Beyond Single Slot: Joint Optimization for Multi-Slot Guaranteed Display Advertising

Zhaoqi Zhang
Nanyang Technological University
Singapore, Singapore
Meituan
Beijing, China
zhaoqi001@e.ntu.edu.sg

Jiaming Deng
Meituan
Beijing, China
dengjiaming02@meituan.com

Miao Xie
China Agricultural University
Beijing, China
xiemiao@cau.edu.cn

Linyou Cai
Meituan
Beijing, China
cailinyou@meituan.com

Qianlong Xie
Meituan
Beijing, China
xieqianlong@meituan.com

Xingxing Wang
Meituan
Beijing, China
wangxingxing04@meituan.com

Siqiang Luo
Nanyang Technological University
Singapore, Singapore
siqiang.luo@ntu.edu.sg

Gao Cong
Nanyang Technological University
Singapore, Singapore
gaocong@ntu.edu.sg

Abstract

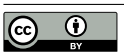
Guaranteed display advertising is crucial for platform monetization, yet existing methods often operate under a single-slot assumption, limiting their ability to optimize allocation across multi-slot page views. In this paper, we propose a novel joint optimization framework for multi-slot GD allocation, addressing key challenges such as slot-level redundancy, contract imbalance, and exposure concentration. Our approach formulates the allocation as an offline bipartite matching problem with a contract roulette mechanism for slot exclusivity and Page View constraints for impression control, and incorporates a scalable allocation optimization algorithm for efficient large-scale deployment. Extensive online tests on the Meituan advertising platform demonstrate that our method significantly improves merchant ROI, platform revenue efficiency, and contract fulfillment robustness. Specifically, online A/B tests show a **28.99%** increase in Average Revenue Per User under 70% traffic, and DID analysis further indicates improved contract stability, demonstrating the strong applicability and effectiveness of our framework in real-world advertising deployments.

CCS Concepts

• **Information systems → Computational advertising**;

Keywords

Guaranteed Display Advertising, Constrained Optimization, Contract Roulette



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808398

ACM Reference Format:
Zhaoqi Zhang, Jiaming Deng, Miao Xie, Linyou Cai, Qianlong Xie, Xingxing Wang, Siqiang Luo, and Gao Cong. 2026. Beyond Single Slot: Joint Optimization for Multi-Slot Guaranteed Display Advertising. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3805712.3808398

1 Introduction

In guaranteed display (GD) advertising, mainstream approaches are predominantly built upon a single-slot modeling assumption, where each ad contract is independently optimized for a specific slot to maximize engagement metrics such as Click-Through Rate(CTR) and Payment Conversion Rate(CVR) [6–8]. While effective in early-stage settings with simpler ad placements, such approaches are increasingly inadequate for modern platforms such as Meituan and Taobao, where a single page may contain multiple ad slots, each page view can trigger simultaneous exposures across several slots, rendering independent slot-wise optimization insufficient for capturing complex allocation dynamics.

Although recent works [10, 11] introduce coarse supply-side constraints to prevent over-delivery, they lack fine-grained slot-level control, allowing a few high-priority contracts to dominate premium positions. Moreover, most systems rely on online greedy allocation [4, 9], which makes slot-wise decisions in isolation. Consequently, current methods remain insufficient to ensure fairness, exposure diversity, and stable delivery across multiple slots.

Despite their effectiveness in improving fulfillment and click performance, mainstream GD methods such as AUAF [3] exhibit critical limitations in real-world multi-slot environments: **(1) Lack of coordination across ad slots:** Most existing methods adopt a local modeling approach based on a single-slot view, where each optimization step considers only the allocation between one ad slot and several contracts. This localized strategy may lead to the overuse of popular slots and the underuse of others, reducing overall

delivery efficiency. **(2) No upper limit on contract impressions:** While most methods ensure minimum delivery for each contract, they do not restrict the maximum number of impressions a contract can receive. As a result, high-priority contracts may monopolize premium exposure, leading to unfair and imbalanced allocations. **(3) Redundant exposures on the same page:** Request-level exclusivity fails to prevent the same contract from appearing in multiple slots within a single page, causing duplicate impressions and degraded user experience.

