CONFIDENCE-AWARE PARETO-EVOLVING AUTO-BIDDING VIA UNCERTAINTY-REGULARIZED COUNTERFACTUAL REGRET

Anonymous authors
Paper under double-blind review

ABSTRACT

Constrained auto-bidding requires maximizing marketing value while strictly satisfying efficiency constraints such as Target Cost-Per-Action (CPA). Recent Decision-Transformer-based methods introduce a dual-stream constraint context and a counterfactual regret loss to actively push the policy toward the Pareto frontier. However, we identify two fundamental gaps that cap their performance: (i) the Pareto frontier is constructed once from historical data and remains static, anchoring the learning target at the *empirical* rather than the *theoretical* optimum; and (ii) the outcome predictor used for counterfactual evaluation is a deterministic point estimator, so out-of-distribution counterfactuals can hallucinate high utility and mislead policy updates. We propose **CoPE-Bid**, a *Confidence-aware Pareto-Evolving* auto-bidding framework built on two synergistic mechanisms: (i) **Uncertainty-Regularized Counterfactual Regret (UR-CRO)** replaces the point-estimate predictor with a lightweight ensemble and evaluates counterfactuals via a lower-confidence-bound utility, so predictor uncertainty automatically discounts hallucinated regret; and (ii) **Frontier-Anchored Self-Improvement (FASI)** periodically performs confidence-gated imagination rollouts from the current policy and admits only high-confidence, Pareto-improving trajectories into the frontier, letting the sampling anchor co-evolve with the policy. The two modules are coupled through a shared uncertainty signal—UR-CRO tells FASI what to trust, and FASI tells UR-CRO what to aim for. Extensive experiments on AuctionNet, its sparse variant, and online A/B tests demonstrate that CoPE-Bid consistently outperforms state-of-the-art baselines in both value acquisition and constraint satisfaction, with particularly large gains in the noise-heavy sparse regime.

1 INTRODUCTION

Real-Time Bidding (RTB) requires a bidder to allocate impressions over a finite horizon and limited budget while satisfying efficiency constraints such as CPC, CPA, and ROAS (Zhao et al., 2018; Balseiro et al., 2021; Zhu et al., 2017; Nazerzadeh et al., 2008; Deng et al., 2023). The policy must therefore seek valuable traffic without exhausting budget or violating campaign-level targets (Li et al., 2024; Aggarwal et al., 2024; Zhang et al., 2025).

Unsafe online exploration motivates learning from historical logs with offline RL and sequence modeling (Korenkevych et al., 2024; Li et al., 2024; Jiang et al., 2025; Peng et al., 2025). Decision Transformers model bidding as autoregressive trajectory generation, but a Return-to-Go-only context ignores remaining cost and action regression imitates average logged behavior instead of the constraint-tight boundary (Chen et al., 2021; Li et al., 2025; Gao et al., 2025; Asadpour et al., 2019).

CDPR+CRO (Wu et al., 2026) addresses the first DT weakness with a Cost-to-Go stream, Pareto-prioritized sampling, and counterfactual regret. However, its frontier is extracted once from historical data, and its point-estimate outcome predictor may extrapolate optimistically on unsupported actions. Thus, the method can remain capped by the empirical frontier or update the policy toward hallucinated utility.

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.

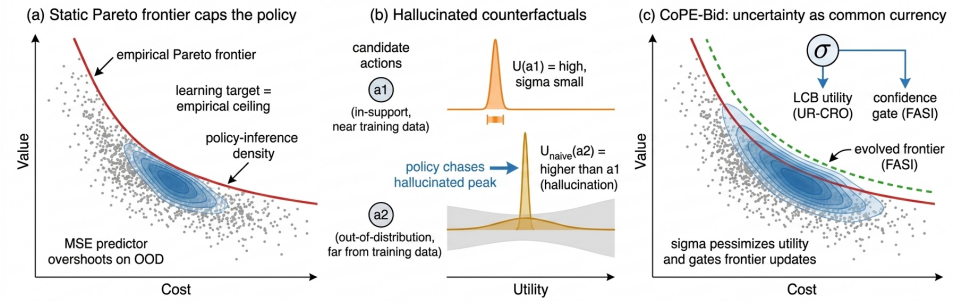


Figure 1: CoPE-Bid uses shared uncertainty to suppress hallucinated regret and gate Pareto-frontier evolution.

This motivates the central question of this paper: *How can an offline constrained bidder improve beyond the empirical Pareto frontier while avoiding policy updates driven by unreliable counterfactual predictions?*

These two limitations interact rather than appearing independently. An optimistic predictor can make unsupported actions look attractive, while a static frontier cannot absorb a genuinely useful improvement even when its outcome estimate is reliable. A practical solution therefore needs both pessimistic counterfactual evaluation and a guarded mechanism for updating the target, so that either component cannot silently amplify the other's errors.

We propose **CoPE-Bid** (**Co**nfidence-aware **P**areto-**E**volving **Bid**ding), which couples two mechanisms. **UR-CRO** uses a Gaussian ensemble and a lower-confidence-bound utility to down-weight uncertain counterfactual regret. **FASI** performs confidence-gated imagination rollouts and admits only compliant, Pareto-improving trajectories into the sampling frontier. The shared uncertainty signal makes the frontier evolve without turning unsupported predictions into training labels.

Figure 1 summarizes the design. We evaluate CoPE-Bid on AuctionNet, AuctionNet-Sparse, controlled robustness studies, and an online A/B test. Our contributions are: (i) diagnosing static-frontier and hallucinated-counterfactual failure modes; (ii) introducing the coupled UR-CRO/FASI framework; and (iii) demonstrating gains in value acquisition and constraint compliance across dense, sparse, noisy, and online settings.

2 Related Work

2.1 Offline Reinforcement Learning

Offline RL avoids costly online exploration but must control extrapolation outside the logging support. Support constraints (Fujimoto et al., 2019), conservative objectives (Kumar et al., 2020), expectile learning (Kostrikov et al., 2021), and diffusion policies (Hansen-Estruch et al., 2023) address this issue. Decision Transformer instead avoids Bellman bootstrapping by generating return-conditioned trajectories (Chen et al., 2021); constrained variants add safety information to the sequence (Liu et al., 2023).

2.2 Auto-Bidding

Auto-bidding translates campaign goals into per-request bids under budget and efficiency constraints (Deng et al., 2021; Aggarwal et al., 2024). Controllers (Chen et al., 2011; Yu et al., 2017) are interpretable but depend on stationarity, while deep/offline RL (Cai et al., 2017; He et al., 2021; Fujimoto et al., 2019) and generative bidders (Guo et al., 2024; Peng et al., 2025; Li et al., 2025; Gao et al., 2025) learn directly from logs. PRO-Bid (Wu et al., 2026) is the closest predecessor: it combines dual-stream context with counterfactual regret, but uses a static frontier and deterministic predictor.

