Bidding for Influence: Auction-Driven Diffusion Image Generation

Anonymous Authors[1]

Abstract

We introduce the first generative image auction, utilizing diffusion models to express multi-agent preferences. Our mechanism first composes advertisers' bids and text prompts inside the diffusion model's reverse process through a dynamic score composition. Then, it implements Monte Carlo sampling to calculate VCG-based payments and select high-welfare images. Extensive experiments on a two-agent benchmark demonstrate three core properties: (1) bid monotonicity, (2) efficiency improvement of up to 21% higher welfare than a single-winner VCG baseline, and (3) approximate incentive compatibility, with average regret achieving below 10% when deviating from truthful bidding. Crucially, these benefits are achieved while preserving high image quality. Our study establishes a principled bridge between auction theory and controllable image diffusion, laying a foundation for economically-aligned, multi-stakeholder image generation in advertising and beyond.

1. Introduction

Auctions are widely used for resource allocation, aiming to achieve objectives like welfare or revenue maximization through carefully designed allocation and price rules (Vickrey, 1961). Online advertising auctions typically allocate discrete slots to single winners. However, generative AI, particularly for image synthesis, enables new auction designs where outcomes are not restricted to single items and multiple entities can influence a single generated output. This potential has been explored for Large Language Models (LLMs) (Dütting et al., 2024; Hajiaghayi et al., 2024; Feizi et al., 2024), motivating us to consider auctions where (i) the outcome is generated content, (ii) bidders express preferences over this content, and (iii) the auction decides winners, preference aggregation, and payments.

Our focus is on auctions using Denoising Diffusion Probabilistic Models (DDPMs) (Song et al., 2021b), known for high-quality image synthesis (Nichol and Dhariwal, 2021). Given that diffusion model outputs can be steered to incorporate specific concepts (Ho and Salimans, 2022), we explore multi-preference image generation within an auction framework. This leads to our central question: How can we design an effective, generative image auction enabling multiple advertisers to bid for influencing the generated image's content?

Our main contributions are:

- **Pioneering Generative Image Auctions:** We introduce generative image auctions and propose a mechanism based on score guidance and the Vickrey-Clarke-Groves (VCG) mechanism. Bidder value is their private value multiplied by prompt-image alignment (e.g., CLIP-based).
- **Expressive Score Composition:** We propose a bid-dependent score composition for nuanced control during reverse diffusion.
- **VCG-Inspired Allocation & Pricing:** We use Monte Carlo sampling with the composed scores to select a high welfare-maximizing image (per bids) and to compute VCG-inspired payments.
- **Desirable Auction Qualities:** We demonstrate bid monotonicity, welfare improvement over a single-winner baseline, approximate truthfulness (incentive compatibility), and preservation of image quality.

This work establishes a foundational framework for generative image auctions, aiming to spur advancements in controllable, economically-efficient diffusion models.

2. Methodology

Our proposed generative image auction aims to aggregate preferences from multiple agents, each defined by a text prompt (c_i) and a private value (m_i), into a single synthesized image. An agent's valuation of an image I is $m_i \cdot \alpha_i(I)$, where $\alpha_i(I)$ is the alignment (e.g., CLIP similarity) between their prompt and the image. The auction conditions on a base prompt c and agent inputs (prompts and bids $\mathbf{b}$). To

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

combine preferences, we introduce a novel score composition technique within the diffusion model's reverse process, weighting agent influence by their bids. This composed score guides the generation of k candidate images. Inspired by VCG, the image maximizing aggregate welfare (based on submitted bids) is selected. Payments are also VCG-inspired, approximating the externality an agent imposes on others, calculated using Monte Carlo estimates from rerunning the auction without that agent. The detailed methodology, including the agent model, social welfare objective, and specific formulations for the aggregation and payment functions, is provided in Appendix D.

3. Results

3.1. Experimental Setup

To evaluate how effectively our auction blends multiple advertiser concepts within a single generated image, we construct a controlled test set of five base prompts. Each base prompt is paired with two advertiser prompts, along with a joint prompt that combines both agents' preferences.

