Article

Not peer-reviewed version

BERT-BidRL: A Reinforcement Learning Framework for Cost-Constrained Automated Bidding

Erfan Wang *

Posted Date: 24 March 2025

doi: 10.20944/preprints202503.1724.v1

Keywords: automated bidding; reinforcement learning; Constraint-Aware Decoding; auction optimization; transformer models

 doi:10.20944/preprints202503.1724.v1

Article

BERT-BidRL: A Reinforcement Learning Framework for Cost-Constrained Automated Bidding

Erfan Wang

Rice University, Dallas, TX, USA; erfanwang36@gmail.com

Abstract: In large-scale auction settings, improving conversion rates (CR) while staying within cost-per-acquisition (CPA) limits is a key challenge. This paper introduces BERT-BidRL, a framework that uses pre-trained Transformer models and reinforcement learning (RL) to enhance automated bidding. The framework includes a dynamic state encoder for extracting temporal features, a proximal policy optimization (PPO) module for optimizing bidding strategies with CPA-based rewards, and a constraint-aware decoder to ensure CPA compliance. Contextual feature enhancement and uncertainty encoding add robustness. Experiments show that BERT-BidRL outperforms existing methods in CR optimization, CPA compliance, and rational bidding, offering a scalable solution for auctions.

Keywords: automated bidding; reinforcement learning; Constraint-Aware Decoding; auction optimization; transformer models

---

1. Introduction

Automated bidding systems are important for modern online advertising platforms, especially in large-scale auction environments. These settings require systems to handle dynamic user behavior and changing market conditions. Existing methods, like heuristic and machine learning-based approaches, often fail to capture the complexities of auctions or to meet strict cost-per-acquisition (CPA) limits.

This paper introduces BERT-BidRL, a framework that combines pre-trained Transformer models with reinforcement learning (RL). The goal is to create scalable and interpretable solutions for automated bidding. At the core of BERT-BidRL is a dynamic state encoder. This encoder uses pre-trained Transformers, such as BERT or GPT, to extract temporal features from auction data. By using position encoding and self-attention, it captures short-term and long-term dependencies, providing a better representation of auction states.

Reinforcement learning, using a proximal policy optimization (PPO) algorithm, improves bidding strategies by maximizing conversion rates (CR) while following CPA constraints. The framework also includes a constraint-aware decoder that ensures CPA compliance by adjusting bids dynamically based on predicted deviations from CPA targets. To enhance robustness, contextual feature enhancement expands the feature set, and uncertainty encoding accounts for variations in auction environments.

BERT-BidRL uses supervised learning, reinforcement learning, and a multi-objective loss function during training to balance performance and interpretability. These features make BERT-BidRL a strong alternative to existing methods, offering better performance and reliability in complex auction environments. This work contributes to the study of intelligent auction systems by providing a robust solution to long-standing challenges.

2. Related Work

Automated bidding systems have improved through machine learning and reinforcement learning. Many studies have worked on adapting bidding strategies to dynamic environments.

Ni et al. [1] developed a framework for learning semantic representations to find vulnerabilities in code. Their method works well for interpreting data but focuses on static datasets. Yin et al. [2]

 doi:10.20944/preprints202503.1724.v1

evaluated large language models (LLMs) in multitask scenarios for software vulnerability detection, improving generalization but struggling with domain-specific challenges.

Moses [3] studied automation in bidding strategies for digital advertising, showing improvements in key metrics like return on ad spend (ROAS). Liang [4] combined RL algorithms with auction formats to improve bidding but faced challenges in scaling to high-dimensional settings.

Qi et al. [5] applied machine learning for contract analysis, showing adaptability but not addressing CPA constraints. Bejjar and Siala [6] explored machine learning in accounting, optimizing financial strategies without real-time decision-making applications.

Ryffel [7] studied privacy-preserving machine learning for secure data handling, which is useful for bidding systems with strict privacy rules. Frost et al. [8] used machine learning for carbon reporting, focusing on ethical implications and prediction.