2.3 Uncertainty and Model-Based Imagination

Deep ensembles support uncertainty-aware pessimism in offline RL (Lakshminarayanan et al., 2017; An et al., 2021; Yu et al., 2020) and confidence control for imagined rollouts (Janner et al., 2019; Hafner et al., 2020). CoPE-Bid transfers both ideas to constrained generative auto-bidding and uses one uncertainty signal for counterfactual utility and frontier admission.

3 Preliminaries

3.1 Problem Formulation

A campaign presents n candidate auctions over a fixed horizon; each impression i has valuation v_i and, if won, cost c_i, with $x_i \in \{0,1\}$ indicating whether the bidder acquires it. The advertiser solves

$$\begin{aligned} \max_{\{x_i\}} \quad & \sum_{i=1}^{n} x_i v_i \\ \text{s.t.} \quad & \sum_i c_i x_i \le B, \\ & \frac{\sum_i c_{ij} x_i}{\sum_i p_{ij} x_i} \le C_j, \quad x_i \in \{0,1\}. \end{aligned} \tag{1}$$

where B is the total budget cap and $\{C_j\}_{j=1}^{J}$ are J ratio-based efficiency targets (e.g., CPA is realized cost over realized conversions). Solving Eq. (1) combinatorially at auction speed is infeasible; primal-dual analysis (He et al., 2021) yields a closed-form optimum linear in v_i, which under CPA-constrained conversion maximization reduces to the proportional rule $b_i^* = \lambda^* v_i$. The campaign thus collapses onto picking a scalar pacing coefficient λ_t per step t.

3.2 Generative Bidding via Decision Transformer

Decision Transformer (Chen et al., 2021) casts control as autoregressive sequence generation. An episode is tokenized as

$$\tau = (R_1, s_1, a_1, \ldots, R_T, s_T, a_T), \tag{2}$$

with s_t the campaign feature vector, $a_t = \lambda_t \in \mathbb{R}^+$, and $R_t = \sum_{t'=t}^{T} r_{t'}$ the remaining reward budget. A causal transformer π_θ is fit by MSE regression on the offline log $\mathcal{D}$:

$$\mathcal{L}_{DT} = \tfrac{1}{T} \sum_{t=1}^{T} (a_t - \pi_\theta(\hat{a}_t \mid R_{\le t}, s_{\le t}, a_{<t}))^2. \tag{3}$$

Under hard ratio constraints this suffers two well-known problems: R_t-only conditioning is cost-blind so two states with identical remaining value but different remaining cost look alike (*state aliasing*), and MSE regression pins π_θ to the population mean of the log with no gradient toward the constraint-tight boundary.

4 The CoPE-Bid Framework

CoPE-Bid builds on a DT-based CDPR + CRO backbone and adds two coupled mechanisms designed to close the gaps identified in Sec. 1. Sec. 4.1 briefly recaps the backbone and pinpoints the gaps. Sec. 4.2 introduces UR-CRO to suppress hallucinated regret; Sec. 4.3 introduces FASI to make the frontier co-evolve with the policy. Sec. 4.4 presents the full objective and algorithm. Figure 2 illustrates the overall architecture.

4.1 Backbone: CDPR + CRO in a Nutshell

Dual-stream constraint context. To resolve state aliasing, CDPR augments the trajectory with a Cost-to-Go stream $C_t = \sum_{t'=t}^{T} c_{t'}$ mirroring $R_t = \sum_{t'=t}^{T} r_{t'}$, giving

$$\tau_{aug} = (R_1, C_1, s_1, a_1, \ldots, R_T, C_T, s_T, a_T), \tag{4}$$

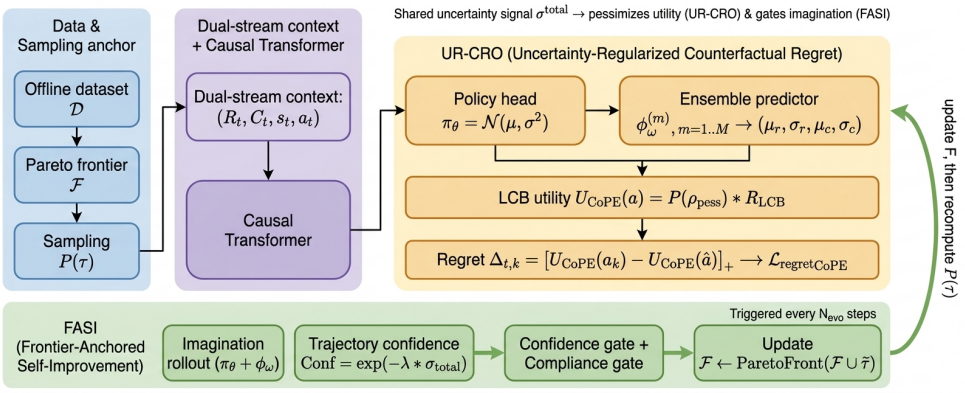


Figure 2: CoPE-Bid couples uncertainty-regularized counterfactual regret with confidence-gated frontier evolution.

with recursive updates $R_{t+1}=R_t-r_t$, $C_{t+1}=C_t-c_t$. At inference $C_1=B$ and $R_1=B/\rho_{\text{tgt}}$.

Pareto-prioritized sampling. Each $\tau_i \in \mathcal{D}$ is a point in $(\mathcal{C}(\tau_i), \mathcal{R}(\tau_i))$, and an empirical Pareto frontier $\mathcal{F}$ is the non-dominated set (Van Moffaert & Nowé, 2014). A quality score

$$\mathcal{Q}_i = \underbrace{e^{-\kappa d(\tau_i,\mathcal{F})}}_{S_{\text{eff}}} \cdot \underbrace{S_{\text{com}}^{(i)}}_{\text{compliance}} \cdot \underbrace{T_i/T_{\max}}_{S_{\text{len}}} \tag{5}$$

combines Pareto-distance efficiency, compliance ($S_{\text{com}}=1$ if $\rho_i \le \rho_{\text{tgt}}$, else $(\rho_{\text{tgt}}/\rho_i)^\omega$), and richness, giving $\mathbb{P}(\tau_i) \propto \mathcal{Q}_i$.