This setup allows us to systematically vary each agent's bid weight while holding prompts constant, allowing a full range of score compositions to generate a set of final images. We vary Agent 1's bid weight from 0.0 to 1.0 in increments of 0.1, and Agent 2's bid weight from 0.0 to 1.0 in increments of 0.25. We conduct Monte Carlo sampling $k \in \{25, 50\}$ images per bidding combination, resulting in a total of 7,825 generated images across all experiments. All images are generated using the `FLUX.1-schnell` diffusion model with 5 inference steps and a guidance scale of 10. We use CLIP (Radford et al., 2021) to compute alignment $\alpha_i = \text{CLIP}(I, c_i)$ as the cosine similarity between the CLIP image embedding and text embedding, which lie in a shared embedding space. Prompt texts are detailed in Appendix E.

3.2. Bid Monotonicity: Prompt Alignment Increases with Bid

One auction trait that we aim to show is bid monotonicity, in which agents who bid more should receive higher valuations. Then, an advertiser's prompt should be more prominent in the generated image as their bid increases. To evaluate this, we increase agent 1's normalized bid values, which decreases agent 2's normalized bids, across all five prompts. For each bid pair, we generate $k = 25$ images using our score composition method and compute the average prompt alignment scores to test whether higher bids yield stronger semantic alignment.

Results. Figure 1 demonstrates that as agent 1's bid increases, its prompt alignment improves monotonically from 0.25 to 0.32, reflecting a higher realized valuation. Conversely, as agent 2's bid decreases, its prompt alignment degrades monotonically from 0.34 to 0.22, indicating reduced influence over the generated content. These results validate that as advertisers bid more, they exert more influence on the generated image, as visualized in Figure 5

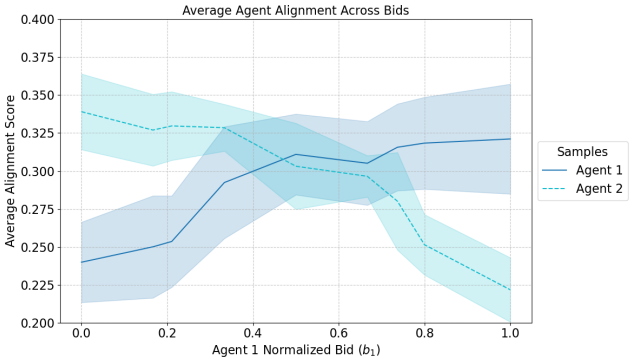
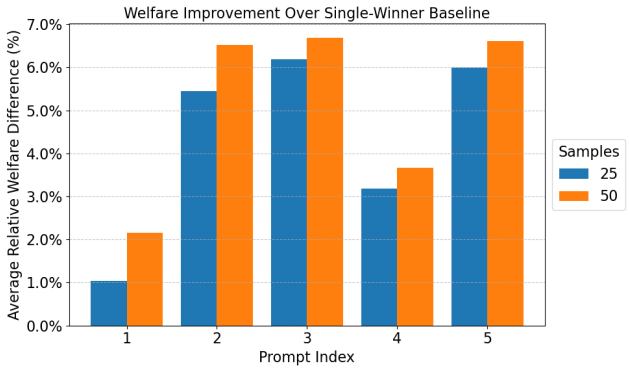


Figure 1. Mechanism satisfies bid monotonicity, with agent's CLIP alignment scores increasing as agent bid increases. Normalized agent 1 bid values: $b_1 \in \{0, 0.17, 0.21, 0.33, 0.5, 0.67, 0.74, 0.8, 1\}$.