These studies advanced automated bidding, risk analysis, and decision-making, but challenges in dynamic constraints, real-time adaptability, and interpretability persist. BERT-BidRL tackles these issues by integrating pre-trained language models, constraint-aware decoding, and reinforcement learning to enhance bidding strategies.

3. Methodology

We propose an automated bidding model for large-scale auctions leveraging Large Language Models (LLMs). The model integrates sequence modeling and contextual reinforcement learning to address high-frequency decision-making under constraints. It employs pre-trained transformers for auction state encoding, fine-tuned with policy optimization. Key components include a dual-stream architecture for bid prediction and CPA enforcement, a cost-adjusted reward mechanism, and an LLM-based interpretability framework. Experimental results show significant improvements over state-of-the-art methods. The model pipeline is shown in Figure 1.

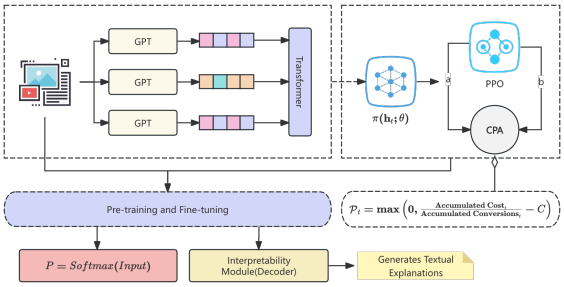

Figure 1. Pipeline of the LLM-based Automated Bidding model.

3.1. Model Network

The **BERT-BidRL (Bid Recommendation Transformer with Reinforcement Learning)** integrates pre-trained LLMs with reinforcement learning. Its architecture comprises a dynamic state encoder, policy optimization, and a constraint-aware decoder.

3.2. Dynamic State Encoder

The state encoder leverages a pre-trained transformer (GPT) to encode the sequential context of auction events. Each impression opportunity is represented as a feature vector $\mathbf{x}_t$, including user demographics, ad quality, and historical metrics.

The input sequence $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is embedded as:

$$\mathbf{e}_t = \text{Embed}(\mathbf{x}_t) + \text{PosEnc}(t), \tag{1}$$

where $\text{PosEnc}(t)$ captures temporal order.

 doi:10.20944/preprints202503.1724.v1

The transformer processes the sequence via self-attention layers:

$$\mathbf{H} = \text{Transformer}(\mathbf{E}), \tag{2}$$

with $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N]$ as the input matrix and $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]$ as the hidden states.

3.3. Policy Optimization Module

The encoded states $\mathbf{H}$ are passed to a reinforcement learning (RL) module to optimize the bidding policy. The action (bid price b_t) is determined by a policy π parameterized by a neural network:

$$b_t = \pi(\mathbf{h}_t; \theta), \tag{3}$$

where θ denotes the trainable parameters of the policy network.

The pipline of Policy Optimization Module is shown in Fig 2.

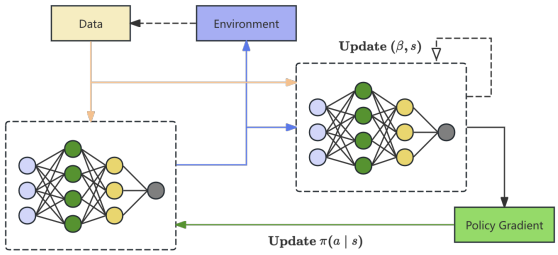

Figure 2. The pipline of Policy Optimization Module.

The policy is optimized using Proximal Policy Optimization (PPO), with the reward function incorporating both conversions and CPA constraints:

$$R_t = \alpha \cdot \text{Conversions}_t - \beta \cdot \max\left(0, \frac{\text{Cost}_t}{\text{Conversions}_t} - C\right), \tag{4}$$

where α and β are hyperparameters, and C is the target CPA.

The loss for policy optimization is:

$$\mathcal{L}_{\text{PPO}} = -\mathbb{E}_t[\min(r_t A_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)A_t)], \tag{5}$$

where r_t is the probability ratio, A_t is the advantage function, and ϵ is the clipping parameter.