Counterfactual regret. A Gaussian policy $\pi_\theta(a_t \mid \cdot) = \mathcal{N}(\mu_\theta(h_t), \sigma_\theta^2(h_t))$ draws K counterfactuals $\{a_t^{(k)}\}$. A shared-backbone predictor ϕ_ω estimates $(\hat{r}_{t+1:T}, \hat{c}_{t+1:T})$, and the constraint-aware utility is

$$U(a) = \mathcal{P}(\rho(a); \rho_{\text{tgt}}) \cdot \mathcal{R}_{\text{total}}(a), \tag{6}$$

with $\mathcal{R}_{\text{total}}(a)=H_t^r+\hat{r}_{t+1:T}$, $\mathcal{C}_{\text{total}}(a)=H_t^c+\hat{c}_{t+1:T}$, $\rho(a)=\mathcal{C}_{\text{total}}(a)/(\mathcal{R}_{\text{total}}(a)+\epsilon)$, and $\mathcal{P}(\rho;\rho_{\text{tgt}}) = \min\{(\rho_{\text{tgt}}/\rho)^\gamma, 1\}$. The regret of omitting $a_t^{(k)}$ w.r.t. baseline $\hat{a}_t = \mu_\theta(h_t)$ is $\Delta_{t,k} = [U(a_t^{(k)}) - U(\hat{a}_t)]^+$, converted to Boltzmann weights $w_{t,k} \propto \exp(\Delta_{t,k}/\tau) \cdot \mathbb{I}[\Delta_{t,k}>0]$, and the policy mean is pulled toward high-utility counterfactuals via

$$\mathcal{L}_{\text{regret}} = \frac{1}{T}\sum_{t,k} w_{t,k} \|\hat{a}_t - \text{sg}(a_t^{(k)})\|^2. \tag{7}$$

Two gaps. This backbone leaves two issues untreated: (G1) $\mathcal{F}$ is fixed, so $\mathbb{P}(\tau_i)$ cannot track policy improvement; (G2) ϕ_ω is a point estimator, so $U(a)$ inherits its OOD extrapolation errors and $\mathcal{L}_{\text{regret}}$ can chase hallucinations. CoPE-Bid targets G2 via UR-CRO and G1 via FASI.

4.2 Uncertainty-Regularized Counterfactual Regret

Ensemble outcome predictor. We replace the single predictor with an M-head Gaussian ensemble sharing the transformer backbone and branching at the last layer. Each head outputs a Gaussian prediction over the future accumulated value and cost:

$$\phi_\omega^{(m)}(\cdot) \to \big(\mu_r^{(m)}, \log(\sigma_r^{(m)})^2, \mu_c^{(m)}, \log(\sigma_c^{(m)})^2\big), \tag{8}$$

trained by Gaussian negative log-likelihood (Lakshminarayanan et al., 2017):

$$\mathcal{L}_{\text{pred}} = \frac{1}{MT}\sum_{m,t}\Bigg[\frac{(r_{t+1:T}-\mu_r^{(m)})^2}{2(\sigma_r^{(m)})^2} + \frac{1}{2}\log(\sigma_r^{(m)})^2 + (\text{cost term})\Bigg]. \tag{9}$$

Ensemble aggregation yields the mean and *total* uncertainty, decomposed into aleatoric and epistemic components:

$$\bar{\mu}_r = \tfrac{1}{M}\sum_m \mu_r^{(m)},$$
$$(\sigma_r^{\text{tot}})^2 = \underbrace{\tfrac{1}{M}\sum_m (\sigma_r^{(m)})^2}_{\text{aleatoric}} + \underbrace{\text{Var}_m[\mu_r^{(m)}]}_{\text{epistemic}}, \tag{10}$$

and analogously for cost. Crucially, σ^{tot} is elevated precisely on counterfactuals lying outside the training support—where the naive point-estimate predictor is most prone to hallucinate.

Lower-confidence-bound utility. We build a pessimistic utility that treats future value as its LCB and future cost as its UCB:

$$\mathcal{R}_{\text{total}}^{\text{LCB}}(a) = H_t^r + \bar{\mu}_r(a) - \kappa_u\, \sigma_r^{\text{tot}}(a), \tag{11}$$
$$\mathcal{C}_{\text{total}}^{\text{UCB}}(a) = H_t^c + \bar{\mu}_c(a) + \kappa_u\, \sigma_c^{\text{tot}}(a),$$

with $\rho^{\text{pess}}(a) = \mathcal{C}_{\text{total}}^{\text{UCB}}(a)/(\mathcal{R}_{\text{total}}^{\text{LCB}}(a) + \epsilon)$, and confidence coefficient $\kappa_u \geq 0$. The CoPE utility replaces Eq. (6):

$$U^{\text{CoPE}}(a) = \mathcal{P}(\rho^{\text{pess}}(a); \rho_{\text{tgt}}) \cdot \mathcal{R}_{\text{total}}^{\text{LCB}}(a). \tag{12}$$

When σ^{tot} is small, U^{CoPE} recovers U; when it is large, the utility is aggressively discounted so the corresponding regret vanishes and the policy is not pulled toward the hallucinated action. The regret-weighted regression form of Eq. (7) is reused with U^{CoPE} replacing U: we define the CoPE regret gap $\Delta_{t,k} = [U^{\text{CoPE}}(a_t^{(k)}) - U^{\text{CoPE}}(\hat{a}_t)]^+$ and Boltzmann weight $w_{t,k} \propto \exp(\Delta_{t,k}/\tau) \cdot \mathbb{I}[\Delta_{t,k} > 0]$, yielding

$$\mathcal{L}_{\text{regret}}^{\text{CoPE}} = \tfrac{1}{T}\sum_{t,k} w_{t,k} \|\hat{a}_t - \text{sg}(a_t^{(k)})\|^2. \tag{13}$$

4.3 Frontier-Anchored Self-Improvement

FASI lets the frontier $\mathcal{F}$ co-evolve with the policy, gated by the very same uncertainty signal produced by UR-CRO.

Warm-up. For the first N_w steps FASI is disabled, so the ensemble first learns to produce well-calibrated σ^{tot} before it is trusted to gate frontier updates. In practice $N_w \approx 0.2N_{\max}$.

Confidence-gated imagination rollout. Every N_{evo} steps after warm-up, we (i) sample N_v initial contexts from $\mathcal{D}$; (ii) autoregressively unroll virtual trajectories $\{\tilde{\tau}^{(v)}\}$ using π_θ for actions and the ensemble mean $(\bar{\mu}_r, \bar{\mu}_c)$ for step transitions; (iii) score each trajectory with a confidence

$$\text{Conf}(\tilde{\tau}) = \exp(-\lambda_c\, \bar{\sigma}^{\text{tot}}(\tilde{\tau})), \tag{14}$$

where $\bar{\sigma}^{\text{tot}}$ is time-averaged along $\tilde{\tau}$; (iv) apply a *confidence gate* ($\text{Conf} > \tau_{\text{conf}}$) and a *compliance gate* ($\rho(\tilde{\tau}) \leq \rho_{\text{tgt}}$); (v) admit into $\mathcal{F}$ only those surviving virtual trajectories that strictly Pareto-dominate an existing frontier point or extend the frontier:

$$\mathcal{F} \leftarrow \text{ParetoFront}(\mathcal{F} \cup \{\tilde{\tau} : \text{both gates pass}\})\,; \tag{15}$$

(vi) recompute $\mathcal{Q}_i$ and $\mathbb{P}(\tau_i)$ for all $\tau_i \in \mathcal{D}$ against the updated $\mathcal{F}$.

Robustness by construction. Two conservative choices guard FASI against imagination errors: (a) *Anchor-only, not label*: virtual trajectories only re-weight the sampling distribution over real data; they never enter minibatches as regression targets. (b) *Monotone frontier*: FASI only adds to $\mathcal{F}$, never removes empirical points, so if the policy generates no confident improvement, $\mathcal{F}$ stays put and CoPE-Bid gracefully degenerates to the CDPR baseline.

4.4 Total Training Objective

The total loss integrates four components—policy anchoring, uncertainty-regularized improvement, ensemble prediction, and exploration:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{NLL}}}_{\text{stability}} + \alpha \underbrace{\mathcal{L}_{\text{regret}}^{\text{CoPE}}}_{\text{safe improvement}} + \beta \underbrace{\mathcal{L}_{\text{pred}}}_{\text{predictor}} - \eta \underbrace{\mathcal{H}(\pi_\theta)}_{\text{exploration}}, \tag{16}$$

Algorithm 1 Training Procedure of CoPE-Bid

Input: Offline dataset $\mathcal{D}$; ensemble size M; warm-up N_w; interval N_{evo}; gate τ_{conf}.
Output: Policy θ; ensemble $\{\omega^{(m)}\}$.

```
Build initial F from D; compute P(τ) via Eq. (5).
for step = 1, ..., N_max do
    Sample B ~ P(τ); augment via Eq. (4).
    Sample K counterfactuals {a^(k)} from π_θ.
    Predict (μ̄, σ^tot) via ensemble (Eqs. (8)–(10)).
    Compute U^CoPE (Eq. (12)) and L^CoPE_regret (Eq. (13)).
    Update (θ, {ω^(m)}) by minimizing L_total (Eq. (16)).
    if step > N_w and step mod N_evo = 0 then
        Confidence-gated rollout → update F via Eq. (15); recompute P(τ).
    end if
end for
```

with $\mathcal{L}_{\text{NLL}} = -\log \pi_\theta(a_t \mid \cdot)$ anchoring the mean to demonstrated actions and $\mathcal{H}(\pi_\theta)$ the Shannon entropy (Haarnoja et al., 2018). The full procedure is summarized in Algorithm 1.

4.5 Discussion: Uncertainty as a Common Currency

The two mechanisms share σ^{tot}: UR-CRO uses it to pessimistically discount uncertain counterfactuals (Eq. (11)), while FASI uses it to reject uncertain imagined trajectories (Eq. (14)). This coupling prevents both failure modes: UR-CRO alone is safe but capped by a static frontier, whereas FASI without uncertainty can admit hallucinated frontier points. The joint method therefore improves only inside the ensemble's trusted region, extending LCB-based offline RL from value pessimism to both counterfactual utility and imagined-data admission (Yu et al., 2020; An et al., 2021).

4.6 Practical Considerations

UR-CRO evaluates K counterfactuals with an M-head ensemble, giving an amortized training overhead proportional to $O(MK)$ per decision step. FASI is invoked only every N_{evo} updates and uses N_v virtual rollouts, so its cost is amortized over training. Neither module changes the serving interface: at deployment, the bidder emits the policy mean $\mu_\theta(h_t)$, while ensemble evaluation and Gaussian sampling are used only during training. Virtual trajectories reweight real logged data but never become regression labels; the monotone frontier update retains empirical points and falls back to the original CDPR anchor when no trajectory passes both gates.

Two details make the gates conservative. We use total uncertainty, including both ensemble disagreement and predicted observation noise, because either source can make a counterfactual unsafe for a ratio constraint. Confidence is averaged over the whole imagined trajectory, so one highly uncertain segment can reject an otherwise attractive rollout. The warm-up period lets the ensemble calibrate before FASI is enabled, and strict Pareto admission prevents small noisy improvements from repeatedly moving the anchor. At serving time, the Cost-to-Go and Return-to-Go tokens are initialized from the campaign budget and target and refreshed with realized outcomes; the uncertainty machinery remains a training-time safety regularizer.

Gradient flow and update ordering. Each ensemble head is fitted with the observed-transition prediction loss, so its targets never depend on imagined rewards or costs. UR-CRO uses the ensemble statistics to reweight the policy improvement term, while the predictor continues to learn from logged outcomes. At an FASI refresh, rollouts are generated with the current policy, filtered by the confidence and compliance gates, and used only to recompute sampling probabilities. The next minibatch then updates the policy and predictors on real transitions. This ordering prevents a high-utility hallucination from reinforcing itself through either the predictor or the frontier.

Table 1: Score comparison across budget settings. Boldface denotes the highest score; underline the best baseline. "*" indicates statistically significant improvements over the best baseline ($p < 0.05$, two-sided t-test).

Dataset	Bud.	BCQ	CQL	IQL	Diff	DT	CDT	GAS	GAVE	PRO	CoPE	Imp.
AuctionNet	50%	152.8	171.9	178.2	77.1	187.1	175.1	188.4	195.4	204.2	**213.8***	4.70%
	75%	210.3	244.3	246.3	128.4	272.1	251.1	278.6	283.4	291.4	**303.6***	4.19%
	100%	267.4	289.6	314.1	189.8	325.8	328.0	350.9	362.6	371.9	**388.7***	4.52%
	125%	313.7	335.7	351.7	193.5	394.6	384.5	404.1	413.9	426.3	**444.9***	4.36%
	150%	354.5	363.2	384.6	234.3	454.9	437.7	453.4	459.6	470.6	**489.1***	3.93%
AuctionNet-Sparse	50%	14.77	16.94	18.16	12.01	14.88	17.72	18.52	19.76	20.95	**22.34***	6.63%
	75%	21.34	24.27	24.23	18.63	22.31	26.39	27.32	28.02	28.76	**30.65***	6.57%
	100%	26.32	28.37	29.16	23.72	29.96	32.64	35.72	36.98	38.10	**40.87***	7.27%
	125%	32.15	33.20	36.89	29.78	34.27	36.36	39.96	42.54	43.74	**46.98***	7.41%
	150%	37.52	37.06	42.19	37.47	43.73	44.56	46.57	47.46	49.87	**53.63***	7.54%

5 Offline Experiment

We evaluate CoPE-Bid on two public datasets for overall performance, module contributions, robustness, and frontier evolution.

5.1 Experimental Setup

Datasets. We use AuctionNet (Su et al., 2024) and its sparse variant, where lower conversion rates and wider CPA ranges amplify both failure modes. Dataset details are in the Appendix.

Metrics. *Value* $V = \sum_i x_i v_i$; *AR* = $\text{CPA}_{\text{realized}}/\text{CPA}_{\text{target}}$ (lower$\leq$1 is better); *ER* = fraction of test episodes with $AR > 1$; *Score* = $V \cdot \min(1, (1/AR)^{\gamma})$ with γ=2.

Baselines. We compare with offline RL (BCQ, CQL, IQL), diffusion (DiffBid), transformer bidders (DT, CDT, GAS, GAVE), and PRO-Bid (Wu et al., 2026). Implementation details are in the Appendix.

5.2 Overall Performance (RQ1)

Table 1 shows that CoPE-Bid achieves the highest Score at every budget on both datasets, improving over the strongest baseline by 3.93–4.70% on AuctionNet and 6.57–7.54% on AuctionNet-Sparse. The full Value/AR/ER breakdown is in the Appendix; CoPE-Bid improves value while reducing violations, with larger gains in the sparse regime.

5.3 Ablation Study (RQ2)

Table 2: Ablation study at 100% budget.

AuctionNet

Method	Conv	AR	ER	Score
CoPE-Bid	**405**	**0.83**	**0.20**	**388.7**
w/o UR-CRO	412	0.90	0.31	364.0
w/o FASI	398	0.84	0.22	376.0
w/o both	390	0.86	0.23	371.9

AuctionNet-Sparse

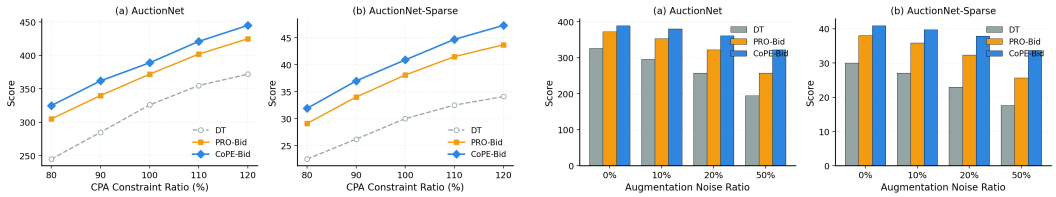
Method	Conv	AR	ER	Score
CoPE-Bid	**42.9**	**0.78**	**0.18**	**40.87**
w/o UR-CRO	43.5	0.89	0.29	37.50
w/o FASI	41.2	0.80	0.20	39.60
w/o both	40.5	0.81	0.21	38.10

Table 2 confirms the complementarity of the two modules. Removing UR-CRO causes the largest ER increase (0.20→0.31 on AuctionNet; 0.18→0.29 on Sparse), while removing FASI mainly lowers Score (388.7→376; 40.87→39.60). Removing both recovers the CDPR+CRO backbone. Uncertainty gating is therefore necessary for safely evolving the frontier.

5.4 Robustness and Frontier Analysis (RQ3, RQ4)

UR-CRO robustness. Figure 3a shows that CoPE-Bid remains strong as the CPA target varies from tight to loose, while Figure 3b shows a slower degradation under injected OOD noise. The LCB utility prevents uncertain counterfactuals from dominating updates.

FASI frontier evolution. Figure 4 shows CoPE-Bid trajectories beyond the empirical frontier, whereas removing FASI collapses the outer shell back toward the historical frontier.



(a) CPA constraint. (b) Training noise.

Figure 3: Robustness to constraint tightness and training-data noise.

5.5 Sensitivity Analysis

We vary the LCB coefficient κ_u and ensemble size M at 100% budget on AuctionNet (Table 3). Setting κ_u=0 recovers point-estimate CRO and increases ER, whereas κ_u=2 becomes mildly over-conservative; κ_u=1 gives the best Score–ER trade-off. Increasing M beyond 3 yields only marginal gains at substantially higher predictor cost, so we use M=3 in the main experiments.

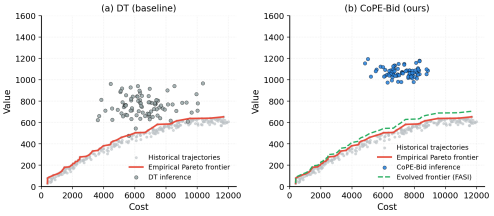


Figure 4: Inference trajectories and the evolved frontier.

Table 3: Sensitivity to LCB coefficient κ_u and ensemble size M at 100% budget on AuctionNet.

κ_u	Score	AR	ER	M	Score	AR	ER
0.0	380.1	0.88	0.26	1	376.0	0.87	0.24
0.5	385.4	0.85	0.22	3	**388.7**	**0.83**	**0.20**
1.0	**388.7**	**0.83**	**0.20**	5	390.1	0.83	0.19
2.0	383.5	0.82	0.19	7	390.4	0.82	0.19

5.6 Counterfactual Reliability Analysis

To examine the mechanism behind UR-CRO, Table 4 compares three counterfactual actions sampled on AuctionNet-Sparse. A naive point predictor assigns the highest utility to a_2, an out-of-distribution action. The ensemble instead reports a large disagreement for a_2, reducing its LCB utility to the lowest value and its regret weight to 3.1%. Supported action a_1 and the low-uncertainty safe action a_3 retain meaningful weights. The same uncertainty signal is used by FASI, so a rollout near a_2 is also rejected before it can redefine the frontier. The full case study and sampling details are provided in the Appendix.

At 100% budget, the safety–value improvement is also visible in the four-metric comparison with the strongest baseline (Table 5). CoPE-Bid raises Value and Score while lowering both the average ratio and episode violation rate on both datasets; the complete baseline comparison remains in the Appendix.

Table 4: Representative counterfactuals on AuctionNet-Sparse.

	a_1 (in)	a_2 (OOD)	a_3 (safe)
U_{naive}	205.3	**245.6**	191.8
σ_r^{tot}	5.2	**45.1**	7.8
U^{CoPE}	<u>197.9</u>	128.5	179.6
$w_{t,k}^{\text{CoPE}}$ (%)	63.4	3.1	33.5

Table 5: Value and constraint metrics at 100% budget.

Dataset	Method	Value	AR	ER	Score
AuctionNet	PRO-Bid	390.1	0.862	0.229	371.9
	CoPE-Bid	**405.2**	**0.831**	**0.198**	**388.7**
Sparse	PRO-Bid	40.52	0.813	0.208	38.10
	CoPE-Bid	**42.87**	**0.782**	**0.176**	**40.87**

6 Online A/B Test

Under the Multiple Constraint Bidding protocol, a seven-day random traffic split showed gains of 8.94% in GMV, 11.30% in clicks, 9.42% in ROI, and 7.65% in constraint compliance. Deployment details are in the Appendix.

6.1 Reproducibility and Scope

Unless stated otherwise, offline results are averages over 10 seeds with independent training and evaluation splits. We train for $200{,}000$ optimizer steps with batch size 128 and AdamW at learning rate 1×10^{-5}; the default CoPE-Bid configuration uses $M{=}3$ ensemble heads, $K{=}5$ counterfactual actions, a $40{,}000$-step warm-up, a $5{,}000$-step FASI interval, and 200 virtual rollouts per refresh. Baselines use author-recommended settings, with mild retuning applied uniformly when it improves validation performance. These choices, together with the feature processing and deployment schedule, are listed in the Appendix so that the reported comparisons can be reproduced without relying on hidden implementation choices.

