CONTEXT-AWARE CRITERIA GENERATION WITH VLMS FOR ADVERTISEMENT RANKING UNDER DATA SCARCITY

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]Kim Jaechul Graduate School of AI, KAIST
[2]Department of Industrial Engineering, 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

∗ Equal contribution

ABSTRACT

Vision-Language Models (VLMs) perform strongly on generic multimodal reasoning tasks, yet real-world business decisions require reasoning that depends on the specific business objective under very limited data and labels. We introduce a new task, *brand-specific advertisement ranking*, which aims to rank ads for a target brand under these constraints. This task requires capturing what makes an ad effective for that particular brand, rather than relying on generic visual-textual cues or large-scale supervision. To tackle this task, we propose ADVISOR, a novel approach that derives explicit brand-aware decision criteria with careful use of VLMs. ADVISOR also augments the limited brand-specific context with ads from similar brands, and uses the generated criteria for reflection-driven scoring to rank ads. Experiments on real-world advertising data from 10 brands, which include performance labels for evaluation, collected from actual ad campaigns, show that ADVISOR outperforms the strongest baseline by up to 7.2% in ranking performance. ADVISOR also performs strongly in online A/B testing, even when compared with human experts. In addition, both qualitative and quantitative analyses confirm that the generated criteria capture effective brand-specific evaluation standards rather than generic ones. Our code is available at https://github.com/K-Kyungho/ADvisor.

1 INTRODUCTION

Vision-Language Models (VLMs) have demonstrated remarkable capabilities across a wide range of general multimodal tasks, such as image understanding, captioning, and visual dialog (Radford et al., 2021; Li et al., 2022; Liu et al., 2023).Yet 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.

Two challenges naturally arise in real-world business decision-making. **(C1) Data scarcity:** Data directly tied to business decisions is often extremely limited, since it can only be obtained through costly deployment or manual evaluation. **(C2) Lack of decision criteria:** Even with multimodal inputs, VLMs are not given explicit guidance on what constitutes a good decision for the target objective, leading them to rely on generic cues instead of the factors that define a good decision for the target objective (Li et al., 2023; Yu et al., 2022; Yuan et al., 2023; Zhou et al., 2024).

In this work, we introduce a new real-world task, *brand-specific advertisement (ad) ranking*, where the goal is to rank new ads for a target brand before deployment. Due to the high cost of ad production and campaign execution, each brand has only a small number of past ads and even fewer performance labels. This task directly reflects the challenges above: decisions must be made under severe data scarcity, without clear criteria for what makes an ad effective for a particular brand.

To address these challenges, we propose ADVISOR, whose key idea is to explicitly generate brand-aware evaluation criteria from brand information and a small set of 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 metrics collected from actual ad campaigns, which are used as evaluation labels. Results show that ADVISOR outperforms the strongest baseline by up to 7.2% in ranking performance. Moreover, ADVISOR also performs strongly in online A/B testing, even when compared to human experts, and ablation studies further demonstrate the effectiveness of each component. Both numerical comparisons and case studies confirm that the generated criteria capture meaningful brand-specific evaluation standards rather than generic cues.

Our contributions are summarized as follows:

- **New problem:** We introduce *brand-specific advertisement ranking*, a practical business task that requires ranking ads for a target brand. It reflects common challenges in business decision-making: severe brand-specific data scarcity and unspecified decision criteria.
- **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.
- **Empirical validation:** We validate the effectiveness of ADVISOR through extensive experiments on real-world advertising data from 10 brands, including case studies and online A/B testing.

2 RELATED WORK

Below, we review prior machine learning approaches to problems related to advertisement ranking.

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 fine-grained personalization critically 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 individual 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, and, especially, they do not have access to fine-grained user-ad interaction logs. 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, measured by the number of views or likes (Wu et al., 2017; Lin & Lee, 2024; Xu et al., 2025; Zhuang et al., 2025). Most approaches to this task are supervised, relying on large-scale, general-domain social media datasets. In contrast, our ad ranking problem aims to predict performance metrics directly tied to business decision-making, such as click-through rate, cost per click, and cost per mille. Unlike simple popularity metrics, these metrics are rare and, in most cases, not publicly accessible. Combined with the brand-specific nature of our task, this intensifies data scarcity and makes supervised learning ineffective. Moreover, due to these differences, using encoders trained for social media popularity prediction is suboptimal for ad ranking, as empirically validated in Section 5.2.