To address the aforementioned limitations, we propose an offline bipartite matching-based joint optimization framework tailored for multi-slot GD allocation. Our method fundamentally moves beyond the conventional single-slot paradigm by modeling the allocation between ad contracts and multiple ad slots at the page view level through a global offline optimization process, enabling coordinated decisions that enhance balance, fairness, and diversity beyond online greedy approaches. We further introduce Page View (PV) constraints to cap per-slot impressions, preventing head-slot overuse and promoting balanced traffic allocation. In addition, we incorporate a Contract Roulette–based exclusivity mechanism that ensures that each contract appears in at most one slot per page, reducing redundant exposures and improving user experience. Our key contributions are as follows:

- We propose a unified joint optimization framework that formulates the contract-slot allocation problem as an offline bipartite matching task at the page view level, achieving globally coordinated and efficient fulfillment.
- We introduce two practical modules for industrial deployment: (i) a PV constraints for fine-grained traffic balance, and (ii) a Contract Roulette-Based Selection Mechanism for one-to-many assignment conflicts and reducing redundant exposures.
- Our proposed framework has been deployed in the GD advertising system of Meituan. Extensive online A/B testing validates its effectiveness and efficiency, where the Average Revenue Per User (ARPU) increased by **28.17%**, and the Fulfillment Rate improved by **2.12%** compared with the previous production baseline.

2 Related Work

Research on GD Advertising has evolved significantly across both modeling objectives and solution strategies. Early approaches focused on fulfilling contractual guarantees through offline optimization. SHALE [1] formulates GD advertising as a scalable quadratic program to meet impression delivery targets, while the dual-based method [2] further considers advertiser-side utility, bridging guarantee fulfillment and business effectiveness. Nearline control strategies were introduced to align delivery with real-time platform dynamics better. XShale [6] adjusts delivery pacing based on short-term feedback, optimizing advertiser outcomes. RAP [13] further incorporates platform-level concerns such as traffic efficiency and delivery fairness, marking an early attempt toward multi-objective coordination in nearline settings. Recent work has formulated GDA as a sequential decision problem under complex constraints. AUAF [3] introduces slot-level mutual exclusivity in an optimal control framework, improving contract delivery precision. CONFLUX [12] and FACC [5] jointly optimize delivery, advertiser utility, and platform-wide objectives using multi-stage control mechanisms.

Despite their effectiveness, existing methods lack fine-grained page-view-level traffic control and explicit modeling of multi-slot coordination, limiting their ability to prevent overexposure and ensure fair delivery across diverse ad positions.

3 Methodology

We formulate the multi-slot ad allocation problem as an offline bipartite matching task between ad requests and contracts, where each page view spans multiple slot positions. Our proposed framework, as illustrated in Figure 1, consists of two key components: (1) Page View-Constrained Allocation Module, which computes globally coordinated allocation probabilities while enforcing a novel Page View constraint to limit per-slot exposure and balance traffic across positions; and (2) Contract Roulette-Based Selection Module, which resolves one-to-many matching conflicts by probabilistically selecting a winning contract per request.

3.1 Page View-Constrained Allocation Module

To coordinate ad delivery across multiple slots within the same page view, we formulate the allocation task as a constrained optimization problem. This formulation captures not only contract fulfillment and user engagement but also fine-grained control over slot-level exposure. In particular, we introduce a novel Page View constraint to limit the number of impressions each slot can serve, thereby preventing over-exposure of high-traffic positions and promoting a more balanced traffic allocation across all ad slots. The detailed objective function and constraints are defined as follows, which consists of three parts. The first term is a smoothness regularization term that penalizes the deviation between the actual allocation x_{ij} and the normalized target delivery ratio θ_j, which helps stabilize contract delivery and improve allocation fairness. The second term is a priority-aware reward term that encourages the system to allocate more impressions to higher-priority contracts. The third term is an interest-aware matching term that promotes assignments with higher estimated user-interest relevance.