Scope and limitations. The safety behavior is conditional on the ensemble being calibrated on the logged support. If traffic moves to a region where all heads extrapolate similarly, disagreement alone can underestimate epistemic risk; conversely, a conservative ensemble can reject a genuinely useful improvement. FASI therefore uses strict confidence, compliance, and Pareto gates, which favors reliable gains at the cost of possibly missing rare improvements. Our offline evaluation covers two AuctionNet variants and controlled training-data noise, while the online result uses one seven-day randomized split. Broader shifts and calibration under policy-induced distribution shift remain future work. At serving time the bidder emits the policy mean, so the ensemble is a training-time safeguard rather than a hard runtime constraint.

Statistical interpretation. Table 1 marks improvements over the best baseline using a two-sided t-test at $p < 0.05$, and all offline tables report the same seed aggregation. The online gains are traffic-level measurements under the stated random split; they support practical effectiveness but do not by themselves establish invariance across advertisers or seasons.

Failure handling. FASI treats each frontier refresh as a gated transaction: if no imagined trajectory satisfies confidence, compliance, and strict Pareto improvement, the frontier and the sampling distribution remain unchanged. This safe no-op preserves the empirical support and reduces sensitivity to anomalous minibatches. The same fallback applies when the policy produces no reliable counterfactual improvement, so the method can revert to the CDPR anchor without changing the serving interface.

7 Conclusion

CoPE-Bid couples UR-CRO and FASI to address static-frontier anchoring and hallucinated counterfactuals. Across offline and online settings, it improves value while tightening compliance; confidence and compliance gates keep imagined rollouts from becoming unreliable training labels. Results across five budget levels, ablations, robustness tests, and the counterfactual case study show that the gains come from complementary roles: UR-CRO filters unreliable improvements, while FASI expands the trusted frontier. Calibrating this shared uncertainty signal under larger policy-induced distribution shifts remains an important direction for future work, especially for longer-running and more diverse advertiser campaigns.

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.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified Q-ensemble. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7436–7447, 2021.

Arash Asadpour, MohammadHossein Bateni, Kshipra Bhawalkar, and Vahab Mirrokni. Concise bid optimization strategies with multiple budget constraints. *Management Science*, 65(12):5785–5812, 2019.

Santiago Balseiro, Yuan Deng, Jieming Mao, Vahab Mirrokni, and Song Zuo. Robust auction design in the auto-bidding world. *Advances in Neural Information Processing Systems*, 34:17777–17788, 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, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Ye Chen, Pavel Berkhin, Bo Anderson, and Nikhil R Devanur. Real-time bidding algorithms for performance-based display ad allocation. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1307–1315, 2011.

Yuan Deng, Jieming Mao, Vahab Mirrokni, and Song Zuo. Towards efficient auctions in an auto-bidding world. In *Proceedings of the Web Conference 2021*, pp. 3965–3973, 2021.

Yuan Deng, Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang, and Vahab Mirrokni. Multi-channel autobidding with budget and roi constraints. In *International Conference on Machine Learning*, pp. 7617–7644. PMLR, 2023.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Jingtong Gao, Yewen Li, Shuai Mao, Peng Jiang, Nan Jiang, Yejing Wang, Qingpeng Cai, Fei Pan, Peng Jiang, Kun Gai, et al. Generative auto-bidding with value-guided explorations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 244–254, 2025.

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*, pp. 5038–5049, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

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

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

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*, pp. 2993–3001, 2021.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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.

Dmytro Korenkevych, Frank Cheng, Artsiom Balakir, Alex Nikulkov, Lingnan Gao, Zhihao Cen, Zuobing Xu, and Zheqing Zhu. Offline reinforcement learning for optimizing production bidding policies. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5251–5259, 2024.

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.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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, 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, 2025.

Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. In *International conference on machine learning*, pp. 21611–21630. PMLR, 2023.

Hamid Nazerzadeh, Amin Saberi, and Rakesh Vohra. Dynamic cost-per-action mechanisms and applications to online advertising. In *Proceedings of the 17th international conference on World Wide Web*, pp. 179–188, 2008.

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. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5963–5970, 2025.

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

Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.

Binglin Wu, Yingyi Zhang, Xianneng Li, Ruyue Deng, Chuan Yue, Weiru Zhang, and Xiaoyi Zeng. Constraint-aware generative auto-bidding via pareto-prioritized regret optimization. *arXiv preprint arXiv:2602.08261*, 2026.

Hao Yu, Michael Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Leping Zhang, Xiao Zhang, Yichao Wang, Xuan Li, Zhenhua Dong, and Jun Xu. Adapting constrained markov decision process for ocpc bidding with delayed conversions. *ACM Transactions on Information Systems*, 43(2):1–29, 2025.

Jun Zhao, Guang Qiu, Ziyu Guan, Wei Zhao, and Xiaofei He. Deep reinforcement learning for sponsored search real-time bidding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1021–1030, 2018.

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*, pp. 2191–2200, 2017.

A Dataset Statistics

AuctionNet is a public multi-agent auction simulator designed to reproduce the competitive dynamics of a real ad exchange. Each release contains 9,987 separate delivery periods, and every period is discretized into 48 decision steps, so the full corpus offers 479,376 bidding trajectories. Because reward sparsity is one of the harder regimes for offline bidders, AuctionNet ships in two variants: a dense split with regular conversion signals, and a sparse split (*AuctionNet-Sparse*) that has much lower conversion rates and a substantially wider CPA target range. The sparse split is intentionally more punishing for learning, and this is where hallucination-related failure modes (G2 in Sec. 1) tend to dominate empirical results. Every trajectory shares the same interface: a 16-dimensional real-valued state, a scalar bidding action, and a scalar return-to-go token. The full dictionary of state features and the parameter ranges of the two variants are collected in Table 6. In detail, the state coordinates provided by the benchmark are:

- time_left: number of decision slots still ahead in the current campaign period.
- budget_left: unspent budget available at the start of the current slot.
- historical_bid_mean: cumulative average of the bidder's own submitted prices up to the current step.
- last_three_bid_mean: same quantity but restricted to the three most recent slots.
- historical_LeastWinningCost_mean: cumulative average of the minimum price needed to win, observed from past auctions.
- last_three_LeastWinningCost_mean: the same clearing-price average, restricted to the last three slots.
- historical_pValues_mean: cumulative mean of the platform-provided predicted conversion probabilities.
- last_three_pValues_mean: predicted-conversion mean restricted to the last three slots.
- current_pValues_mean: average pCVR over the impressions available in the current slot.
- historical_conversion_mean: mean realized conversion count accumulated before the current slot.
- last_three_conversion_mean: conversion-count mean over the previous three slots.
- historical_xi_mean: fraction of participated auctions historically won by the bidder.
- last_three_xi_mean: same winning-rate signal restricted to the last three slots.
- current_pv_num: number of impression opportunities that just arrived.
- last_three_pv_num_total: total impressions served during the previous three slots.
- historical_pv_num_total: total impressions served across the entire history so far.

Table 6: AuctionNet corpus statistics used in the offline evaluation.