Advertisement understanding and evaluation. Malakouti et al. (2024) introduced benchmarks that frame ad understanding as a reasoning task, examining whether vision–language models

(VLMs) can infer persuasive intent or unusual visual relationships in advertisements. Their results show that VLMs often struggle with such high-level reasoning and require large language models (LLMs) for textual reasoning as compensation. More recently, Yang et al. (2025) studied ad performance prediction combined with post-hoc explanation, where supervised deep models estimate engagement metrics and LLMs translate intermediate model outputs (e.g., attention maps) into human-interpretable recommendations for marketers. In contrast, our work addresses brand-specific ad ranking, which goes beyond general (i.e., brand-agnostic) content analysis, interpretation, or evaluation of ads. This brand-specific nature worsens data scarcity, making reliance only on supervised training suboptimal, as explored in Section 5.2.

3 Problem Definition

In this section, we formalize the problem of brand-specific advertisement ranking and discuss its practical relevance and challenges.

Problem definition. Let $\mathcal{B}$ denote the set of brands. For each brand $b \in \mathcal{B}$, we are given (i) 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 click-through rate (CTR), cost per click (CPC), or cost per mille (CPM) (see Appendix D for details on these labels). For a target brand b with description d_b and a set of new ads $\mathcal{H}_{test} = \{(v'_j, t'_j)\}_{j=1}^{M}$ without performance labels, the objective is to rank these ads according to their marketing performance prior to deployment. Notably, this problem setting allows the use of labeled data from other brands, which are few in number in practice, reflecting the real-world scenario described below. Our approach, however, remains applicable and effective even when only target-brand data is available, as shown empirically in Section 5.2.

Practical relevance. This problem directly captures a core task faced by advertisers. From the advertiser's perspective, deciding which ads to deploy is a high-stakes responsibility that directly affects marketing outcomes. Producing ads and running campaigns incur substantial costs, and replacing underperforming ads after deployment often requires additional budget and time. Thus, reliable pre-deployment ad ranking is a critical task. In practice, professional advertising agencies that operate independently from brands often work with multiple brands. As a result, labeled data across brands can be accessible within the same agency, which aligns with the cross-brand data availability considered in our problem definition.

Challenges. This problem naturally presents two fundamental challenges:

- **(C1) Data scarcity:** Both producing ads and obtaining performance feedback are costly processes. Creating ads requires substantial time, budget, and professional effort, and performance labels can only be obtained after deployment. This inherently limits the amount of available data, even when data from a small number of other brands is accessible.
- **(C2) Lack of decision criteria:** What makes an ad effective varies significantly across brands, as each brand follows its own identity, target audience, and marketing strategy. Thus, there is often no explicit or well-defined decision criterion for what makes an ad effective for a particular brand.

4 Proposed Method

In this section, we present ADVISOR, our proposed approach for brand-specific advertisement 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, and **detailed input prompts for each step are given in Appendix A.**

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

Due to Challenge (C2), naively applying a vision–language model (VLM) to directly rank ads is ineffective, even with few-shot demonstrations (see Section 5.2). In the absence of explicit guidance

