CRITERION-AWARE FACTORIAL OPTIMIZATION WITH APPLICATIONS IN BLACK-BOX TEXT-TO-IMAGE SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt optimization is commonly formulated as free-form search over instructions using a scalar reward, which obscures both criterion-level failures and the effects of individual prompt operations. We instead formulate prompt optimization as a finite factorial design problem. A transformation is composed of interpretable two-level factors, and each generated output is evaluated using multiple binary criteria. We introduce Factorial Upper Confidence Bound Algorithm (FACT-UCB), which fits criterion-specific factorial logistic models to share information across configurations, estimate main and interaction effects, and enforce explicit pass-probability constraints. The method combines uncertainty-aware adaptive sampling with forced exploration to maintain identifiability. We further establish fixed-confidence recovery of the best feasible transformation and an anytime-valid stopping rule. We conduct a text-to-image case study with seven factors, 128 configurations, nine criteria, and 999 adaptive queries, the fitted model suggests concise, Markdown-structured, positively phrased instructions on an industrial 3D pre-generation dataset. Together, these results provide an explicit framework for interpretable, criterion-aware optimization of black-box generative systems.

1 INTRODUCTION

Generative models are highly sensitive to how a request prompt is expressed. Small changes in instruction order, formatting, level of detail, examples, or constraint wording can substantially alter the generated output. Automatic prompt optimization addresses this sensitivity by searching over natural-language instructions (Zhou et al., 2023; Pryzant et al., 2023; Yang et al., 2024); text-to-image methods similarly learn prompt rewriters or optimize prompt conditioning (Hao et al., 2023; Mo et al., 2024). Most existing approaches, however, search over a rich but effectively unstructured prompt space and select prompts using a scalar reward. For operational applications, a single score can conceal important failure modes. For example, a visually attractive image may crop the target, contain multiple objects, depict the wrong count or orientation, or otherwise violate a requirement imposed by a pipeline. Maximizing average reward therefore need not produce a prompt transformation that satisfies all relevant criteria.

We study prompt optimization as a factorized sequential decision problem. A practitioner first specifies m interpretable two-level *factors*, each representing a reusable prompt operation. A configuration of the m factors, namely $\delta \in \mathcal{D} \subseteq \{0,1\}^m$ selects one level of every factor, where $\mathcal{D}$ is the feasible set. Evaluating the frozen system produces a (possibly random) vector

$$A(\delta) = \big(A_1(\delta), \ldots, A_Q(\delta)\big) \in \{0,1\}^Q$$

of binary criterion outcomes, where there are totally Q of binary evaluation criteria, rather than only a scalar preference score. We model the criterion-wise pass probabilities

$$p_q(\delta) = \mathbb{P}\left(A_q(\delta) = 1\right), \qquad q \in [Q],$$

using factorial logistic response surfaces with main and pairwise effects. The resulting models share information across configurations while retaining the criterion-level structure needed to diagnose

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.

trade-offs and impose hard requirements. They support both an aggregate objective,

$$V(\delta) = \sum_{q=1}^{Q} w_q p_q(\delta),$$

and optional feasibility constraints of the form $p_q(\delta) \geq \tau_q$.

In industrial applications, for example, this formulation is particularly natural when generated images serve as references for 3D model generation. Systems such as DreamFusion, Magic3D, Zero-1-to-3, and TripoSR rely on two-dimensional priors or reference images (Poole et al., 2023; Lin et al., 2023; Liu et al., 2023; Tochilkin et al., 2024). A visually pleasing image is not necessarily a useful 3D reference: complete visibility, a clean background, a single object, correct count and orientation, and coherent style are distinct operational properties. We therefore optimize their pass probabilities separately instead of collapsing them immediately into an undifferentiated reward. Importantly, the present case study measures *VLM-judged proxies* for 3D suitability; it does not establish that the selected transformation improves downstream 3D reconstruction. Although our object $V(\delta)$ has a form of reward function, it has a fundamental difference between the common reward functions. We parameterize $p_q(\delta)$, modeling the effectiveness of different factors and their interactions, which is elaborated in Section 3.

Prompt selection has also been formulated as a pure-exploration bandit problem (Shi et al., 2024), including recent work on Pareto-prompt recovery and best-feasible prompt identification (Li et al., 2026). Our setting is complementary but more structured. Complete prompts are not treated as unrelated arms: each arm is a compositional combination of interpretable factors, each observation contains multiple binary criterion labels, and a generalized-linear factorial model transfers information through shared main and pairwise effects. This structure provides effect-level interpretation and can improve statistical efficiency, but it also introduces explicit model-specification, excitation, and evaluator-reliability assumptions that must be examined.

We summarize our contributions as follows.

