DIFFAUCTION: AN AUCTION-DRIVEN FRAMEWORK FOR GENERATIVE IMAGE ADVERTISING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models create an opportunity for advertising in generated visual content, with multiple advertisers competing for exposure within a shared image. Yet existing approaches are computationally costly when generating images under varying target-share configurations, and it remains unclear how the resulting outcomes should be evaluated and priced. To address these gaps, we propose DiffAuction, an auction framework connecting controllable generation, click-intent evaluation, and auction design. First, our training-free Residual Share Calibration (RSC) decomposes joint-attention updates into product-specific residuals and rescales them according to target shares. Because a fixed prompt tends to produce similar viewpoints that favor the same products regardless of their target shares, we further introduce Share-Conditioned Prompting (SCP) to adapt the viewpoint to each desired allocation. Second, we develop an evaluator that uses a VLM to extract visual evidence and an LLM to assess Click-Through Intent (CTI), a proxy for click propensity, along four dimensions: Product Prominence, Brand Recognizability, Scene Integration, and Contextual Value Communication. We validate its assessments against judgments from 10 trained annotators on 1,000 images. Third, we formulate this advertising setting as an auction over jointly feasible generation configurations. Each configuration has a public expected CTI vector, while advertisers hold configuration-specific private values per unit of CTI that map click intent into commercial value. We adopt a learnable affine maximizer auction (AMA), which jointly selects a shared generation configuration and determines advertiser-specific payments to maximize expected platform revenue. The resulting allocation and payment rules guarantee dominant-strategy incentive compatibility (DSIC) and individual rationality (IR). Experiments show that our generation method strengthens the monotonic relationship between target shares and CTI relative to baselines, while the learned AMA achieves higher platform revenue in the resulting auction setting.

1 INTRODUCTION

Online advertising monetizes user attention by allocating exposure through auctions (Edelman et al., 2007; Varian, 2007). Whereas conventional auctions assign predefined ad slots, a growing body of work on LLM-native advertising has begun to auction influence over generated content by incorporating advertisers' bids and preferences into model responses (Duetting et al., 2024; Hajiaghayi et al., 2024; Soumalias et al., 2025; Zhao et al., 2025; Yun et al., 2026; Zhang et al., 2026a;b; Liu et al., 2026; Turner-Smith et al., 2026). These advances demonstrate the potential of treating generated content itself as advertising inventory.

Diffusion models extend this opportunity to visual advertising, allowing advertisers to compete for exposure within a shared generated image. Compared with LLM-based ad auctions, this direction remains underexplored. Constructing an auction over generated images requires three interdependent capabilities. First, the generation process must translate the auction's selected target-share configuration into corresponding changes in the generated image. However, diffusion models provide limited prompt-based responsiveness to changes in target shares. Second, the effectiveness of each resulting exposure must be measurable. Click intent provides a suitable evaluation criterion. However, unlike predefined ad slots with historical click-through rates (CTR), each generated image is a new outcome whose product-level click intent must be assessed from the content itself.

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.

Third, the mechanism must combine the public expected CTI of each configuration with advertisers' private commercial values per CTI unit, select one jointly feasible configuration, and determine bidder-specific payments under truthful reporting.

To address the first challenge, Sun et al. (2025) introduce bid-weighted score composition, combining diffusion predictions conditioned on nested subsets of product prompts. This formulation requires multiple conditional model evaluations per timestep, increasing computational cost as the number of products grows. To accelerate generation, we condition the model jointly on all product prompts and incorporate the target-share specification within a single forward pass per timestep. We implement this approach through Residual Share Calibration (RSC), a training-free method that decomposes the residual output of joint attention into product-specific update directions induced by different product tokens. These residuals represent how each product shifts the latent image features during generation. RSC then calibrates the magnitudes of these product-specific updates according to the target exposure shares. However, a fixed prompt can favor similar viewpoints even as target shares change, leaving certain products consistently obscured or overly prominent. We therefore introduce Share-Conditioned Prompting (SCP) that adjusts the viewing perspective according to target shares to better support the desired exposure allocation.

To address the second challenge, we introduce a product-level *Click-Through Intent* (CTI) proxy based on four interpretable dimensions: product prominence, brand recognizability, scene integration, and contextual value communication. A VLM extracts visual evidence for each product, and an LLM scores that evidence using a fixed rubric. The aggregate provides a content-based estimate of product-level click intent, which we use to evaluate the effectiveness of the exposure realized under each generation configuration. We assess this evaluator on a dataset of 1,000 images annotated by ten annotators, demonstrating strong correlation with human ratings.

To address the third challenge, we formulate exposure trading in generative image advertising as an auction over generation configurations subject to joint feasibility constraints. For a given scene and set of competing advertisers, the platform offers candidate generation configurations, each specifying an exposure-share vector and the expected CTI for each advertiser. However, public CTI captures anticipated click intent but does not fully represent commercial returns: advertisers may draw on private conversion data and campaign objectives to value the same click intent differently across configurations. We therefore allow each advertiser to hold multiple private valuations, one for each configuration. Each valuation represents how much one unit of CTI is worth to the advertiser under that configuration, thereby mapping the public estimate of click intent into advertiser-specific commercial value. Building on this formulation, we adopt a learnable affine maximizer auction (AMA) (Curry et al., 2023) to optimize expected platform revenue under truthful reporting. Using a joint scoring function with learnable advertiser weights and configuration-specific offsets, the mechanism selects the highest-scoring configuration and constructs advertiser-specific price menus that depend only on the other advertisers' reports. The allocation rule makes the selected configuration utility-maximizing for each advertiser within its own menu, while a shared tie-breaking rule ensures a common choice. The mechanism guarantees dominant-strategy incentive compatibility (DSIC) and individual rationality (IR). Experiments with simulated private valuations show that the learned AMA achieves higher platform revenue than the baselines in this auction setting.

In summary, our contributions are threefold:

- We introduce RSC, a training-free method that calibrates product-specific residual contributions within a single forward pass per timestep. SCP complements this control by adapting the viewing perspective to the desired exposure allocation.
- We develop a VLM–LLM evaluator that assesses product-level CTI along four interpretable dimensions and construct a dataset of 1,000 images annotated by 10 annotators to evaluate its agreement with human judgments.
- We formulate generative image advertising as an auction over jointly feasible generation configurations with multiple configuration-specific private values per CTI unit. We instantiate this formulation with a learnable affine maximizer auction (AMA) that optimizes expected platform revenue while guaranteeing DSIC and IR.

2 Related Work

Auction-based generative advertising. Generative advertising treats model-generated content as advertising inventory, allowing advertisers to bid for influence beyond predefined ad slots. Existing LLM-native mechanisms realize this idea at different levels: token-level auctions allow bids to influence individual tokens (Duetting et al., 2024); RAG-based segment auctions allocate ads to discourse segments according to bids and relevance (Hajiaghayi et al., 2024); MOSAIC aggregates advertisers' preferences over generated responses without fine-tuning the LLM (Soumalias et al., 2025); and LLM-Auction jointly models auction allocation and response generation (Zhao et al., 2025). Neuron Auctions shift the auction object from surface text to internal FFN representations and treat continuous intervention budgets as auctionable commodities (Yun et al., 2026). These studies motivate our formulation of product exposure in a generated image as an auctionable resource.

However, auction-driven generative image advertising remains relatively underexplored. Sun et al. (2025) provide an initial study. Their target-share-conditioned generation requires multiple conditional model evaluations per denoising timestep, causing computational cost to grow with the number of products. Moreover, they use CLIP similarity as a proxy for product value, but this simple proxy lacks interpretability and direct validation. They also do not consider the more complex setting in which advertisers hold multiple configuration-specific private valuations, or integrate image generation and auction design into a unified framework for optimizing platform revenue. Therefore, systematic work is still needed to fill these research gaps.

Training-free controllable image generation. A broad line of research controls pretrained diffusion models at inference time without additional model training. Existing approaches intervene at different stages of the generation process. Attention-based methods manipulate token–image interactions to preserve, amplify, or suppress specific concepts; for example, Prompt-to-Prompt edits cross-attention maps and supports token-level attention reweighting (Hertz et al., 2022), while Rich Text introduces explicit token and region-aware guidance (Ge et al., 2023). Guidance-based approaches modify denoising directions to strengthen or suppress semantic concepts, as exemplified by SEGA (Brack et al., 2023). More recent methods directly edit internal representations of diffusion Transformers; FluxSpace, for instance, identifies editable directions in FLUX feature spaces (Dalva et al., 2025). These methods demonstrate the effectiveness of inference-time intervention for semantic and compositional control. However, our setting differs from conventional concept editing: multiple advertised products coexist in the same jointly generated scene, and the objective is not merely to include or emphasize a concept, but to regulate their relative exposure according to a prescribed share vector. Sun et al. (2025) address this problem by combining generation directions obtained from multiple conditional model evaluations, resulting in increased computational cost as the number of products grows. In contrast, the combination of Share-Conditioned Prompting (SCP) and Residual Share Calibration (RSC) supports target-share-conditioned generation with only a single model forward pass per denoising timestep.

3 Method

As illustrated in Figure 1, target shares, serving as control inputs to the generative model, determine the intended exposure allocation; CTI estimates the product-level click intent associated with the resulting exposure, and the auction combines the expected CTI of each candidate configuration with advertisers' reported values per CTI unit for allocation and pricing. We describe each component in the following sections.

3.1 Residual Share Calibration and Share-Conditioned Prompting

Given a scene prompt and N product prompts, we use a target-share vector $\mathbf{s} = (s_1, \ldots, s_N) \in \Delta^{N-1}$ to condition the relative exposure of the products. We propose Residual Share Calibration (RSC), a training-free method that decomposes the image-feature update from joint attention into product-specific residuals and rescales them so that their shares of the summed product-wise Frobenius norms match the target vector.

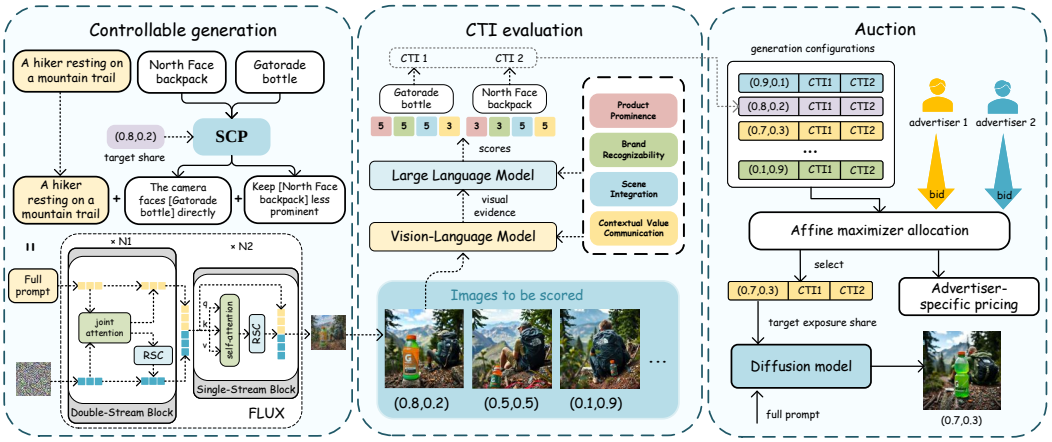


Figure 1: Overview of DiffAuction. DiffAuction connects generation, CTI evaluation, and auction. Given target shares, RSC and SCP control the exposure of products during generation. The resulting images are evaluated offline by a VLM–LLM pipeline that estimates product-level CTI. During the auction, a learnable affine maximizer combines the expected CTI of each configuration with advertisers' reported values per CTI unit to select a shared configuration and determine advertiser-specific payments. The selected exposure shares are used to generate the advertisement image.