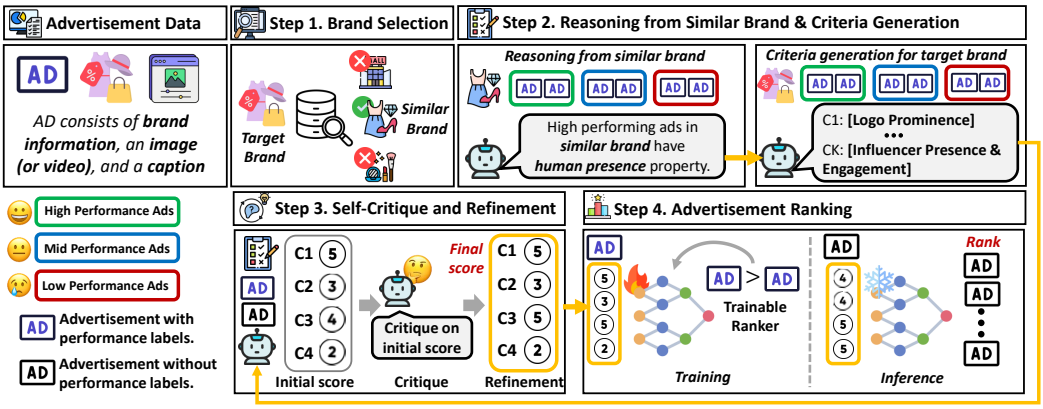


Figure 1: Overview of ADVISOR. It consists of three steps: (1) brand-specific criteria generation, (2) self-critique and refinement, and (3) brand-specific advertisement ranking.

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, which often fail to reflect brand-specific objectives.

The key idea of ADVISOR is to make such *brand-specific decision criteria* explicit by constructing a set of brand-aware evaluation criteria with the aid of VLMs. To elicit these criteria, we first build a contrastive few-shot context from the target brand's past ads. Specifically, we partition the ads into *high-, medium-, and low-performance* groups based on their performance labels, and when the number of ads is too large to be processed jointly by a VLM, we sample a fixed number from each group. By structuring the context to reflect contrasts across performance levels, this contrastive setup encourages the VLM to identify visual or semantic attributes that distinguish successful ads from less effective ones.

However, due to Challenge **(C1)**, 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 further 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 $\mathbf{z}_b$ [1] using brand descriptions of each brand, and then define the set of similar brands as those with cosine similarity greater than a threshold τ, as follows:

$$\mathcal{S}(b) = \Big\{ b' \in \mathcal{B} \setminus \{b\} \;\Big|\; \cos(\mathbf{z}_b, \mathbf{z}_{b'}) \geq \tau \Big\}. \tag{1}$$

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 samples for the target brand. Instead, the VLM reasons over these samples to extract high-level, performance-relevant insights, which are subsequently used as auxiliary guidance (see Appendix C for examples of extracted insights). Formally, the brand-specific criteria generation process is defined as

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

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, d_b denotes the brand description that specifies the brand's identity and positioning, $\mathcal{E}_b$ 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, which is used only for high-level reasoning rather than as direct examples. Examples of the brand-specific evaluation criteria can be found in Section 5.4 and Appendix B.

4.2 SELF-CRITIQUE AND REFINEMENT FOR SCORING

Given the generated brand-specific evaluation criteria set $\mathcal{C}_b$, the VLM assigns scores to each ad based on its visual and textual content. Since single-pass evaluation may produce inconsistent or

[1] We use text-embedding-3-large from OpenAI as our embedding model in our experiments.

weakly grounded judgments across criteria, especially when multiple modalities should be considered simultaneously, we formulate scoring through self-critique and refinement, in which initial assessments are revisited and refined. Specifically, 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 evaluation criterion. This step produces an initial score vector along with an explicit textual rationale:

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

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, to examine and refine potential errors in the initial scoring, we introduce another VLM that serves as a critic. The critic performs a reflective examination of the initial assessment by verifying whether each criterion-specific score is sufficiently 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

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

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:

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

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 $\mathbf{e}_i$ for each ad 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. To account for the scarcity of labeled ads, we use a lightweight model as a ranker.

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

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

Notably, the MLP is trained in a brand-specific manner using only the past ads of the target brand, i.e., $\mathcal{H}_b$ for brand b, to better capture brand-specific characteristics.

The ranking of ads in the test set is obtained by sorting them according to their final relevance scores.

Training loss: For training a brand-specific ranker, we use 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:

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

The final loss is obtained by summing $\mathcal{L}_{ij}$ over all such ad pairs.