Params	AuctionNet	AuctionNet-Sparse
Trajectories	479,376	479,376
Delivery Periods	9,987	9,987
Time steps in a trajectory	48	48
State dimension	16	16
Action dimension	1	1
Return-To-Go Dimension	1	1
Budget range	$[450, 11850]$	$[2000, 4850]$
Action range	$[0, 493]$	$[0, 589]$
Impression's value range	$[0, 1]$	$[0, 1]$
CPA range	$[6, 12]$	$[60, 130]$
Total conversion range	$[0, 1512]$	$[0, 57]$

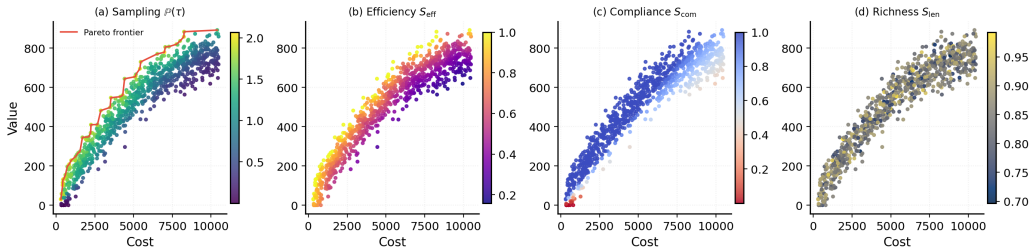


Figure 5: Decomposition of CDPR sampling weights on AuctionNet.

B Implementation Details

All runs use a cluster of NVIDIA L20 GPUs. Each training job goes through 200,000 optimizer steps at batch size 128, using AdamW at learning rate 1×10^{-5}. The causal transformer backbone shared between the policy head and the ensemble outcome predictor is instantiated with 8 layers and 16 attention heads, matching the capacity used by strong DT-based bidders in the benchmark protocol. CoPE-Bid-specific hyperparameters take the following default values: ensemble size M=3; LCB coefficient κ_u=1.0; warm-up length N_w=40,000 steps; FASI evolution interval N_{evo}=5,000; per-evolution virtual rollout count N_v=200; trajectory-confidence decay λ_c=0.5; and confidence gate τ_{conf}=0.6. The regret temperature τ, counterfactual count K=5, and loss weights (α, β, η) follow the PRO-Bid backbone (Wu et al., 2026), isolating the effect of the uncertainty and frontier-evolution modules. For competing baselines, we start from author-recommended settings and report the better result when mild retuning improves performance. All reported numbers are averages across 10 seeds with independent training/evaluation splits.

C Additional Metric Comparison

Table 7 reports the full four-metric breakdown at 100% budget across all baselines. CoPE-Bid achieves the highest Value *and* the lowest AR/ER on both AuctionNet and AuctionNet-Sparse, confirming that the safety-vs-value trade-off is not resolved by sacrificing one to gain the other. Compared with the strongest baseline, CoPE-Bid tightens ER by 13% (0.229 → 0.198) on AuctionNet and 15% (0.208 → 0.176) on Sparse, while simultaneously raising Value.

Table 7: Full four-metric comparison at 100% budget. Arrows indicate the desired direction.

Dataset	Metric	BCQ	CQL	IQL	DiffBid	DT	CDT	GAS	GAVE	PRO-Bid	CoPE-Bid
AuctionNet	Value ↑	325.1	335.3	328.8	297.6	367.1	357.4	378.6	399.7	390.1	**405.2**
	AR ↓	1.020	0.978	0.871	1.434	1.003	0.968	0.966	1.040	0.862	**0.831**
	ER ↓	0.479	0.354	0.292	0.729	0.458	0.396	0.375	0.479	0.229	**0.198**
	Score ↑	267.4	289.6	314.1	189.8	325.8	328.0	350.9	362.6	371.9	**388.7**
AuctionNet-Sparse	Value ↑	30.40	30.28	32.17	31.40	36.27	35.31	38.76	41.43	40.52	**42.87**
	AR ↓	0.942	0.922	0.834	1.358	0.976	0.857	0.882	1.014	0.813	**0.782**
	ER ↓	0.417	0.312	0.292	0.688	0.438	0.322	0.396	0.458	0.208	**0.176**
	Score ↑	26.32	28.37	29.16	23.72	29.96	32.64	35.72	36.98	38.10	**40.87**

D Hallucination Case Study

To provide an explicit illustration of how UR-CRO suppresses predictor hallucinations, we examine three representative counterfactual actions sampled during training on AuctionNet-Sparse, spanning the full distributional-shift spectrum (Table 8). Actions are annotated by their distance to the training support: a_1 is a well-supported action near the empirical density, a_2 is a highly out-of-distribution action, and a_3 is an OOD but genuinely safe action (rare in the offline logs but consistent with the ensemble's agreement).

Interpretation. A vanilla point-estimate predictor is deceived by a_2, assigning it the highest utility of the three actions (245.6) and thus generating a strong positive regret signal—a naive single-

Table 8: Illustrative counterfactual actions during training on AuctionNet-Sparse. U_{naive} is the utility a naive single-model predictor would assign; $\bar{\mu}_r$ and σ_r^{tot} are the ensemble mean and total uncertainty for the future value; U^{CoPE} is the LCB utility under κ_u=1. The last row shows the resulting regret weight $w_{t,k}^{\text{CoPE}}$.

Counterfactual action	a_1 (in-support)	a_2 (OOD)	a_3 (OOD, safe)
U_{naive} (naive utility)	205.3	**245.6**†	191.8
Ensemble mean $\bar{\mu}_r$	208.1	218.4	195.2
Ensemble std σ_r^{tot}	5.2	**45.1**	7.8
LCB utility U^{CoPE}	<u>197.9</u>	128.5	179.6
Regret weight $w_{t,k}^{\text{CoPE}}$ (%)	63.4	3.1	33.5

† Highest under the naive utility—a naive CRO would learn from this hallucinated action.

model CRO would confidently pull the policy toward this hallucinated over-bidding action. The ensemble, however, records a large disagreement (epistemic σ_r^{tot}=45.1) precisely because different heads extrapolate differently on a_2; the LCB utility (Eq. (12)) discounts a_2's effective utility to 128.5, which is now the lowest of the three, and consequently a_2 contributes near-zero regret weight $w_{t,k}^{\text{CoPE}}$. Meanwhile a_1 and a_3, both of which enjoy low ensemble disagreement, retain most of their utility, and the policy update is dominated by whichever of them provides a real improvement.

The same σ^{tot} is subsequently reused by FASI: if the imagination rollout starting near a_2 inherits similarly high variance, the confidence gate (Eq. (14)) rejects the entire imagined trajectory before it can enter $\mathcal{F}$. This single shared uncertainty signal thus achieves two-sided protection: it prevents CRO from chasing hallucinated regret, and it prevents FASI from admitting hallucinated frontier points.

