ACL 2026
SAN DIEGO JULY 2-7

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

Kyungho Kim*, Yeonje Choi*, Gyurim Hwang, Sejin Chung,

Hongseok Lee, Myeong Ho Song, Yeongho Kim, Sunwoo Kim,

Jongha Lee, Juyeon Kim, and Kijung Shin

KAIST AI
Kim Jaechul Graduate School

MADUP
For your innovative growth

Outline

1. **Introduction.**
2. **Proposed method.**
3. **Experiments.**
4. **Conclusion.**

Task Definition: Brand-specific Advertisement Ranking

(Example) Target brand:

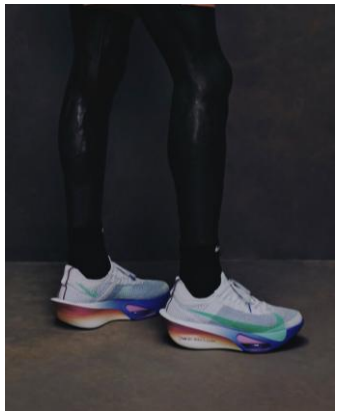
[Given: Nike ads from social media platform]

A MIND-ALTERING …

AD #1

Light work. Sabrina 4 …

AD #2



“My dream is to make this …

AD #3

[Predict: Ranking of ads]

AD #1 > AD #3 > AD #2

Reference: Nike Official Instagram (@nike)

Task Definition: Brand-specific Advertisement Ranking

- **Given:** A target brand and a set of deployed ads with visual image and caption.
- **Goal:** Ranking new ads for the target brand before deployment.
- **Performance Labels:** CTR, CPC, CPM
  - **CTR** (Click-Through-Rate) ↑ : clicks / impressions
  - **CPC** (Cost Per Click) ↓ : spend / clicks
  - **CPM** (Cost Per Mille) ↓ : 1000 × spend / impressions

Practical Relevance

- Our task represents the core task of real-world **advertising agencies**.
- We aim to enable **effective pre-deployment ad selection** under limited data, reducing costly trial-and-error in real-world ad campaigns.

Challenges

- There are two main challenges in solving our problem:

1) <u>Data scarcity</u>

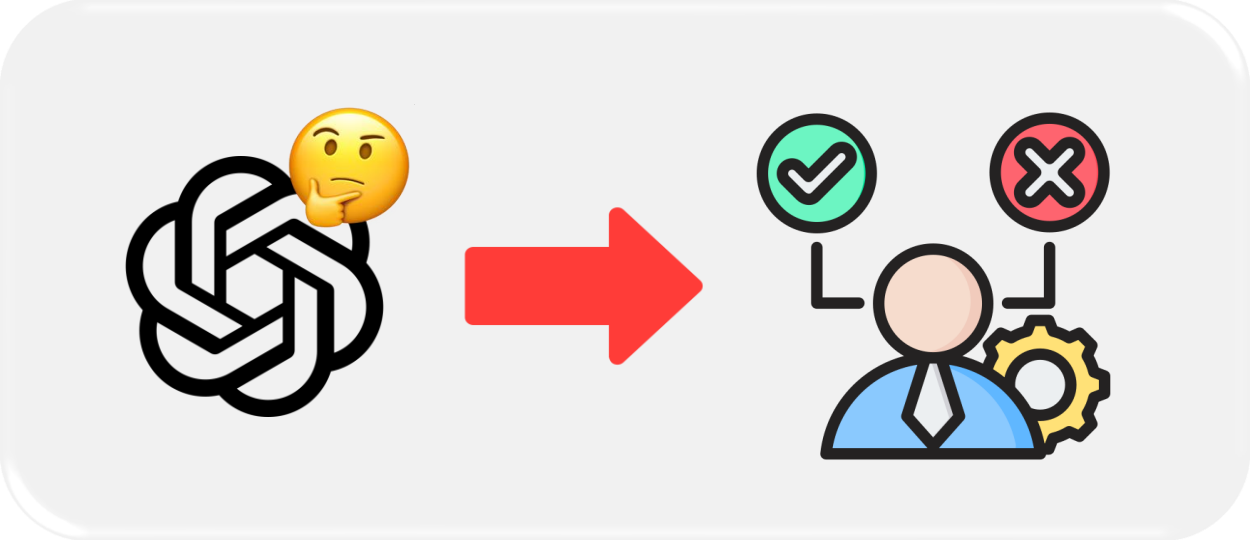
Limited ads and performance labels for each brand.

2) <u>Lack of decision criteria</u>

No clear definition of an effective ad for specific brands.

VLMs: A Potential but Suboptimal Solution

- While VLMs excel at general image understanding, they struggle with real-world business decisions.
- Thus, naïve VLMs are insufficient and require a task-aware framework.