$$\arg\min_{x_{ij}} \quad \frac{1}{2}\sum_j \sum_{i\in\Gamma(j)} s_i \frac{V_j}{\theta_j}(x_{ij}-\theta_j)^2 - \sum_j w_j \sum_{i\in\Gamma(j)} s_i x_{ij} \quad (1)$$
$$- \sum_j \lambda_j \sum_{i\in\Gamma(j)} s_i x_{ij} c_{ij}$$
$$\text{s.t.} \quad \sum_{i\in\Gamma(j)} s_i x_{ij} \le d_j, \quad \forall j \quad (2)$$
$$\sum_{j\in\Gamma(i)} x_{ij} \le 1, \quad \forall i \quad (3)$$
$$x_{ij} \ge 0, \quad \forall i,j \quad (4)$$
$$s_i x_{ij} \le pv_i, \quad \forall i,j \quad (5)$$

where x_{ij} denotes the allocation probability from request i (i.e., supply) to contract j (i.e., demand), and c_{ij} represents the user interest between them; s_i denotes the capacity of request i, and d_j represents the required number of impressions for contract j; λ_j reflects the importance of user interest for demand j. Distinct from prior work, we introduce pv_i to denote the page-view constraint of request i, enabling explicit control of slot-level exposure and balanced traffic allocation, which constitutes a key novelty of our formulation. Let $\Gamma(j)$ denote the set of requests that can serve

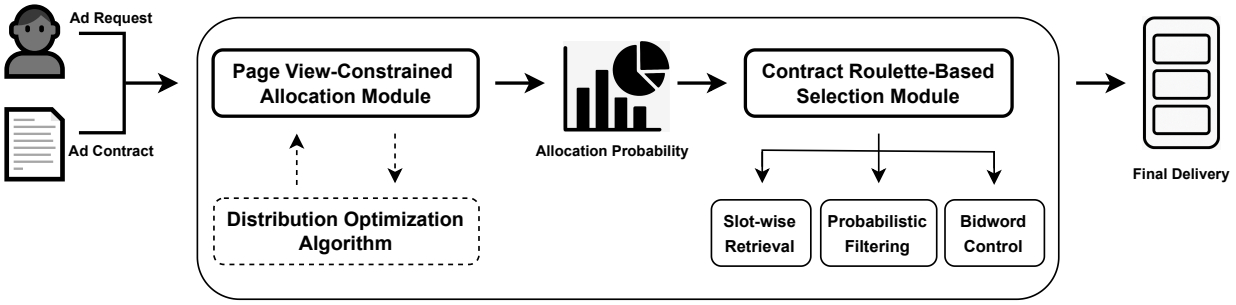


Figure 1: Overview of proposed framework.

contract j, and let $\Gamma(i)$ denote the set of contracts that request i is eligible to serve.

$$\begin{aligned} L(\alpha,\beta,\gamma,\delta) =& \frac{1}{2}\sum_j \sum_{i\in\Gamma(j)} s_i \frac{V_j}{\theta_j}(x_{ij}-\theta_j)^2 - \sum_j w_j \sum_{i\in\Gamma(j)} s_i x_{ij} \\ & - \sum_j \lambda_j \sum_{i\in\Gamma(j)} s_i x_{ij} c_{ij} + \sum_j \alpha_j \left(\sum_{i\in\Gamma(j)} s_i x_{ij} - d_j \right) \\ & + \sum_i \beta_i \left(\sum_{j\in\Gamma(i)} x_{ij} - 1 \right) - \sum_j \sum_{i\in\Gamma(j)} \gamma_{ij} x_{ij} \\ & + \sum_j \sum_{i\in\Gamma(j)} \delta_{ij} \left(s_i x_{ij} - p v_i \right) \end{aligned} \tag{6}$$