5 EXPERIMENTS

In this section, we aim to answer the following research questions:

- **RQ1. Performance comparison:** Does ADVISOR yield more accurate rankings than 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?
- **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

Datasets. We evaluate ADVISOR on real-world ads from 10 brands on Instagram, a social media platform. The brands span three categories: *platform* (1 brand), *fashion* (3 brands), and *beauty* (6 brands). Each brand is associated with a set of unique ad posts, with an average of 35 ads per brand. Each ad is associated with three standard marketing performance metrics–*click-through rate* (CTR), *cost per click* (CPC), and *cost per mille* (CPM)–used as performance labels. The metrics were collected through the actual deployment of the ads on Instagram. For the definitions of these metrics and dataset statistics, refer to Appendix D. For evaluation, we adopt a temporal split. For each brand, the most recent 10 ads are used as the test set, while the others are used for training.

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

- **Basic multimodal baselines:** This group consists of simple ranking models with three input configurations: (i) text only (T), (ii) visual only (V), and (iii) text + visual (T+V). For MLP-based rankers, denoted as MLP, we first extract modality-specific representations using a pretrained multimodal encoder [2] and train a 3-layer MLP [3] to predict the scores for ranking from these representations. For VLM-based rankers, denoted by VLM, raw text and/or visual inputs are fed directly into a VLM [4], and the VLM directly produces rankings without explicit scoring or reflection. We consider two versions of the rankers: one with randomly sampled few-shot ad examples from the target brand (few-shot) and one without them (zero-shot).
- **Social media popularity prediction baselines:** This group includes three representative methods originally proposed for social media popularity prediction (see Section 2): DEVL (Wu et al., 2022), ECSF (Mao et al., 2023b), and MMF (Lin & 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.

Evaluation metrics. To evaluate ranking quality, 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). NDCG measures how well the predicted ranking aligns with the ground-truth ordering, placing greater emphasis on correctly ranking top-ranked ads.

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 $\tau = 0.6$. The effects of both parameters are analyzed in Section 5.6.

Refer to Appendix D for further details on the experimental setup.

5.2 RQ1. PERFORMANCE COMPARISON

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

As shown in Table 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

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

[3] We use a 3-layer MLP with dimensions $256 \to 128 \to 64 \to 1$ and ReLU activations.

[4] We use GPT-4.1-mini as the VLM.

Table 1: **(RQ1)** Performance on brand-specific advertisement ranking. All 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. ADVISOR achieves the best overall ranking performance.

Measure	Metric	MLP			VLM (Zero-shot)			VLM (Few-shot)			DEVL	ECSF	MMF	ADVISOR (Ours)
		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**

Table 2: **(RQ2)** Ablations study of ADVISOR. All reported results are averaged over three independent runs and scaled by 100 for readability. The best performance is highlighted in **bold**, and the second-best performance is underlined. ADVISOR outperforms its variants in most cases, validating the contributions of its individual components.

Method	NDCG@1			NDCG@3			NDCG@5			Avg
	CTR	CPC	CPM	CTR	CPC	CPM	CTR	CPC	CPM	
ADVISOR-CB	42.55	60.94	65.27	48.96	65.16	71.95	59.19	68.80	75.47	62.03
ADVISOR-AB	38.85	56.08	67.38	49.78	64.00	67.17	58.20	67.11	70.08	59.85
ADVISOR-RE	38.37	44.79	49.50	45.20	50.79	58.12	53.11	58.79	63.85	51.39
ADVISOR-RA	45.40	48.12	60.35	53.05	59.28	63.35	61.32	62.08	66.65	57.73
ADVISOR	**52.32**	**68.55**	**70.79**	**56.00**	**67.23**	**73.21**	**63.84**	**70.23**	**76.45**	**66.51**

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 (defined in Appendix D), as CTR is more affected by user-level variability and is therefore harder to predict.

5.3 RQ2. ABLATION STUDY.

To evaluate the effectiveness of the individual components in ADVISOR, we examine the following four variants of ADVISOR:

- **ADVISOR-CB:** Criteria are generated using only the target brand's description and ads, without cross-brand context augmentation.
- **ADVISOR-AB:** Criteria are generated using cross-brand context augmentation using *all* available brands without selection.
- **ADVISOR-RE:** Ads are evaluated using the generated evaluation criteria through single-pass VLM inference, without critique or refinement stages.