Related Work: CTR Prediction

- Conventional CTR (Click-Through-Rate) prediction relies on **user-ad interaction logs** available only on the **platform-side**.
- Our task targets **advertiser-side** ad ranking under limited brand-specific data that does not have user-ad interaction logs.
- Our task requires a different approach from CTR prediction.

Outline

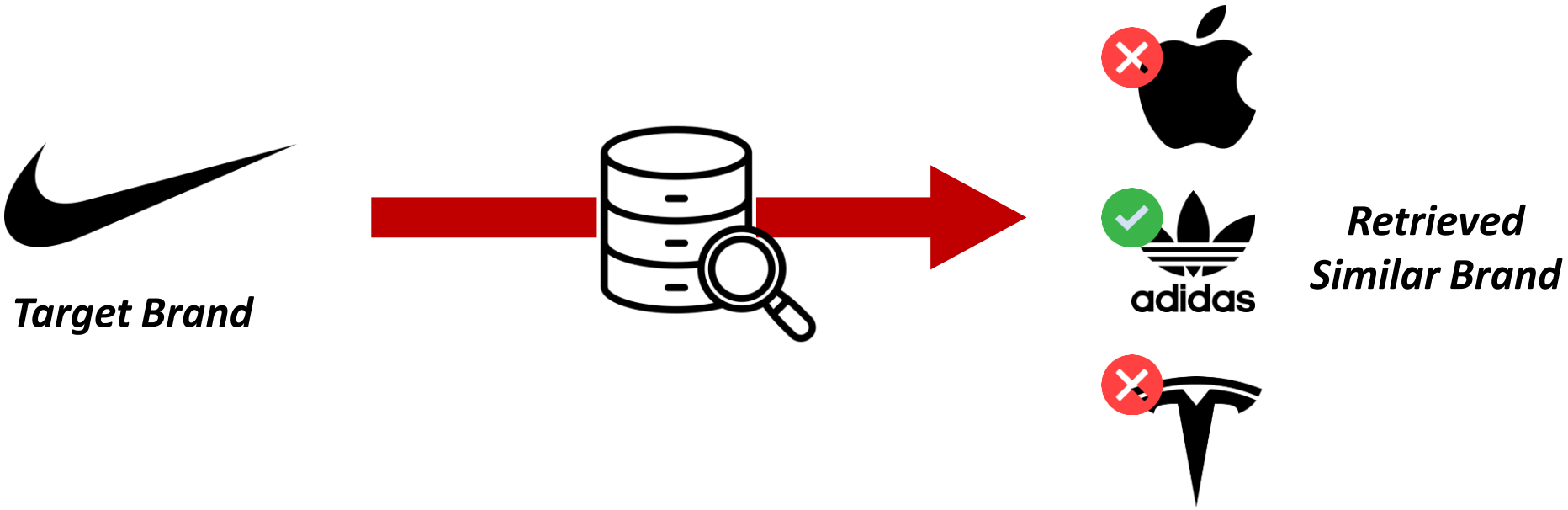
1. Introduction.
2. Proposed method.
3. Experiments.
4. Conclusion.

Proposed Method: ADvisor

- **ADvisor** is a VLM-based framework for brand-specific advertisement ranking.
- **ADvisor** consists of three main components.

