1.98	2106.66 ± 168.05	Safe Action Rep (SAR)	23.86 ± 1.72	26.91 ± 1.17	2716.35 ± 73.58
JointExploration (LICRA)	9.50 ± 1.31	18.78 ± 0.89	2899.95 ± 0.01	JointExploration (LICRA)	7.57 ± 0.91	17.53 ± 0.78	2899.95 ± 0.00
Ours (AIGB-ReAL Hill)	$\mathbf{23.41 \pm 2.26}$	24.06 ± 2.17	1992.20 ± 170.40	Ours (AIGB-ReAL Hill)	$\mathbf{23.25 \pm 1.84}$	26.34 ± 1.35	2553.90 ± 101.10
$T = 144$				$T = 288$			
Method	Score	Conv	Cost	Method	Score	Conv	Cost
Random	23.29 ± 1.92	27.24 ± 1.30	2790.61 ± 37.56	Random	26.09 ± 1.59	28.92 ± 1.17	2759.12 ± 45.44
Uniform (TTC)	24.44 ± 1.90	27.93 ± 1.26	2814.59 ± 36.75	Uniform (TTC)	27.17 ± 1.74	29.72 ± 1.29	2779.92 ± 46.17
WaitWatch (ETC)	25.50 ± 2.04	27.00 ± 1.73	2327.31 ± 109.45	WaitWatch (ETC)	25.28 ± 1.84	26.31 ± 1.69	2066.56 ± 125.83
Interval-RL (DAR)	23.45 ± 2.01	27.16 ± 1.30	2808.04 ± 44.24	Interval-RL (DAR)	25.08 ± 2.11	28.22 ± 1.60	2607.86 ± 84.70
Safe Action Rep (SAR)	24.54 ± 2.13	28.22 ± 1.48	2791.36 ± 70.88	Safe Action Rep (SAR)	25.52 ± 1.95	28.66 ± 1.38	2790.54 ± 56.94
JointExploration (LICRA)	8.09 ± 1.11	17.91 ± 0.78	2899.96 ± 0.00	JointExploration (LICRA)	7.34 ± 1.15	17.25 ± 0.86	2899.96 ± 0.01
Ours (AIGB-ReAL Hill)	$\mathbf{27.35 \pm 2.10}$	29.44 ± 1.64	2641.10 ± 81.10	Ours (AIGB-ReAL Hill)	$\mathbf{27.64 \pm 2.08}$	29.78 ± 1.62	2652.40 ± 68.90

E Generalization and Structural Verification: The Generalized Hill Formulation

In the main text, we instantiate the geometric curve decoder utilizing an exponential formulation to capture macroscopic market dynamics. However, the theoretical guarantees of the AIGB-ReAL scheduling framework, particularly the computable bounds derived for the merge regret, are fundamentally adaptable to other parametric families, provided they strictly adhere to the mathematical conditions defined in Theorem 1. To empirically validate this structural universality, we augment our evaluation by substituting the exponential representation with a generalized 6-parameter surrogate: the Hill function.

The Hill function is extensively utilized in econometrics, systems biology, and advertising dynamics to model saturation effects and diminishing marginal returns. For our continuous bidding horizon, we parameterize the expected GMV and Cost responses under the bidding multiplier α as:

$$G(\alpha) = \frac{A_g\, \alpha^{a_g}}{K_g^{a_g} + \alpha^{a_g}}, \qquad C(\alpha) = \frac{A_c\, \alpha^{a_c}}{K_c^{a_c} + \alpha^{a_c}}, \tag{32}$$

where the parameter space is strictly positive ($A_g, a_g, K_g > 0$ and $A_c, a_c, K_c > 0$). These parameters naturally govern the maximum capacity asymptotes (A_g, A_c), the nonlinear slope steepness (a_g, a_c), and the sensitivity thresholds (K_g, K_c).

We rigorously verify that this generalized 6-parameter formulation strictly satisfies the structural assumptions required by Theorem 1 on a compact domain $[\alpha_{\min}, \alpha_{\max}]$ with $\alpha_{\min} > 0$.