- We formulate prompt optimization for (frozen) black-box generative systems as a criterion-aware factorial model, allowing a weighted objective together with explicit criterion-wise feasibility constraints.
- For our model, we propose the Factorial Upper Confidence Bound algorithm (FACT-UCB), which combines criterion-specific factorial logistic models with uncertainty-aware allocation, optional evaluator calibration, and a forced-exploration mechanism that preserves information across the factorial design.
- Under proper assumptions, we establish fixed-confidence recovery of the best feasible configuration and provide an anytime-valid stopping rule.
- As an industry-level application, we conduct an exploratory text-to-image study with seven factors, $2^7 = 128$ configurations, 9 binary criteria, and 999 adaptive queries. We further conduct post-experiment tests, robustness experiments, and benchmark comparisons. We give the optimal factor configurations in the experimental setting that could improve the effectiveness of such industrial application.

Related works. Our work is most closely related to bandit-based prompt selection and online factorial experimental design. TRIPLE (Shi et al., 2024) formulates prompt optimization as fixed-budget best-arm identification over a finite set of candidate prompts, while Li et al. (2026) extend this perspective to multi-objective pure exploration, including Pareto-prompt recovery and best-feasible prompt identification. These methods primarily treat complete prompts as arms and seek to estimate or compare their reward vectors. In contrast, we construct each transformation as a combination of interpretable two-level factors and fit a separate logistic response surface for each binary criterion, thereby sharing information across configurations while estimating criterion-specific main and interaction effects. On the experimental-design side, Haizler & Steinberg (2021) combine two-level factorial structure, Bayesian probit modeling, and bandit allocation for online experiments, but focus on balancing effect estimation with cumulative reward rather than certifying the best configuration subject to multiple criterion-wise constraints. Similarly, Fiez et al. (2019) instantiate a factorial arm set with intercept, main-effect, and pairwise-interaction features in transductive linear

best-arm identification, but assume a single linear response and do not address binary criterion vectors, feasibility thresholds, or evaluator uncertainty. Thus, our contribution is not the generic use of UCB, factorial features, or multi-objective optimization, but their integration into a criterion-aware prompt-optimization framework with forced exploration, interpretable probability-scale effects, fixed-confidence best-feasible recovery, an anytime-valid stopping rule, and post-experiment inference.

2 PROBLEM FORMULATION

Let $x \sim \mathcal{D}_{\text{prompt}}$ be a base prompt and $\xi \sim \mathcal{D}_\xi$ be the generation randomness. A (frozen) black-box model M receives the transformed prompt $T_\delta(x, c)$, where c is fixed context (or constraints of generation) and $\delta \in \mathcal{D} \subseteq \{0,1\}^m$ is the operations we conduct to the original prompt. Its output is denoted as

$$I(x, \delta, \xi) = M(T_\delta(x, c); \xi)\,. \tag{1}$$

For criterion $q \in [Q]$, the latent outcome $A_q := A_q(I(x, \delta, \xi)) \in \{0, 1\}$ indicates whether the output passes a prespecified question. Define

$$p_q(\delta) = \mathbb{E}_{x,\xi}[A_q(x, \delta, \xi)], \qquad V(\delta) = \sum_{q=1}^{Q} w_q p_q(\delta), \quad w_q \geq 0, \tag{2}$$

with $\sum_q w_q = 1$. If $C \subseteq [Q]$ contains mandatory criteria with thresholds τ_q, the population-feasible set and target are

$$\mathcal{D}_0 = \{\delta \in \mathcal{D} : p_q(\delta) \geq \tau_q \; \forall q \in C\}, \qquad \delta^\star = \arg\max_{\delta \in \mathcal{D}_0} V(\delta). \tag{3}$$

Each round selects δ_t from past information, then samples a fresh x_t and ξ_t, generates I_t, and observes all Q evaluations.

3 METHOD

For two-level factors, define the intercept, main effects, and pairwise interactions

$$z(\delta) = \big(1, \delta_1, \ldots, \delta_m, \{\delta_j \delta_k : j < k\}\big)^\top \in \mathbb{R}^d, \quad d = 1 + m + \binom{m}{2}. \tag{4}$$

For each criterion, let

$$p_q(\delta) = \sigma\big(z(\delta)^\top \beta_q^\star\big) = \sigma\left(\beta_0 + \sum_{j=1}^{m} \delta_j \beta_j + \sum_{i<j} \beta_{ij} \delta_i \delta_j\right), \qquad \sigma(u) = (1 + e^{-u})^{-1} \tag{5}$$

The coefficients identify which factors affect which criteria and where pairs act synergistically or antagonistically, modeling the interactions of the pairs. This low-order model is a testable approximation, not a consequence of the black-box generator. A main-effect-only model and a more flexible surrogate are therefore essential ablations.

With exact labels, after t observations we estimate

$$\widehat{\beta}_{q,t} \in \arg\min_{\beta \in \Theta} \frac{1}{t} \sum_{s=1}^{t} \Big[\log(1 + e^{z_s^\top \beta}) - A_{sq} z_s^\top \beta\Big] + \frac{\lambda_{q,t}}{2} \|\beta_{-0}\|_2^2, \tag{6}$$

where $z_s = z(\delta_s)$ and $\beta_{-0} := (\beta_1, \cdots, \beta_{d-1}) \in \mathbb{R}^{d-1}$ (the intercept is unpenalized). In practice, one may choose other penalization terms according to different practical problems. We summarize the complete algorithm in Algorithm 1.

Ideally, A_q is evaluated accurately. For example, one may obtain A_q by human experts. However, if one uses visual language models or other substitute models for evaluation to get a higher evaluating speed or lower evaluating budget, one may encounter an estimation error. Such error is nevertheless

recoverable. Specifically, let $\widetilde{A}_q$ be the label of an biased evaluator. On an independent human-audited set, estimate the false-positive and false-negative rates

$$a_q = \mathbb{P}(\widetilde{A}_q = 1 \mid A_q = 0), \qquad b_q = \mathbb{P}(\widetilde{A}_q = 0 \mid A_q = 1).$$

If errors are conditionally class-dependent and $a_q + b_q < 1$, the observed-label probability is

$$m_q(\delta) = a_q + (1 - a_q - b_q)\sigma(z(\delta)^\top \beta_q). \tag{7}$$

Replacing the Bernoulli mean in (6) by m_q targets the latent probability. Calibration-set uncertainty must be propagated, for example by a prompt-level bootstrap that re-estimates (a_q, b_q, β_q). Our theoretical analysis assumes exact labels, which does not certify an unaudited VLM.

4 Fixed-confidence analysis of FACT-UCB

We consider $C = \varnothing$ and retain $w_q \geq 0$ and $\sum_q w_q = 1$. For each criterion, use the ridge estimator

$$\widehat{\beta}_{q,n} \in \arg\min_{\beta \in \mathbb{R}^d} \left\{ \sum_{s=1}^{n} \left[\log(1 + e^{z_s^\top \beta}) - A_{sq} z_s^\top \beta \right] + \frac{\rho_q}{2} \|\beta_{-0}\|_2^2 \right\}, \tag{8}$$

where $0 \leq \rho_q \leq \rho < \infty$ are fixed. Thus, the penalty coefficient in the average-loss convention is $\lambda_{q,n} = \rho_q / n$. The intercept is unpenalized. If a finite minimizer does not exist, set $\widehat{\beta}_{q,n} = 0$ for that round.

Let $\widehat{p}_{q,n}(\delta) = \sigma(z(\delta)^\top \widehat{\beta}_{q,n})$ and $\widehat{V}_n(\delta) = \sum_q w_q \widehat{p}_{q,n}(\delta)$. On every round not reserved for forced exploration, FACT-UCB selects

$$\delta_{n+1} \in \arg\max_{\delta \in \mathcal{D}} F_n(\delta), \qquad F_n(\delta) = \widehat{V}_n(\delta) + \kappa_n \widehat{s}_{V,n}(\delta), \qquad \kappa_n = \sqrt{2 \log(n+1)}. \tag{9}$$

We make the following assumptions.

(A1) **Response model.** The design δ_s is measurable with respect to the information $\mathcal{F}_{s-1}$ available before round s, and

$$\mathbb{E}[A_{sq} \mid \mathcal{F}_{s-1}] = p_q(\delta_s) = \sigma(z_s^\top \beta_q^\star), \qquad q \in [Q].$$

For a finite constant H,

$$\max_{\delta \in \mathcal{D},\, q \in [Q]} |z(\delta)^\top \beta_q^\star| \leq H.$$

The initial observations satisfy the same condition. The labels from the same output may be dependent.

(A2) **Forced exploration.** The fixed exploration set $E \subseteq \mathcal{D}$ satisfies

$$G_E = \frac{1}{|E|} \sum_{\delta \in E} z(\delta) z(\delta)^\top \succ 0.$$

For deterministic $\gamma > 0$ and $n_{\mathrm{e}} \geq 1$,

$$G_n := \sum_{s=1}^{n} z_s z_s^\top \succeq n\gamma G_E \quad \text{almost surely for all } n \geq n_{\mathrm{e}}.$$

The forced-exploration schedule is deterministic. After n_{e}, every interval of k rounds contains at most $\varepsilon k + 1$ forced rounds, where $0 < \varepsilon < 1$.

(A3) **Uncertainty scale.** There are deterministic constants $0 < \underline{c} \leq \overline{c} < \infty$ such that, for $n \geq n_{\mathrm{e}}$, whenever

$$\max_{\delta \in \mathcal{D},\, q \in [Q]} |z(\delta)^\top \widehat{\beta}_{q,n}| \leq H + 1,$$

the uncertainty used in (9) satisfies

$$\underline{c} \sqrt{z(\delta)^\top G_n^{-1} z(\delta)} \leq \widehat{s}_{V,n}(\delta) \leq \overline{c} \sqrt{z(\delta)^\top G_n^{-1} z(\delta)} \qquad \text{for every } \delta \in \mathcal{D}. \tag{10}$$

Algorithm 1 FACT-UCB: criterion-aware factorial UCB with forced exploration

Require: Design space $\mathcal{D}$, exploration set E, budget B, initial data $\mathcal{H}_0$, weights w, constrained criteria C, thresholds $\tau = (\tau_q)_{q \in C}$, exploration fraction ε, and exploration multipliers $(\kappa_t)_{t \geq 0}$

Ensure: A certified design, or "not certified" together with the best possibly feasible design

$\mathcal{H} \leftarrow \mathcal{H}_0$
Initialize a balanced cyclic ordering of the designs in E
for $t = |\mathcal{H}_0|, \ldots, B-1$ **do**
    Fit the criterion-wise models using $\mathcal{H}$
    for $q = 1, \ldots, Q$ **do**
        Compute the estimated pass probability $\widehat{p}_{q,t}(\delta)$ and confidence bounds $L_{q,t}(\delta), U_{q,t}(\delta)$ for every $\delta \in \mathcal{D}$
    end for
    Compute the estimated objective $\widehat{V}_t(\delta)$ and its standard error $\widehat{s}_{V,t}(\delta)$ for every $\delta \in \mathcal{D}$
    Construct the certified-feasible and possibly-feasible sets

$$\widehat{\mathcal{D}}_t^- \leftarrow \{\delta \in \mathcal{D} : L_{q,t}(\delta) \geq \tau_q \text{ for every } q \in C\},$$

$$\widehat{\mathcal{D}}_t^+ \leftarrow \{\delta \in \mathcal{D} : U_{q,t}(\delta) \geq \tau_q \text{ for every } q \in C\}.$$

    if round $t+1$ is reserved for forced exploration **then**
        $\delta_{t+1} \leftarrow$ next design in the balanced cycle through E
    else if $\widehat{\mathcal{D}}_t^+ \neq \varnothing$ **then**
        Select

$$\delta_{t+1} \in \arg\max_{\delta \in \widehat{\mathcal{D}}_t^+} \left\{ \widehat{V}_t(\delta) + \kappa_t \widehat{s}_{V,t}(\delta) \right\}.$$

    else
        Select

$$\delta_{t+1} \in \arg\max_{\delta \in \mathcal{D}} \min_{q \in C} \{U_{q,t}(\delta) - \tau_q\}.$$

    end if
    Independently sample a fresh input x_{t+1}
    Generate the output using δ_{t+1} according to (1)
    Observe all criterion labels $A_{t+1,1}, \ldots, A_{t+1,Q}$
    Update

$$\mathcal{H} \leftarrow \mathcal{H} \cup \left\{ \left(x_{t+1}, \delta_{t+1}, A_{t+1,1}, \ldots, A_{t+1,Q} \right) \right\}.$$

end for
Refit the criterion-wise models using the final data $\mathcal{H}_B \equiv \mathcal{H}$
Compute $\widehat{V}_B$, $\widehat{\mathcal{D}}_B^-$, and $\widehat{\mathcal{D}}_B^+$
if $\widehat{\mathcal{D}}_B^- \neq \varnothing$ **then**
    return

$$\widehat{\delta}_B \in \arg\max_{\delta \in \widehat{\mathcal{D}}_B^-} \widehat{V}_B(\delta).$$

else
    Choose

$$\widetilde{\delta}_B \in \arg\max_{\delta \in \widehat{\mathcal{D}}_B^+} \widehat{V}_B(\delta),$$

    with the convention that, if $\widehat{\mathcal{D}}_B^+ = \varnothing$, then

$$\widetilde{\delta}_B \in \arg\max_{\delta \in \mathcal{D}} \min_{q \in C} \{U_{q,B}(\delta) - \tau_q\}.$$

    return "not certified" and $\widetilde{\delta}_B$
end if

Assumption (A2) is supplied by a balanced forced-exploration schedule on a spanning set. Assumption (A3) concerns the scale of the uncertainty computed by FACT-UCB; it does not assume confidence

coverage. For a criterion-wise Laplace calculation, it follows from bounds on the fitted logistic Hessians, as verified in the appendix.

Write

$$P = \mathrm{diag}(0, 1, \ldots, 1), \qquad \mu_H = \sigma(H)\{1 - \sigma(H)\}, \qquad L_E = \max_{\delta \in \mathcal{D}} \sqrt{z(\delta)^\top G_E^{-1} z(\delta)}.$$

Choose $R_E \geq \max_q \|P\beta_q^\star\|_{G_E^{-1}}$. One valid choice is $R_E = H\|G_E^{-1/2} P G_E^{-1/2}\|_{\mathrm{op}}$.

Fix $\alpha \in (0, 1)$. For $n \geq 1$, define

$$h_n(\alpha) = \log\left\{\frac{\pi^2 Q(1 + \log_2 n)^2}{6\alpha}\right\} + \frac{d}{2}\log\left(1 + \frac{2L_E^2}{d\gamma}\right), \tag{11}$$

$$a_n(\alpha) = \sqrt{h_n(\alpha)} + \frac{\rho R_E}{\sqrt{n\gamma}}. \tag{12}$$

Choose a deterministic $n_0 \geq n_{\mathrm{e}}$ such that, for every $n \geq n_0$,

$$\frac{L_E a_n(\alpha)}{\mu_H \sqrt{n\gamma}} \leq \frac{1}{2}, \qquad \frac{a_n(\alpha)}{2\mu_H} \leq \underline{c}\,\kappa_n. \tag{13}$$

Such an n_0 exists because $a_n(\alpha) = O(\sqrt{\log\log n})$, whereas $\kappa_n = \sqrt{2\log(n+1)}$. FACT-UCB may run before n_0; this index only specifies when certification begins.

For fixed-confidence recommendation, let

$$\widehat{\delta}_n \in \arg\max_{\delta \in \mathcal{D}} \widehat{V}_n(\delta), \qquad L_{V,n}(\delta) = \widehat{V}_n(\delta) - \kappa_n \widehat{s}_{V,n}(\delta).$$

For a prescribed tolerance $\eta > 0$, stop at

$$\tau_\eta = \inf\left\{n \geq n_0 : \max_{\delta \in \mathcal{D} \setminus \{\widehat{\delta}_n\}} F_n(\delta) - L_{V,n}(\widehat{\delta}_n) \leq \eta\right\}, \tag{14}$$

and return $\widehat{\delta}_{\tau_\eta}$. We use $\max \varnothing = -\infty$.

Theorem 1 (Fixed-confidence guarantee for FACT-UCB)**.** *Suppose (A1)–(A3) hold, and let n_0 satisfy* (13)*. For any deterministic $N > n_0$ satisfying*

$$(1 - \varepsilon)(N - n_0) - 1 > \frac{d\log\left(1 + \dfrac{(N - n_0)L_E^2}{dn_0\gamma}\right)}{\log\left(1 + \dfrac{\eta^2}{4\overline{c}^{\,2}\kappa_N^2}\right)}, \tag{15}$$

FACT-UCB satisfies

$$\mathbb{P}\left(\tau_\eta \leq N, \quad V(\delta^\star) - V(\widehat{\delta}_{\tau_\eta}) \leq \eta\right) \geq 1 - \alpha.$$

A finite N satisfying (15) *exists.*

More precisely, on a common event of probability at least $1 - \alpha$, every non-forced round $n + 1$ with $n_0 \leq n < \tau_\eta$ satisfies

$$\kappa_n \widehat{s}_{V,n}(\delta_{n+1}) > \eta/2 \tag{16}$$

and

$$\log\frac{\det G_{n+1}}{\det G_n} > \log\left(1 + \frac{\eta^2}{4\overline{c}^{\,2}\kappa_n^2}\right). \tag{17}$$

The theorem uses the maximization in (9). The idea of proof mainly refers to Faury et al. (2020). Before certification, FACT-UCB must select a design with a large uncertainty bonus. Every such selection increases the information matrix by the amount in (17). The sample-size bound therefore counts the non-forced UCB observations, rather than relying only on forced exploration.

When $\eta \leq 2\overline{c}\,\kappa_N$, a simpler sufficient condition for (15) is

$$(1 - \varepsilon)(N - n_0) - 1 > \frac{16\overline{c}^{\,2} d\log(N + 1)}{\eta^2}\log\left(1 + \frac{(N - n_0)L_E^2}{dn_0\gamma}\right). \tag{18}$$

For a fixed problem and confidence level, the right-hand side grows as $O(\eta^{-2}\log^2 N)$.

Table 1: Seven two-level prompt factors in the reported case study.

Factor	Level 0	Level 1
Position	Subject, then constraints	Constraints, then subject
Format	Natural-language paragraphs	Markdown headings/bullets
Density	Long and detailed	Short and concise
Exemplar	Include worked example	Instructions only
Camera/light	3/4 front; default light; no shadow	Orthographic; flat studio light
Schema	Separate suffix block	Constraints integrated inline
Polarity	State what to avoid	State desired properties positively

Corollary 1 (Exact recovery)**.** *Under the assumptions of Theorem 1, suppose $\delta^\star$ is unique and*

$$\Delta = V(\delta^\star) - \max_{\delta \in \mathcal{D}\setminus\{\delta^\star\}} V(\delta) > 0.$$

For any $0 < \eta < \Delta$, a budget N satisfying (15) *guarantees*

$$\mathbb{P}\left(\tau_\eta \le N, \quad \widehat{\delta}_{\tau_\eta} = \delta^\star\right) \ge 1 - \alpha.$$

In particular, a known lower bound $0 < \underline{\Delta} \le \Delta$ permits the choice $\eta = \underline{\Delta}/2$.

5 Experiment: Pre-3D Generation case study

5.1 Experiment Setup

The case study targets optimizing text-to-image models intended as references for 3D game assets, which is a real-world industry-level dataset. Its pipeline samples a base object prompt, rewrites it with GPT-4o according to δ, generates an image with `google/gemini-2.5-flash-image`, and asks GPT-4o to return nine defect indicators, which are inverted to passes. The available draft names six rubric dimensions—complete visibility, clean background, a single object, count, orientation, and style—but does not state all nine questions verbatim. Releasing the full rubric is necessary for reproducibility. The aggregate uses $w_q = 1/9$ and no hard constraints, so this run is the special case $C = \varnothing$.

The prompt pool contains 10,123 base prompts: 4,096 synthetic prompts (608 simple, 608 LLM-enriched, and 2,880 mixed-concept) and 6,027 production-derived prompts whose natural syntax is retained. We take a subset of 999 queries of prompts. There are $m = 7$ factors and $2^7 = 128$ configurations. Table 1 names both levels; a zero is a reference level, not necessarily "no transformation," and $\delta = 0$ is therefore not the raw-prompt baseline. When $\delta = 1$, it means we take transformation on that factor.

The fitted surrogate has $d = 29$ terms per criterion and uses ridge penalty 10^{-2}, a Laplace covariance approximation, and $\kappa_t = \sqrt{2\log(t+1)}$. A random warm start is followed by UCB for a total of 999 simulations, with checkpoints every 50 iterations.

5.2 Reported discovery result

After 999 queries, the highest fitted configuration is

$$\widehat{\delta} = (1,1,1,0,0,1,1), \qquad \widehat{V}(\widehat{\delta}) = 0.9345, \quad \widehat{s}(\widehat{\delta}) = 0.0076.$$

It uses constraints first, Markdown, concise wording, a worked example, the 3/4-front camera/light level, inline constraints, and positive phrasing. The reference-level vector has fitted value 0.746, so the surrogate difference is 0.1885 (25.3% relative to 0.746). This is not a paired held-out estimate. It is susceptible to winner's curse because the same adaptive data selected and evaluated the winner.

The top two differ only in position and their fitted values differ by 0.0002, far less than either marginal standard error. The available data therefore support a near-optimal set more naturally than a unique empirical winner. Markdown, concise wording, positive polarity, and camera/light level 0 recur among the top configurations, but recurrence in a selected top list is not a marginal causal-effect estimate.

Table 2: Top fitted configurations after the single reported discovery run. The uncertainty column is the manuscript's model-based standard error, not a held-out confidence interval.

Rank	Configuration δ	$\widehat{V}(\delta)$	$\widehat{s}(\delta)$
1	$(1, 1, 1, 0, 0, 1, 1)$	0.9345	0.0076
2	$(0, 1, 1, 0, 0, 1, 1)$	0.9343	0.0078
3	$(1, 1, 1, 1, 0, 0, 1)$	0.9322	0.0081
4	$(1, 1, 1, 0, 0, 0, 1)$	0.9317	0.0083
5	$(1, 1, 1, 1, 0, 1, 1)$	0.9295	0.0090
–	Reference levels $(0, 0, 0, 0, 0, 0, 0)$	0.7460	-

Factor-specific conclusions should come from model coefficients with simultaneous intervals or prespecified factorial contrasts.

Post-experiment Test. To test whether the ranking's ordering survives on freshly rendered images, we ran a confirmatory comparison between the optimal configuration and a challenger drawn uniformly at random from the other 127 designs under a logged seed. The draw returned the rank-10 design, so the ranking separates the two by only 0.0230 and they differ in just two of the seven operations, `format` and `camera_light` (Table 3) — a considerably harder test than a random draw would typically produce. The design is a randomized complete block on the prompt: 60 prompts, each rendered under both configurations, two renders per cell sharing a single temperature-0 rewrite, arm order shuffled within every block, and all images judged by `google/gemini-2.5-flash`. Of the 240 planned images 238 were collected (two image-generation failures in one cell), leaving 59 complete blocks and zero judge parse failures. The paired difference is $+0.0132$ (95% CI -0.0169 to $+0.0433$; paired t, $p = 0.38$; Wilcoxon signed-rank, $p = 0.66$; $d_z = 0.11$), with 22 blocks favouring the optimum, 14 favouring the challenger and 23 tied exactly (Figure 1b). The measured gap carries the sign the ranking predicted and is of the same order as the predicted 0.0230, but it is not separable from zero; equivalence cannot be declared either, since the 90% CI (-0.0120 to $+0.0383$) crosses the pre-registered ± 0.03 margin (Figure 1a). Against the observed within-block difference SD of 0.1155, resolving a gap of the predicted size at 80% power would need roughly 198 blocks rather than 59. Figure 2 gives an example of the optimal configuration outperforms the rank-10 configuration.

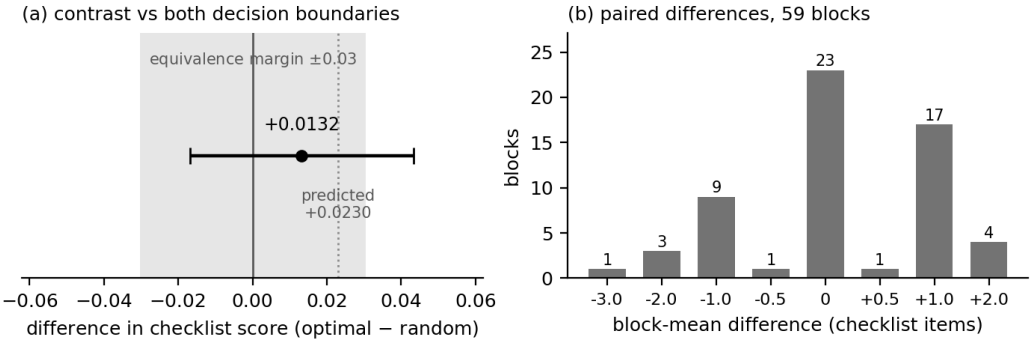


Figure 1: (a) The paired difference in checklist score between the optimal and the randomly drawn configuration, with its 95% confidence interval, shown against both decision boundaries: zero and the pre-registered ± 0.03 equivalence margin (shaded). The interval crosses both, which is why the result is unresolved rather than negative. The dotted line marks the difference the UCB ranking predicted. (b) Distribution of the block-mean differences across the 59 complete blocks, in units of one checklist item ($1/9$ of the scale).

6 Conclusion

In this paper, we formulate criterion-aware prompt optimization as a finite factorial design problem. We propose Fact-UCB, which combines criterion-specific logistic models with adaptive sampling

Table 3: The two configurations. Bits follow the flag order (`position`, `format`, `density`, `exemplar`, `camera_light`, `schema`, `polarity`); the configurations differ only in `format` ($1 \to 0$) and `camera_light` ($0 \to 1$). Measured quantities are over the 59 complete blocks, with the standard error of the block means in parentheses.

	optimal	rank 10
Bits	`1111011`	`1011111`
Rank in the FACT-UCB ranking	1 / 128	10 / 128
Predicted $\widehat{V}$	0.8793	0.8562
Pulls during the FACT-UCB run	75	15
Mean checklist score	0.8701 (0.0139)	0.8569 (0.0161)
All-nine-pass rate	0.271	0.322

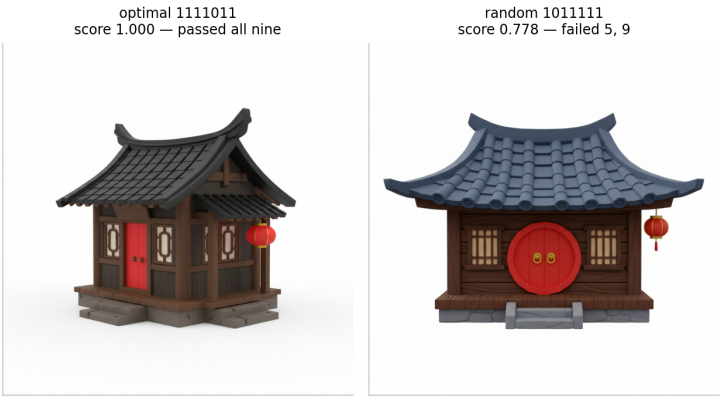


Figure 2: One block rendered under both configurations from the same base prompt (a xianxia cultivator cottage). The optimal configuration produces a genuine three-quarter view with two faces of the building visible; the challenger, which has `camera_light` enabled, produces a flat head-on elevation and the judge flagged both the viewing-angle and the prompt-match checks.

to share information across prompt configurations. The factorial representation allows us to estimate the effects of individual prompt operations and their pairwise interactions, while the criterion-specific formulation supports optimization subject to prescribed feasibility requirements. Under correct model specification and sufficient exploration, we establish finite-sample uniform confidence bounds for ridge logistic estimation and derive guarantees for feasible and near-optimal recommendation. We further provide an anytime stopping rule with almost-sure finite stopping under the stated regularization and margin conditions. Our algorithm applies to general prompt tuning algorithms, such as text/image-to-text/image, 3D generation, video generation, etc., with given criterion evaluations.

Experiments on black-box text-to-image generation illustrate the use of this framework for identifying promising prompt configurations and examining how prompt operations affect different evaluation criteria. The independent evaluations also highlight the uncertainty in fitted rankings and the importance of validating selected configurations on fresh observations. The theoretical guarantees and empirical results together support factorial modeling as a structured approach to criterion-aware prompt optimization, while leaving several questions for further investigation. Future work includes extending the analysis to model misspecification and evaluator errors, developing tighter confidence bounds for practical query budgets, and evaluating the method across additional generation models and downstream tasks.

AI USE STATEMENT

We used Generative AI tools to assist with manuscript drafting and revision, formulating and refining mathematical statements and proof arguments, checking internal consistency and references, and providing feedback on experimental methodology and the interpretation of reported results. The authors take responsibility for the final content of the paper, including its mathematical claims, proofs, experimental results, references, and other materials prepared with AI assistance.

ETHICS STATEMENT

This work studies criterion-aware prompt optimization for black-box text-to-image systems. Optimizing automated evaluation scores may amplify biases or systematic errors in the generation model and evaluator, and may favor outputs that satisfy the specified rubric without meeting broader requirements for quality, fairness, or safety. The reported results should therefore be interpreted within the scope of the stated evaluation criteria and experimental setting. Applications of the method require appropriate safeguards against harmful or misleading content. The experiments use company-confidential data and proprietary algorithms that cannot be publicly disclosed. These restrictions protect confidential company information but also limit external auditing of the empirical findings.

REPRODUCIBILITY STATEMENT

The problem formulation and proposed methodology are described in the main text, and the assumptions and proofs supporting the theoretical results are provided in the theoretical section and appendix. The experimental section and appendix describe the evaluation setting and report the experimental results. The underlying company data, proprietary algorithms, and associated source code are subject to confidentiality restrictions and cannot be publicly released. Consequently, the reported company experiments cannot be fully reproduced from the publicly available materials. The theoretical claims can be assessed from the stated assumptions and proofs, while independent reproduction and verification of results that depend on the confidential resources remain limited.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020.

Tanner Fiez, Lalit Jain, Kevin Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8ba6c657b03fc7c8dd4dff8e45defcd2-Paper.pdf.

Tamar Haizler and David M Steinberg. Factorial designs for online experiments. *Technometrics*, 63 (1):1–12, 2021.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66923–66939, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d346d91999074dd8d6073d4c3b13733b-Abstract-Conference.html.

Donghao Li, Chengshuai Shi, Weijuan Ou, Cong Shen, and Jing Yang. Efficient multi-objective prompt optimization via pure-exploration bandits. In *International Conference on Learning Representations*, pp. 45274–45299, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4b0713b4f15946acaecb0676389715ed-Abstract-Conference.html.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300–309, June 2023.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9298–9309, October 2023.

Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. Dynamic prompt optimizing for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26627–26636, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Mo_Dynamic_Prompt_Optimizing_for_Text-to-Image_Generation_CVPR_2024_paper.html.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FjNys5c7VyY.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7957–7968, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.494. URL https://aclanthology.org/2023.emnlp-main.494/.

Chengshuai Shi, Kun Yang, Zihan Chen, Jundong Li, Jing Yang, and Cong Shen. Efficient prompt optimization through the lens of best arm identification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 99646–99685. Curran Associates, Inc., 2024. doi: 10.52202/079017-3161. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b46bc1449205888e1883f692aff1a252-Abstract-Conference.html.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. TripoSR: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. doi: 10.48550/arXiv.2403.02151. URL https://arxiv.org/abs/2403.02151. Technical report.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3339f19c5fcee3ad74502947a32be9e6-Abstract-Conference.html.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=92gvk82DE-.