Product-wise residual decomposition. Consider one joint-attention block and omit its denoising-step and block indices for clarity. Let $\mathbf{A} \in \mathbb{R}^{H\times I\times T}$ denote the image-to-text attention tensor and $\mathbf{V} \in \mathbb{R}^{H\times T\times d_h}$ the corresponding text-value tensor. Here, H is the number of attention heads, I is the number of image tokens, T is the number of text tokens, and d_h is the feature dimension of each head. For product i, let $\mathbb{T}_i \subseteq \{1,\dots,T\}$ be its text-token index set; these sets are mutually disjoint across products. Once $\mathbf{A}$ has been computed, value aggregation and the output projection are linear with respect to these token sets. The residual matrix of product i is therefore

$$\boldsymbol{R}_i = \left(\mathrm{Concat}_{h=1}^{H}\left[\sum_{j\in\mathbb{T}_i}\mathbf{A}_{h,:,j}\mathbf{V}_{h,j,:}^{\top}\right]\right)\boldsymbol{W}_O, \qquad \boldsymbol{Y} = \boldsymbol{U} + \sum_{i=1}^{N}\boldsymbol{R}_i. \tag{1}$$

Here, $\boldsymbol{R}_i$ is the additive image-feature update induced by product i's text tokens. The matrix $\boldsymbol{W}_O$ denotes the output projection without its bias, while $\boldsymbol{U}$ collects the unchanged terms, including the image-to-image contribution, non-product text residuals, and output bias. All product residuals are thus obtained from the same attention computation and share a consistent scene context.

Target-share reweighting. Directly multiplying each residual $\boldsymbol{R}_i$ by its target share s_i generally fails to achieve the desired proportions, since the residuals may differ in magnitude. We therefore define each product's current share of the total residual magnitude as

$$p_i = \frac{\|\boldsymbol{R}_i\|_F}{\sum_{j=1}^{N}\|\boldsymbol{R}_j\|_F}. \tag{2}$$

We then compensate for this imbalance before combining the updates:

$$\widetilde{\boldsymbol{R}}_i = \frac{s_i}{p_i}\boldsymbol{R}_i, \qquad \widetilde{\boldsymbol{Y}} = \boldsymbol{U} + \sum_{i=1}^{N}\widetilde{\boldsymbol{R}}_i. \tag{3}$$

Provided every product with a positive target share has a nonzero residual, this rescaling makes each product's share of the summed residual norms equal to s_i, while preserving the total residual magnitude and leaving $\boldsymbol{U}$ unchanged.

Share-Conditioned Prompting (SCP). RSC can modulate the relative exposure of different products, but in some cases, the viewpoint and layout induced by a fixed prompt may leave high-share

products insufficiently visible or low-share products overexposed. We therefore augment each product description with prominence instructions based on its target share relative to equal allocation:

$$r_i = \frac{s_i}{1/N} = N s_i. \tag{4}$$

Here, N is the number of products, and r_i measures how much exposure product i is intended to receive relative to an equal share. We map r_i to one of five ordered prominence levels, each associated with a shared descriptive template that encourages more favorable views at higher levels and limits excessive exposure at lower levels. Because SCP maps target shares to five ordered levels, it provides discrete prompt conditioning; RSC separately uses the real-valued target shares to calibrate residual-norm contributions. We combine these descriptions with the scene prompt to guide composition, while RSC calibrates residual-norm shares for each target-share setting. Appendix B.3 provides the detailed prompt templates. Our ablation study further demonstrates the complementary effects of SCP and RSC in improving share–CTI monotonicity.

3.2 Evaluating Product-Level Click-Through Intent

We assess the effectiveness of product exposure in terms of its potential to attract further user interest, using click intent as a practical criterion. While conventional ad placements can draw on historical click-through rates, newly generated images require a content-based estimate of this intent. We therefore introduce a product-level *click-through intent* (CTI) proxy based on four dimensions: *product prominence*, *brand recognizability*, *scene integration*, and *contextual value communication*.

Evaluation dimensions. We focus on image-observable factors that may influence CTI while abstracting away subjective product preferences. To characterize these factors, we define four complementary dimensions: *product prominence*, which captures visual salience; *brand recognizability*, which reflects the visibility of brand cues; *scene integration*, which measures how naturally the product fits into the scene; and *contextual value communication*, which assesses whether the scene conveys the product's use, relevance, or value. Further details are provided in Appendix F.

Image-grounded CTI evaluation. Given an image I, its scene description, and the advertised products, a vision–language model (VLM) extracts a structured caption documenting each product's observable evidence across the four dimensions. Using these captions and the same scene and product information, a separate language model (LLM) assigns scores $z_{id} \in \{1, \dots, 5\}$ to product i on dimension d under a fixed rubric; higher scores are better. Both prompts appear in Appendix H. The equally weighted mean is linearly rescaled to $[0, 1]$:

$$\mathrm{CTI}_i(I) = \frac{1}{16} \sum_{d=1}^{4} (z_{id} - 1) \in [0, 1]. \tag{5}$$

3.3 Learnable Affine-Maximizer Auction Mechanism

We formulate exposure trading as an auction over generation configurations subject to joint feasibility constraints: all advertisers receive exposure through one shared configuration rather than independently assigned outcomes. We adopt a learnable affine maximizer auction (AMA) (Curry et al., 2023) to optimize expected platform revenue. The mechanism selects a single common configuration using learnable advertiser weights and configuration offsets and determines advertiser-specific payments according to the affine-maximizer payment rule.

Configurations and private valuations. The public configuration set is indexed by $\mathcal{K} = \{0, 1, \dots, K\}$. Each paid configuration $k \geq 1$ specifies a target share vector $\mathbf{s}_k \in \Delta^{n-1}$ and an expected CTI vector estimated from a historical scene set $\mathcal{X}$:

$$q_{ik} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathrm{CTI}_i(I_k(x)), \qquad \mathbf{q}_k = (q_{1k}, \dots, q_{nk}) \in [0, 1]^n. \tag{6}$$

Here $I_k(x)$ is a historical image generated under configuration k. These share and CTI vectors are fixed before bidding; they are not obtained by generating every candidate for the current auction. Configuration 0 denotes no allocation, with $\mathbf{s}_0 = \mathbf{q}_0 = \mathbf{0}$.

The expected CTI q_{ik} captures the anticipated click intent for advertiser i under configuration k, but the same CTI may have different commercial value across advertisers and configurations. Drawing on private conversion data and campaign objectives, advertisers may value different configurations differently. We therefore represent advertiser i's private values per CTI unit by a vector $\mathbf{v}_i = (v_{i1}, \ldots, v_{iK}) \in \mathbb{R}_+^K$, with $v_{i0} = 0$, rather than using a single coefficient shared across configurations. The resulting total expected value of configuration k to advertiser i is $V_i(k) = q_{ik} v_{ik}$. We assume that the platform has access to the joint distribution F of advertisers' private valuations.

Selecting a common configuration. Let $\phi = (\mathbf{w}, \boldsymbol{\beta})$ denote the mechanism parameters, where $w_i > 0$ is an advertiser weight and β_k is a configuration offset, with $\beta_0 = 0$. Advertiser i reports its value per CTI unit for each configuration, denoted by b_{ik}, so that $q_{ik} b_{ik}$ is its reported total value. We score and select configurations by

$$S_\phi(k; \mathbf{b}) = \beta_k + \sum_{i=1}^{n} w_i q_{ik} b_{ik}, \qquad k_\phi(\mathbf{b}) \in \arg\max_{k \in \mathcal{K}} S_\phi(k; \mathbf{b}). \tag{7}$$

The platform learns advertiser weights and configuration offsets offline and fixes them before bidding. During the auction, it selects the highest-scoring configuration, breaking ties by choosing the first maximizer in a fixed order. If configuration 0 is selected, no allocation is made and all realized payments are zero, as shown in Appendix E.4.

Advertiser-specific prices and truthful incentives. To price advertiser i's menu, the platform removes its own reported contribution from each configuration's joint score, retaining only the offset and the other advertisers' weighted reported values. Denote this remaining score by C_{ik} and its maximum across configurations by H_i:

$$C_{ik}(\mathbf{b}_{-i}) = \beta_k + \sum_{j \neq i} w_j q_{jk} b_{jk}, \qquad H_i(\mathbf{b}_{-i}) = \max_{\ell \in \mathcal{K}} C_{i\ell}(\mathbf{b}_{-i}). \tag{8}$$

The platform assigns advertiser i a price for each configuration and charges only the price of the selected configuration:

$$P_{ik}(\mathbf{b}_{-i}) = \frac{H_i(\mathbf{b}_{-i}) - C_{ik}(\mathbf{b}_{-i})}{w_i}, \qquad p_i(\mathbf{b}) = P_{i,k_\phi(\mathbf{b})}(\mathbf{b}_{-i}). \tag{9}$$

Because these prices depend only on the other advertisers' reports, advertiser i cannot change its own price menu by changing its report. Substituting the prices into its reported utility gives

$$\begin{aligned} q_{ik} b_{ik} - P_{ik}(\mathbf{b}_{-i}) &= \frac{S_\phi(k; \mathbf{b}) - H_i(\mathbf{b}_{-i})}{w_i}, \\ \arg\max_{k \in \mathcal{K}} \{q_{ik} b_{ik} - P_{ik}(\mathbf{b}_{-i})\} &= \arg\max_{k \in \mathcal{K}} S_\phi(k; \mathbf{b}). \end{aligned} \tag{10}$$

Consequently, the highest-scoring configuration maximizes each advertiser's reported utility, while a common tie-breaking rule ensures a consistent selection. Since each advertiser's induced price menu is independent of its own report, truthful reporting maximizes its true utility, establishing DSIC. Furthermore, every menu contains at least one zero-priced entry with nonnegative value, guaranteeing nonnegative utility under truthful reporting and hence IR.

Revenue-oriented learning. We learn weights and offsets to maximize expected revenue under truthful reporting:

$$\max_{\phi} R(\phi), \qquad R(\phi) = \mathbb{E}_{\mathbf{v} \sim F}\left[\sum_{i=1}^{n} p_i(\mathbf{v})\right], \tag{11}$$

where F is the joint distribution of private value vectors. To obtain gradients through discrete selection, we introduce

$$\pi_\phi^\tau(k \mid \mathbf{v}) = \frac{\exp(S_\phi(k; \mathbf{v})/\tau)}{\sum_{\ell \in \mathcal{K}} \exp(S_\phi(\ell; \mathbf{v})/\tau)} \tag{12}$$

and use the minibatch revenue surrogate

$$\mathcal{L}_{\mathrm{rev}}(\phi) = -\frac{1}{B} \sum_{m=1}^{B} \sum_{k=1}^{K} \pi_\phi^\tau(k \mid \mathbf{v}^{(m)}) \sum_{i=1}^{n} P_{ik}(\mathbf{v}_{-i}^{(m)}). \tag{13}$$

4 EXPERIMENTS

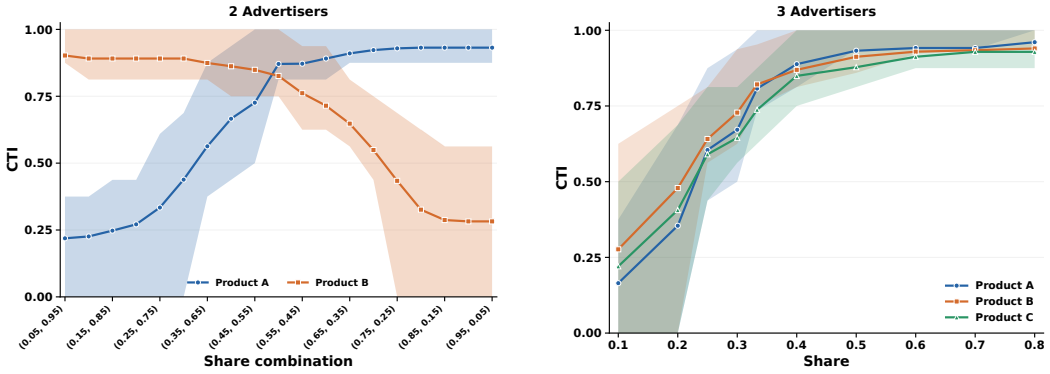
4.1 MONOTONICITY BETWEEN TARGET EXPOSURE SHARES AND CTI

Effectiveness of our method. Effective exposure control should translate higher target shares into more effective exposure in generated images, as measured by CTI. A stronger monotonic relationship therefore indicates that the generation method more consistently realizes the intended allocation. Therefore, we evaluate this relationship on 1,000 two-product scenes covering 400 distinct products, generating each scene under 19 share configurations for a total of 19,000 images. For our method, we use FLUX.1-schnell to generate images at 1024×1024 resolution with 10 denoising steps, applying RSC during the first 8 steps. We compare against Sun et al. (2025), which weights separately computed generation directions by bids, and prompt-only FLUX, which specifies exposure shares directly in the prompt. Our method performs best across all five metrics (Table 1), whose definitions are provided in Appendix B.2.

Table 1: Monotonicity between target exposure shares and CTI.

Method	Pearson ↑	Spearman ↑	Pairwise ↑	Kendall's τ_b ↑	Isotonic R^2 ↑
FLUX (Prompt-only)	0.2181	0.2344	0.5727	0.1837	0.2802
Sun et al. (2025)	0.6675	0.6630	0.7343	0.5332	0.7708
Ours	**0.7222**	**0.7370**	**0.7603**	**0.6027**	**0.8126**

We further visualize the average CTI under different target-share configurations for both the two- and three-advertiser settings in Figure 2. The results are aggregated over 1,000 scenarios. The curves show a clear increasing trend as each product's target share increases. These results show that varying the target shares produces consistent changes in the effectiveness of the resulting product exposure, as measured by CTI, thereby yielding a controllable set of generation configurations for the subsequent auction.



(a) Two advertisers.

(b) Three advertisers.

Figure 2: Monotonicity between target share and CTI.

Ablation study of our method. We evaluate the complementarity of RSC and SCP across 200 scenes, each with 19 share configurations ranging from $(0.95, 0.05)$ to $(0.05, 0.95)$, as shown in Table 2. The results show that RSC and SCP are complementary, with their combination achieving the best performance across all metrics. Moreover, RSC + SCP outperforms the baseline equipped with the same SCP conditioning, indicating that the improvement cannot be attributed to share-conditioned prompting alone.

Efficiency Comparison. We compare the efficiency of our method with the baseline (Sun et al., 2025) under generation settings involving different numbers of products. We evaluate both methods

Table 2: Ablation of RSC and SCP and comparison with the SCP-enhanced baseline on 200 scenes.

Method	Pearson ↑	Spearman ↑	Pairwise ↑	Kendall's τ_b ↑	Isotonic R^2 ↑
SCP only	0.3039	0.3172	0.5989	0.2430	0.4897
RSC only	0.3730	0.3774	0.6236	0.2990	0.4964
Baseline + SCP	0.7651	0.7837	0.7797	0.6430	0.8880
RSC + SCP	**0.8120**	**0.8346**	**0.8024**	**0.6981**	**0.9265**

using the same set of 200 prompts and measure their inference time per denoising timestep. All inference times are measured on one NVIDIA A800 GPU and reported as mean ± standard deviation over ten runs. As reported in Table 3, our method is consistently faster across all product counts, with its efficiency advantage becoming increasingly pronounced as the number of products grows.

Table 3: Per-timestep inference-time comparison.

# Products	Ours (s) ↓	Sun et al. (2025) (s) ↓	Speedup ↑	Time reduction ↑
2	0.6654 ± 0.0002	1.2734 ± 0.0006	$1.91\times$	47.74%
4	0.7317 ± 0.0002	1.6992 ± 0.0004	$2.32\times$	56.94%
6	0.7985 ± 0.0018	2.5485 ± 0.0001	$3.19\times$	68.67%
8	0.8623 ± 0.0003	3.3980 ± 0.0017	$3.94\times$	74.62%
10	0.9289 ± 0.0026	4.2571 ± 0.0012	$4.58\times$	78.18%

4.2 Validation of the CTI Evaluator

The CTI evaluator provides the product-level estimates used to construct the public expected-CTI table. The auction combines this table with advertisers' private values per CTI unit for allocation and pricing. We evaluate its agreement with human judgments and assess the reliability of the human annotations.

Human annotation protocol. We recruited 10 trained annotators to evaluate 1,000 generated images using the same four-dimensional rubric as our CTI evaluator, with each annotator rating 100 images. Product order was randomized, and target exposure shares were hidden. Annotators scored each product based on the generated image, product descriptions, and scene context. Using these human annotations as a reference, we compare two evaluation pipelines: VLM, which directly scores images with InternVL3-14B (Zhu et al., 2025), and VLM–LLM, which uses InternVL3-14B to extract visual evidence and Qwen2.5-7B-Instruct (Yang et al., 2024) to score it. Human and predicted CTI use the same aggregation and normalization. As shown in Table 4, VLM–LLM achieves higher agreement with human judgments across all three metrics, supporting its use as our CTI evaluator.

Table 4: Agreement between predicted CTI and human judgments on 1,000 images from 200 scenes. Results are averaged over the two product positions. Higher values are better.

Evaluator	Pearson ↑	Spearman ↑	ICC(2,1) ↑
VLM only	0.662	0.574	0.610
VLM + LLM (ours)	**0.669**	**0.603**	**0.653**

Inter-annotator agreement. We assess agreement among five annotators on 20 scenes to examine the reliability of human ratings. Ordinal Krippendorff's α measures chance-corrected agreement while accounting for rating order. Exact agreement requires all five annotators to assign the same score. MAE measures the average absolute difference between annotator pairs, with lower values indicating closer agreement.

Table 5: Inter-annotator agreement among five annotators on 20 scenes. PP, BR, SI, and CVC denote four dimensions.

Measure	PP	BR	SI	CVC	CTI
Ordinal α	0.8181	0.9654	0.8342	0.8419	0.9179
Exact (%)	78.0	82.0	66.0	68.0	42.0
MAE	0.1600	0.1400	0.2933	0.2733	0.0450

4.3 Evaluating the Economic Performance of the Auction Mechanism

We evaluate whether learning advertiser weights and configuration offsets improves platform revenue under different structures of private preferences. We conduct separate experiments with two and three advertisers, each using a fixed set of 19 paid configurations and a no-allocation option. Expected CTI is averaged across historical scenes and fixed across valuation profiles, with total configuration value given by $q_{ik}v_{ik}$.

Valuation distributions and baselines. We consider two synthetic valuation distributions. In the *correlated* setting, we independently sample two endpoint coefficients from $\mathrm{Uniform}[0, 1]$ for each advertiser and interpolate its values per CTI unit linearly over exposure shares. This models preferences that vary smoothly across nearby exposure configurations. In the *independent* setting, we sample every v_{ik} independently from $\mathrm{Uniform}[0, 1]$, allowing configuration-specific preferences without the interpolation constraint. We train AMA separately for each setting and select among three training seeds using validation revenue. Revenue is the mean sum of actual payments under truthful reporting and hard configuration selection. We compare against VCG (Vickrey, 1961; Clarke, 1971; Groves, 1973) and a Myerson-based baseline (Myerson, 1981). VCG maximizes total reported value and uses pivot payments. The Myerson-based baseline uses unit advertiser weights and fixed offsets $\beta_k = -\frac{1}{2}\sum_i q_{ik}$ with AMA payments, retaining the uniform single-parameter Myerson scoring rule without claiming optimality in the present multidimensional setting.

Table 6: Mean platform revenue under truthful reporting.

Advertisers	Distribution	VCG	Myerson-based	AMA
2	Correlated	0.1592	0.4036	**0.4242**
2	Independent	0.3699	0.5648	**0.8535**
3	Correlated	0.2019	0.5238	**0.5462**
3	Independent	0.3498	0.6031	**0.9086**

Results. Table 6 shows that learned AMA achieves the highest revenue in all four settings. Under correlated valuations, it improves over the Myerson-based baseline by 5.10% and 4.27% for two and three advertisers, respectively; under independent valuations, the gains reach 51.12% and 50.66%. It also consistently outperforms VCG in platform revenue. These results demonstrate the benefit of adapting auction parameters to configuration-dependent private valuations, with larger gains over the fixed baselines in the independent setting.

5 Conclusion

We presented DiffAuction, a unified framework integrating controllable image generation, CTI evaluation, and truthful auction design. RSC and SCP enable continuous exposure control with one model forward pass per denoising step, while the CTI evaluator provides interpretable product-level click-intent estimates. A learnable affine maximizer combines expected CTI with advertisers' private values to select feasible configurations and determine payments while ensuring DSIC and IR. Experiments demonstrate improved share–CTI monotonicity, generation efficiency, human agreement, and auction revenue.

AI USE STATEMENT

Generative AI tools were used to enhance the clarity, readability, and language quality of the manuscript. All AI-assisted edits were reviewed and verified by the authors. In addition, we used large language models to generate textual data; further details are provided in the Appendix. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide corresponding experimental evaluations for the methods introduced in the main paper. The appendix further provides implementation and evaluation details, including share-control settings, definitions of the monotonicity metrics, share-conditioned prompting strategies, and the complete prompts used for CTI evaluation. We will release the source code and relevant evaluation resources to facilitate reproducibility and future research.

REFERENCES

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.

Peter Pin-Shan Chen. The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems*, 1(1):9–36, 1976.

Edward H. Clarke. Multipart pricing of public goods. *Public Choice*, 11:17–33, 1971. doi: 10.1007/BF01726210.

David Collier, Jody LaPorte, and Jason Seawright. Putting typologies to work: Concept formation, measurement, and analytic rigor. *Political research quarterly*, 65(1):217–232, 2012.

Michael Curry, Tuomas Sandholm, and John Dickerson. Differentiable economics for randomized affine maximizer auctions. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 2633–2641, 2023. doi: 10.24963/ijcai.2023/293.

Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. Fluxspace: Disentangled semantic editing in rectified flow models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13083–13092. IEEE, 2025.

Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pp. 144–155, 2024.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review*, 97(1):242–259, 2007.

Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7511–7522. IEEE, 2023.

Theodore Groves. Incentives in teams. *Econometrica*, 41(4):617–631, 1973. doi: 10.2307/1914085.

MohammadTaghi Hajiaghayi, Sébastien Lahaie, Keivan Rezaei, and Suho Shin. Ad auctions for llms via retrieval augmented generation. *Advances in Neural Information Processing Systems*, 37:18445–18480, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3668–3678, 2015.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7.

Jiayuan Liu, Barry Wang, Jiarui Gan, Tonghan Wang, Leon Xie, Mingyu Guo, and Vincent Conitzer. Incentive-aware multi-fidelity optimization for generative advertising in large language models. In *Proceedings of the Third Conference on Language Modeling*, 2026.

Deborah J MacInnis and Bernard J Jaworski. Information processing from advertisements: Toward an integrative framework. *Journal of marketing*, 53(4):1–23, 1989.

Deborah J. MacInnis, Christine Moorman, and Bernard J. Jaworski. Enhancing and measuring consumers' motivation, opportunity, and ability to process brand information from ads. *Journal of Marketing*, 55(4):32–53, 1991.

Edward F. McQuarrie and David Glen Mick. Visual rhetoric in advertising: Text-interpretive, experimental, and reader-response analyses. *Journal of Consumer Research*, 26(1):37–54, 1999.

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

Thomas J. Reynolds, Charles E. Gengler, and Daniel J. Howard. A means-end analysis of brand persuasion through advertising. *International Journal of Research in Marketing*, 12(3):257–266, 1995.

Cristel Antonia Russell. Investigating the effectiveness of product placements in television shows: The role of modality and plot connection congruence on brand memory and attitude. *Journal of Consumer Research*, 29(3):306–318, 2002.

Linda M. Scott. Images in advertising: The need for a theory of visual rhetoric. *Journal of Consumer Research*, 21(2):252–273, 1994.

Ermis Soumalias, Michael Curry, and Sven Seuken. Truthful aggregation of llms with an application to online advertising. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 145262–145308. Curran Associates, Inc., 2025. doi: 10.52202/085713-4862. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d60b6b7f0ba6bf07d975b3bbdacea702-Paper-Conference.pdf.

Lillian Sun, Henry Huang, Fucheng Warren Zhu, Giannis Daras, and Constantinos Costis Daskalakis. Bidding for influence: Auction-driven diffusion image generation. In *ICML 2025 Workshop on Collaborative and Federated Agentic Workflows*, 2025. URL https://openreview.net/forum?id=v5tF1THvt9.

John L. Turner-Smith, Zimeng Huang, Yuhan Fu, Yihang Zhang, and Tonghan Wang. Evaluating and pricing advertisements in ai-generated responses, 2026.

Hal R Varian. Position auctions. *international Journal of industrial Organization*, 25(6):1163–1178, 2007.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961. doi: 10.1111/j.1540-6261.1961.tb02789.x.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Peiran Yun, Wenxin Xu, Jiayuan Liu, Yihang Zhang, Liang Zeng, Lingkai Kong, and Tonghan Wang. Llm advertisement based on neuron auctions. *arXiv preprint arXiv:2605.08326*, 2026.

Yihang Zhang, Zimeng Huang, Ren Zhai, Yipeng Kang, and Tonghan Wang. Naiad: Initiate data-driven research for llm advertising. *arXiv preprint arXiv:2605.09918*, 2026a.

Zhaowei Zhang, Yuhan Fu, Yihang Zhang, Xiaohan Liu, Ceyao Zhang, Xiaoyuan Zhang, Yipeng Kang, Tonghan Wang, and Yaodong Yang. Pila: Plug-and-play insertion for llm-native advertising. *arXiv preprint arXiv:2607.25590*, 2026b.

Chujie Zhao, Qun Hu, Shiping Song, Dagui Chen, Han Zhu, Jian Xu, and Bo Zheng. Llm-auction: Generative auction towards llm-native advertising. *arXiv preprint arXiv:2512.10551*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

A FURTHER EXPERIMENTS ON THE MONOTONICITY BETWEEN TARGET EXPOSURE SHARE AND CTI

We further evaluate monotonicity in the three-advertiser setting and compare our method with the baseline, as shown in Table 7. The results show that our method continues to outperform the baseline while maintaining strong monotonicity.

Table 7: Monotonicity between target exposure shares and CTI in the three-advertiser setting.

Method	Pearson ↑	Spearman ↑	Pairwise ↑	Kendall's τ_b ↑	Isotonic R^2 ↑
Sun et al. (2025)	0.5861	0.6579	0.7530	0.5449	0.6143
Ours	**0.6839**	**0.7932**	**0.8171**	**0.6685**	**0.6922**

We further increase the number of advertisers to five and ten to examine whether the monotonic relationship can still be maintained. We use 100 scene prompts and generate 2,000 images for evaluation. The results are reported in Table 8 and Figure 3. The results show that even when a larger number of products compete for exposure simultaneously, our method still preserves a clear monotonic relationship between target shares and the resulting CTI, demonstrating its scalability to settings with more advertisers.

Table 8: Monotonicity under increasing numbers of advertised products.

Setting	Pearson ↑	Spearman ↑	Pairwise ↑	Kendall's τ_b ↑	Isotonic R^2 ↑
5 products	0.7084	0.7324	0.7773	0.6037	0.6992
10 products	0.5928	0.6004	0.7162	0.4981	0.5689

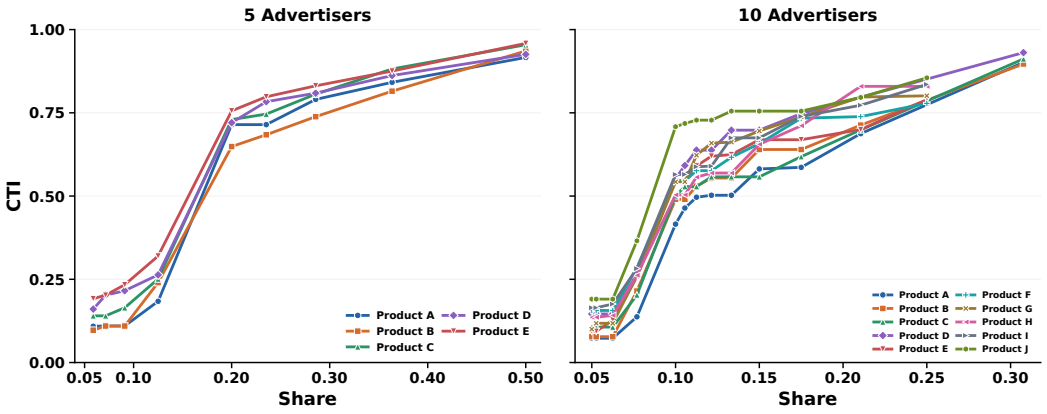


Figure 3: Monotonicity between target exposure share and CTI under extreme numbers of competing products.

In addition, we investigate whether our method preserves monotonicity when the competing products belong to the same category, such as a Luckin coffee cup and a Starbucks coffee cup. We construct 50 scenarios involving competition between products from the same category and generate 1,000 images in total, with two advertisers competing in each image. We then evaluate the monotonicity between CTI and exposure share. The results are shown in Table 9.

Table 9: Share–CTI monotonicity for competing products.

Method	Pearson ↑	Spearman ↑	Pairwise ↑	Kendall's τ_b ↑	Isotonic R^2 ↑
Sun et al. (2025)	0.4794	0.4858	0.6597	0.4007	0.6687
Ours	**0.5600**	**0.5961**	**0.7073**	**0.4964**	**0.6755**

B Additional Details on Share Control

B.1 Further Experiments on RSC

Number of controlled denoising steps. We vary the number of denoising steps during which RSC is applied to study the trade-off between control effectiveness and computational cost. For each setting, we report the monotonicity between the target exposure share and CTI. As shown in Table 10, increasing the control duration from two to eight steps consistently improves all five metrics, indicating that sustained residual calibration strengthens share–CTI monotonicity over the evaluated configurations. The best performance is obtained with eight controlled steps. Extending RSC to all ten denoising steps leads to a slight degradation, suggesting that intervention during the final refinement stage may over-constrain local appearance and disrupt the previously established prominence ordering. Therefore, controlling the first eight steps provides the best balance between monotonicity and computational overhead.

Table 10: Effect of the number of controlled denoising steps on share–CTI monotonicity and inference time. Runtime is averaged over three runs on an NVIDIA A800 GPU.

Control Steps	Pearson ↑	Spearman ↑	Pairwise ↑	Kendall's τ_b ↑	Isotonic R^2 ↑	Runtime (s) ↓
2	0.7581	0.7839	0.7788	0.6413	0.8604	$\mathbf{5.448 \pm 0.003}$
4	0.7875	0.7925	0.7823	0.6522	0.8972	6.422 ± 0.002
6	0.7892	0.8074	0.7920	0.6717	0.8979	7.401 ± 0.002
8	**0.8120**	**0.8346**	**0.8024**	**0.6981**	**0.9265**	8.378 ± 0.006
10	0.8083	0.8249	0.7987	0.6893	0.9184	9.344 ± 0.001

Intervention location. We further investigate where RSC should be applied within FLUX. Specifically, we compare three variants: RSC applied only to the double-stream blocks, only to the single-stream blocks, and to both types of blocks. This analysis examines how the intervention stage affects the resulting share–CTI monotonicity. All variants are evaluated on the same 200 scenes. As shown in Table 11, applying RSC to both the double-stream and single-stream blocks achieves the best performance across all metrics. This is because the double-stream blocks establish early interactions between product tokens and image features, whereas the subsequent single-stream blocks further integrate and refine these multimodal representations. Applying RSC throughout both stages therefore preserves the intended product-wise residual allocation during early alignment and reinforces it during later feature refinement, leading to more consistent share control.

Table 11: Effect of applying residual control to different FLUX transformer streams.

Controlled Blocks	Pearson ↑	Spearman ↑	Pairwise ↑	Kendall's τ_b ↑	Isotonic R^2 ↑
Double-stream only	0.6878	0.7021	0.7444	0.5649	0.7861
Single-stream only	0.6478	0.6605	0.7297	0.5291	0.7588
Double- and single-stream	**0.8120**	**0.8346**	**0.8024**	**0.6981**	**0.9265**

Ablation of residual reweighting. We examine whether normalizing product-specific residual magnitudes improves share control compared with directly scaling residuals by their target shares. Direct scaling uses $\widetilde{R}_i = s_i R_i$, whereas RSC applies $\widetilde{R}_i = (s_i/p_i)R_i$, where $p_i = \|R_i\|_F / \sum_j \|R_j\|_F$. The latter redistributes residual magnitudes according to the target shares while

preserving the sum of their Frobenius norms. Table 12 compares the two strategies in terms of share–CTI monotonicity. The results show that RSC achieves better performance. This is primarily because direct scaling does not ensure that each product's share of the summed residual norms matches its target share. It also reduces the sum of the product-specific residual norms, which may weaken product visibility. RSC compensates for differences in residual magnitudes while preserving the sum of their Frobenius norms.

Table 12: Comparison of direct residual scaling and RSC.

Method	Pearson ↑	Spearman ↑	Pairwise ↑	Kendall's τ_b ↑	Isotonic R^2 ↑
Direct scaling	0.7792	0.7705	0.7687	0.6232	0.8393
RSC	**0.8120**	**0.8346**	**0.8024**	**0.6981**	**0.9265**

B.2 Definitions of Correlation and Monotonicity Metrics

In this section, we introduce five metrics for evaluating the monotonicity between exposure share and CTI. We describe their meanings and calculation methods below.

Pearson correlation. Pearson correlation measures the strength of the *linear relationship* between target share and CTI:

$$r = \frac{\sum_k (s_k - \bar{s})(y_k - \bar{y})}{\sqrt{\sum_k (s_k - \bar{s})^2 \sum_k (y_k - \bar{y})^2}}, \tag{14}$$

where $\bar{s}$ and $\bar{y}$ are the means within the current scene–product curve. A larger value indicates a stronger positive linear association.

Spearman correlation. Spearman correlation measures *rank consistency*, allowing the share–CTI relationship to be nonlinear:

$$\rho = \mathrm{Corr}\big(\mathrm{rank}(\mathbf{s}), \mathrm{rank}(\mathbf{y})\big). \tag{15}$$

Tied CTI values receive their average rank. A larger value indicates that configurations with higher target shares generally receive higher CTI.

Pairwise monotonicity. This metric measures how frequently a higher target share corresponds to higher CTI. With $M = \binom{K}{2}$, we compute

$$P = \frac{1}{M} \sum_{k<\ell} \left[\mathbb{I}(y_\ell > y_k) + \frac{1}{2} \mathbb{I}(y_\ell = y_k) \right]. \tag{16}$$

An increase receives one point, a tie receives half a point, and a decrease receives zero. Thus, $P = 1$ indicates strict increase across all pairs, whereas a constant CTI curve receives $P = 0.5$.

Kendall's τ_b. Kendall's τ_b measures the monotonic association between target share and CTI while accounting for ties. For each pair of configurations with $s_\ell > s_k$, we define the pair as concordant if $y_\ell > y_k$, discordant if $y_\ell < y_k$, and tied if $y_\ell = y_k$. Let C, D, and T denote the numbers of concordant, discordant, and CTI-tied pairs, respectively. Since the target shares contain no ties, Kendall's τ_b reduces to

$$\tau_b = \frac{C - D}{\sqrt{(C + D)(C + D + T)}}. \tag{17}$$

A larger τ_b indicates a stronger monotonically increasing relationship.

Isotonic R^2. Isotonic R^2 measures how closely the CTI curve can be approximated by a nondecreasing sequence:

$$\widehat{\mathbf{y}} = \arg \min_{z_1 \le \cdots \le z_K} \sum_{k=1}^{K} (y_k - z_k)^2, \qquad R^2_{\mathrm{iso}} = 1 - \frac{\sum_k (y_k - \widehat{y}_k)^2}{\sum_k (y_k - \bar{y})^2}. \tag{18}$$

We obtain the fit using the pool-adjacent-violators algorithm. This metric accounts for the *magnitude of departures* from a nondecreasing curve relative to the curve's total variation. A nonconstant, nondecreasing curve achieves one, including curves with plateaus; therefore, a high value does not by itself imply fine-grained share responsiveness.

B.3 SHARE-CONDITIONED PROMPTS

We augment product descriptions with share-conditioned prominence instructions to address limited exposure control under a fixed viewpoint and layout. For each product i, we compute its target share relative to equal allocation:

$$r_i = \frac{s_i}{1/N} = N s_i, \tag{19}$$

where N is the number of products and $\sum_i s_i = 1$. Thus, $r_i = 1$ corresponds to equal allocation, providing a common reference across different numbers of products.

We map each positive relative share to one of five prominence levels:

$$L(r_i) = \begin{cases} 1, & 0 < r_i < 1 - \alpha, \\ 2, & 1 - \alpha \le r_i < 1 - \delta, \\ 3, & 1 - \delta \le r_i < 1 + \delta, \\ 4, & 1 + \delta \le r_i \le 1 + \alpha, \\ 5, & r_i > 1 + \alpha, \end{cases} \tag{20}$$

where $0 < \delta < \alpha < 1$. We use $\alpha = 0.5$ and $\delta = 0.2$, yielding boundaries of 0.5, 0.8, 1.2, and 1.5.

Each level uses the following template, with `[product]` replaced by the corresponding product description:

- **Level 1:** Keep [product] visible but substantially less prominent, with a smaller visual footprint and a more peripheral or background position.
- **Level 2:** Keep [product] less prominent but visible.
- **Level 3:** Keep [product] visible.
- **Level 4:** Frame [product] somewhat more prominently than other products.
- **Level 5:** The camera should face [product] directly and present it as the primary focal product.

These share-conditioned instructions encourage more favorable views of higher-share products while limiting excessive exposure of lower-share products. They complement residual control by enabling viewpoint and layout changes that residual reweighting alone may not achieve under a fixed prompt.

Analysis of threshold values. We further investigate how α and δ affect the resulting monotonicity. Intuitively, α controls the boundary between mild and strong prominence preferences, while δ determines the neutral region around equal allocation. We select three parameter configurations and evaluate their effects on the monotonicity metrics using the same 50 scenes.

Table 13: Hyperparameter search for threshold values.

Advertisers	α	δ	Pearson ↑	Spearman ↑	Pairwise ↑	Kendall's τ_b ↑	Isotonic R^2 ↑
2	0.40	0.15	0.7716	0.7750	0.7802	0.6431	0.8788
2	0.50	0.20	**0.7747**	**0.7829**	**0.7826**	**0.6462**	0.8840
2	0.60	0.25	0.7626	0.7706	0.7758	0.6333	**0.8886**
3	0.40	0.15	0.6899	0.7590	0.7956	0.6370	0.6922
3	0.50	0.20	**0.7042**	**0.7817**	**0.8071**	**0.6579**	**0.7048**
3	0.60	0.25	0.6783	0.7560	0.7920	0.6340	0.6760

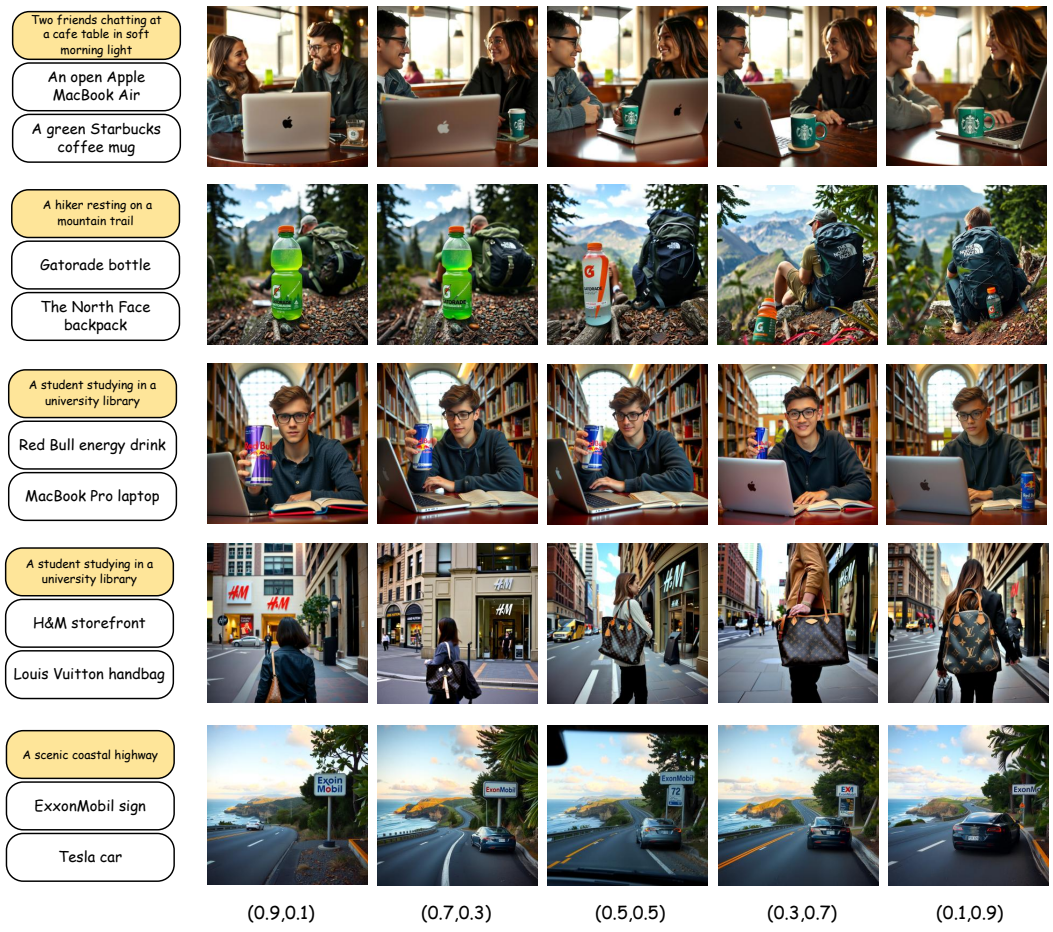


Figure 4: Visualization of exposure-share control.

B.4 Visualization of exposure-share control

To illustrate the effectiveness of our method, we render five scenes, each under five evenly spaced exposure-share configurations, as shown in Figure 4. As the target share of a product increases, its exposure in the generated image also increases accordingly.

C Further Experiments on CTI evaluation

Section 4.2 compares human and model scores under the same four-dimensional rubric. To further examine the relationship between CTI and human click intent, we recruited five additional annotators who had not participated in the earlier rubric-based evaluation. Without access to the four-dimensional rubric, they rated their willingness to click on the products depicted in the images.

The evaluation set comprises 570 images from 30 scenes, with 19 generated images per scene. The images were presented in shuffled order, and each annotator independently rated the complete set. For each evaluated item, we averaged the five annotators' ratings and compared the resulting mean click-intent score with model-predicted CTI.

Our results show that model-predicted CTI achieves Pearson and Spearman correlations of 0.8526 and 0.8339, respectively, with the mean human click-intent ratings, indicating strong linear and rank associations. The ICC(2,1) between model-predicted CTI and the mean human ratings is 0.7694. These results provide further support for CTI as a proxy for average self-reported click intent within the evaluated sample.

D Additional Details on the Auction Mechanism

We train a separate AMA for each advertiser count and valuation distribution, initializing advertiser weights to $w_i = 1$ and paid-configuration offsets to $\beta_k = -\frac{1}{2}\sum_i q_{ik}$ as in the Myerson-based baseline, and subsequently learning both. We parameterize weights as $w_i = \exp(a_i)$ and constrain a_i to $[-1, 1]$ and paid-configuration offsets to $[-2, 1]$ after each update, while fixing the no-allocation offset at zero. We use Adam with a learning rate of 0.003, a batch size of 512, and gradient norm clipping at 10 for $5{,}000$ updates. The softmax temperature decreases exponentially from 0.04 to 0.0005. The no-allocation outcome is included when computing selection probabilities but contributes zero revenue.

During training, we periodically evaluate the model on a fixed validation set of $16{,}384$ valuation profiles. For each profile, we select the highest-scoring configuration and compute platform revenue using the AMA payment rule, retaining the model with the highest mean validation revenue. We evaluate the selected model on $131{,}072$ independent test profiles, shared by all mechanisms within each setting. All reported revenues are computed from payments for the selected configurations, rather than the softmax approximation used during training.

E Proofs of DSIC and IR

We establish dominant-strategy incentive compatibility (DSIC) and individual rationality (IR) for the affine maximizer auction defined in equation 7–equation 9. The guarantees apply to any number of advertisers and any finite configuration set.

Setting and assumptions. Fix the public configuration set $\mathcal{K} = \{0, 1, \ldots, K\}$, the public expected-CTI table $q_{ik} \in [0, 1]$, and the mechanism parameters $\phi = (\mathbf{w}, \boldsymbol{\beta})$, where $w_i > 0$, $\beta_0 = 0$, and $q_{i0} = 0$ for every advertiser. The configuration set, CTI table, and mechanism parameters are fixed independently of the current reports. All score ties are resolved using the same predetermined order over configurations.

Advertiser i has a multidimensional private type $\mathbf{v}_i = (v_{i1}, \ldots, v_{iK}) \in \mathbb{R}_+^K$, where v_{ik} denotes its true value per CTI unit under configuration k. Its total expected value for that configuration is modeled as $V_i(k) = q_{ik}v_{ik}$. The advertiser submits a report $\mathbf{b}_i = (b_{i1}, \ldots, b_{iK}) \in \mathbb{R}_+^K$. For convenience, set $v_{i0} = b_{i0} = 0$. Utilities are quasi-linear: selecting configuration k and charging payment p_i gives utility $q_{ik}v_{ik} - p_i$.

The components of $\mathbf{v}_i$ need not be equal, independent, or monotone across configurations. The proof also applies when admissible type and report vectors belong to a restricted subset of $\mathbb{R}_+^K$, provided truthful reporting is admissible. No distributional assumption on private types is required.

E.1 Report-Independent Menus and Common Optimality

Lemma 1 (Common menu optimality). For every report profile $\mathbf{b}$ and advertiser i, the price menu $\{P_{ik}(\mathbf{b}_{-i})\}_{k\in\mathcal{K}}$ is independent of $\mathbf{b}_i$, and

$$\arg\max_{k\in\mathcal{K}} \{q_{ik}b_{ik} - P_{ik}(\mathbf{b}_{-i})\} = \arg\max_{k\in\mathcal{K}} S_\phi(k; \mathbf{b}). \tag{21}$$

Consequently, applying the shared tie-breaking rule to each advertiser's reported utility-maximizing entries selects the same configuration.

Proof. Fix advertiser i and the other reports $\mathbf{b}_{-i}$. Recall that

$$C_{ik}(\mathbf{b}_{-i}) = \beta_k + \sum_{j\neq i} w_j q_{jk} b_{jk}, \qquad H_i(\mathbf{b}_{-i}) = \max_{\ell\in\mathcal{K}} C_{i\ell}(\mathbf{b}_{-i}).$$

Both quantities are independent of $\mathbf{b}_i$. Since w_i is fixed before bidding, the price

$$P_{ik}(\mathbf{b}_{-i}) = \frac{H_i(\mathbf{b}_{-i}) - C_{ik}(\mathbf{b}_{-i})}{w_i}$$

is also independent of advertiser i's own report.

For every configuration k,

$$q_{ik}b_{ik} - P_{ik}(\mathbf{b}_{-i}) = \frac{w_i q_{ik} b_{ik} + C_{ik}(\mathbf{b}_{-i}) - H_i(\mathbf{b}_{-i})}{w_i} = \frac{S_\phi(k;\mathbf{b}) - H_i(\mathbf{b}_{-i})}{w_i}. \tag{22}$$

The term $H_i(\mathbf{b}_{-i})$ is constant across configurations, and $w_i > 0$. Subtracting this constant and dividing by w_i therefore preserve the ordering of configurations, including ties. This proves equation 21. The common tie-breaking rule selects the same element from these identical maximizer sets.

E.2 Dominant-Strategy Incentive Compatibility

Proposition 1 (DSIC). Under the allocation and payment rules in Equations equation 7 and equation 9, truthful reporting is a weakly dominant strategy. For every advertiser i, true type $\mathbf{v}_i$, alternative report $\mathbf{r}_i$, and profile of other reports $\mathbf{b}_{-i}$, define

$$k^{\mathrm{T}} = k_\phi(\mathbf{v}_i, \mathbf{b}_{-i}), \qquad k^{\mathrm{R}} = k_\phi(\mathbf{r}_i, \mathbf{b}_{-i}).$$

Then

$$\begin{aligned} &q_{i,k^{\mathrm{T}}} v_{i,k^{\mathrm{T}}} - p_i(\mathbf{v}_i, \mathbf{b}_{-i}) \\ &\quad \geq q_{i,k^{\mathrm{R}}} v_{i,k^{\mathrm{R}}} - p_i(\mathbf{r}_i, \mathbf{b}_{-i}). \end{aligned} \tag{23}$$

Proof. Fix i, $\mathbf{v}_i$, and $\mathbf{b}_{-i}$. By Lemma 1, reporting $\mathbf{b}_i = \mathbf{v}_i$ selects a configuration that maximizes advertiser i's true utility over its entire menu:

$$k^{\mathrm{T}} \in \arg\max_{k\in\mathcal{K}} \{q_{ik}v_{ik} - P_{ik}(\mathbf{b}_{-i})\}.$$

A deviation to $\mathbf{r}_i$ may change the selected configuration to k^{R}, but does not change any menu price. Since $k^{\mathrm{R}} \in \mathcal{K}$, optimality of k^{T} implies

$$q_{i,k^{\mathrm{T}}} v_{i,k^{\mathrm{T}}} - P_{i,k^{\mathrm{T}}}(\mathbf{b}_{-i}) \geq q_{i,k^{\mathrm{R}}} v_{i,k^{\mathrm{R}}} - P_{i,k^{\mathrm{R}}}(\mathbf{b}_{-i}).$$

The mechanism charges the price of the selected entry, so this is precisely equation 23.

The deviation $\mathbf{r}_i$ may change arbitrarily many coordinates simultaneously. The argument does not assume that the other advertisers report truthfully. Multiple maximizing configurations do not affect DSIC: they yield the same true utility, so changing between them cannot produce a strict gain.

E.3 Individual Rationality

Proposition 2 (IR). For every advertiser i, nonnegative true type $\mathbf{v}_i$, and profile of other reports $\mathbf{b}_{-i}$, truthful participation yields nonnegative utility:

$$q_{i,k^{\mathrm{T}}} v_{i,k^{\mathrm{T}}} - p_i(\mathbf{v}_i, \mathbf{b}_{-i}) \geq 0, \qquad k^{\mathrm{T}} = k_\phi(\mathbf{v}_i, \mathbf{b}_{-i}). \tag{24}$$

Proof. Fix i, $\mathbf{v}_i$, and $\mathbf{b}_{-i}$. Since $q_{ik} \geq 0$, $v_{ik} \geq 0$, and $w_i > 0$, for every configuration k we have

$$S_\phi(k;\mathbf{v}_i,\mathbf{b}_{-i}) = C_{ik}(\mathbf{b}_{-i}) + w_i q_{ik} v_{ik} \geq C_{ik}(\mathbf{b}_{-i}).$$

Taking maxima gives

$$\max_{k\in\mathcal{K}} S_\phi(k;\mathbf{v}_i,\mathbf{b}_{-i}) \geq \max_{k\in\mathcal{K}} C_{ik}(\mathbf{b}_{-i}) = H_i(\mathbf{b}_{-i}). \tag{25}$$

Using the selected configuration's optimality and equation 22, truthful utility is

$$\begin{aligned} q_{i,k^{\mathrm{T}}} v_{i,k^{\mathrm{T}}} - p_i(\mathbf{v}_i, \mathbf{b}_{-i}) &= \frac{S_\phi(k^{\mathrm{T}};\mathbf{v}_i,\mathbf{b}_{-i}) - H_i(\mathbf{b}_{-i})}{w_i} \\ &= \frac{\max_{k\in\mathcal{K}} S_\phi(k;\mathbf{v}_i,\mathbf{b}_{-i}) - H_i(\mathbf{b}_{-i})}{w_i} \geq 0. \end{aligned}$$

This proves equation 24.

Equivalently, any $\ell_i \in \arg\max_{\ell\in\mathcal{K}} C_{i\ell}(\mathbf{b}_{-i})$ has menu price $P_{i\ell_i}(\mathbf{b}_{-i}) = 0$ and nonnegative true value. Truthful reporting maximizes utility over all entries and therefore performs at least as well as this entry. The zero-priced entry may differ across advertisers; it need not be configuration 0.

E.4 Nonnegative Payments and No Transaction

Proposition 3. All menu prices and realized payments are nonnegative. If configuration 0 is selected, every advertiser's realized payment is zero.

Proof. By definition, $H_i(\mathbf{b}_{-i}) \geq C_{ik}(\mathbf{b}_{-i})$ for every k. Together with $w_i > 0$, this implies

$$P_{ik}(\mathbf{b}_{-i}) = \frac{H_i(\mathbf{b}_{-i}) - C_{ik}(\mathbf{b}_{-i})}{w_i} \geq 0.$$

Hence the selected-entry payment $p_i(\mathbf{b})$ is nonnegative.

Now suppose $k_\phi(\mathbf{b}) = 0$. Because $\beta_0 = 0$ and $q_{i0} = 0$ for every i, we have $S_\phi(0; \mathbf{b}) = 0$. Optimality of configuration 0 therefore gives $S_\phi(k; \mathbf{b}) \leq 0$ for all k. For every advertiser i and configuration k, nonnegative reports and CTI imply

$$C_{ik}(\mathbf{b}_{-i}) = S_\phi(k; \mathbf{b}) - w_i q_{ik} b_{ik} \leq 0.$$

Since $C_{i0}(\mathbf{b}_{-i}) = 0$, we obtain $H_i(\mathbf{b}_{-i}) = 0$. Thus

$$p_i(\mathbf{b}) = P_{i0}(\mathbf{b}_{-i}) = \frac{H_i(\mathbf{b}_{-i}) - C_{i0}(\mathbf{b}_{-i})}{w_i} = 0.$$

For a general profile of other reports, however, the no-allocation entry has menu price $P_{i0}(\mathbf{b}_{-i}) = H_i(\mathbf{b}_{-i})/w_i$, which need not be zero when that entry is not selected. The IR proof does not require configuration 0 to have zero price in every menu.

Learning and deployment. The guarantees hold for every fixed parameter setting with strictly positive advertiser weights. They do not require revenue training to reach an optimum, independence of private types, or monotonicity of CTI in the target exposure shares. The soft selection used during training is an optimization surrogate. The guarantees concern the deployed hard-max allocation and the associated payments in equation 9.

Valuation-model scope. The guarantees are conditional on the valuation model $V_i(k) = q_{ik} v_{ik}$ and apply pointwise in private types and other advertisers' reports. If $V_i(k)$ equals advertiser i's true expected value of the image generated under configuration k, the results imply DSIC and IR in expectation over image generation. A fixed historical CTI estimate alone does not establish this equality. The results do not guarantee nonnegative realized utility for every generated image or identify the CTI proxy with an observed click-through rate.

F Theoretical Basis of Placement-Specific Evaluation Dimensions

This section provides the theoretical basis for our four evaluation dimensions. We first define the scope of factors considered in our framework, then identify two complementary evaluation perspectives within this scope and further divide each perspective into two categories, forming a 2×2 conceptual space. The four evaluation metrics are then positioned as the corresponding operationalizations of the four conceptual cells.

Research scope. We focus on evaluation factors tailored to AIGC-based product placement rather than directly adopting criteria from conventional advertising evaluation. Specifically, we consider factors that are relevant to embedded product advertising and can be reliably assessed from generated images. Highly subjective or abstract judgments, such as overall aesthetics or generic attractiveness, are excluded because they are difficult to define consistently and to evaluate robustly with automated models.

Two-dimensional conceptual space. We adopt the logic of *conceptual typology*, which organizes a complex construct through multiple dimensions and their combinations into a structured property space (Collier et al., 2012). Following this methodology, we characterize placement-specific influencing factors from two complementary perspectives: *Stimulus* and *Processing*. This choice is

grounded in classical advertising information-processing theory, which distinguishes the advertising stimuli or executional cues presented to consumers from the subsequent processing through which these stimuli influence consumer responses (MacInnis & Jaworski, 1989; MacInnis et al., 1991).

Formally, we define

$$\mathcal{D} = \mathcal{S} \times \mathcal{P}, \tag{26}$$

where $\mathcal{S}$ and $\mathcal{P}$ denote the Stimulus and Processing dimensions, respectively.

Stimulus: Product vs. Product–Scene Relation. In our framework, *Stimulus* refers to the visual advertising content from which a placement-specific influence originates. We divide this dimension into the inserted product itself and the relation between the product and its surrounding scene. This distinction is supported by established structural and advertising theories: the Entity–Relationship model treats entities and their relationships as different primitives (Chen, 1976), scene-graph representations similarly separate objects, object attributes, and inter-object relationships in visual scenes (Johnson et al., 2015; Krishna et al., 2017), and product-placement research further shows that how a brand is presented and how it is connected to host content can have different effects on brand memory and attitude (Russell, 2002). Together, these findings support treating product properties and product–scene relational properties as two distinct sources of placement-specific visual influence.

Accordingly, we define

$$\mathcal{S} = \{\text{Product}, \text{Product-Scene Relation}\}. \tag{27}$$

Other visual elements are outside our research scope unless they interact with the inserted product; such effects are subsumed under the *Product–Scene Relation* category.

Processing: Explicit vs. Implicit. In our framework, *Processing* describes how visual advertising content is interpreted by consumers. We distinguish between *Explicit* processing, which relies primarily on directly observable visual configurations such as product size, position, and color, and *Implicit* processing, which requires further interpretation to derive higher-level commercial meanings such as brand identity, product functionality, and perceived value.

Accordingly, we define

$$\mathcal{P} = \{\text{Explicit}, \text{Implicit}\}. \tag{28}$$

This distinction is supported by advertising and consumer-processing theories: Scott (1994) characterizes advertising images as symbolic systems capable of conveying meanings beyond their literal appearance; visual-rhetoric research further shows that consumers construct higher-level meanings from visual configurations (McQuarrie & Mick, 1999); and means–end theory describes a progression from concrete product attributes to increasingly abstract consequences and values (Reynolds et al., 1995).

Four conceptual cells. Crossing the two dimensions yields four conceptual cells, each corresponding to a distinct source and mode of placement-specific influence. We instantiate these cells with four evaluation dimensions, as summarized in Table 14.

Table 14: The 2×2 conceptual organization of placement-specific visual evaluation.

Stimulus	**Processing**	
	Explicit	**Implicit**
Product	Product Prominence	Brand Recognizability
Product–Scene Relation	Scene Integration	Contextual Value Communication

Product Prominence corresponds to the *Product*×*Explicit* cell because it measures the directly observable visual presence of the inserted product, reflected by factors such as its size, position, and color.

Brand Recognizability represents the *Product*×*Implicit* cell. Although brand cues are physically present on the product, their commercial meaning emerges only when visual features such as logos, text, colors, or distinctive product designs are interpreted as a particular brand identity.

Scene Integration corresponds to the *Product–Scene Relation*×*Explicit* cell because it evaluates directly observable compatibility between the product and its surrounding scene, including perspective, scale, illumination, contact, and occlusion.

Contextual Value Communication represents the *Product–Scene Relation*×*Implicit* cell. It captures whether the relationship between the product and its context conveys higher-level commercial meanings, such as product usage, functional benefits, lifestyle associations, or perceived value.

Scope of the completeness claim. We do not claim that the four metrics constitute all determinants of click-through intent, nor that they exhaust every possible property within their corresponding cells. Rather, our completeness claim concerns the placement-specific conceptual space defined above. Within this space, many plausible alternative metrics substantially overlap with the four selected constructs, while others fall outside our scope because they are either not particularly salient to AIGC product placement or cannot be evaluated reliably from generated images by automated models. The four metrics therefore provide compact and operationalizable coverage of the major conceptual aspects relevant to our setting.

Case-based analysis of alternative evaluation criteria. To assess the conceptual completeness of the four metrics within the scope of this study, we examine whether any candidate criterion satisfies both of the following conditions: (i) it falls within our scope of evaluating product placement in generated images; and (ii) its primary evaluation content or focus differs substantially from those of the four existing metrics, warranting its inclusion as a separate dimension. To this end, we analyze representative candidate criteria proposed by advertising experts, together with common behavioral and ad-delivery metrics. For candidates within the research scope, we examine whether they introduce distinct evaluation content or substantially overlap conceptually with the existing metrics. Table 15 summarizes the results. For brevity, PP, BR, SI, and CVC denote Product Prominence, Brand Recognizability, Scene Integration, and Contextual Value Communication, respectively.

The analysis indicates that, among the candidates examined, those within the research scope substantially overlap conceptually with the four existing metrics; the remaining candidates do not belong to this scope because they concern different evaluation targets or require different information. These findings support the adequacy of the four proposed dimensions in capturing the principal aspects of product-placement evaluation within the scope of this study.

Table 15: Selected alternative evaluation criteria and their relationships with the proposed metrics.

Candidate criterion	**Case-based analysis**	**Relation to framework**
Attention Retention	Sustained attention is a temporal behavioral response. PP describes visual prominence but does not directly measure how long viewers continue attending to a product.	Behavioral outcome
Product Recognizability	Identifying a product category or specific item is related to identifying its brand, but these tasks differ. BR concerns brand identity and does not fully represent broader product recognizability.	Related to BR
Viewpoint Favorability	Viewpoint affects product visibility, the visibility of brand cues, and the plausibility of the product's presentation in the scene. These aspects relate to PP, BR, and SI.	PP + BR + SI
Brand–Scene Congruence	Visual compatibility is addressed by SI, while the relevance of the depicted context to product use or commercial meaning is addressed by CVC.	SI + CVC
Brand–Plot Connection	In static images, the observable counterpart of plot connection is the product's meaningful participation in the depicted situation, which is addressed by CVC.	CVC

Candidate criterion	Case-based analysis	Relation to framework
Placement Intrusiveness	Perceived interruption may depend on prominence, integration, and viewers' goals. PP and SI describe relevant visual factors but do not directly measure whether a placement feels intrusive.	Consumer response
Ad–Product Relevance	Whether the surrounding content supports the product's use, benefits, or positioning falls within the contextual meaning evaluated by CVC.	CVC
Informativeness	Depicted uses and benefits overlap with CVC. Other information, such as visible prices or specifications, extends beyond this contextual criterion and is not directly evaluated.	Partial overlap: CVC
Processing Fluency	Clear presentation, identifiable brand cues, and coherent composition may facilitate processing. PP, BR, and SI describe these factors but do not directly measure processing fluency.	Processing response
Attitude toward the Brand	Brand attitude reflects consumer evaluations that may predate the image and vary with experience and preference. Such consumer-specific information is outside our evaluation scope.	Consumer-specific
Brand Recall / Memorability	Retention and retrieval of brand information require post-exposure assessment. BR evaluates identification from the current image, rather than subsequent memory performance.	Behavioral outcome
Overall Aesthetics	Overall aesthetics concerns the composition, style, and artistic quality of the entire image. It may influence advertising responses but is not directly evaluated by our placement-specific metrics.	Out of scope
Creativity / Novelty	Creativity and novelty concern the originality of a placement relative to prior examples and expectations. They are distinct design qualities not directly measured by the four dimensions.	Out of scope
Persuasiveness	Persuasiveness concerns changes in consumer beliefs, attitudes, or intentions. It depends on message interpretation and consumer characteristics beyond the placement properties assessed here.	Consumer response
Purchase Intention	Purchase intention depends on product need, price, trust, and purchasing constraints in addition to visual presentation. These factors extend beyond our content-based evaluation.	Consumer response
User–Product Relevance	Relevance to a particular user depends on that user's interests, needs, and preferences. It requires audience information beyond the depicted product–scene relationship.	Consumer-specific
Historical CTR	Observed CTR requires impression and click records and reflects both image content and delivery conditions. Our CTI score is a content-based proxy, not an observed click rate.	Behavioral outcome
Platform Placement	Feed position, device, screen size, and neighboring content characterize the delivery environment. These factors are outside the generated image's product-level presentation.	Delivery context

G SCENARIO DATA SYNTHESIS

We use GPT-5.6-Luna to synthesize both scene descriptions and product descriptions. The synthesis procedure begins by generating 1,500 candidate scene descriptions. These descriptions are normalized and deduplicated, after which 1,000 distinct scenes are retained. Let the resulting scene set be

$$\mathcal{S} = \{s_1, s_2, \ldots, s_{1000}\}.$$

Products are generated separately according to a predefined set of product categories. The same number of candidate products is generated for each category to maintain a balanced category distribution. After normalization and deduplication, 400 distinct products are retained while preserving this balance. The resulting product set is denoted by

$$\mathcal{P} = \{p_1, p_2, \ldots, p_{400}\}.$$

Each scene is assigned two products. GPT-5.6-Luna determines whether a product is compatible with a scene by considering whether the product can naturally appear in the location, activity, and context described by that scene. The two selected products must also be able to appear together naturally within the same scene. Products that introduce an evident semantic or contextual conflict are excluded from the assignment.