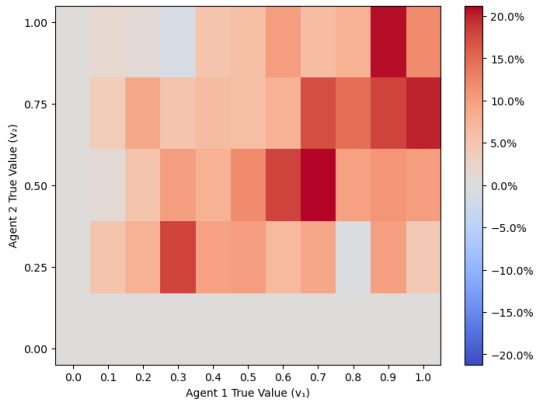
3.3. Efficiency: Improved Welfare Compared to Single-Winner Baseline

Unlike single-winner auctions, our mechanism enables multiple agents to be represented simultaneously in the same generated output. A key goal is to show that this compositional capability leads to higher total welfare. To evaluate this, we compare our method against a single-winner VCG auction baseline that generates images using only one agent's prompt. In this baseline, $2k$ images are sampled, k images using each agent's prompt appended to the base prompt. The final output is the image that optimizes overall welfare. This baseline serves as a direct comparison without agent preference compositionality.

Results. Figure 2a shows that our method consistently outperforms the single-winner baseline in terms of average welfare across prompts. Our diffusion image auction improves average relative welfare by up to 6.5% over the baseline. In addition, as the number of Monte Carlo samples increases, we show that average relative welfare also increases, suggesting that more sampling better captures high-welfare outcomes and further boosts the benefits of our mechanism. We present bid-specific welfare improvements in Figure 2b for individual prompt 5, displaying welfare improvements of up to 21%. These results show that by aggregating agent preferences, our auction approach effectively produces efficient outcomes.



(a) Average welfare improvement of up to 6.5% over the single-winner baseline.



(b) Bid-specific welfare improvement of up to 21% for prompt 5 with $k = 50$ samples.

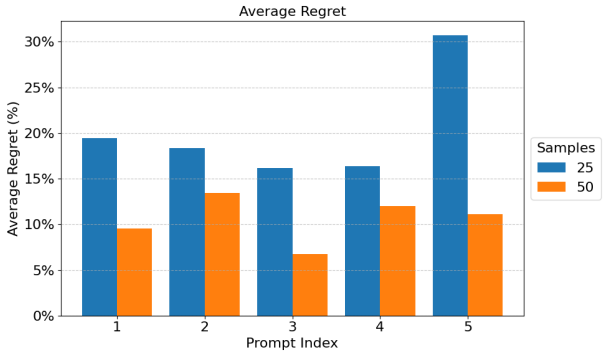
Figure 2. Our diffusion image auction improves average relative welfare over the single-winner VCG auction baseline. Increasing Monte Carlo sample size further improves welfare.

3.4. Incentive Compatibility: Truthful Bidding is Approximately Optimal

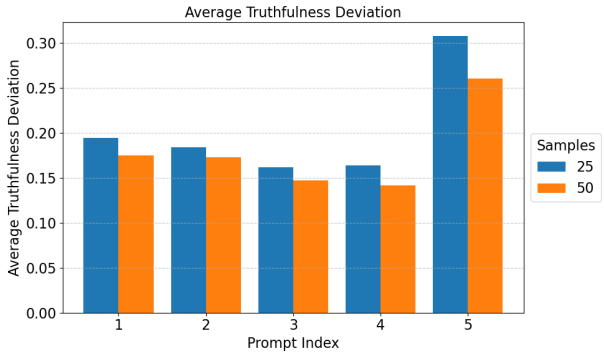
A central goal of our VCG-based mechanism is to incentivize truthful bidding, where each agent maximizes their utility by bidding their true value. This means an advertiser's optimal bid b_i should match their true value m_i. To test this, we simulate a range of bidding scenarios. For every combination of true values (m_1, m_2), we fix agent 2's bidding to be truthful and allow agent 1 to bid a range of values. We sweep Agent 1's true value from 0.0 to 1.0 in steps of 0.1 and Agent 2's true value from 0.0 to 1.0 in steps of 0.25, covering a wide range of value scenarios. For each combination of submitted bid and true values, we evaluate the average regret, the relative utility loss from truthful bidding compared to the optimal bid. In particular, we calculate it as the raw utility difference divided by its utility under the truthful bid. Additionally, we compute the absolute difference between agent 1's truthful bid value and the utility-maximizing bid.