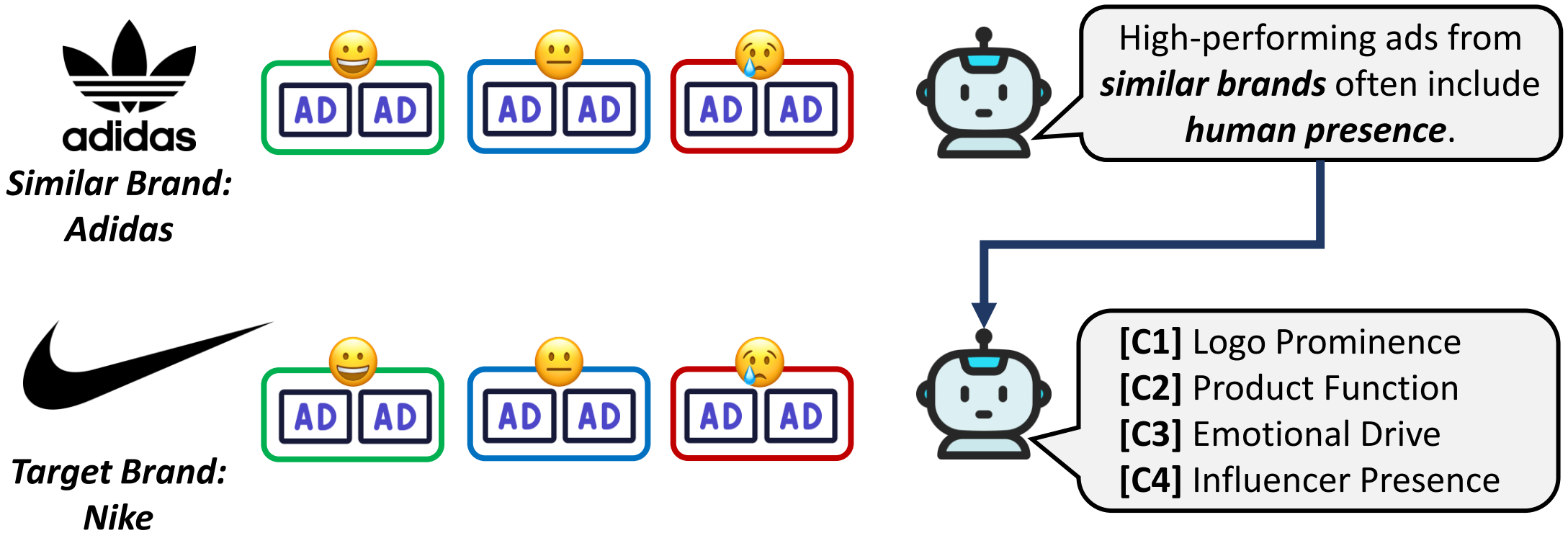
ADvisor: [C1] Brand-specific Criteria Generation

- **Goal:** To address lack of decision criteria under data scarcity.
  - **Step 1:** Similar brand selection.



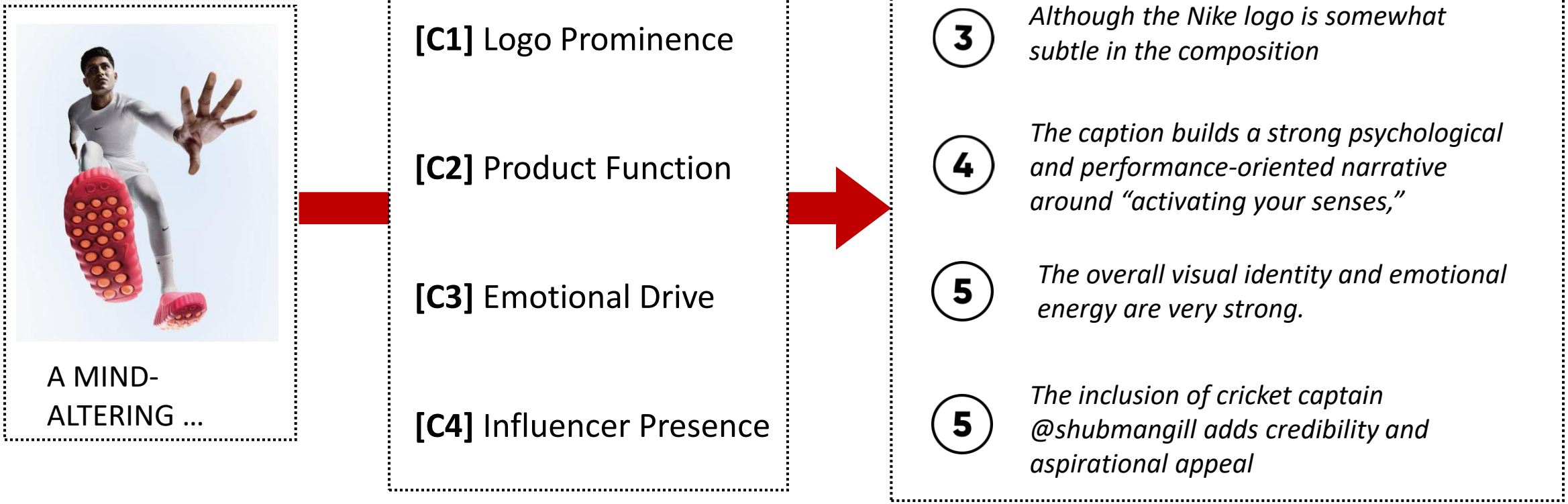
ADvisor: [C1] Brand-specific Criteria Generation

- **Goal:** To address lack of decision criteria under data scarcity.
  - **Step 2:** Reasoning from similar brand & criteria generation.
    - Construct **few-shot examples** (high-medium-low performance ads are grouped).



ADvisor: [C2] Reflection-based Scoring

- **Goal:** To improve reliability and consistency of VLM scoring.
  - **Step 1:** Initial scoring.