E Deploy Details of Online A/B Test

For online deployment we retrain CoPE-Bid daily on a sliding fourteen-day window of anonymized production logs, totaling roughly 300,000 trajectories. Because the raw features span several orders of magnitude and exhibit heavy-tailed distributions, we apply per-family logarithmic normalization before feeding them into the transformer; this keeps the ensemble variance estimates well-conditioned and, empirically, reduces predictor collapse on the very high-cost tail. The utility function inside UR-CRO is reweighted at deployment time to reflect the business-level trade-off between value acquisition and constraint slack—compliant trajectories receive their raw utility, whereas overspending trajectories are down-weighted by the same LCB mechanism described in the main paper.

At the start of each campaign we initialize the Cost-to-Go token with the full daily budget and the Return-to-Go token by rescaling the daily budget with the advertiser-specified efficiency target. Both tokens are refreshed after every filled impression using the realized cost and value, so the dual-stream representation acts as a live pacing signal rather than a static prior. Because real-time bidding tolerates only single-digit millisecond latencies, we bypass Gaussian sampling at serving time and issue the mean of the predicted action distribution as the final λ_t; UR-CRO's uncertainty penalty is applied only during training, since it is a training-time regularizer on which counterfactuals to trust, not a runtime randomization. The empirically observed uplift in GMV and ROI (Sec. 5) confirms that this configuration preserves the value-seeking and pacing-safety properties of CoPE-Bid in production traffic.

F Visualization of Sampling Weights

Figure 5 decomposes the CDPR sampling probability $\mathbb{P}(\tau)$ on the AuctionNet corpus into its three constituent factors, plotted in $(\mathcal{C}, \mathcal{R})$ space. Each dot is one recorded trajectory, and its color encodes the factor being visualized.

Panel (b) shows that the efficiency term S_{eff} upweights trajectories that lie near the empirical Pareto frontier and quickly attenuates otherwise, matching the exp-of-negative-distance formulation of Eq. (5). Panel (c) shows that the compliance term S_{com} imposes a sharp gradient against the region

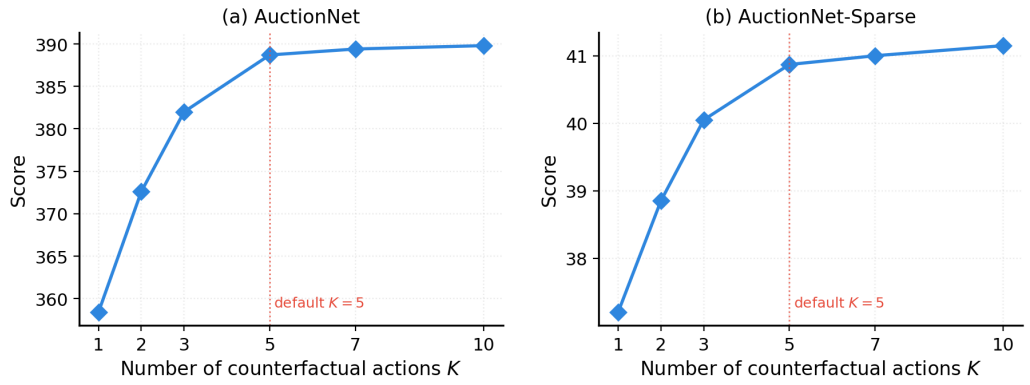


Figure 6: Sensitivity to the number of sampled counterfactual actions.

of low value and high cost—these are the ratio-violating trajectories that the frontier alone would not exclude. Panel (a) plots the resulting sampling probability $\mathbb{P}(\tau)$; probability mass concentrates on a narrow strip along the Pareto frontier, producing the desired "high-signal, compliant, non-trivial" training distribution. Compared with uniform sampling from the raw log, this reweighting reduces effective noise contamination of the training minibatch while keeping the distribution well-spread along the feasible boundary.

G Effect of the Number of Counterfactual Actions

The number of counterfactual actions K sampled per step is the main knob that controls the granularity of the CoPE regret signal. Figure 6 plots Score against K on both benchmarks. When K is very small the search neighborhood around the policy mean is too narrow to catch high-utility deviations, so the regret loss $\mathcal{L}_{\text{regret}}^{\text{CoPE}}$ is nearly degenerate and CoPE-Bid drops toward the base backbone. Increasing K initially yields fast gains: the ensemble now sees enough alternatives to identify genuinely superior counterfactuals, and—crucially—UR-CRO's LCB gating keeps the added candidates safe even when a fraction of them are OOD. Past a moderate $K \approx 5$ the curve flattens, since further samples mostly re-cover the same high-confidence region already reached, while linearly inflating per-step compute. We adopt $K{=}5$ throughout as a good speed/quality trade-off.

H Case Study

Figure 7 illustrates a single serving-time trajectory. The left panel plots the pacing coefficient λ_t alongside the remaining budget rate, and the right panel does the same for the cumulative reward. Two behaviors stand out. First, when the remaining budget rate falls faster than the schedule allows (steps 15–22), the CDPR dual-stream context transmits the tightened Cost-to-Go signal to the policy, which in response cuts λ_t sharply until the pacing curve is back on trend. Second, once the budget consumption stabilizes, the model progressively lifts λ_t again to keep chasing valuable impressions; how aggressively it does so is bounded by UR-CRO's LCB utility, so the policy does not over-shoot even though the budget headroom would allow it. The cumulative-reward panel shows the resulting behavior: reward growth accelerates during the aggressive windows and coasts during the throttled ones, closing the day near the constraint boundary rather than colliding into it. This is the empirical fingerprint of the shared-uncertainty coupling described in Sec. 3—FASI's evolving anchor gives the policy a higher ceiling to aim for while UR-CRO's pessimistic utility prevents that ambition from spilling over into constraint violations.

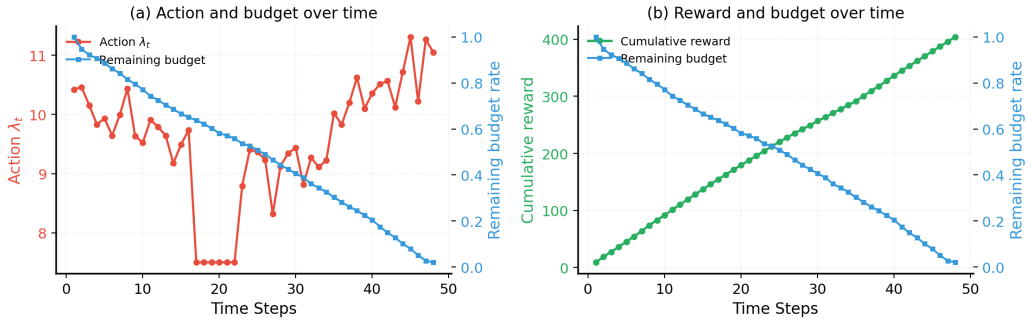


Figure 7: Serving-time pacing trajectory for one representative campaign.