Results. Figure 3a shows the average regret experienced by Agent 1 across the same set of prompts and sample sizes. We present results with $k \in \{25, 50\}$ Monte Carlo samples and demonstrate that average regret consistently decreases with increased sampling. For $k = 50$, average regret remains below 15% across all prompts, reaching as low as 7% for individual prompts. This trend indicates that agents have reduced incentive to deviate from truthful bidding, further supporting the robustness of our mechanism under sampling-based generation.

Figure 3b reports the average deviation across all bidding combinations between Agent 1's true bid value and the utility-maximizing bid across five prompts. Across all prompts, the average truthfulness deviation consistently decreases as k increases. For $k = 50$, the average truthfulness deviation decreases to as low as 0.14. These findings provide empirical support that our mechanism preserves the truthfulness property of VCG in an approximate sense. Despite relying on Monte Carlo estimates of welfare during image generation, the learned auction dynamics can still incentivize agents to bid in alignment with their true values.



(a) Average regret down to 7% with $k = 50$ Monte Carlo samples.



(b) Average deviation down to 0.14 between true and optimal bid when $k = 50$.

Figure 3. Increasing Monte Carlo sample size improves truthfulness. Average regret and truthfulness deviation decrease from $k = 25$ to $k = 50$ samples, indicating truthfulness is approximately optimal.

3.5. Image Quality Preservation

When combining multiple advertiser prompts, a fundamental concern is whether image quality degrades due to conflicting objectives. To ensure our mechanism maintains visual quality, we evaluate the alignment between generated images and the prompt "`High quality photo`". Figure 4 reports the mean and standard deviation of CLIP quality alignment scores across prompts. For all bid combinations, average image quality remains stable above 0.505, standard deviations under 0.005 (shaded in Figure 4). No statistically significant differences are observed between a single-winner outcome ($b_1 = 0, 1$) and a multi-winner outcome (e.g. $b_1 = 0.5$). These results indicate that our mechanism can blend multiple advertiser inputs without compromising image quality.

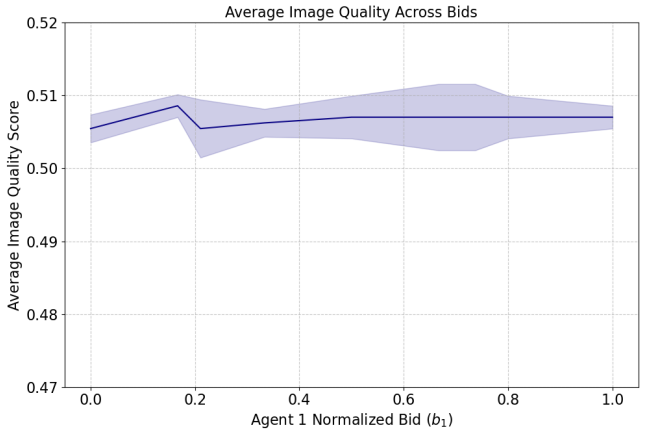


Figure 4. Expressing multiple agent preferences does not degrade image quality.

4. Discussion

Our experiments compellingly demonstrate the effectiveness of our auction-driven image generation mechanism. We found a strong positive link between advertiser bids and the visual prominence of their desired content, confirming bid monotonicity. Moreover, our method significantly improved social welfare compared to a single-winner baseline, with greater gains from increased sampling. This highlights the efficiency of our approach for integrating multiple preferences in a single image, valuable for multi-winner advertising. Importantly, the mechanism showed approximate truthfulness, as more samples incentivized agents to bid their true values. To the best of our knowledge, this work pioneers the design of auctions specifically for diffusion-based image generation, offering a novel framework with significant promise for future applications in advertising and beyond.