KKT conditions:

$$s_i \frac{V_j}{\theta_j}(x_{ij}-\theta_j) - w_j s_i - \lambda_j s_i c_{ij} + \alpha_j s_i + \beta_i s_i - \gamma_{ij} + \delta_{ij} s_i = 0 \tag{7}$$

$$\alpha_j \left(\sum_{i\in\Gamma(j)} s_i x_{ij} - d_j \right) = 0 \tag{8}$$

$$\beta_i \left(\sum_{j\in\Gamma(i)} x_{ij} - 1 \right) = 0 \tag{9}$$

$$\gamma_{ij} x_{ij} = 0 \tag{10}$$

$$\delta_{ij} \left(s_i x_{ij} - p v_i \right) = 0 \tag{11}$$

$$\alpha_j \geq 0, \quad \beta_i \geq 0, \quad \gamma_{ij} \geq 0, \quad \delta_{ij} \geq 0 \tag{12}$$

where α_j, β_i, γ_{ij} and δ_{ij} are the Lagrangian multipliers of constraints Equation 8, 9, 10 and 11 respectively. According to the KKT conditions, the optimal allocation probability x_{ij} can be derived as:

$$x_{ij} = \max\left\{0,\ \theta_j \left(1 + \frac{w_j + \lambda_j c_{ij} - \alpha_j - \beta_i - \delta_{ij}}{V_j}\right)\right\} \tag{13}$$

To efficiently solve the PV-constrained allocation problem, we adopt an online distributed strategy that updates the dual variables α_j, β_i, and δ_{ij} based on the KKT conditions. These dual variables correspond to the contract demand, request capacity, and slot-level page view constraints, respectively. Instead of solving the primal problem directly, which is computationally expensive in large-scale settings, we leverage the closed-form expression of x_{ij} derived from the KKT optimality condition and iteratively adjust the dual variables using projected gradient steps.

At each iteration, α_j is increased if the total allocated impressions to contract j exceed its demand d_j, while β_i is updated to ensure the total allocation from request i remains within its capacity. Similarly, δ_{ij} is adjusted to enforce the fine-grained page view constraint for each ad slot. Each contract j is characterized by a priority weight w_j, a smoothness parameter V_j, and a normalized target delivery ratio $\theta_j = \frac{d_j}{\sum_{i\in\Gamma(j)} s_i}$, where V_j controls the strength of the fairness regularization for the j-th contract. The gradient of the objective with respect to x_{ij} is used to guide the iterative updates. These operations are executed in parallel across all (i, j) pairs, ensuring scalability and real-time adaptability. The overall procedure is summarized in Algorithm 1.

Algorithm 1: PV Constrained Allocation Algorithm

Input: $s_i, d_j, \lambda_j, V_j, w_j, c_{ij}$
Output: $\alpha_j, \beta_i, \delta_{ij}, x_{ij}$

Step 1: Initialize. Set $\alpha_j = w_j + \lambda_j c_{ij}, \theta_j$, calculate x_{ij}, and gradient $grad_j$
Step 2:
for *iteration = 1 to n* **do**
  update α_j with Equation: $\alpha_j^{t+1} = \alpha_j^t - V_j(1 - \frac{d_j(\alpha^t)}{d_j})$
  solve β_i: $\beta_i \leftarrow \max\left(0, \beta_i + \eta_\beta \left(\sum_{j\in\Gamma(i)} x_{ij} - 1\right)\right)$
  calculate δ_{ij}: $\delta_{ij} \leftarrow \max\left(0, \delta_{ij} + \eta_\delta \left(s_i x_{ij} - p v_i\right)\right)$
  calculate x_{ij}: $x_{ij} = \max\left\{0,\ \theta_j \left(1 + \frac{w_j + \lambda_j c_{ij} - \alpha_j - \beta_i - \delta_{ij}}{V_j}\right)\right\}$
  update $grad_j$: $grad_j = \sum_{i\in\Gamma(j)} s_i x_{ij} - d_j$
end

3.2 Contract Roulette-Based Selection Module

To ensure that each contract advertisement appears at most once per page and to maximize delivery diversity across ad slots, we design a contract roulette-based selection mechanism that integrates efficient candidate generation, probabilistic filtering, and adaptive bidword control under online latency constraints.