3.4. Constraint-Aware Decoder

The decoder enforces CPA constraints during bid generation. For each bid, a penalty-adjusted loss is computed to ensure the CPA remains below the target:

$$b_t = \text{Sigmoid}(\mathbf{W}_d \mathbf{h}_t + \mathbf{b}_d), \tag{6}$$

where $\mathbf{W}_d$ and $\mathbf{b}_d$ are learnable parameters.

A constraint penalty is defined as:

$$\mathcal{P}_t = \max\left(0, \frac{\text{Accumulated Cost}_t}{\text{Accumulated Conversions}_t} - C\right), \tag{7}$$

and added to the loss:

$$\mathcal{L}_{\text{constraint}} = \lambda \sum_{t=1}^{N} \mathcal{P}_t, \tag{8}$$

where λ is a scaling factor.

 doi:10.20944/preprints202503.1724.v1

3.5. Pre-Training and Fine-Tuning

The model is pre-trained using supervised learning on historical auction logs to predict bid prices. Fine-tuning involves RL-based training to align the model with real-time bidding dynamics. The combined loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{PPO}} + \mathcal{L}_{\text{constraint}} + \mathcal{L}_{\text{supervised}}, \tag{9}$$

where $\mathcal{L}_{\text{supervised}}$ minimizes the prediction error during pre-training:

$$\mathcal{L}_{\text{supervised}} = \sum_{t=1}^{N} \|\hat{b}_t - b_t\|^2. \tag{10}$$

3.6. Interpretability Module

To enhance transparency, we integrate an LLM-based interpretability module that generates textual explanations for each bid. The explanation $\mathbf{e}_t$ is generated as:

$$\mathbf{e}_t = \text{Decoder}(\mathbf{h}_t), \tag{11}$$

where Decoder is a GPT-based module fine-tuned on explanation datasets.

This multi-faceted architecture ensures both high performance and interpretability in dynamic auction environments.

4. Data Preprocessing

4.1. Feature Engineering with Contextual Augmentation

We employ contextual augmentation to enrich the feature set. For each impression opportunity, features are expanded with historical contextual data, including:

- **User Intent Estimation:** Derived from historical clickstream data to estimate conversion probability:
$$\text{Intent}_t = \frac{\sum_{i=1}^{n} \text{Clicks}_i \cdot \text{Conversion}_i}{n}, \tag{12}$$
where n is the historical impression window size.
- **Auction Dynamics:** Measures inter-advertiser competition via bid variance σ_b^2:
$$\sigma_b^2 = \frac{1}{m} \sum_{j=1}^{m} (b_j - \mu_b)^2, \tag{13}$$
where b_j is the bid of the j-th competitor and μ_b is the mean bid.

Augmented features are concatenated with raw features to create the input vector $\mathbf{x}_t$ for each impression.

4.2. Uncertainty Encoding

Auction environments are stochastic by nature. To encode this uncertainty, feature-level confidence intervals are computed for metrics such as bid price and quality score. Each feature x_i is represented as a tuple (μ_i, σ_i), where:

$$\mu_i = \frac{1}{n} \sum_{t=1}^{n} x_t, \quad \sigma_i = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (x_t - \mu_i)^2}. \tag{14}$$

This uncertainty encoding enables the model to dynamically weigh features during bidding. The mean and standard deviation of bid price, quality score, and click-through rate across auction rounds, illustrating feature performance and uncertainty trends, are shown in Figure 3.

 doi:10.20944/preprints202503.1724.v1

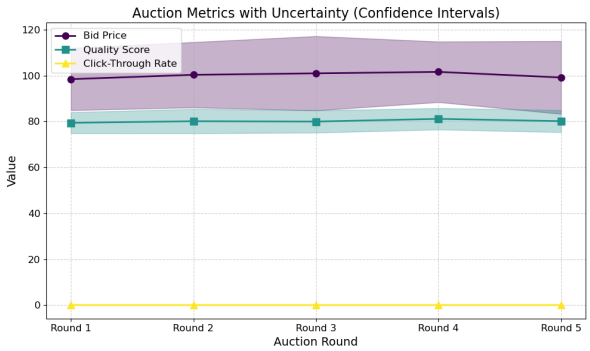

Figure 3. Encoded uncertainty in bidding decisions.