Cost Monotonicity. The first derivative of the cost function is derived algebraically:

$$C'(\alpha) = \frac{A_c a_c K_c^{a_c} \alpha^{a_c - 1}}{(K_c^{a_c} + \alpha^{a_c})^2}. \tag{33}$$

Because the multiplier domain is strictly positive ($\alpha \geq \alpha_{\min} > 0$) and all geometric parameters are positive, $C'(\alpha) > 0$ holds inherently. This ensures the cost structure is continuous and strictly increasing.

Bounded Curvature. The second derivatives $G''(\alpha)$ and $C''(\alpha)$ are smooth, continuous rational functions. Since the denominators $(K_g^{a_g} + \alpha^{a_g})^3$ and $(K_c^{a_c} + \alpha^{a_c})^3$ are strictly positive and bounded away from zero on the compact domain $[\alpha_{\min}, \alpha_{\max}]$, the absolute values of these second derivatives are bounded by finite constants. This limits the second derivative of the Lagrange function, satisfying the condition $|V''(\alpha)| \leq M$.

Strictly Bounded Marginal Efficiency Slope. To prove the existence of a non-zero lower bound L for the absolute slope of the marginal efficiency curve $|\rho'(\alpha)|$, we define the marginal efficiency

$\rho(\alpha) = G'(\alpha)/C'(\alpha)$. Substituting the derivatives yields:

$$\rho(\alpha) = \left(\frac{A_g a_g K_g^{a_g}}{A_c a_c K_c^{a_c}}\right) \alpha^{a_g - a_c} \left(\frac{K_c^{a_c} + \alpha^{a_c}}{K_g^{a_g} + \alpha^{a_g}}\right)^2. \tag{34}$$

Let $M_0 = \frac{A_g a_g K_g^{a_g}}{A_c a_c K_c^{a_c}} > 0$. To analyze the monotonicity cleanly without excessive polynomial expansion, we apply logarithmic differentiation:

$$\ln \rho(\alpha) = \ln M_0 + (a_g - a_c) \ln \alpha + 2 \ln(K_c^{a_c} + \alpha^{a_c}) - 2 \ln(K_g^{a_g} + \alpha^{a_g}). \tag{35}$$

Taking the derivative with respect to α and multiplying by $\rho(\alpha)$, we obtain:

$$\rho'(\alpha) = \rho(\alpha) \left[\frac{a_g - a_c}{\alpha} + \frac{2a_c \alpha^{a_c - 1}}{K_c^{a_c} + \alpha^{a_c}} - \frac{2a_g \alpha^{a_g - 1}}{K_g^{a_g} + \alpha^{a_g}}\right]. \tag{36}$$

In practical auto-bidding environments, the return on investment (ROI) inherently saturates, dictating that marginal efficiency must strictly decrease as the bid multiplier increases ($\rho'(\alpha) < 0$). To guarantee this theoretical constraint, our AIGB curve decoder restricts the output parameter space during inference (e.g., enforcing $a_g \leq a_c$ and controlling threshold bounds) to ensure the bracketed term remains strictly negative. Under this architecturally enforced constraint, the derivative $\rho'(\alpha)$ never crosses zero on the compact domain. Thus, there exists a strictly positive constant $L > 0$ such that $|\rho'(\alpha)| \geq L$, completely fulfilling Theorem 1.

Empirical Integration. We adapt the curve decoder to regress the full six parameters $\hat{\theta}_t = (\hat{A}_g, \hat{a}_g, \hat{K}_g, \hat{A}_c, \hat{a}_c, \hat{K}_c)$. Empirical results in Table 5 reveal that the AIGB-ReAL scheduler exhibits highly consistent trigger patterns when deploying the 6-parameter Hill formulation. This confirms that our dual-phase scheduling architecture successfully extracts underlying temporal boundaries regardless of the underlying functional complexity, demonstrating the profound robustness of the proposed framework across disparate parametric geometries.