Applications The framework we present offers strong potential for real-world applications, particularly in online advertising. Consider a search engine displaying banner ads on web pages. Using our auction, it can generate visually compelling, contextually relevant images that reflect the preferences of multiple advertisers. Our finding—that this generative auction yields higher welfare than simple single-winner ad auction—provides a compelling incentive for platforms to adopt such methods. Advertisers, motivated by visibility and influence aligned with their bids, benefit from having their branding or products meaningfully integrated into the generated content. Beyond static images, our approach extends naturally to dynamic media. For example, a video platform could use our mechanism to embed branded elements, enabling advertisers to bid for subtle, context-aware product placement. This opens new possibilities for non-intrusive, personalized advertising across multimedia formats.

5. Conclusion

Our work presents the first generative image auction, pioneering the novel concept of utilizing diffusion models to enable multi-winner image generation reflecting aggregated participant preferences. We introduce an expressive score composition technique that enables nuanced, bid-dependent control over generated content, in conjunction with Monte Carlo VCG-based payments designed to ensure incentive compatibility. Our exploratory empirical analysis provides compelling evidence for key desirable properties, including bid monotonicity, improved social welfare compared to single-winner auction baselines, and approximate truthfulness. As a foundational step, our research lays the groundwork for future investigation in controllable, incentive-aligned image generation, offering a new lens through which to design and evaluate generative systems.

6. Impact Statement

Our work pioneers a new paradigm for economically-aligned content creation, offering significant positive societal impacts by enabling multiple stakeholders to collaboratively shape AI-generated visuals through a dynamic auction system. This innovation allows for more efficient and nuanced brand integration in digital spaces, moving beyond traditional, often intrusive, advertising to create richer, more contextually relevant experiences for users. By fostering a mechanism for shared influence, we envision a future where diverse preferences are reflected in generated media, leading to more personalized and engaging content for everyone. However, this new mechanism also requires careful consideration: its ability to blend commercial interests seamlessly into images should come with robust transparency and ethical guidelines from platforms to ensure that content remains authentic and trustworthy, and to prevent the over-amplification of well-resourced entities.

References

Edward H. Clarke. Multipart pricing of public goods. *Public Choice*, 11:17–33, 1971.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Kumar Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. Auctions with llm summaries, 2024. URL https://arxiv.org/abs/2404.08126.

Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of WWW 2024*, 2024. URL https://arxiv.org/abs/2310.10826. arXiv:2310.10826.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. Online advertisements with llms: Opportunities and challenges, 2024. URL https://arxiv.org/abs/2311.07601.

Theodore Groves. Incentives in teams. *Econometrica*, 41: 617–631, 1973.

MohammadTaghi Hajiaghayi, Sébastien Lahaie, Keivan Rezaei, and Suho Shin. Ad auctions for llms via retrieval augmented generation, 2024. URL https://arxiv.org/abs/2406.09459.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. URL https://arxiv.org/abs/2112.10741.

Alexander Q Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

Alec Radford, Jong Wook Kim, et al. Learning transferable visual models from natural language supervision. In *ICML 2021*, 2021. URL https://arxiv.org/abs/2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *CoRR*, abs/1907.05600, 2019. URL http://arxiv.org/abs/1907.05600.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, et al. Scorebased generative modeling through stochastic differential equations. In *ICLR*, 2021b. URL https://arxiv.org/abs/2011.13456.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance*, 16(1):8–37, 1961.

A. Related Work

Our work draws from auction theory, especially the Vickrey-Clarke-Groves (VCG) mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1973), known for social welfare maximization and incentive compatibility. Applying VCG to AI-generated content poses new challenges and opportunities.

Recent research intersects auctions and LLMs (Feizi et al., 2024). Dütting et al. (2024) introduced token auctions for fine-grained control over LLM output. For ad placement, Hajiaghayi et al. (2024) proposed auctions using retrieval-augmented generation (RAG), and Dubey et al. (2024) focused on ad allocation in LLM summaries. Our work differs by developing aggregation techniques for image generation, addressing multi-agent image preference aggregation.