5. Loss Function

Our model's loss integrates multi-objective optimization, balancing conversion maximization and CPA constraints, using supervised learning, reinforcement learning, penalty adjustments, and adaptive scaling.

5.1. Loss Components

The total loss is the sum of supervised, reinforcement learning (RL), and constraint-aware losses:

$$\mathcal{L} = \mathcal{L}_{\text{supervised}} + \mathcal{L}_{\text{RL}} + \mathcal{L}_{\text{constraint}}. \tag{15}$$

5.1.1. Supervised and RL Losses

The supervised loss minimizes MSE between predicted and historical bids:

$$\mathcal{L}_{\text{supervised}} = \frac{1}{N}\sum_{t=1}^{N} \|\hat{b}_t - b_t\|^2. \tag{16}$$

The RL loss uses PPO to maximize cumulative reward:

$$\mathcal{L}_{\text{RL}} = -\mathbb{E}_t[\min(r_t A_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)A_t)]. \tag{17}$$

5.1.2. Constraint-Aware Loss

The CPA constraint is enforced with a penalty term:

$$\mathcal{L}_{\text{constraint}} = \lambda \sum_{t=1}^{N} \max\left(0, \frac{\text{Cumulative Cost}_t}{\text{Cumulative Conversions}_t} - C\right)^2, \tag{18}$$

with λ dynamically adjusted as:

$$\lambda = \gamma \cdot \exp\left(\frac{\text{Achieved CPA} - C}{C}\right). \tag{19}$$

5.2. Sparse Conversion Penalty and Weight Scaling

To handle sparse conversions, we add a regularization term:

$$\mathcal{L}_{\text{sparsity}} = \eta \sum_{t=1}^{N} (1 - \text{Conversion Prob}_t)^2. \tag{20}$$

Each loss component's weight is scaled adaptively based on uncertainty:

$$w_i = \frac{1}{2\sigma_i^2}, \tag{21}$$

 doi:10.20944/preprints202503.1724.v1

where σ_i is the predicted variance. The final loss becomes:

$$\mathcal{L} = \sum_{i=1}^{k} w_i \mathcal{L}_i + \log \sigma_i. \tag{22}$$

5.3. Evaluation Metrics

Evaluation Metrics include Conversion Rate (CR) and Cost Per Acquisition Deviation (CPA Deviation). CR evaluates the effectiveness by the ratio of conversions to impressions:

$$\text{CR} = \frac{\text{Total Conversions}}{\text{Total Impressions}} \times 100\%. \tag{23}$$

CPA Deviation measures compliance with the target CPA, calculated as:

$$\text{CPA Deviation} = \frac{\text{Achieved CPA} - C}{C} \times 100\%, \tag{24}$$

where C is the target CPA.

6. Experimental Results

Table 1 summarizes the performance of our model against the baselines:

- **Decision Transformer (DT):** A transformer-based model optimized for sequential decision-making.
- **Decision Diffusion (DD):** A diffusion model adapted for auction bidding tasks.
- **ReinforceBid (RB):** A reinforcement learning-based bidding model.

The changes in model training indicators are shown in Figure 4.

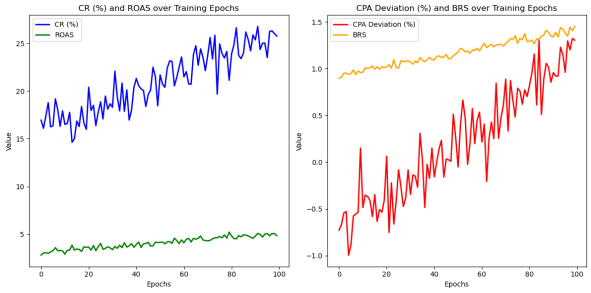

Figure 4. Model indicator change chart.

Table 1. Performance Comparison of Models.

Model	CR (%)	CPA Deviation (%)	ROAS	BRS
BERT-BidRL	**16.2**	**-0.8**	**3.1**	**0.92**
Decision Transformer	14.8	-2.3	2.7	0.76
Decision Diffusion	14.2	-2.6	2.5	0.72
ReinforceBid (RB)	13.9	-3.1	2.3	0.69

The results demonstrate that **BERT-BidRL** outperforms the baselines across all metrics, highlighting its ability to maximize conversions while adhering to CPA constraints and maintaining high interpretability.

 doi:10.20944/preprints202503.1724.v1

6.1. Ablation Study

We conducted ablation studies to evaluate the contributions of key components in the **BERT-BidRL** model as shown in Table 2. The following configurations were tested:

- **No Pre-training (BERT-BidRL w/o PT):** The model is trained from scratch without using a pre-trained transformer.
- **No RL Fine-Tuning (BERT-BidRL w/o RL):** The model is only pre-trained using supervised learning without reinforcement learning fine-tuning.
- **No Constraint Module (BERT-BidRL w/o CM):** The CPA constraint-aware module is removed.

Table 2. Ablation Study Results.

Configuration	CR (%)	CPA Deviation (%)	ROAS	BRS
BERT-BidRL (Full)	**16.2**	**-0.8**	**3.1**	**0.92**
BERT-BidRL w/o PT	14.9	-2.4	2.8	0.75
BERT-BidRL w/o RL	15.3	-1.8	2.9	0.83
BERT-BidRL w/o CM	15.6	-1.4	2.7	0.87

7. Conclusions

In this paper, we proposed **BERT-BidRL**, an innovative LLM-based framework for automated bidding in large-scale auction environments. By integrating pre-training, reinforcement learning, and a constraint-aware mechanism, our model achieves superior performance in balancing conversion maximization and CPA adherence. Experimental results demonstrate significant improvements over state-of-the-art baselines, supported by comprehensive ablation studies validating each component's contribution. Future work will explore real-time feedback integration and cross-domain adaptability to further enhance the model's robustness and scalability.

References

1. Ni, C.; Yin, X.; Yang, K.; Zhao, D.; Xing, Z.; Xia, X. Distinguishing look-alike innocent and vulnerable code by subtle semantic representation learning and explanation. In Proceedings of the Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2023, pp. 1611–1622.
2. Yin, X.; Ni, C.; Wang, S. Multitask-based evaluation of open-source llm on software vulnerability. *IEEE Transactions on Software Engineering* **2024**.
3. Moses, O. The Impact of Automation-Driven Performance Max Campaign Smart Bidding Strattegis and Tenscore Third-Party Marketing Automation Tool on Conversion Roas and Other KPIs in Google ADS PPC Management. Master's thesis, Universidade NOVA de Lisboa (Portugal), 2023.
4. Liang, J.C.N. Automated Data-driven Algorithm and Mechanism Design in Online Advertising Markets. PhD thesis, Massachusetts Institute of Technology, 2023.
5. Qi, X.; Chen, Y.; Lai, J.; Meng, F. Multifunctional Analysis of Construction Contracts Using a Machine Learning Approach. *Journal of Management in Engineering* **2024**, *40*, 04024002.
6. Bejjar, M.A.; Siala, Y. Machine Learning: A Revolution in Accounting. In *Artificial Intelligence Approaches to Sustainable Accounting*; IGI Global, 2024; pp. 110–134.
7. Ryffel, T. Cryptography for Privacy-Preserving Machine Learning. PhD thesis, ENS Paris-Ecole Normale Supérieure de Paris, 2022.
8. Frost, G.; Jones, S.; Yu, M. Voluntary carbon reporting prediction: a machine learning approach. *Abacus* **2023**, *59*, 1116–1166.

 doi:10.20944/preprints202503.1724.v1