F Instability of Joint RL Baselines

Figure 7 plots the reward of LICRA across all four temporal granularities ($T \in \{48, 96, 144, 288\}$). In every setting the reward exhibits large update-to-update oscillations around an essentially flat mean and never develops a rising, converging trend; the mean level even degrades slightly as the horizon becomes finer, while the oscillation amplitude grows (most severe at T=288). LICRA optimizes its surrogate objectives yet struggles to *learn a converged, high-quality solution* for the trigger-timing problem. This is corroborated by the offline evaluation (Table 1), where LICRA obtains the lowest score at every granularity while saturating the budget (cost pinned at the ceiling), indicating convergence to a poor local solution.

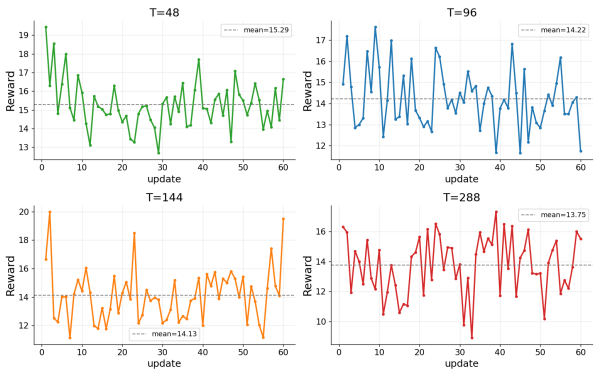

Figure 7: Reward of the LICRA baseline across varying temporal granularities ($T \in \{48, 96, 144, 288\}$). The dashed lines denote the mean reward level over the updates, illustrating both a lack of convergence and performance degradation as the horizon expands.

G Algorithm of AIGB-ReAL

In this section, we present the complete algorithmic workflow of the proposed Regret-Aware Adaptive Scheduling (AIGB-ReAL) framework, as detailed in Algorithm 1.

During the **Offline Phase** (Lines 3–14), the system processes historical observations to reconstruct the sequential geometric landscapes. Starting with the finest granularity (per timestep), the scheduler greedily merges adjacent intervals by strictly minimizing the theoretically bounded Merge Regret. This offline compression continues until the total number of partitions strictly satisfies the

pre-defined computational budget K. The resulting optimal schedule naturally yields a geometric threshold τ^*, which bounds the maximum tolerable divergence before an update becomes structurally necessary.

During the **Online Phase** (Lines 15–25), this derived threshold τ^* is deployed as a highly efficient observation-dependent trigger. At each incoming timestep, the lightweight curve decoder estimates the current market geometry. The system continuously evaluates the functional divergence d_t against the static anchor point. An active parameter recomputation ($g_t = 1$) is triggered if the threshold τ^* is breached.

Algorithm 1 AIGB-ReAL (**Re**gret-Aware **A**daptive Schedu**L**ing)

Input: Historical observations $\{s_t\}_{t=1}^{T}$, decision budget K
Output: Online trigger policy
Offline Phase:
for $t = 1, \ldots, T$ **do**
  $z_t \leftarrow \mathrm{Encoder}(s_{1:t})$
  $\hat{\theta}_t \leftarrow \mathrm{Decoder}(z_t)$
end for
Initialize $\mathcal{P} \leftarrow \{\{1\}, \ldots, \{T\}\}$.
while $|\mathcal{P}| > K$ **do**
  Compute Merge Regret for all adjacent intervals.
  $(P_i, P_{i+1}) \leftarrow \arg\min \mathrm{MR}(P_i, P_{i+1})$.
  Merge P_i and P_{i+1}.
end while
Derive optimal geometric threshold τ^* from the resulting schedule.
Online Phase:
Initialize temporal anchor $\tau \leftarrow 1$ and compute b_τ.
for each incoming timestep $t > \tau$ **do**
  $\hat{\theta}_t \leftarrow \mathrm{Decoder}(\mathrm{Encoder}(s_{1:t}))$.
  Evaluate functional divergence $d_t = D(\hat{\theta}_t, \hat{\theta}_\tau)$.
  if $d_t > \tau^*$ **then**
    $g_t \leftarrow 1$; compute new bid b_t; update anchor $\tau \leftarrow t$.
  else
    $g_t \leftarrow 0$; suppress action $b_t \leftarrow b_{t-1}$.
  end if
end for

H Online Experiment: Bidding Policy via BidDreamer

To enable the bidding policy adapt to the dynamic temporal boundaries established by the trigger, we formulate a joint control model that integrates the bidding policy with the temporal triggering mechanism.