Diffusion models (Ho et al., 2020; Song and Ermon, 2019; Rombach et al., 2022; Dhariwal and Nichol, 2021) excel in image synthesis by reversing a noising process (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021a). Recent methods like Rectified Flow (Liu et al., 2022; Esser et al., 2024) and FLUX.1 aim for faster, direct mappings from noise to data.

Classifier-free guidance (Ho and Salimans, 2022) is a key technique for controlling diffusion model generation without separate classifiers, effective for high-fidelity image and text-to-image synthesis (Saharia et al., 2022; Nichol et al., 2022). Our work leverages diffusion models and classifier-free guidance to integrate multiple preferences into a single image.

B. Limitations

A fundamental limitation of our approach arises from the inherent probabilistic nature of denoising diffusion models. This stochasticity, coupled with the potential for variability in how text prompts are interpreted by the diffusion model, means that we cannot analytically optimize the generated image to perfectly maximize welfare for a given set of bids. Our utilization of Monte Carlo sampling to estimate welfare and VCG payments is a direct consequence of this intractability. While increasing the number of samples leads to improved welfare and closer adherence to truthfulness, this comes at the cost of increased computational resources and generation time, creating a trade-off between accuracy and efficiency.

C. Future work

While we provide a comprehensive empirical evaluation of the two-agent auction setting, a future research direction involves empirically evaluating agent behavior in auctions with more than two agents. Investigating the performance and economic properties of our auction with many agents will be crucial for understanding its applicability in more complex real-world scenarios.

D. Methodology

Our goal is to design auctions capable of generating images that effectively aggregate the preferences of multiple agents. In our stylized setting, each agent is characterized by a (public) text prompt, describing their desired visual content, e.g. naming branded objects or concepts, and a scalar (private) value; we assume that the value that an agent derives from an image is their private value times the alignment between the image and their prompt. Thus each agent wants to increase the presence of its desired visual content (as described by their prompt) in the image generated by the auction. The auction seeks to output a single image, conditioning on a base prompt — describing the content that would be generated in the absence of the bidders — and the bidders' inputs (i.e. their prompts and bids), such that (a) each agent's influence on the final image increases with their bid, (b) welfare improves over a baseline auction that only chooses a single winner and gives them the opportunity to influence the generated image, (c) agents are incentivized to bid their true values (truthfulness), and (d) the overall image remains high quality and faithful to the base prompt. We proceed to formalize our setting.

D.1. Overall Setting

Our setting involves three ingredients:

1. **Base Prompt.** There is a base text prompt c. In the absence of any bidders, the generated image would be sampled by a diffusion model p_0 conditioning on this prompt, i.e. the generated image would be $I \sim p_0(\cdot \mid c)$.

2. **Agents.** A finite set of agents $N = \{1, \ldots, n\}$, each of which wishes to steer the generated image to include their desired content. Each agent i is characterized by a text prompt c_i describing their desired content, and a scalar value $m_i \in \mathbb{R}_{\geq 0}$. Together these determine the value $v_i(I)$ that the agent derives from an image I, depending on the alignment between I and their prompt, as described in Section D.2.

3. **Auction.** The bidders are asked to submit bids to the auction. Given a vector of bids $\mathbf{b} = (b_1, \ldots, b_n)$ the mechanism computes a (randomized) aggregation function $p_{\mathbf{b}}$, and generates an image $I \sim p_{\mathbf{b}}(\cdot \mid c)$. The intent of the aggregation function is to combine the agents' preferences, as determined by their prompts and bids, with the base prompt in the generated image. The mechanism also computes (randomized) payment rule $q_{\mathbf{b},i}$ for each agent i, sampling for this agent a price $c_i \sim q_{\mathbf{b},i}(\cdot|c, I)$.

Our goal in this paper is to maximize the total welfare attained by running the auction, as described in Section D.3. Towards this goal we propose an aggregation funciton $p_{\mathbf{p}}$ and price functions $q_{\mathbf{b},i}$ inspired by the celebrated VCG mechanism.

D.2. Agent Model

Each agent i is characterized by a public prompt c_i and a private value m_i. We assume that their value $v_i(I)$, for a given image I, equals m_i multiplied by the alignment between I and their prompt, denoted by $\alpha_i(I)$, where $\alpha_i(\cdot)$ is an agent-specific (public) alignment function, e.g. $\alpha_i(I)$ could be the cosine similarity between the CLIP embeddings of I and c_i. In particular,

$$v_i(I) \;=\; m_i \cdot \alpha_i(I). \tag{1}$$

Intuitively, $\alpha_i(I)$ reflects how well-represented in I is the agent's preferred content as described by c_i. The multiplier m_i measures how much the agent values a single unit of image-prompt alignment. As either alignment or multiplier increases, so does the value that agent i derives from the image.

D.3. Social Welfare Optimization

The goal of the auction is to generate an image I that maximizes social welfare, i.e. $\sum_{i=1}^{m} v_i(I)$.

Suppose that the auction uses an aggregation function $p_{\mathbf{b}}$. Then, if the bidders submit a bid vector $\mathbf{b}$, the expected welfare of the auction is

$$\mathbb{E}_{I \sim p_{\mathbf{b}}(\cdot|c)} \left[\sum_{i=1}^{m} v_i(I) \right]. \tag{2}$$

Our goal is to choose an aggregation and a price rule for the auction so as to incentivize the bidders, who think strategically, to bid in a way that induces an expected welfare (2) that is close to the maximum social welfare. In the next couple of sections we propose an aggregation and a price rule towards this goal, inspired by the celebrated VCG auction.

D.4. Proposed Aggregation Function

We assume that we have access to a denoising diffusion model that can be guided via text prompts. We denote by $s_t(x) = \nabla_x \log p_t(x)$ the score of the unconditional diffusion model at each noise level t, and by $s_t(x|c) = \nabla_x \log p_t(x|c)$ the score of the conditional diffusion model given the prompt. If we know these score functions at all noise levels t we can run the reverse diffusion process to generate samples from $p_0(x)$ and, respectively, $p_0(x|c)$.

To aggregate the agent preferences in the image generation process, we introduce a score composition technique. We present our score composition technique for two agents, postponing its obvious multi-agent generalization to Appendix F. Suppose that the prompts of the agents are c_1, c_2 and their bids are b_1, b_2, and assume that $b_1 \geq b_2$, without loss of generality. Take $b_1^{(1)} = \frac{b_1}{b_1+b_2}$ and $b_2^{(1)} = \frac{b_2}{b_1+b_2}$ to be the normalized bids and $c_{1,2} = \{c_1, c_2\}$ to be the agents' combined prompts. Further, define the normalized weight $w^{(1)} = 2b_1^{(1)} - 1$. We propose a score function composition, $s_t^{(1,2)}(x)$ as follows:

$$s_t^{(1,2)}(x) = (1 - w^{(1)})s_t(x|c, c_{1,2}) + (w^{(1)})s_t(x|c, c_1). \tag{3}$$

Intuitively, this score function composition enables both agents' preferences to be expressed through the score of the joint prompt $c_{1,2}$. It controls for how dominant agent 1 is by adding the additional score conditional on c_1, scaled by $w^{(1)}$, which is proportional to normalized bid $b_1^{(1)}$. In one extreme, when the normalized bids of the agents are 0 and 1, the composed score function is equal to the score function conditional on the dominant agent's prompt and the base prompt. In the other extreme, when the normalized bids are 0.5 and 0.5, then the composed score function becomes only the score function conditional on both agents' prompts and the base prompt. Therefore, this score composition allows for a continuous gradient of weightings based on the agent bids.

Given the composed score, we run a reverse diffusion process to sample an image I. We repeat k times, for some hyperparameter k, and output the image that maximizes the total welfare, as computed using the bids submitted by the agents (since we do not know their true values). This is in the spirit of the VCG allocation rule, which would choose the exact maximizer of the welfare (according to the bids). Instead, we use the reverse diffusion process defined by the composed score to sample k images and output the one that has the highest welfare (according to the bids).