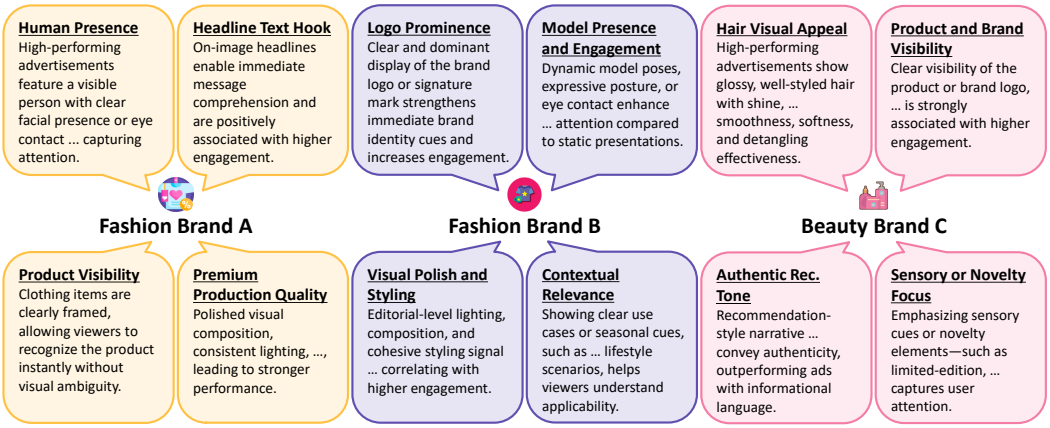


Figure 2: (**RQ3**) Brand-specific evaluation criteria generated by ADVISOR.

- **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 2, ADVISOR outperforms the variants in nine out of ten cases, confirming the effectiveness of each component of ADVISOR. Among the variants, **ADVISOR-RE** exhibits the largest performance drop, demonstrating the critical role of self-critique and refinement in stabilizing VLM-based evaluation and mitigating errors or inconsistencies that commonly arise from single-pass inference. Notably, **ADVISOR-AB**, which leverages information from all brands, underperforms **ADVISOR-CB**, which does not use any other brand information. This finding suggests that, without careful selection, information from other brands may introduce noise or confounding factors, ultimately harming brand-specific ranking performance.

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. The first provides a qualitative analysis of the generated criteria, while the second evaluates their impact on quantitative ranking performance.

Case study 1: Qualitative analysis of brand-specific evaluation criteria. Figure 2 presents the brand-specific criteria for three brands: fashion brand A, fashion brand B, and beauty brand C. Results for additional brands are provided in Appendix B. The generated criteria exhibit clear differences across brands, reflecting their distinct identities.

Specifically, for fashion brand A, favored by trend-conscious consumers, the criteria prioritize "headline text hooks" and "premium production quality," highlighting a strategy centered on polished presentation and attention capture. For fashion brand B, which targets everyday wear, "logo prominence" and "contextual relevance" are emphasized, highlighting brand identity expression and everyday usage scenarios. For beauty brand C, which specializes in hair damage care, the generated criteria emphasize "hair visual appeal" and "product and brand visibility," aligning with the brand's focus on demonstrating functional benefits and maintaining product visibility.

These qualitative differences indicate that the generated criteria adapt to brand-specific positioning and marketing strategies, even among brands within the same brand category.

Case study 2: Quantitative analysis of brand-specific criteria generation. In this case study, we evaluate the extent to which the generated brand-specific criteria contribute to ranking performance, particularly in comparison with criteria generated for other brands, both within the same category and across different categories. Specifically, we consider three evaluation settings: **(i) brand-matched criteria**, where ads are evaluated using criteria generated for the same brand, **(ii) brand-mismatched but category-matched criteria**, where criteria generated for a different brand within the same brand category are applied, and **(iii) brand-mismatched and category-mismatched criteria**, where criteria generated for a brand from a different brand category are used. Note that the first setting corresponds to the intended use of our approach, while the second and third serve as baselines for examining the relative effectiveness of the generated criteria.