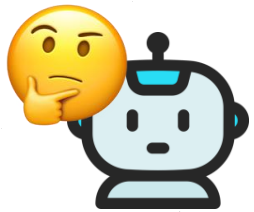
ADvisor: [C2] Reflection-based Scoring

- **Goal:** To improve reliability and consistency of VLM scoring.
  - **Step 2:** Critique on initial score.

A MIND-ALTERING …

[Criteria & Initial scores]

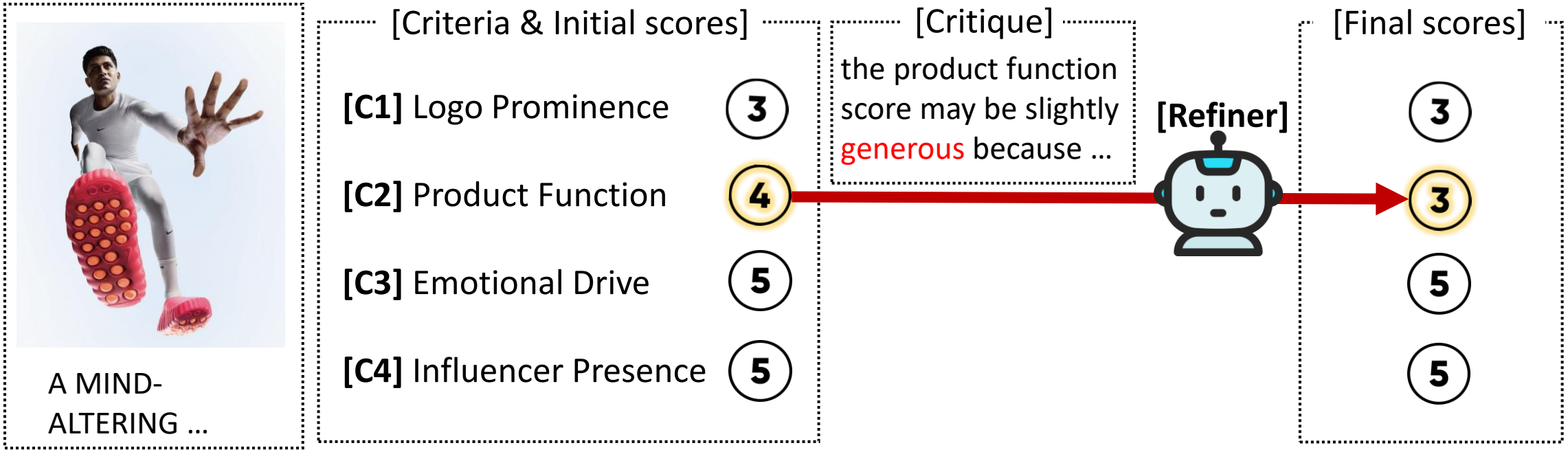
[C1] Logo Prominence	3
[C2] Product Function	4
[C3] Emotional Drive	5
[C4] Influencer Presence	5



Critique: … However, the product function score may be slightly generous because the ad communicates mindset and emotional readiness more clearly *than concrete shoe functionality* or technical benefits, and the Nike logo remains relatively understated.

ADvisor: [C2] Reflection-based Scoring

- **Goal:** To improve reliability and consistency of VLM scoring.
  - **Step 3:** Refinement.



ADvisor: [C3] Ad Ranking

- **Goal:** To learn brand-specific ranking under limited labeled data.
  - Use refined scores as input features and train a brand-specific lightweight ranker (MLP).

[Generated criteria]

[C1] Logo Prominence

[C2] Product Function

[C3] Emotional Drive

[C4] Influencer Presence

[Ads & Refined scores]

A MIND-ALTERING …

3 3 5 5

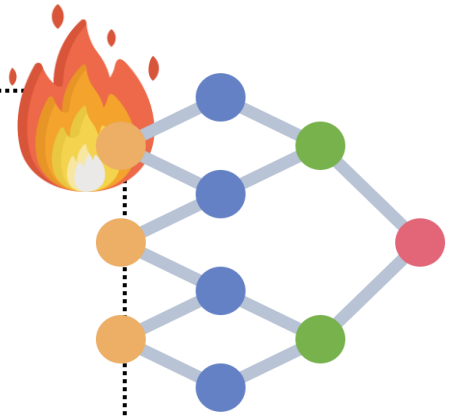
CTR: 0.08

Light work. Sabrina 4 …

5 3 3 1

CTR: 0.04

…



Outline

1. Introduction.
2. Proposed method.
3. Experiments.
4. Conclusion.

Experimental Settings

Datasets

- **Real-world Instagram ads**
- 10 brands (Beauty / Fashion / Platform)
- Avg. 35 ads per brand

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

Experimental Settings

Datasets

- **Real-world Instagram ads**
- 10 brands (Beauty / Fashion / