Pre-Deployment Advertisement Ranking under Data Scarcity via Context-Aware Criteria Generation with VLMs

Kyungho Kim[1*] **Yeonje Choi**[1*] **Gyurim Hwang**[1] **Sejin Chung**[2]
Hongseok Lee[3†] **Myeong Ho Song**[3] **Yeongho Kim**[1] **Sunwoo Kim**[1]
Jongha Lee[1] **Juyeon Kim**[1] **Kijung Shin**[1†]
[1]KAIST [2]Yonsei University [3]Madup Inc.
[1]{kkyungho, yeonjechoi, rbflaaa, yeongho, kswoo97,
jhsk777, juyeonkim, kijungs}@kaist.ac.kr
[2]{chungsj1462}@yonsei.ac.kr [3]{hs.lee, mhsong}@madup.com

Abstract

Vision-Language Models (VLMs) perform well on general multimodal tasks, yet applying them to real-world advertisement (ad) evaluation is challenging due to strong brand specificity and limited labeled data. We introduce a new practical task, *brand-specific ad ranking*, which aims to rank ads for a target brand prior to deployment by modeling brand-specific effectiveness. To this end, we propose ADVISOR, which derives explicit brand-aware decision criteria using VLMs, augments limited brand context with ads from similar brands, and applies reflection-based scoring for ranking. Experiments on real-world advertising data from 10 brands, collected from actual ad campaigns, show that ADVISOR outperforms strong baselines by up to 7.2%. Further analyses show the generated criteria capture meaningful brand specificity, and ADVISOR also performs strongly in online A/B testing. Our code is available at https://github.com/K-Kyungho/ADvisor.

1 Introduction

While VLMs achieve strong performance on general multimodal tasks (Radford et al., 2021; Li et al., 2022; Liu et al., 2023; Anthropic, 2025; Google, 2023), applying them to real-world business decisions is nontrivial, as such decisions are guided by specific business-related objectives rather than generic visual–textual understanding.

In this context, we introduce a new real-world task, *brand-specific ad ranking*, where the goal is to rank new ads for a target brand before deployment, under two key challenges: **(C1) Data scarcity**: creating ads requires substantial time, budget, and professional effort, and performance labels are only available after running campaigns, resulting in limited brand-specific ads and even fewer labeled examples. **(C2) Lack of decision criteria**: Ad effectiveness is highly brand-specific, shaped by distinct brand identities, target audiences, and marketing goals, requiring brand-specific decision criteria beyond general visual–textual relevance, yet such criteria are rarely explicitly defined.

To address these challenges, we propose ADVISOR, which explicitly generates brand-aware evaluation criteria from brand information and a few sample ads using VLMs, directly addressing **(C2)**. These criteria make explicit what constitutes a good ad for the target brand and guide the scoring of new ads. To mitigate **(C1)**, ADVISOR augments the limited context for the target brand with sample ads from similar brands during the criteria generation. The resulting criteria are then used for reflection-based scoring, and the resulting scores are used as input features to produce the final ranking of ads.

We evaluate ADVISOR on real-world advertising data collected from 10 brands, where each ad is associated with performance labels collected from actual ad campaigns. Results show that ADVISOR outperforms strong baselines by up to 7.2% in ranking performance, and it also performs strongly in online A/B testing, even compared to human experts. Both numerical comparisons and case studies confirm that the generated criteria capture meaningful brand-specific evaluation standards.

Our contributions are summarized as follows:

- **New problem:** We introduce *brand-specific ad ranking*, a practical business task that requires ranking ads for a target brand.
- **Novel solution:** We propose ADVISOR, which leverages VLMs to generate brand-aware evaluation criteria by augmenting the context with ads from similar brands, and applies reflection-based scoring for final ranking.
- **Empirical validation:** We validate the effectiveness of ADVISOR through extensive experiments

[*]Equal contribution. [†]Corresponding authors. A preliminary version of this work was presented at the ICLR 2026 Workshop on Navigating and Addressing Data Problems for Foundation Models, which is non-archival.

on real-world advertising data, including case studies and online A/B testing.

2 Related Works

Click-through rate prediction. Click-through rate (CTR) prediction aims to estimate the probability that a target user clicks on a given ad (or item) (He et al., 2014; Cheng et al., 2016). Such personalization depends on large-scale user–ad interaction logs (Guo et al., 2017; Mao et al., 2023a; Zhang et al., 2025; Li et al., 2025), which are typically available only to platform owners (e.g., Google, Meta) that serve ads to users. In contrast, our problem is motivated by advertisers who create or select ads for target brands or products. For advertisers, the available data is very limited, often consisting of only a small number of past ads with performance labels. This requires a fundamentally different approach from CTR prediction.

Social media popularity prediction. Social media popularity prediction aims to forecast the future popularity of posts, such as the number of views or likes (Wu et al., 2017; Lin and Lee, 2024; Xu et al., 2025; Zhuang et al., 2025). Most approaches are supervised, relying on large-scale, general-domain social media datasets. In contrast, our problem predicts business-critical performance metrics, such as CTR, cost per click (CPC), and cost per mille (CPM), which are typically not publicly accessible. Combined with the brand-specific nature of our task, this leads to severe data scarcity and makes supervised learning ineffective. Moreover, encoders trained for social media popularity prediction are suboptimal for ad ranking, as shown in Section 5.2.

Advertisement understanding and evaluation. Prior work frames ad understanding as a reasoning task, showing that VLMs struggle with high-level persuasive intent (Malakouti et al., 2024), and studies supervised ad performance prediction with LLM-based post-hoc explanations (Yang et al., 2025). In contrast, we study brand-specific real-world ad ranking beyond brand-agnostic analysis, where data scarcity is severe and supervised training becomes suboptimal, as shown in Section 5.2.

3 Task Formulation

Let $\mathcal{B}$ denote the set of brands. For each brand $b \in \mathcal{B}$, we are given a description d_b of the brand, and a labeled dataset $\mathcal{H}_b = \{(v_i, t_i, y_i)\}_{i=1}^{N_b}$, containing N_b tuples, each corresponding to an ad of the brand. Each tuple consists of multimodal components: (i) v_i, the visual content (e.g., images or video frames), (ii) t_i, the textual content (e.g., captions and headlines), and (iii) y_i, the marketing performance metric (i.e., performance label), such as CTR, CPC, and CPM. Given a target brand b and a set of *new* ads $\mathcal{H}_{test} = \{(v'_j, t'_j)\}_{j=1}^{M}$ without performance labels, the goal of *brand-specific ad ranking* is to rank these ads by their expected marketing performance prior to deployment.

Practical relevance. This problem directly captures a core task faced by real-world advertising agencies. Deciding which ads to deploy is a high-stakes business decision: ad production and campaign execution are costly, and replacing underperforming ads after deployment adds further cost and delay. Thus, reliable pre-deployment ranking is critical. Since agencies typically manage campaigns for multiple brands, labeled data across brands are naturally available within the same agency, motivating our cross-brand setting.

4 Method

In this section, we propose ADVISOR for brand-specific ad ranking. ADVISOR consists of three steps: (i) brand-specific criteria generation with cross-brand context augmentation, (ii) self-critique and refinement for scoring, and (iii) brand-specific ranking. ADVISOR is outlined in Figure 1. **Input prompts are given in Appendix A.**

4.1 Brand-specific Criteria Generation with Cross-Brand Context Augmentation

Due to Challenge **C2** in Section 1, naively applying a VLM to directly rank ads is ineffective, even with few-shot demonstrations (see Section 5.2). Without explicit guidance on what defines an effective ad for a particular brand, a VLM tends to rely on generic visual or textual cues, such as overall aesthetics or sentiment, failing to capture brand specificity.

The key idea of ADVISOR is to make such *brand-specific decision criteria* explicit by constructing a set of brand-aware evaluation criteria using VLMs. To elicit these criteria, we first build a contrastive few-shot context from the target brand's past ads by partitioning them into *high-, medium-, and low-performance* groups based on their performance labels.* This contrastive setup encourages the VLM

*When the number of ads is too large to be processed jointly by a VLM, we sample a fixed number from each group.

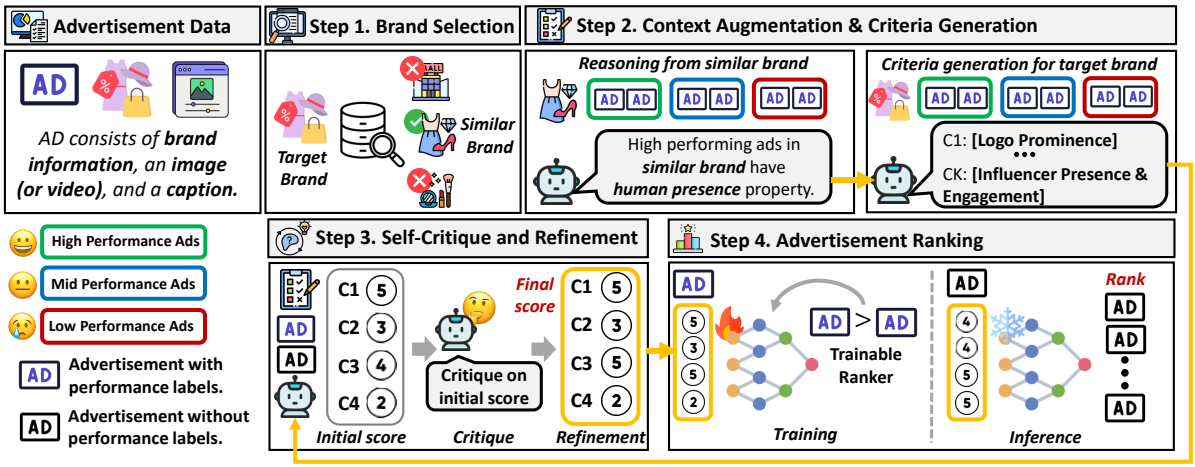


Figure 1: Overview of ADVISOR. It consists of three steps: (1) brand-specific criteria generation, (2) self-critique and refinement, and (3) brand-specific advertisement ranking.

to identify visual or semantic attributes that distinguish successful ads from less effective ones.

However, due to Challenge **C1** in Section 1, relying solely on the target brand often yields too few ads with performance labels to induce stable and diverse decision criteria. To mitigate this challenge, ADVISOR incorporates labeled ads from brands that are similar to the target brand, a process we call *cross-brand context augmentation*. To identify similar brands, we compute a brand embedding $\mathbf{z}_b$ [†] using brand descriptions (d_b), and define the set of similar brands as those with cosine similarity greater than a threshold τ, as follows:

$$\mathcal{S}(b) = \left\{ b' \in \mathcal{B} \setminus \{b\} \;\middle|\; \cos(\mathbf{z}_b, \mathbf{z}_{b'}) \geq \tau \right\} \quad (1)$$

Among the similar brands, we select up to n of the most similar brands, denoted by $\mathcal{S}_n(b)$, and construct their sample ads in the same manner as for the target brand. Importantly, these ads are not directly provided as few-shot examples for the target brand. Instead, the VLM reasons over them to extract high-level, performance-relevant insights, which are subsequently used as auxiliary guidance (see Appendix C.3 for examples of extracted insights). Formally, the brand-specific criteria generation process is defined as

$$\mathcal{C}_b = \mathrm{VLM}\Big(\mathcal{T}_{\text{gen}},\ d_b,\ \mathcal{E}_b,\ \{\phi(\mathcal{E}_{b^*})\}_{b^* \in \mathcal{S}_n(b)}\Big), \quad (2)$$

where the output $\mathcal{C}_b = \{c_1, \ldots, c_k\}$ denotes the generated set of brand-specific evaluation criteria, $\mathcal{T}_{\text{gen}}$ is a task description specifying the objectives of the ad ranking task and criteria generation, $\mathcal{E}_b$ represents few-shot examples sampled from past ads of the target brand, and $\phi(\mathcal{E}_{b^*})$ provides auxiliary context derived from each brand b^* similar to the target brand. Examples of the brand-specific evaluation criteria can be found in Section 5.4.

[†]We use text-embedding-3-large as our embedding model.

4.2 Self-critique and Refinement for Scoring

Given the generated criteria $\mathcal{C}_b$, the VLM scores each ad based on its visual and textual content. As single-pass evaluation may produce inconsistent or weakly grounded judgments, especially when multiple modalities are considered jointly (Zhong et al., 2024; Sarkar et al., 2025), we formulate scoring through self-critique and refinement (Li et al., 2024; Zhang et al., 2024), where initial assessments are revisited and refined. The scoring pipeline consists of three sub-steps: *(i) initial scoring*, *(ii) self-critique*, and *(iii) final refinement*.

Sub-step 1: Initial scoring. For each ad i, we first prompt a VLM to evaluate the ad by assigning an integer score from 1 to 5 with respect to each generated criterion. This step produces an initial score vector along with an explicit textual rationale:

$$\mathbf{e}_i^{\text{init}},\ \mathbf{r}_i^{\text{init}} = \mathrm{VLM}\big(\mathcal{T}_{\text{init}}, \mathcal{C}_b, v_i, t_i\big), \quad (3)$$

where $\mathbf{e}_i^{\text{init}} \in \mathbb{R}^K$ denotes the initial scores of advertisement i, $\mathbf{r}_i^{\text{init}}$ denotes the corresponding rationale, and $\mathcal{T}_{\text{init}}$ denotes the role specification and evaluation instructions provided to the VLM.

Sub-step 2: Self-critique. Next, we introduce a critic VLM that verifies the validity of the initial scoring by assessing whether each criterion-specific score is properly grounded in the observed visual and textual evidence. Specifically, the critic checks for internal inconsistencies, missing visual grounding, or over-interpretation. Formally, this process is defined as

$$\mathbf{r}_i^{\text{crit}} = \mathrm{VLM}\big(\mathcal{T}_{\text{cri}}, \mathbf{e}_i^{\text{init}}, \mathbf{r}_i^{\text{init}}, \mathcal{C}_b, v_i, t_i\big), \quad (4)$$

where $\mathbf{r}_i^{\text{crit}}$ represents the critique of the initial scores and rationales, and $\mathcal{T}_{\text{cri}}$ denotes the critic role specification provided to the VLM.

Sub-step 3: Final refinement. Lastly, the VLM performs self-refinement by explicitly reflecting on the critique feedback $\mathbf{r}_i^{\text{crit}}$ and revising the initial score accordingly as follows:

$$\mathbf{e}_i = \text{VLM}\big(\mathcal{T}_{\text{final}}, \mathbf{e}_i^{\text{init}}, \mathbf{r}_i^{\text{crit}}, \mathcal{C}_b, v_i, t_i\big), \quad (5)$$

where the output $\mathbf{e}_i \in \mathbb{R}^K$ represents the final refined score vector for ad i after refinement, which serves as features for the downstream ranker. Notably, these features, together with the generated evaluation criteria, are human-interpretable.

4.3 Brand-Specific Ranking

Given the score vector $\mathbf{e}_i$, ADVISOR performs brand-specific ad ranking using a trainable ranker. While using a VLM as the final ranker is a possible alternative, it yields lower overall performance (see Section 5.3), indicating that although VLMs can provide high-level, performance-related scores, learning fine-grained ranking functions still benefits from supervised learning. Considering label scarcity, we use a lightweight model as a ranker.

Specifically, for each brand b, we train a 3-layer MLP so that it maps the generated score vector of each ad i to its final relevance score s_i as follows:

$$s_i = \text{MLP}_b(\mathbf{e}_i). \quad (6)$$

Training loss. For training, we adopt a pairwise ranking loss. Let y_i denote the target performance label (e.g., CTR, CPC, and CPM) of each ad i. For each ad pair (i, j) of the target brand such that $y_i > y_j$, the corresponding loss term is defined as:

$$\mathcal{L}_{ij} = \log\big(1 + \exp(-(s_i - s_j))\big). \quad (7)$$

The final loss sums $\mathcal{L}_{ij}$ over all such ad pairs.

5 Experimental Results

In this section, we review our experiments to answer the following questions:
RQ1. Performance comparison: Does ADVISOR yield more accurate brand-specific advertisement rankings than the baselines?
RQ2. Ablation study: How does each component of ADVISOR contribute to performance?
RQ3. Case study: Does ADVISOR generate reasonable and effective brand-specific criteria?

Table 1: Dataset statistics.

Category	**Beauty**						**Fashion**			**Platform**
Brand	A	B	C	D	E	F	A	B	C	A
# train	8	5	15	10	6	7	33	99	33	34
# test	10	10	10	10	10	10	10	10	10	10
Total	18	15	25	20	16	17	43	109	43	44

RQ4. Online A/B testing: How does ADVISOR perform in practice compared to human experts?
RQ5. Hyperparameter analysis: How do the hyperparameters of ADVISOR affect performance?

5.1 Experimental Settings

We first describe the experimental settings.

Real-World advertising data. We evaluate ADVISOR on real-world ads from 10 brands across beauty, fashion, and platform categories (avg. 35 ads per brand). Each ad is labeled with three performance metrics–*click-through rate* (CTR), *cost per click* (CPC), and *cost per mille* (CPM)–collected from actual deployment of the ads on Instagram. For each brand, the 10 most recent ads are held out for testing and the rest for training, with their counts reported in Table 1.

Performance labels. We use the following standard metrics as performance labels for ad ranking:

- **Click-Through Rate (CTR):** This user engagement measure is defined as the ratio of clicks to impressions (i.e., clicks/impressions). Higher CTR indicates stronger user interest in the ads.
- **Cost Per Click (CPC):** This cost efficiency measure is defined as the average cost per click (i.e., spend/clicks). Lower CPC values indicate more cost-efficient ads.
- **Cost Per Mille (CPM):** This exposure efficiency measure is defined as the cost per 1,000 impressions (i.e., 1,000×spend/impressions). Lower CPM values indicate more efficient exposure.

In online A/B testing, we use an additional metric, **Return on Ad Spend (ROAS)**. This return-based performance measure is defined as the ratio of advertising revenue to advertising cost (i.e., 100 × revenue/cost). A higher ROAS indicates more efficient revenue generation per unit of ad spend.

Baselines. We compare ADVISOR against twelve baselines, categorized into two groups.

Table 2: **(RQ1)** Performance on brand-specific advertisement ranking. All reported results are averaged over three runs and scaled by 100 for readability. The best results are in **bold**, and the second-best results are underlined.

Measure	Metric	MLP			VLM (Zero-shot)			VLM (Few-shot)			DEVL	ECSF	MMF	ADVISOR
		T	V	T+V	T	V	T+V	T	V	T+V				
NDCG @1	CTR	42.78	52.56	41.98	31.89	36.09	41.76	36.90	35.15	39.36	39.17	24.54	39.98	52.32
	CPC	55.68	61.33	50.72	49.39	62.12	62.80	55.11	66.05	65.02	64.28	48.53	63.20	68.55
	CPM	57.93	60.96	66.22	63.88	62.83	63.00	68.11	69.00	65.62	72.83	75.43	69.26	70.79
	Avg	52.13	58.29	52.97	48.38	53.68	55.85	53.37	56.74	56.67	58.76	49.50	57.48	**63.89**
NDCG @3	CTR	55.95	52.58	51.64	44.88	46.88	50.30	45.51	48.14	49.26	47.59	43.43	48.38	56.00
	CPC	60.94	63.03	57.80	56.30	66.93	65.19	59.20	65.25	67.68	62.86	64.50	66.85	67.23
	CPM	66.16	66.58	69.19	70.68	65.85	64.62	71.08	73.55	69.96	72.85	76.70	70.23	73.21
	Avg	61.01	60.73	59.55	57.29	59.89	60.04	58.59	62.31	62.30	61.10	61.54	61.82	**65.48**
NDCG @5	CTR	63.50	56.92	56.66	53.91	58.09	54.46	53.94	57.19	53.56	54.37	52.14	57.86	63.84
	CPC	63.58	66.68	64.31	63.16	68.70	67.67	64.68	67.70	67.40	69.76	67.12	70.03	70.23
	CPM	73.58	74.50	73.29	71.93	69.98	67.88	72.16	73.57	73.02	73.22	79.40	74.91	76.45
	Avg	66.89	66.03	64.75	63.00	65.56	63.33	63.60	66.16	64.66	65.78	66.22	67.60	**70.17**
Total	**Avg**	60.01	61.68	59.09	56.22	59.71	59.74	58.02	61.74	61.21	61.88	59.09	62.03	**66.51**

- **Basic multimodal baselines:** We consider simple ranking models under three input configurations: (i) text only (T), (ii) visual only (V), and (iii) text+visual (T+V). MLP-based rankers use modality-specific representations[‡] as inputs to a 3-layer MLP. VLM-based rankers directly infer rankings from raw textual and/or visual inputs. For VLM-based rankers, we evaluate zero-shot and few-shot variants that include randomly sampled examples from the target brand.

- **Social media popularity prediction baselines:** We include three representative methods proposed for social media popularity prediction (see Section 2): DEVL (Wu et al., 2022), ECSF (Mao et al., 2023b), and MMF (Lin and Lee, 2024). We obtain multimodal representations of ads using their pretrained encoders, without updates, and train a separate ranker that maps these representations to ranking scores. As the ranker, we use either a 3-layer MLP or LightGBM (Ke et al., 2017), choosing the option with better overall performance for each method.

We use GPT-4.1-mini as the default backbone for ADVISOR and VLM-based rankers. We test ADVISOR with more backbone VLMs in Appendix D.2.

Evaluation metrics. For evaluation, we report normalized discounted cumulative gain (NDCG) at cutoff levels $k \in \{1, 3, 5\}$ with respect to each performance label (CTR, CPC, and CPM).

[‡]We use CLIP VIT-B/32 model as pretrained encoder.

Refer to Appendix B for implementation details and hyperparameter settings.

5.2 RQ1. Performance comparison

We compare ADVISOR with baselines on the task of brand-specific advertisement ranking. Table 2 reports the average performance across all brands, and results for individual brand categories are provided in Appendix D.1.

ADVISOR achieves the best overall ranking performance, with an average improvement of 7.2% over the strongest baseline. In particular, its superiority over zero-shot and few-shot VLM-based rankers indicates that the advanced use of VLMs by ADVISOR is essential beyond basic usage.

Comparisons among baseline methods reveal several interesting observations. First, comparisons among VLM-based ranker variants demonstrate the importance of visual features over textual ones in this task, as well as the benefit of few-shot demonstrations. Second, VLM-based rankers do not consistently outperform MLP-based counterparts, indicating that this task remains challenging for VLMs, especially without advanced utilization.

Notably, all methods perform better on CPM and CPC than on CTR, as CTR is more affected by user-level variability and is harder to predict.

5.3 RQ2. Ablation study.

To evaluate the effectiveness of the individual components, we examine five variants of ADVISOR:
(i) ADVISOR-AB: Criteria are generated with context augmentation using *all* brands.

Table 3: **(RQ2)** Ablation study of ADVISOR. Results are averaged across performance metrics and three runs, and scaled by 100 for readability. The best results are in **bold**, and the second-best results are underlined.

Method	NDCG@1	NDCG@3	NDCG@5	Avg
ADVISOR	**63.89**	**65.48**	**70.17**	**66.51**
-AB	54.10	60.32	65.13	59.85
-CB	56.25	62.02	67.82	62.03
-PR	57.78	64.20	68.49	63.49
-RE	44.22	51.37	58.58	51.39
-RA	51.29	58.56	63.35	57.73

(iii) ADVISOR-CB: Criteria are generated using only the target brand's description and ads, without cross-brand context augmentation.
(ii) ADVISOR-PR: Criteria are generated via cross-brand context augmentation; however, instead of using private performance labels from similar brands, augmentation relies solely on public data (i.e., brand descriptions and ads).
(iv) ADVISOR-RE: Ads are scored via single-pass VLM inference, without critique and refinement.
(v) ADVISOR-RA: A VLM performs ranking based on the generated criterion-specific scores, without a trainable ranker (i.e., an MLP).

As shown in Table 3, ADVISOR outperforms all variants, showing the effectiveness of its individual components. The largest drop occurs in **ADVISOR-RE**, showing the critical role of self-critique and refinement. **ADVISOR-AB** underperforms **ADVISOR-CB**, suggesting that, without careful selection, information from other brands can introduce noise. **ADVISOR-PR** outperforms **ADVISOR-CB**, indicating that even without private ad performance labels, context augmentation using public information from similar brands remains effective.

5.4 RQ3. Case study.

To examine whether the brand-specific evaluation criteria generated by ADVISOR are truly brand-specific and effective, we conduct two case studies.

Case study 1: Qualitative analysis of brand-specific criteria. As shown in Figure 2, the generated criteria for fashion brand A, favored by trend-conscious female consumers, emphasize *human presence* and *headline text hook*, reflecting a strategy focused on visual appeal and attention capture. In contrast, for fashion brand B, a more widely recognized everyday-wear brand, the criteria emphasize *logo prominence* and *contextual relevance* (e.g., seasonal fit), prioritizing visible brand identity and relatable daily usage contexts. This demonstrates that the generated criteria indeed capture brand specificity rather than being generic. Additional cases are provided in Appendix C.1.

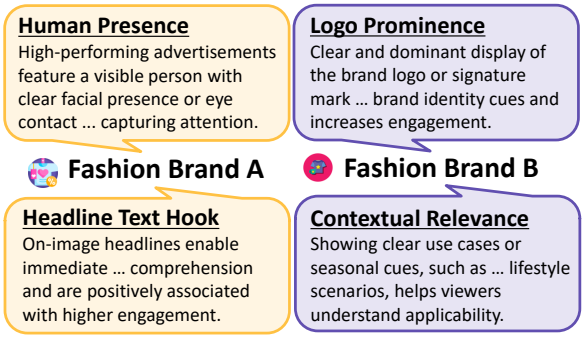


Figure 2: (**RQ3**) Brand-specific evaluation criteria for fashion brand A and B, generated by ADVISOR.

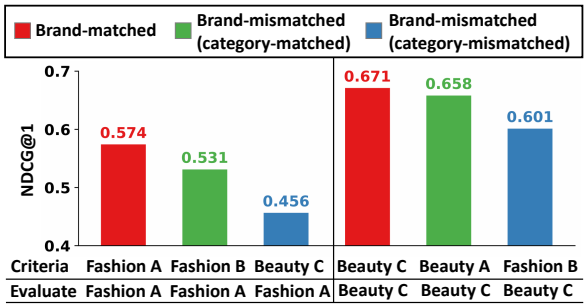


Figure 3: (**RQ3**) Effectiveness of brand-specific criteria, compared with criteria generated for other brands within the same category or from different categories. Results are averaged across performance metrics and three runs.

Case study 2: Quantitative analysis of brand-specific criteria. We evaluate the effectiveness of the generated brand-specific criteria by comparing them with criteria generated for other brands under three settings: **(i) brand-matched**, where ads are evaluated using criteria generated for the target brand, **(ii) brand-mismatched but category-matched**, where criteria generated for a different brand within the same brand category are applied, and **(iii) brand-mismatched and category-mismatched**, where criteria generated for a brand from a different brand category are used. The first setting is the intended use of ADVISOR, while the second and third serve as baselines.

Figure 3 shows performance results for fashion brand A and beauty brand C. In both cases, brand-matched criteria achieve the best performance; same-category criteria degrade performance, and mismatched-category criteria cause a larger drop. This brand specificity and effectiveness of the generated criteria are consistently observed across other evaluation metrics (see Appendix C.2).

Table 4: **(RQ4)** Online A/B testing results for fashion brand A. The best performance is in **bold**.

Method	CTR ↑	CPC ↓	ROAS ↑
Human Marketers	8.37%	428	1,070%
ADVISOR	**10.14%**	**231**	**1,219%**
Improvement (%)	21.15%	46.03%	13.93%

5.5 RQ4. Online A/B testing.

To examine effectiveness in practice, we present online A/B testing results on Instagram. Unlike prior results based on historical data, online A/B testing more strictly controls external factors (e.g., platform-level delivery mechanisms) affecting ad performance. Note that, due to cost and the need for brand approval, we conduct the comparison in a focused setting, evaluating the top-2 ads selected by ADVISOR and professional human marketers (see Appendix B for detailed settings).

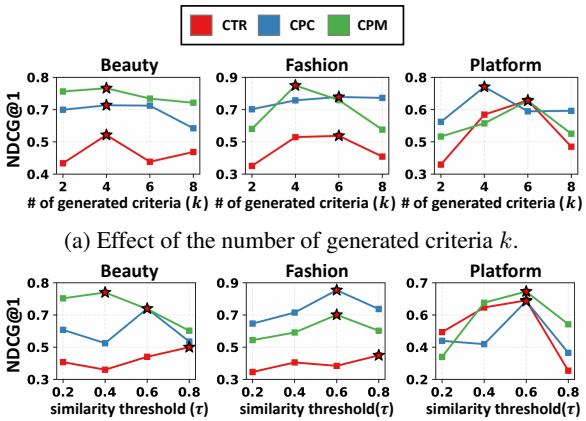
As shown in Table 4, ADVISOR shows consistent gains over human selection across CTR, CPC, and ROAS, with an average improvement of 27.04%. Note that we report ROAS instead of CPM, as CPM is not provided in A/B testing. This indicates benefits in real ad campaigns beyond offline evaluation.

5.6 RQ5. Hyperparameter analyses.

We examine how key hyperparameter choices affect the ranking performance of ADVISOR, focusing on (i) the number of generated criteria k, and (ii) the brand similarity threshold τ.

Effect of the number of generated criteria. To investigate the effect of the number of generated criteria k on ranking performance, we conduct experiments with $k \in \{2, 4, 6, 8\}$ while keeping all other settings fixed. Figure 4a presents results across different brand categories. We observe that performance does not increase monotonically with k, but instead peaks at moderate values of k (4 or 6), supporting our default choice of using $k = 4$ in the main experiments.

Effect of brand similarity threshold. The brand similarity threshold τ specifies the minimum cosine similarity required for another brand to be used as cross-brand context for a target brand. Figure 4b shows the ranking performance under different τ values. Performances tend to peak at a moderate threshold value ($\tau = 0.6$) across all categories. When τ is too low ($\tau = 0.2$), weakly related brands are used for context augmentation, introducing noisy information that degrades performance.



(a) Effect of the number of generated criteria k.

(b) Effect of the brand similarity threshold τ.

Figure 4: (**RQ5**) Effects of key hyperparameters of ADVISOR on ranking performance across brand categories. Stars indicate the best performances.

In contrast, overly high thresholds ($\tau = 0.8$) restrict cross-brand augmentation, forcing the model to rely solely on scarce target-brand data. These results demonstrate that selectively choosing brands for context augmentation is important.

6 Conclusion and Future Directions

In this work, we explored how VLMs can be leveraged for business decision-making under severe data scarcity and the absence of explicit decision criteria. As a concrete instance, we introduced the problem of brand-specific ad ranking and addressed it with ADVISOR, which explicitly generates brand-aware decision criteria through cross-brand context augmentation and applies reflection-based scoring. Experiments on real-world advertising data collected from ad campaigns showed that ADVISOR consistently outperforms strong baselines and remains effective in online A/B testing. Furthermore, case studies confirmed the brand-specificity and effectiveness of the generated criteria.

A limitation of our approach is the need for at least a small set of labeled ads per brand, restricting direct use for entirely new brands. Future work includes relaxing this requirement, for example by advancing cross-brand augmentation. Moreover, shifting the focus from overall ad performance to performance prediction over time would be valuable for time-critical ads and campaign planning.

7 Ethical Statement

Our work adheres strictly to the ethical guidelines throughout the development and deployment. The data used in this work were provided by collaborating brands with explicit approval for the intended

research usage. To protect privacy and confidentiality, all sensitive information was anonymized prior to use. No personal or sensitive user data were collected, inferred, or used in our framework, and all data were processed at the advertisement level.

Acknowledgements

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00438638, EntireDB2AI: Foundations and Software for Comprehensive Deep Representation Learning and Prediction on Entire Relational Databases) (No. RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST)).

References

Anthropic. 2025. Claude 4 model card. https://www.anthropic.com.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, and 1 others. 2016. Wide & deep learning for recommender systems. In *DLRS@RecSys*.

Gemini Team Google. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: A factorization-machine based neural network for ctr prediction. In *IJCAI*.

Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and 1 others. 2014. Practical lessons from predicting clicks on ads at facebook. In *AdKDD@KDD*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *NeurIPS*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Xiangyang Li, Bo Chen, Lu Hou, and Ruiming Tang. 2025. Ctrl: Connect collaborative and language model for ctr prediction. *ACM Transactions on Recommender Systems*, 4(2):1–23.

Yanhong Li, Chenghao Yang, and Allyson Ettinger. 2024. When hindsight is not 20/20: Testing limits on reflective thinking in large language models. In *NAACL*.

Yu-Shi Lin and Anthony J.T. Lee. 2024. Mmf: Winning solution to social media popularity prediction challenge 2024. In *MM*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Sina Malakouti, Aysan Aghazadeh, Ashmit Khandelwal, and Adriana Kovashka. 2024. Benchmarking vlms' reasoning about persuasive atypical images. In *WACV*.

Kelong Mao, Jieming Zhu, Liangcai Su, Guohao Cai, Yuru Li, and Zhenhua Dong. 2023a. Finalmlp: an enhanced two-stream mlp model for ctr prediction. In *AAAI*.

Shijian Mao, Wudong Xi, Lei Yu, Gaotian Lü, Xingxing Xing, Xingchen Zhou, and Wei Wan. 2023b. Enhanced catboost with stacking features for social media prediction. In *MM*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and IIYa Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. 2025. Mitigating hallucinations in vision-language models through image-guided head suppression. In *EMNLP*.

Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Qiushi Huang, Jintao Li, and Tao Mei. 2017. Sequential prediction of social media popularity with deep temporal context networks. *arXiv*.

Jianmin Wu, Liming Zhao, Dangwei Li, Chen-Wei Xie, Siyang Sun, and Yun Zheng. 2022. Deeply exploit visual and language information for social media popularity prediction. In *MM*.

Xovee Xu, Yifan Zhang, Fan Zhou, and Jingkuan Song. 2025. Improving multimodal social media popularity prediction via selective retrieval knowledge augmentation. *AAAI*.

Qi Yang, Aleksandr Farseev, Marlo Ongpin, Alfred Huang, Yu-Yi Chu-Farseeva, Damin You, Kirill Lepikhin, and Sergey Nikolenko. 2025. Fusing predictive and large language models for actionable recommendations in creative marketing. *ACM Transactions on Information Systems*, 43(5):1–31.

Guoxiao Zhang, Yi Wei, Yadong Zhang, Huajian Feng, and Qiang Liu. 2025. Balancing efficiency and effectiveness: An llm-infused approach for optimized ctr prediction. In *WWW*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. In *ACL*.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *ACL*.

Yan Zhuang, Wei Bai, Yanru Zhang, Minhao Liu, Jiawen Deng, and Fuji Ren. 2025. Fame: Fusion-aware multi-modal ensemble for social media popularity prediction. In *MM*.

A Detailed Prompts

In Figures 5, 6, 7, 8, and 9, we provide detailed prompts for each step of ADVISOR.

Cross-Brand Context Augmentation

You are an expert in **Instagram ad performance analysis**. You will analyze **real ad performance data** (caption + image) and **infer scoring criteria**.

Observed Ads from Similar Brand:
`ADS WITH CAPTION + IMAGE + METRIC`

Target Metric:
`[METRIC]`

Instructions:
Based on these examples (both captions AND images), please reason about:
1. What visual/textual characteristics correlate with HIGH vs LOW performance?
2. What specific patterns do you see in the high-performing ads?
Please provide reasoning outputs that could help score similar ads.

Figure 5: VLM prompt template for cross-brand context augmentation and insight extraction.

Criteria Generation

You are an expert in **brand marketing** and **ad performance optimization**. Your task is to generate and prioritize the most important features for evaluating ads.

Few-shot Examples: `[FEWSHOT EXAMPLES]`
Insights from Similar brands: `[INSIGHTS]`

Instructions: Based on the ranking examples above, identify and prioritize the `[NUM_FEATURE]` MOST IMPORTANT features for evaluating ads for this brand. For each feature, provide:
1. Feature name/key
2. Why this feature is critical based on the patterns you observe in the examples
3. How to score it on a 1-5 scale

Format each line as: [feature key | why important | scoring scale]

Figure 6: VLM prompt template for brand-specific criteria generation.

Reflection-based Scoring: Initial Scoring

You are an expert in Instagram ad performance analysis. You will analyze real ad performance data (caption + image) and infer scoring criteria.

Fewshot Examples: `[EXAMPLES from SIMILAR BRAND]`
Based on these examples (both captions AND images), please reason about:
1. What visual/textual characteristics correlate with HIGH vs LOW `[METRIC]`?
2. What specific patterns do you see in the high-performing ads?
Please provide 3-4 sentences of reasoning that could help score similar ads.

Figure 7: VLM prompt template for cross-brand insight generation.

Reflection-based Scoring: Self-critique

You are the second-stage critic. You receive initial reasoning and 1-5 scores for each ad across multiple features. Review the initial scorer's reasoning and detect inconsistencies, scale collapse, bias, or missing penalties.
RULES:
- First, for each ad, provide your critique reasoning (1-2 sentences explaining what the initial scorer got right or wrong).
- Then suggest corrected scores (1-5 integers).
- Keep the SAME features and scale (1-5 integers).
- If initial scoring looks reasonable, keep the same score but still explain why.
- If you adjust, stay within 1-5 and avoid inflating everything.

Features: `[FEATURES]`
Initial Score: `[INITIAL SCORE]`
Initial Reasoning: `[INITIAL REASONING]`

Ads: `[ADS]`

Format your response as:
Ad [ID] Score
Critique Reasoning

Figure 8: VLM prompt template for self-critique.

Reflection-based Scoring: Final Scoring

You are the third-stage arbiter. You see both the initial scorer's scores and critic's reasoning. Decide the FINAL scores (1-5 integers) for each ad, feature-wise.
RULES:
- First, for each ad, provide your final reasoning (1-2 sentences explaining your decision).
- Prefer critic adjustments when they fix scale compression, bias, or obvious errors.
- If critic over-corrects or seems inconsistent with evidence, keep the initial value.
- Preserve full-scale usage; avoid all ads ending 4-5.

Features: `[FEATURES]`
Initial Score: `[INITIAL SCORE]`
Critique Reasoning: `[CRITIQUE REASONING]`

Ads: `[ADS]`

Format your response as:
Ad [ID] Final Score
Final Reasoning

Figure 9: VLM prompt template for final scoring.

B Detailed Experimental Settings

Baseline details. For MLP-based rankers, we use a 3-layer MLP model with dimensions $256 \to 128 \to 64 \to 1$ and ReLU activations, as the ranker. For VLM-based rankers, we use GPT-4.1-mini as the backbone VLM, and the VLM directly produces rankings without explicit scoring or reflection. We set the temperature to 0 for all VLM calls.

Implementation details. We use GPT-4.1-mini as VLMs and text-embedding-3-large from OpenAI as the embedding model for identifying similar brands. For cross-brand context augmentation, we select only the most similar brand ($n = 1$) for each target brand, as incorporating additional brands leads to performance degradation (see Section 5.3). Unless otherwise specified, the number of generated evaluation criteria is fixed to $k = 4$, and the brand similarity threshold is fixed to $\tau = 0.6$. Refer to Section 5.6 for the effects of both parameters.

Online A/B test settings. We conducted an online A/B test for fashion brand A over a 7-day period (Nov 26 to Dec 2, 2025) on Instagram, a popular online advertising platform. The target audience was set to female users aged 25 and above, excluding those who had purchased from the brand in the previous seven days. ADVISOR and human marketers selected two ads prior to deployment.

C Additional Case Studies

In this section, we present full case study results.

C.1 Qualitative analysis of brand-specific criteria.

In this subsection, we extend the qualitative analysis in Case Study 1 (Section 5.4) by presenting the results for all brands. Figures 10, 11, and 12 show the brand-specific evaluation criteria generated by ADVISOR for brands in beauty, fashion, and platform categories, respectively. Each criterion is accompanied by a short description to provide its semantic meaning and its association with high-performing advertisements.

Generated Criteria for Beauty Brands

Beauty Brand A
Face Close-up. Tightly framed, vertical face close-ups resembling creator-style tutorials dominate the visual field and outperform distant or full-body shots.
Expressive Face Engagement. Candid expressions with direct eye contact convey authenticity, whereas neutral or overly posed faces are associated with lower performance.
On-screen Problem–Benefit Text. Problem-to-solution framing facilitates rapid comprehension and is linked to higher engagement rates.
Product In-use Visibility. Explicit visualization of product application steps signals effectiveness and instructional value.

Beauty Brand B
Human Subject Presence. High-performing advertisements consistently include a visible human subject (e.g., face or hands), providing social context and salient attention cues. Product-only images tend to underperform.

Expressive Engagement. Expressive facial cues, direct eye contact, or observable actions (e.g., product application) increase immediacy and viewer engagement.
On-screen Hook Text. Clear, curiosity-driven on-image text—such as questions or benefit-oriented phrases—communicates value at a glance and correlates with higher engagement.
UGC Lifestyle Vibe. UGC-style presentations (tutorials, POV shots, hands-on demonstrations) appear more authentic and consistently outperform polished studio imagery.

Beauty Brand C
Hair Visual Appeal. High-performing advertisements prominently show glossy, well-styled hair with visible movement or shine, directly demonstrating functional benefits such as smoothness, softness, and detangling effectiveness.
Product and Brand Visibility. Clear visibility of the product or brand logo—such as packshots, hand-held products, or on-screen brand text—enhances immediate brand recognition and is strongly associated with higher engagement.
Authentic Recommendation Tone. First-person or recommendation-style narratives (e.g., personal routines or daily use) convey authenticity and trustworthiness, outperforming purely informational or promotional language.
Sensory or Novelty Focus. Emphasizing sensory cues or novelty elements—such as fragrance notes, texture descriptions, or limited-edition attributes—captures user attention and differentiates products in crowded feeds.

Beauty Brand D
Face Proximity and Framing. Tight, close-up facial framing creates intimacy and direct engagement, outperforming distant shots.
Product Application Visibility. Clear visualization of product usage reduces uncertainty and increases engagement.
On-screen Text Hook. Concise problem–promise or how-to text effectively communicates value at a glance.
Authentic UGC Feel. Candid UGC-style visuals consistently outperform staged lifestyle shots.

Beauty Brand E
Focal Point Strength. A single dominant focal element (e.g., face or product label) captures attention more effectively than cluttered compositions.
Headline Readability and Hook. Short, bold, readable headlines attract attention, whereas dense or handwritten fonts reduce engagement.
Color Contrast and Lighting. Bright lighting and strong contrast help subjects stand out in feed thumbnails.
Clutter and Overlay Density. Clean layouts outperform designs with excessive stickers or overlays.

Beauty Brand F
Face or Scalp Prominence. Large close-ups of faces or scalp areas draw immediate attention.
Eye Contact and Emotional Expression. Direct eye contact and expressive emotions enhance relatability and engagement.
Curiosity Trigger. Teasing visual or textual cues (e.g., before–after highlights) stimulate curiosity.
Composition and Contrast. Clean, high-contrast compositions with minimal clutter consistently perform better.

Figure 10: Brand-specific evaluation criteria generated by ADVISOR for beauty brands.

Generated Criteria for Fashion Brands

Fashion Brand A
Human Presence. High-performing advertisements consistently feature a visible person with clear facial presence or direct eye contact, enhancing relatability and capturing viewer attention.
Headline Text Hook. Prominent, benefit-oriented on-image headlines—such as occasion cues or readiness messages—immediately communicate relevance and are associated with higher engagement rates.
Product Visibility. Clothing items are clearly framed, either as full outfits or highlighted key pieces, allowing viewers to quickly recognize the product without visual ambiguity.
Premium Production Quality. Polished composition, consistent lighting, and cohesive, on-brand styling convey a premium impression

and consistently outperform low-quality or amateur visuals.

Fashion Brand B
Logo Prominence. Clear and dominant display of the brand logo or signature mark effectively signals brand identity and increases recognition-driven engagement.
Model Presence and Engagement. Strong model poses, expressive posture, or direct eye contact enhance visual impact compared to static or disengaged presentations.
Visual Polish and Styling. Editorial-level lighting, composition, and cohesive styling signal desirability and product quality, correlating with higher engagement rates.
Contextual Relevance. Explicit depiction of use cases or situational cues—such as seasonal context, school settings, or sports-related scenarios—helps viewers immediately understand product relevance.

Fashion Brand C
Event Promotion Clarity. Explicit highlighting of sales events, discounts, or time-limited offers creates urgency and is strongly associated with higher engagement rates.
Human Presence and Engagement. Visible people with expressive gestures, direct eye contact, or pointing actions attract attention more effectively than distant or absent human subjects.
Text Overlay Legibility and Message. Large, clearly legible overlay text that communicates the offer or value proposition enables immediate comprehension in feed-based viewing.
UGC or Creator Tone. Creator-style advertisements with informal filming and personality-driven presentation align with community-oriented audiences and consistently outperform catalog-style creatives.

Figure 11: Brand-specific evaluation criteria generated by ADVISOR for fashion brands.

Generated Criteria for Platform Brand

Platform Brand A
Promotional Offer Strength. Prominent incentives—such as coupons, reward points, or discounts—clearly communicate value and are strongly associated with higher engagement rates.
Engagement Presence. High-performing advertisements include explicit calls-to-action (e.g., comment prompts or simple participation requests) that directly encourage user interaction and engagement.
Relatability and Target Relevance. Depiction of everyday, target-specific scenarios resonates with core users, increasing perceived relevance and viewer attention.
Visual Hook Text Clarity. Concise, bold on-screen headline text effectively communicates the primary hook or question at a glance, enabling immediate comprehension in feed-based viewing.

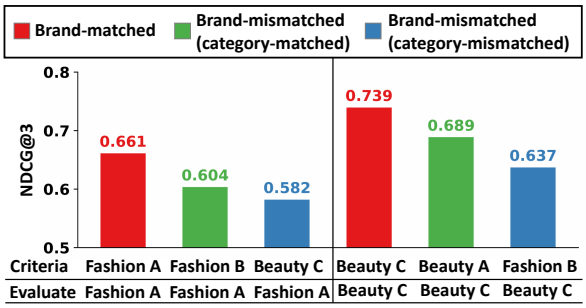
Figure 12: Brand-specific evaluation criteria generated by ADVISOR for platform brand.

C.2 Quantitative analysis of brand-specific criteria

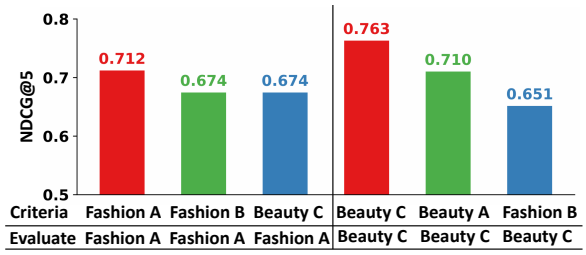
We extend the quantitative analysis in case study 2 (Section 5.4) by reporting additional results on evaluation metrics NDCG@3 and NDCG@5. We compare three settings: **(i) brand-matched**, **(ii) brand-mismatched but category-matched**, and **(iii) brand-mismatched and category-mismatched criteria**, following the same experimental setup in the main paper. We consider the same evaluation target brands, fashion brand A and beauty brand C. Figure 13 reports the results for both evaluation metrics across all target brands under the three criteria settings. For both evaluation targets, using brand-matched criteria yields the best performance. When criteria from another brand within the same category are applied, performance degrades but remains higher than that obtained using criteria from a different category. A severe performance drop is observed in the category-mismatched setting, indicating that criteria generated for different categories are not suitable for evaluating the target brands. These results are consistent with the results for NDCG@1 in the main paper and demonstrate that our approach generates brand-specific evaluation criteria that are not merely plausible, but are empirically effective.

C.3 High-level insights from similar brands for cross-brand context augmentation

In this subsection, we present examples of high-level insights produced by the VLM when reasoning over sample ads from selected similar brands



(a) Effectiveness of brand-specific criteria in terms of NDCG@3.



(b) Effectiveness of brand-specific criteria in terms of NDCG@5.

Figure 13: Effectiveness of brand-specific criteria measured by NDCG@3 and NDCG@5. We compare them with criteria generated for other brands within the same category or from different categories.

(Section 4.1). Recall that ADVISOR uses these insights as auxiliary guidance for evaluation criteria generation, rather than directly providing the sample ads as few-shot demonstrations. Specifically, these insights are used as "Insights from Similar Brands" in the criteria generation prompt shown in Figure 6. Figure 14 presents an example reasoning output for the target brand, beauty brand A, where beauty brand B is selected as the similar brand. Figure 15 presents an example reasoning output for the target brand, beauty brand E, where beauty brand F is selected as the similar brand.

Reasoning Output for *Beauty Brand A*

High-performing advertisements feature close-up, dynamic, and relatable visuals of women actively engaging in skincare routines, often with visible facial expressions or actions (e.g., applying patches or explaining skin concerns). These ads typically include minimal or no text overlay, allowing natural and candid moments to capture viewer attention.

Medium-performing advertisements often include some on-image text overlays combined with product showcases or lifestyle context, but they lack the immediacy and emotional engagement observed in high-performing ads. In contrast, low-performing advertisements tend to rely on posed, static imagery with limited engagement or storytelling cues, frequently lacking clear skincare context or explanatory text, which reduces viewer interest and click motivation.

Based on these observations, advertisements that appear authentic, action-oriented, and emotionally engaging should be prioritized, particularly those with minimal or no text overlay. Overly posed or static visuals without clear skincare relevance or narrative cues should be avoided. When text overlays are used, they should remain concise and directly tied to a relatable skincare problem or routine to sustain medium-to-high engagement.

Figure 14: Cross brand insights for beauty brand A.

Reasoning Output for *Beauty Brand E*

High-engagement advertisements prominently feature close-up, expressive human faces with direct eye contact, often conveying relatable or engaging emotions or actions. These ads typically rely on visual storytelling, with minimal or no caption text outside the video, and incorporate on-screen text naturally within the video content. Simple indoor backgrounds further help focus attention on the person and their expression or action.

In contrast, low-engagement advertisements often consist of product-only images or less engaging visuals that lack a visible human face or emotional connection. These ads tend to be more static and less dynamic, sometimes including text overlays that resemble calls-to-action but fail to convey a strong personal or emotional appeal.

Medium-engagement advertisements fall between these extremes, occasionally featuring people but with less engaging expressions or indirect eye contact, as well as more generic scenes (e.g., outdoor settings). They also tend to include heavier text overlays that may reduce immediacy and visual impact.

Based on these observations, advertisements should be scored higher when they feature close-up human faces with expressive emotions or actions, minimal but well-integrated text, and simple backgrounds that maintain focus on the person. Static product-only shots or overly text-heavy visuals without a strong emotional or personal element should be deprioritized.

Figure 15: Cross brand insights for beauty brand E.

D Additional Experimental Results

D.1 Performance by category

We analyze brand-specific advertisement ranking performance across different brand categories. The results are reported in Tables 5–7. Notably, ADVISOR consistently outperforms all baseline methods in the *beauty* and *platform* categories. In these categories, MLP-based rankers and social-media popularity prediction baselines, which rely on pre-trained multimodal representations, exhibit weak performance overall. In contrast, for the *fashion* category, these baselines achieve strong performance and often even outperform ADVISOR. This difference may arise from category-dependent effectiveness of pre-trained embeddings, particularly

Table 5: Performance on brand-specific advertisement ranking for the **beauty** brands. All reported results are averaged over three independent runs and scaled by 100 for readability. The best results are highlighted in **bold**, and the second-best results are underlined.

Measure	Metric	MLP			VLM (Zero-shot)			VLM (Few-shot)			DEVL	ECSF	MMF	ADVISOR
		T	V	T+V	T	V	T+V	T	V	T+V				
NDCG @1	CTR	32.73	51.89	45.88	31.32	38.93	46.21	43.41	35.46	40.94	31.17	24.83	39.10	50.63
	CPC	49.85	60.25	49.02	47.17	66.93	65.07	51.85	69.95	64.83	61.61	41.64	60.58	70.93
	CPM	59.25	63.57	63.31	59.13	65.83	64.56	71.62	66.95	62.57	68.93	78.49	65.94	73.22
	Avg	47.28	58.57	52.74	45.87	57.23	58.61	55.63	57.45	56.11	53.90	48.32	55.21	**64.92**
NDCG @3	CTR	47.87	47.42	46.24	49.98	51.00	58.96	51.78	52.51	56.25	43.76	41.32	47.74	55.90
	CPC	55.46	66.47	53.49	70.24	67.42	57.32	68.09	68.89	60.17	58.10	61.11	64.65	68.17
	CPM	64.48	64.22	66.16	64.73	65.22	69.38	73.48	64.17	63.54	70.35	72.38	68.33	71.32
	Avg	55.93	59.37	57.59	56.54	61.99	63.87	59.49	64.69	63.10	57.40	58.27	60.24	**65.13**
NDCG @5	CTR	56.86	51.56	52.69	59.85	61.14	61.99	61.12	61.22	60.22	48.76	51.12	55.11	63.28
	CPC	60.24	66.41	64.41	61.71	69.15	68.27	63.94	68.55	68.06	68.49	64.35	68.10	70.21
	CPM	69.55	71.76	70.05	69.52	69.80	68.44	71.57	72.64	68.21	73.44	76.53	72.48	75.22
	Avg	62.22	63.24	62.38	63.69	66.70	66.23	65.55	67.47	65.50	63.56	64.00	65.23	**69.57**
Total	**Avg**	55.14	60.40	57.57	55.37	61.97	62.90	60.22	63.20	61.57	58.29	56.86	60.22	**66.54**

due to better alignment between pre-training data and the visual and textual properties of fashion ads.

D.2 Comparison across VLM backbones

We compare the ranking performances of ADVISOR using seven backbone VLMs, including four open-source models (Qwen-8B, Qwen-30B, Gemma-12B, Gemma-27B) and three proprietary models (Gemini-2.5-flash, GPT-4.1-mini, Claude-Sonnet-4). As shown in Table 8, performance generally improves as model size increases. Proprietary models tend to achieve strong overall performance within our framework, with GPT and Claude demonstrating consistently competitive results across different settings, and large open-source models ($\geq$ 27B) also outperform the strongest baselines.

D.3 Cross-brand ranking performance

We further evaluate a cross-brand ranking setting within each multi-brand category: *beauty* and *fashion*. Unlike our main setting, where ADVISOR generates brand-specific criteria, it constructs category-specific criteria. As shown in Table 9, ADVISOR consistently outperforms the strongest baselines for both categories, demonstrating the robustness of our approach beyond brand-specific settings.

Table 6: Performance on brand-specific advertisement ranking for the **fashion** brands. All reported results are averaged over three independent runs and scaled by 100 for readability. The best results are highlighted in **bold**, and the second-best results are underlined.

Measure	Metric	MLP			VLM (Zero-shot)			VLM (Few-shot)			DEVL	ECSF	MMF	ADVISOR
		T	V	T+V	T	V	T+V	T	V	T+V				
NDCG @1	CTR	74.38	43.87	38.93	28.84	41.58	30.99	33.44	44.24	39.94	63.91	27.02	43.79	50.84
	CPC	84.76	82.79	61.23	57.62	62.26	63.83	76.95	67.40	68.19	89.90	63.25	71.38	63.26
	CPM	72.33	67.23	78.63	62.81	60.97	57.54	85.19	74.59	67.51	94.23	78.42	83.22	71.47
	Avg	77.16	64.63	63.77	59.59	49.76	54.94	50.79	65.19	62.07	**82.68**	56.23	66.13	61.86
NDCG @3	CTR	80.88	64.09	62.72	42.99	39.39	45.07	40.44	41.90	46.69	62.43	50.01	52.92	53.49
	CPC	78.29	64.34	63.94	68.42	64.17	67.77	65.50	69.19	69.44	79.25	78.99	74.85	66.39
	CPM	80.45	77.34	78.42	81.92	68.60	61.78	71.56	81.63	78.96	86.33	87.37	78.48	81.50
	Avg	**79.87**	68.59	68.36	64.44	57.39	58.21	59.17	64.24	65.03	76.00	72.12	68.75	67.12
NDCG @5	CTR	86.57	66.05	66.27	51.98	53.70	51.38	44.26	51.58	51.82	72.19	59.66	66.21	64.29
	CPC	73.27	72.05	72.22	71.78	69.48	72.21	69.81	74.54	72.33	78.35	78.08	77.06	72.73
	CPM	89.98	82.73	83.28	78.82	71.16	64.24	71.81	81.52	78.59	82.43	89.83	83.67	81.19
	Avg	**83.27**	73.61	73.92	67.53	64.78	62.61	61.96	69.21	67.58	77.66	75.86	75.65	72.74
Total	**Avg**	**80.10**	68.94	68.68	63.85	57.31	58.59	57.31	66.21	64.89	78.78	68.07	70.18	67.24

Table 7: Performance on brand-specific advertisement ranking for the **platform** brands. All reported results are averaged over three independent runs and scaled by 100 for readability. The best results are highlighted in **bold**, and the second-best results are underlined.

Measure	Metric	MLP			VLM (Zero-shot)			VLM (Few-shot)			DEVL	ECSF	MMF	ADVISOR
		T	V	T+V	T	V	T+V	T	V	T+V				
NDCG @1	CTR	8.30	82.64	8.30	14.13	40.82	15.60	15.60	38.46	15.25	13.01	15.36	33.76	66.95
	CPC	3.47	3.47	3.47	27.17	46.75	50.86	48.55	9.97	58.96	3.47	45.71	54.37	70.19
	CPM	6.83	26.49	54.33	48.09	44.88	59.69	78.69	32.72	57.05	32.02	48.09	47.32	54.19
	Avg	6.20	37.53	22.03	29.79	44.15	42.05	47.62	27.05	43.75	16.17	36.39	45.15	**63.78**
NDCG @3	CTR	29.59	49.02	50.83	19.93	44.58	14.05	23.08	40.64	15.09	25.98	36.36	38.58	64.18
	CPC	41.77	38.44	27.37	36.85	55.39	44.04	51.59	36.43	55.14	42.25	41.33	56.04	64.12
	CPM	33.37	48.40	56.32	64.13	64.38	69.47	79.82	49.72	77.73	47.40	70.59	56.88	59.64
	Avg	34.91	45.29	44.84	40.31	54.79	42.52	51.50	42.26	49.32	38.54	49.43	50.50	**62.65**
NDCG @5	CTR	34.17	61.74	51.60	24.06	53.02	18.51	39.92	49.84	18.82	34.59	35.73	49.32	65.81
	CPC	54.61	52.18	40.00	45.98	63.67	50.43	53.77	42.06	48.68	51.59	50.89	60.49	62.80
	CPM	48.57	66.21	62.79	65.72	67.56	75.44	76.76	55.36	85.16	44.28	65.32	63.23	69.64
	Avg	45.78	60.04	51.46	45.26	61.42	48.13	56.81	49.09	50.89	43.49	50.65	57.68	**66.08**
Total	**Avg**	28.96	47.62	39.45	38.45	53.45	44.23	51.98	39.47	47.99	32.73	45.49	51.11	**64.17**

Table 8: Performance of ADVISOR with varying VLM backbones on brand-specific advertisement ranking. All results are scaled by 100 for readability. The best results are in **bold**, and the second-best results are underlined.

Measure	Metric	ADVISOR						
		Qwen-8B	**Gemma-12B**	**Gemma-27B**	**Qwen-30B**	**Gemini-2.5**	**GPT-4.1**	**Claude-4**
NDCG @1	CTR	36.68	49.65	47.68	42.70	45.50	52.32	44.18
	CPC	53.93	49.94	71.09	70.30	68.89	68.55	65.92
	CPM	59.76	59.83	66.34	71.75	60.34	70.79	70.72
	Avg	50.12	53.14	61.70	61.58	58.24	**63.89**	60.27
NDCG @3	CTR	48.77	60.77	52.86	50.83	48.44	56.00	55.78
	CPC	53.93	59.59	70.10	68.48	70.24	67.23	72.66
	CPM	59.76	68.59	65.77	70.75	68.58	73.21	72.33
	Avg	50.12	62.98	62.91	63.36	62.42	65.48	**66.92**
NDCG @5	CTR	56.47	65.88	59.52	57.04	54.47	63.84	62.33
	CPC	66.05	63.38	71.35	70.87	73.06	70.23	72.54
	CPM	72.18	73.26	68.94	73.61	72.89	76.45	75.41
	Avg	64.90	67.51	66.60	67.17	66.81	**70.17**	70.09
Total	**Avg**	58.00	61.21	63.74	64.04	62.49	**66.51**	65.76

Table 9: Performance on cross-brand advertisement ranking. All results are scaled by 100. The best results are in **bold**, and the second-best results are underlined.

(a) Beauty

Measure	Metric	DEVL	ECSF	MMF	ADVISOR
NDCG @1	CTR	26.93	26.46	39.10	41.18
	CPC	52.44	56.62	60.58	92.47
	CPM	44.86	59.35	65.94	83.94
	Avg	41.41	47.48	55.21	**72.54**
NDCG @3	CTR	40.10	42.28	47.74	36.60
	CPC	63.45	60.80	64.65	88.95
	CPM	62.19	65.61	68.33	87.50
	Avg	55.25	56.23	60.24	**71.02**
NDCG @5	CTR	51.51	56.62	55.11	35.52
	CPC	66.45	62.56	68.10	89.24
	CPM	66.83	70.95	72.48	82.96
	Avg	61.60	63.37	65.23	**69.24**
Total	**Avg**	52.75	55.69	60.22	**70.93**

(b) Fashion

Measure	Metric	DEVL	ECSF	MMF	ADVISOR
NDCG @1	CTR	41.91	34.96	51.62	30.13
	CPC	44.13	78.71	64.65	70.23
	CPM	94.23	83.19	68.10	98.23
	Avg	60.09	65.62	61.46	**66.20**
NDCG @3	CTR	50.83	37.78	67.56	27.37
	CPC	62.02	82.61	71.25	80.11
	CPM	89.74	84.50	66.79	98.30
	Avg	67.53	68.30	68.54	**68.59**
NDCG @5	CTR	64.56	47.67	70.16	35.12
	CPC	68.47	79.85	75.62	79.89
	CPM	87.71	81.82	71.61	98.46
	Avg	**73.58**	69.78	67.90	71.16
Total	**Avg**	67.07	67.90	67.48	**68.65**