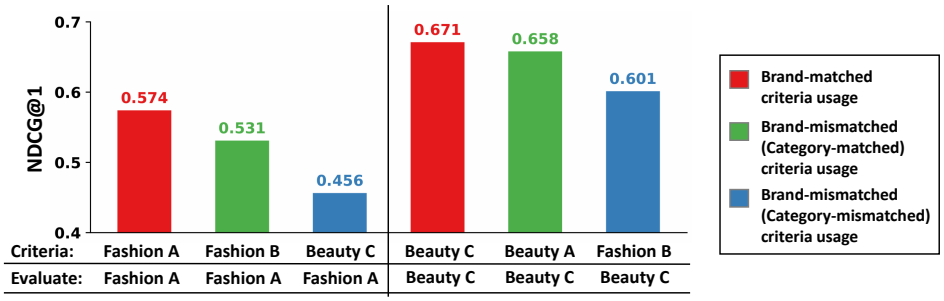


Figure 3: (**RQ3**) Empirical effectiveness of brand-specific evaluation criteria generated by ADVISOR, compared with criteria generated for other brands within the same category or from different categories. Note that criteria generated specifically for the target brand yield the best performance.

Table 3: **(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%

Figure 3 shows the ranking performance results. We consider two evaluation targets: fashion brand A and beauty brand C. 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. Although we report NDCG@1, similar performance trends are observed consistently across other evaluation measures. These results demonstrate that our proposed approach generates brand-specific evaluation criteria that are not merely plausible, but are empirically effective.

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, as described in detail in Appendix D.

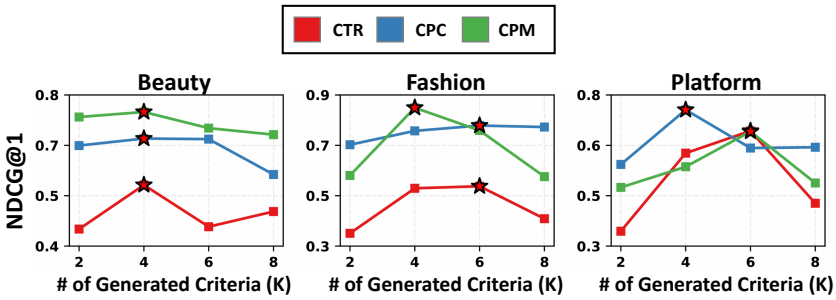
As shown in Table 3, ADVISOR shows consistent gains over human selection across CTR, CPC, and ROAS, with an average improvement of 27.04%. Note that we report ROAS instead of CPM, as CPM is not provided in A/B testing (see Appendix D for metric details). This indicates benefits in real ad campaigns beyond offline evaluation.

5.6 RQ5. HYPERPARAMETER ANALYSES.

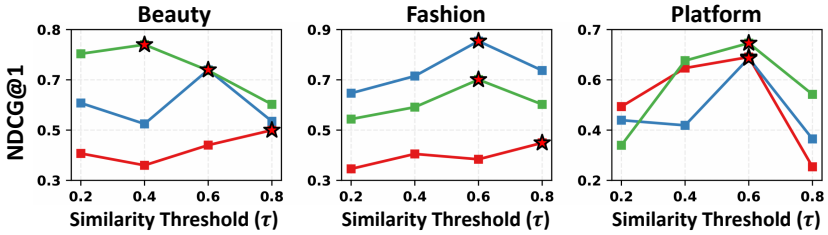
We examine how key hyperparameter choices affect the performance of ADVISOR. Specifically, we analyze the impact of (i) the number of generated criteria k, and (ii) the brand similarity threshold τ.

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. In contrast, overly high thresholds ($\tau = 0.8$) restrict cross-brand augmentation, forcing



(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. Performances peak at moderate values of k (4 or 6) and τ (around 0.6) in most cases.

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 practical instance, we introduced and addressed the problem of brand-specific ad ranking. For this problem, we proposed ADVISOR, whose central idea is to explicitly generate brand-aware decision criteria, enhanced by cross-brand context augmentation to mitigate data scarcity and by reflection-based scoring. By doing so, ADVISOR effectively guides VLMs to make decisions based on task-relevant principles rather than generic visual cues. We conducted experiments on real-world advertising data from 10 brands, with performance labels collected from ad campaigns. The results 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 both qualitatively and quantitatively.