3.2.1 Slot-wise Retrieval. In the recall phase, each ad slot s_i independently retrieves a set of candidate contracts $\mathcal{A}_{s_i}$ based on the current query q and an associated bidword b_i. To maintain fairness and long-tail exposure, a roulette sampling strategy is applied. For each candidate contract $a \in \mathcal{A}_{s_i}$, the probability of being recalled is proportional to its delivery weight $w(a)$:

$$P(a \mid s_i) = \frac{w(a)}{\sum_{a' \in \mathcal{A}_{s_i}} w(a')} \tag{14}$$

This design prioritizes contracts with higher delivery urgency while remaining compatible with budget delivery requirements and real-time demand-supply conditions. It also enforces that each POI-level contract can be recalled for only one slot per page, thus preventing redundant exposures and ensuring slot-level diversity.

3.2.2 Probabilistic Filtering. Once candidates are recalled, a multi-stage filtering process is applied. First, contracts are scored within each slot using a utility function that reflects both click-through rate and post-click conversion rate:

$$\text{Score}(a) = \text{CTR}_a \cdot \text{CVR}_a \tag{15}$$

Candidates with $\text{Score}(a) \geq K_1$ are retained, where K_1 is a system-defined threshold. For contracts appearing in multiple slots, only the one with the best relative position is preserved:

$$s^*(a) = \arg\min_{s_i} \text{Rank}_{s_i}(a) \tag{16}$$

where $\text{Rank}_{s_i}(a)$ denotes the intra-slot rank of contract a in slot s_i. This ensures that each contract is bound to at most one slot, prioritizing the position where it has the strongest competitive advantage.

3.2.3 Adaptive Bidword Control. To support dynamic retrieval across slots, we design an adaptive bidword selection strategy that ensures each slot retrieves relevant candidates while promoting diversity and contextual consistency across the page. For each slot s_i, the selected bidword b_i is determined as:

$$b_i = \begin{cases} \text{RandomSample}(\mathcal{B}_q), & \text{if } i = 1 \\ \text{RandomSample}(\mathcal{B}_q^{\text{avail}}(i)), & \text{if } i > 1 \text{ and } \mathcal{A}_{s_i}(b_{i-1}) = \emptyset \\ b_{i-1}, & \text{if } i > 1 \text{ and } \mathcal{A}_{s_i}(b_{i-1}) \neq \emptyset \end{cases} \tag{17}$$

where $\mathcal{B}_q$ denotes the set of candidate bidwords for query q, and $\mathcal{A}_{s_i}(b)$ is the set of ads retrievable for slot s_i using bidword b. The available bidwords for slot s_i are defined as:

$$\mathcal{B}_q^{\text{avail}}(i) = \left\{ b \in \mathcal{B}_q \mid \mathcal{A}_{s_i}(b) \neq \emptyset \right\} \setminus \{b_1, \ldots, b_{i-1}\} \tag{18}$$

When determining the bidword for the first slot, the system randomly samples from the full bidword pool $\mathcal{B}_q$. Subsequent slots preferentially reuse the previous bidword b_{i-1} if it can still retrieve valid ads; otherwise, the system switches to another eligible bidword from the remaining available pool. In practice, this fallback process is also guided by inventory availability and delivery deficit, especially when certain bidwords are highly supply-constrained. This design ensures relevant ads for each slot while reducing redundancy and maintaining page-level coherence.

4 Experiments

To evaluate the real-world effectiveness of our proposed multi-slot allocation framework, we conducted two rounds of online gray-scale experiments on the Meituan advertising platform, where our method was fully deployed and actively served real traffic. Then we discuss experimental results.