D.5. Proposed Payment Function

Our VCG-inspired payment scheme charges each agent a price that approximates the externality that the presence of this agent causes to the welfare of the other agents.

To compute the price that we charge to agent i we compute the total welfare (according to the bids) that all agents other than i get from the image selected by the aggregation function of Section D.4. We then re-run the aggregation function of Section D.4 pretending that bidder i does not participate in the auction, i.e. we cancel their bid and ignore their prompt. We compute the resulting welfare (according to the bids) that all agents other than i would have gotten in the absence of this bidder.

We charge bidder i the difference between the counter-factual welfare of the others if i had not shown up and the realized welfare of the other now that i did show up.

E. Prompts

Table 1. Prompt Index Table

Index	Base Prompt	Agent 1 Prompt	Agent 2 Prompt
1	A hiker resting on a mountain trail	North Face backpack	Gatorade bottle
2	A professional working at a desk in a bright office	Dell monitor	Deer Park water bottle
3	A swimming pool outdoors	Marriott Bonvoy hotel	Ray-Ban sunglasses
4	Friends gathered for a movie night at home	Doritos chips	Coca-Cola cans
5	A car driving along a scenic coastal highway	Tesla car	ExxonMobil sign

(a) $b_1 = 0, b_2 = 1$

(b) $b_1 = 0.3, b_2 = 0.7$

(c) $b_1 = 0.5, b_2 = 0.5$

(d) $b_1 = 0.7, b_2 = 0.3$

(e) $b_1 = 1, b_2 = 0$

Figure 5. Example auction generations for various normalized agent bids. Presence of the North Face backpack increases as b_1 increases, while presence of the Gatorade bottle increases as b_2 increases.

F. Generalized Score Composition

The aggregation function can be extended to n agents ($s^*_{1,n}(x)$) by iteratively defining score compositions with one less agent ($s^*_{1,n-1}(x)$). Assume, without loss of generality, that agents 1 to n submit bids in decreasing value: $b_1 \geq b_2 \geq \ldots \geq b_n$. Denote $b_j^{(0)} = b_j$. For iteration i, denote $b_j^{(i)} = \frac{b_j^{(i-1)}}{\sum_{k=1}^{n-i+1} b_k^{(i-1)}}$ as the normalized bids for iteration i and $w^{(i)} = \left(2\sum_{j=1}^{n-i} b_j^{(i)} - 1\right)$ as the normalized weights. Denote $c_{1,n} = \{c_1, c_2, \ldots, c_n\}$ as the combined prompt from all agents, and $s^*_{1,1}(x) = s(x|c, c_1)$ as the "base case" with the score conditional on dominant agent 1's prompt and the base prompt. We can recursively define $s^*_{1,n}(x)$, separating the least dominant agent n and weighting the score function for the remaining $n-1$ agents.

$$s^*_{1,n}(x) = (1 - w^{(1)})\, s(x|c, c_{1,n}) + w^{(1)}\, s^*_{1,n-1}(x) \tag{4}$$

$$s^*_{1,j}(x) = (1 - w^{(n-j+1)})\, s(x|c, c_{1,j}) + w^{(n-j+1)}\, s^*_{1,j-1}(x) \tag{5}$$

Then, we derive the closed form equation for $s^*_{1,n}(x)$.

$$s^*_{1,n}(x) = \prod_{i=1}^{n-1} w^{(i)}\, s(x|c_1) + \sum_{i=2}^{n} \left[\left(\prod_{j=1}^{n-i} w^{(j)} \right) (1 - w^{(n-i+1)})\, s(x|c_{1,i}) \right] \tag{6}$$

G. Experimental Details

Models: We use the FLUX.1-schnell model from Black Forest Labs, which is licensed under Apache License 2.0, which we abide by. We used the model stored in HuggingFace, https://huggingface.co/black-forest-labs/FLUX.1-schnell. Additionally, we use the CLIP model from OpenAI, which is licensed under MIT License, which we abide by. We used the model stored in HuggingFace, https://huggingface.co/openai/clip-vit-large-patch14-336.