H.1 World Model Representation Learning

Traditional reinforcement learning approaches struggle to converge in the extreme volatility and data sparsity inherent with 1-minute market modeling. To overcome this limitation, BidDreamer leverages a world model architecture to construct a dense, robust representation of the high-frequency auction environment within a latent space (Hafner et al., 2020). By conducting policy optimization through latent space imagination, the framework efficiently bridges the 1-minute bidding resolution with the trigger module. The environment dynamics are captured through a world model comprising three core components:

- **Causal Transformer Encoder**: The encoder processes the high-dimensional and noisy real observation sequence $s_1, \ldots, s_t$ to extract a dense latent state z_t.

- **State Transition Module**: Operating entirely in the latent space, this module predicts the subsequent latent state z_{t+1} conditioned on the current latent state z_t and the executed bid action a_t.
- **Sequence Decoder**: A Multi-Layer Perceptron (MLP) decoder projects the predicted latent state back to reconstruct the traffic sequence $\hat{s}_{t+1}$.

The entire world model is trained end to end by minimizing a combined objective comprising the latent transition loss and the observation reconstruction loss.

H.2 Policy Optimization via Latent Imagination

While the joint world model successfully captures the environment dynamics and incorporates the temporal trigger, the raw bidding policy is not inherently optimal under these trigger times. To achieve alignment with the temporal structure dictated by the trigger, the bidding policy is optimized using the reinforcement learning post-training.

During the latent imagination phase, the state transition mechanism explicitly incorporates the trigger information g_t alongside the bid a_t to simulate the forward trajectory $z_t \rightarrow z_{t+1} \rightarrow \ldots$. A dedicated Reward Model evaluates these imagined trajectories to compute a cumulative score. This score serves as the reinforcement feedback signal to update the bid head, as detailed in Figure 8.

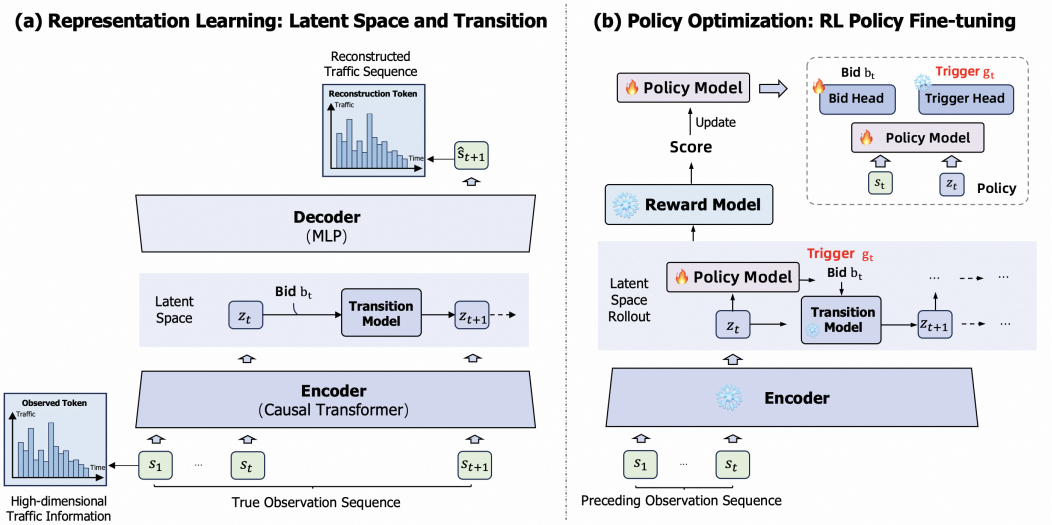

Figure 8: The BidDreamer architecture. (a) Representation Learning via latent space transitions and (b) Policy Optimization via RL fine-tuning.