4.1 Experimental Settings

The experiments were conducted on Meituan's production environment with real online traffic and real-time ad delivery, under 35% and 70% gray-scale settings, where the treatment and control groups were split by corresponding proportions of POIs. Both experiments were evaluated using a combination of the A/A test to validate group-level stability, the A/B test for one-day cross-group comparison, and the DID (Difference-in-Differences) analysis to estimate the net causal effect while controlling for temporal fluctuations. The online experiment followed a progressive gray-scale rollout. For the 35% gray-scale setting, the baseline period was from March 29 to April 2, 2025, and the experimental period was from April 3 to April 7, 2025. For the 70% gray-scale setting, the baseline period was from March 27 to April 1, 2025, and the experimental period was from April 9 to April 14, 2025.

4.2 Evaluation Metrics

To assess the effectiveness of our proposed method, we adopt a comprehensive set of evaluation metrics, grouped into three key dimensions: *Merchant Efficiency*, *Platform Revenue*, and *Contract Fulfillment* to capture merchant-side benefits and platform-level business objectives jointly:

- **Merchant Return on Investment (ROI)**: Measures the return on investment from the merchant's perspective, computed as revenue generated through advertising divided by ad spend.
- **Payment Return on Investment (ROI)**: Focuses on actual purchase behaviour, providing a more direct indicator of commercial effectiveness.
- **Click-Through Rate (CTR)**: Captures user engagement with advertisements, indicating exposure quality.
- **Payment Conversion Rate (CVR)**: The proportion of clicks that result in a completed payment, reflecting the efficacy of the ad content and targeting.
- **ARPU Average Revenue Per User (ARPU)**: Reflects monetization efficiency per user, capturing per-capita revenue yield.
- **Fulfillment Rate**: The ratio of fulfilled to planned contractual objectives to evaluate service quality and delivery reliability.

4.3 Performance Analysis

We conduct controlled A/A and A/B tests under two gray-scale settings (35% and 70%) to rigorously evaluate the effectiveness and robustness of our method. As shown in Tables 1, our method consistently outperforms across key dimensions, merchant efficiency, platform monetization, and contractual delivery quality.

Our proposed method yields substantial gains in merchant-side metrics. In particular, the DID (Difference-in-Differences) results demonstrate a 42.17% improvement in Merchant ROI, 29.13% in Payment ROI, 7.67% in CTR, and 23.35% in Payment CVR, indicating enhanced user engagement and stronger conversion performance

Table 1: 35% and 70% Gray-Scale Experiment Results

Gray Scale	Experiment	Merchant Efficiency				Platform Revenue	Contract Fulfillment
		Merchant ROI	Payment ROI	CTR	Payment CVR	ARPU	Fulfillment Rate
35%	A/A Effect (Treatment)	-33.98%	79.58%	17.78%	25.62%	0.71%	0.22%
	A/A Effect (Control)	-38.63%	38.52%	0.58%	-5.85%	31.48%	1.32%
	A/B Effect	-31.73%	54.96%	26.33%	32.23%	-5.95%	-10.39%
	DID Effect	4.65%	41.06%	17.20%	31.47%	-30.77%	-11.50%
70%	A/A Effect (Treatment)	25.06%	-4.75%	12.10%	13.35%	49.87%	-2.90%
	A/A Effect (Control)	-17.11%	-33.87%	4.42%	-10.01%	21.70%	-5.03%
	A/B Effect	-41.82%	64.76%	0.99%	4.57%	28.99%	-3.95%
	DID Effect	42.17%	29.13%	7.67%	23.35%	28.17%	2.12%

throughout the advertising funnel. These improvements highlight the method's effectiveness in not only attracting user attention but also driving purchase behaviours.

On the platform side, ARPU increases by 28.17%, suggesting better monetization efficiency per user under the multi-slot optimization strategy. While some volatility may still exist in secondary metrics, the overall ARPU trend indicates improved budget utilization and delivery prioritization.

In terms of contract fulfillment, the Fulfillment Rate shows a positive DID effect of 2.12%, suggesting that the proposed allocation framework maintains stable and reliable contract delivery even under expanded exposure. This further confirms the robustness of the system under scaled deployment.

In summary, the experimental results demonstrate that our approach consistently improves merchant return, user interaction quality, and revenue efficiency while maintaining contract integrity and scalability across different traffic levels, demonstrating strong practical value in real-world advertising deployments.

5 Conclusion

This paper presents a unified framework for guaranteed display advertising in multi-slot environments. By modeling the contract-slot assignment as a page-view-level bipartite matching problem, and introducing novel constraints and selection mechanisms, our approach achieves fine-grained traffic control, eliminates redundant exposures, and enhances fairness in ad delivery. The proposed Page View-constrained allocation model enables precise regulation of slot-level traffic, while the contract roulette mechanism ensures diversity and mutual exclusivity. Extensive online experiments on Meituan validate the effectiveness of our framework across key metrics. These results underscore the practical value of coordinated multi-slot optimization in industrial GD systems and provide a foundation for future research on fairness-aware and scalable ad delivery mechanisms.

References

[1] Vijay Bharadwaj, Peiji Chen, Wenjing Ma, Chandrashekhar Nagarajan, John Tomlin, Sergei Vassilvitskii, Erik Vee, and Jian Yang. 2012. Shale: an efficient algorithm for allocation of guaranteed display advertising. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1195–1203.

[2] Ye Chen, Pavel Berkhin, Bo Anderson, and Nikhil R Devanur. 2011. 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*. 1307–1315.

[3] Xiao Cheng, Chuanren Liu, Liang Dai, Peng Zhang, Zhen Fang, and Zhonglin Zu. 2022. An adaptive unified allocation framework for guaranteed display advertising. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 132–140.

[4] Liang Dai, Kejie Lyu, Chengcheng Zhang, Guangming Zhao, Zhonglin Zu, Liang Wang, and Bo Zheng. 2024. Percentile risk-constrained budget pacing for guaranteed display advertising in online optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7987–7994.

[5] Liang Dai, Zhonglin Zu, Hao Wu, Liang Wang, and Bo Zheng. 2023. Fairness-aware guaranteed display advertising allocation under traffic cost constraint. In *Proceedings of the ACM Web Conference 2023*. 3572–3580.

[6] Zhen Fang, Yang Li, Chuanren Liu, Wenxiang Zhu, Yu Zheng, and Wenjun Zhou. 2019. Large-scale personalized delivery for guaranteed display advertising with real-time pacing. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 190–199.

[7] Ali Hojjat, John Turner, Suleyman Cetintas, and Jian Yang. 2014. Delivering guaranteed display ads under reach and frequency requirements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

[8] Hang Lei, Yin Zhao, and Longjun Cai. 2020. Multi-objective optimization for guaranteed delivery in video service platform. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3017–3025.

[9] Yu Lei, Jiayang Zhao, Yilei Zhao, Zhaoqi Zhang, Linyou Cai, Qianlong Xie, and Xingxing Wang. 2025. Generative Large-Scale Pre-trained Models for Automated Ad Bidding Optimization. *arXiv preprint arXiv:2508.02002* (2025).

[10] Yan Li, Yundu Huang, Wuyang Mao, Furong Ye, Xiang He, Zhonglin Zu, and Shaowei Cai. 2024. Bi-Objective Contract Allocation for Guaranteed Delivery Advertising. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1691–1700.

[11] Wuyang Mao, Chuanren Liu, Yundu Huang, Zhonglin Zu, M Harshvardhan, Liang Wang, and Bo Zheng. 2023. End-to-End Inventory Prediction and Contract Allocation for Guaranteed Delivery Advertising. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1677–1686.

[12] XiaoYu Wang, Bin Tan, Yonghui Guo, Tao Yang, Dongbo Huang, Lan Xu, Nikolaos M Freris, Hao Zhou, and Xiang-Yang Li. 2022. CONFLUX: A Request-level Fusion Framework for Impression Allocation via Cascade Distillation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4070–4078.

[13] Hong Zhang, Lan Zhang, Lan Xu, Xiaoyang Ma, Zhengtao Wu, Cong Tang, Wei Xu, and Yiguo Yang. 2020. A request-level guaranteed delivery advertising planning: Forecasting and allocation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2980–2988.