One limitation of our approach is that the generated criteria may remain subjective or ambiguous, which can make them difficult to consistently apply in subsequent evaluation steps. Moreover, our approach assumes the availability of at least a small number of brand-specific ads with performance labels, which limits its direct applicability to new brands without any labeled ads. As future work, our approach can be extended to reduce subjectivity and ambiguity in the generated criteria, and to support application to new brands or even new brand categories without labeled ads by leveraging alternative information sources (e.g., web) for criterion generation. Moreover, shifting the focus from overall ad performance to performance prediction over time would be valuable for time-critical ads and campaign planning.

Acknowledgments

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

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

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

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

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

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

Xiangyang Li, Bo Chen, Lu Hou, and Ruiming Tang. Ctrl: Connect collaborative and language model for ctr prediction. *ACM Transactions on Recommender Systems*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

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

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

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

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

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

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. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

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

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

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

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

Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Tamara Lee Berg, and Zhang Ning. Commercemm: Large-scale commerce multimodal representation learning with omni retrieval. In *KDD*, 2022.

Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin No. Where to go next for recommender systems? id- vs. modality-based recommender models revisited. In *SIGIR*, 2023.

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

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *ICLR*, 2024.

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

APPENDIX

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 the first-stage evaluator for Instagram ads. You will be given a small number of training examples from this brand, where ads are ordered by performance, and a new set of ads (caption + image) to evaluate.
RULES:
- First, for each ad, briefly explain your reasoning (2–3 sentences) about the ad, based on both the caption and the image.

- Then, at the end of the entire response, output a final scores block.
- Scale usage guidelines: Use the FULL 1–5 scale. Do not collapse everything into 3–5.
1 = very poor / unacceptable
2 = below average / needs improvement
3 = acceptable / average
4 = good / above average
5 = excellent / outstanding

Features: [FEATURES]
Ads: [ADS]

Format your response as:
Ad [ID] Score

Figure 7: VLM prompt template for initial ad scoring.

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 Refinement

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 BRAND-SPECIFIC EVALUATION CRITERIA GENERATED BY ADVISOR.

Figures 10, 11, and 12 provide 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. The detailed set-up of ADVISOR can be found in Section 5.1.

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 Brands

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 brands.

C HIGH-LEVEL INSIGHTS FROM SIMILAR BRANDS FOR CROSS-BRAND CONTEXT AUGMENTATION

In this section, we present examples of high-level insights produced by the VLM when reasoning over sample ads from selected similar brands. 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 13 presents an example reasoning output for the target brand, beauty brand A, where beauty brand B is selected as the similar brand. Figure 14 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 13: VLM reasoning outputs 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 14: VLM reasoning outputs for beauty brand E.

Table 4: Dataset statistics.

Split	**Beauty Brands**						**Fashion Brands**			**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

D Detailed Experimental Settings

Dataset statistics. Table 4 summarizes the dataset statistics for each brand. As discussed earlier, we adopt a temporal split for evaluation, using the most recent 10 ads per brand as the test set and all others for training. The number of training samples varies across brands, ranging from 5 to 99. This setup reflects real-world data-scarce scenarios discussed in Section 3.

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.

Performance labels in advertising. We use the following three standard ad performance metrics as performance labels for advertisement ranking:

- **Click-through rate (CTR):** This user engagement measure is defined as the ratio of clicks to impressions (i.e., clicks/impressions). A 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$\times$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 $\times$ revenue / cost). A higher ROAS indicates more efficient revenue generation per unit of ad spend.

E Detailed Experimental Results for Each Brand Category

In this subsection, we analyze brand-specific advertisement ranking performance across different brand categories (beauty, fashion, and platform). The results are summarized 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 due to better alignment between pre-training data and the visual and textual characteristics of fashion ads.

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 (Ours)
		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**

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 (Ours)
		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 (Ours)
		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**