To ensure a uniform product distribution, the maximum number of assignments for every product is set to five:

$$c(p_i) \leq 5, \qquad \forall p_i \in \mathcal{P},$$

where $c(p_i)$ denotes the number of scenes to which product p_i is assigned. Because each of the 1,000 scenes contains two products, the dataset contains exactly

$$1000 \times 2 = 2000$$

product positions. The total assignment capacity of the 400 products is also

$$400 \times 5 = 2000.$$

Therefore, once all product positions have been filled without exceeding the per-product limit, every product necessarily appears exactly five times:

$$c(p_i) = 5, \qquad \forall p_i \in \mathcal{P}.$$

During assignment, the remaining capacity of each product is updated after every selection. For a given scene, two compatible products are selected only from products whose remaining capacities are greater than zero. If no valid product pair can be assigned to a scene, two products that have not yet reached their required numbers of assignments are selected first. GPT-5.6-Luna then generates a new scene in which both products can naturally coexist. The new scene is deduplicated against the retained scene set and replaces the scene for which no compatible assignment could be found.

This procedure continues until all 1,000 scenes have been assigned two compatible products and all 400 products have been assigned exactly five times. Each final data record consists of one scene prompt and two corresponding product prompts.

For scenes containing three or more products, we follow the same procedure. Once the total numbers of scenes and products are specified, the required dataset can be synthesized accordingly.

H PROMPTS FOR CTI EVALUATION

In this section, we provide the prompts used for CTI prediction, including the prompt for the VLM to extract visual evidence from images and the prompt for the LLM to score the extracted visual evidence.

H.1 VISUAL EVIDENCE EXTRACTION PROMPT

Visual-evidence extraction prompt

```
You are a visual analysis expert. Examine the input image and extract
    visual evidence for each target product for subsequent model scoring
    .

Scene/context:
{base_prompt}

Target products:
product_A: {agent1_prompt}
product_B: {agent2_prompt}

General Requirements:

- If a target product is absent, state "not visible." If it cannot be
    reliably matched to an object in the image, state "not clearly
    visible" or "uncertain." Do not infer its brand, location, or
    relationship to the scene.

- Distinguish between product visibility and confirmation of the target
     brand. Describe all logos, text, and patterns that are actually
    visible. If text is illegible, explicitly state so rather than
    reconstructing or completing it.

- If an object that may correspond to a target product displays
    branding inconsistent with the specified target brand, explicitly
    describe the visible branding and the mismatch. Do not attribute
    branding found on other objects to the target product.

- Describe visible facts first, followed only by limited
    interpretations supported by those facts. Do not replace concrete
    evidence with vague conclusions such as "highly congruent," "highly
    recognizable," or "naturally integrated."

Extract evidence for each product along the following four dimensions:

1. Product Prominence

Describe the product's position, depth within the scene, visible area,
    occlusion or cropping, brightness, clarity, and other relevant
    visual properties.

Compare the two target products in terms of visible area, proportion of
     the image occupied, and overall visual prominence. If a comparison
    cannot be made, explain why. Do not equate visibility with
    prominence, infer prominence from size alone, or fabricate precise
    proportions.

2. Brand Recognizability

Record any visible logos, text, patterns, or distinctive design
    features. Explain whether these cues support identification of the
    target brand or only recognition of the general product category. Do
     not confirm a brand solely from the target product name, generic
    colors, or ordinary packaging.

3. Scene Integration

Describe the product's contact with and support from its surroundings,
    as well as its scale, perspective, lighting, shadows, boundaries,
    and occlusion relationships.

Identify any visible floating, interpenetration, ghosting, pasted-on
    appearance, or structural abnormalities. If these properties cannot
```

```
    be assessed, state that they are uncertain. Do not use suitability
    for the usage context as a substitute for physical and visual
    evidence.

4. Contextual Value Communication

Describe whether the scene represents a typical usage context for the
    product, whether it demonstrates the product's use, functional
    benefits, or lifestyle associations, and whether the visible context
     supports the product's promotional communication.

Output Requirements:

- Output valid JSON only.

- If a target product is not visible or cannot be reliably identified,
    state this explicitly in each relevant field without fabricating
    information.

- Do not provide scores, ratings, or conclusions about overall
    advertising effectiveness. Report only visually grounded evidence.

Output Format:

{
  "product_A": {
    "product_prominence_evidence": "",
    "brand_recognizability_evidence": "",
    "scene_integration_evidence": "",
    "contextual_value_communication_evidence": ""
  },
  "product_B": {
    "product_prominence_evidence": "",
    "brand_recognizability_evidence": "",
    "scene_integration_evidence": "",
    "contextual_value_communication_evidence": ""
  }
}
```

H.2 CTI Scoring Prompt

CTI scoring prompt

```
You are a professional advertising effectiveness evaluator.

Below is objective visual evidence extracted by a VLM from a single
    advertisement image. Based only on the provided evidence, assign
    each product an integer score from 1 to 5 on each of the following
    four dimensions, where 5 indicates the best performance and 1
    indicates the worst.

1. Product Prominence
2. Brand Recognizability
3. Scene Integration
4. Contextual Value Communication

Important:

- This is a single-image evaluation task. Do not refer to or compare
    with other images.
- Base your ratings only on the provided VLM visual evidence. Do not
    introduce information beyond the evidence.
- Evaluate each dimension independently.
```

```
- If the target product itself is not visible, do not assign a high
    score merely because the scene contains products from the same
    category, similar products, logos of other brands, or textual
    mentions of the product.
- Do not output explanations, Markdown, code blocks, or your reasoning
    process.

Scene/context:
{Current scene description}

Products to evaluate:
product_A: {Product A description}
product_B: {Product B description}

VLM visual evidence:
{
  "product_A": {
    "product_prominence_evidence": "{VLM-extracted prominence evidence
    for Product A}",
    "brand_recognizability_evidence": "{VLM-extracted brand
    recognizability evidence for Product A}",
    "scene_integration_evidence": "{VLM-extracted scene integration
    evidence for Product A}",
    "contextual_value_communication_evidence": "{VLM-extracted
    contextual value communication evidence for Product A}"
  },
  "product_B": {
    "product_prominence_evidence": "{VLM-extracted prominence evidence
    for Product B}",
    "brand_recognizability_evidence": "{VLM-extracted brand
    recognizability evidence for Product B}",
    "scene_integration_evidence": "{VLM-extracted scene integration
    evidence for Product B}",
    "contextual_value_communication_evidence": "{VLM-extracted
    contextual value communication evidence for Product B}"
  }
}

Scoring criteria:

1. Product Prominence

5. Highly prominent
- The product occupies a large portion of the image and is positioned
    near the visual center or another highly salient location.
- The product exhibits strong contrast with the background in color,
    brightness, texture, or spatial arrangement.
- The product strongly attracts visual attention and serves as a
    primary focus of the image.

4. Relatively prominent
- The product has a clear visual presence, occupying a relatively large
     portion of the image or appearing in a prominent position.
- The product is unobstructed or only slightly occluded.
- The product is easy to notice but is not necessarily the dominant
    focal point of the image.

3. Moderately prominent
- The product is clearly visible, with a moderate size and ordinary
    placement.
- The product has a noticeable visual presence but no clear visual
    advantage over other elements in the image.

2. Low prominence
```

```
- The product is relatively small, positioned away from the visual
    center, or noticeably occluded.
- Other visual elements may distract attention from the product, making
     it unlikely to be noticed immediately.

1. Very low prominence
- The product occupies only a very small portion of the image, is
    severely occluded, or is barely visible.
- The product is difficult for viewers to notice.

2. Brand Recognizability

5. Highly recognizable
- The logo, brand name, packaging, or other distinctive brand cues are
    clear and prominent.
- Brand information is unobstructed and contains no obvious generation
    errors.
- Viewers can quickly and accurately identify the brand.

4. Relatively recognizable
- Clear logos, brand names, packaging, or other brand cues are visible.
- Viewers can identify the specific brand relatively easily, although
    the brand cues are less prominent than at score 5.

3. Moderately recognizable
- Recognizable brand cues are present but may be small, positioned away
     from the visual center, or partially occluded.
- Viewers can still identify the specific brand from the available
    visual information.

2. Low recognizability
- Brand cues are blurred, heavily occluded, too small, or affected by
    noticeable generation errors.
- The product category may be recognizable, but the specific brand is
    difficult to identify reliably.

1. Unrecognizable
- Reliable logos, brand names, packaging, or other distinctive brand
    cues are absent.
- Viewers cannot determine the specific brand based solely on the
    visual information.

3. Scene Integration

5. Highly integrated
- The product is highly consistent with the scene in terms of spatial
    placement, scale, perspective, lighting, shadows, and boundaries.
- Spatial relationships between the product and surrounding objects are
     natural and physically plausible.
- There are no obvious generation artifacts, and the overall appearance
     resembles a real photograph.

4. Well integrated
- The product is naturally integrated into the scene overall.
- Spatial relationships, scale, perspective, and lighting are generally
     plausible, with only minor imperfections.

3. Moderately integrated
- The product is reasonably integrated into the scene but exhibits some
     visible generation artifacts.
- Minor inconsistencies may appear in lighting, boundaries, scale,
    perspective, or local details.

2. Poorly integrated
```

```
- The product is noticeably inconsistent with its surrounding
    environment.
- There may be obvious compositing artifacts, perspective errors,
    abnormal scale, inconsistent lighting, unnatural boundaries, or
    implausible spatial relationships.

1. Not integrated
- The product clearly fails to integrate naturally into the scene.
- Severe artifacts are present, such as floating objects, ghosting,
    structural abnormalities, major scale or perspective errors, or an
    obvious pasted-on appearance.

4. Contextual Value Communication

5. Highly effective
- The scene clearly demonstrates or reinforces the product's use, usage
     scenario, relevance, or core value.
- The product forms clear and meaningful relationships with surrounding
     people, objects, or environmental elements.
- Viewers can quickly understand from the scene why the product is
    useful or appealing.

4. Relatively effective
- The scene is strongly related to the product's use or value.
- The scene helps viewers understand the product's usage context or
    value, although the communication is weaker than at score 5.

3. Moderately effective
- The scene does not conflict with the product and provides some
    context regarding its use or value.
- However, the scene provides only limited reinforcement of the product
    's use, relevance, or value.

2. Low effectiveness
- The scene has only a weak relationship with the product's use or
    value.
- Although the product appears in the scene, the surrounding context
    provides little help in understanding its use, relevance, or value.

1. No effective communication
- The scene fails to communicate the product's use, relevance, or value
    , or clearly conflicts with the intended use conveyed by the product
    .
- Viewers cannot obtain meaningful information about the product's
    value from the scene.

Return only valid JSON. Replace each <score> with an integer from 1 to
    5. Do not output any additional content:

{
  "product_A": {
    "product_prominence": <score>,
    "brand_recognizability": <score>,
    "scene_integration": <score>,
    "contextual_value_communication": <score>
  },
  "product_B": {
    "product_prominence": <score>,
    "brand_recognizability": <score>,
    "scene_integration": <score>,
    "contextual_value_communication": <score>
  }
}
```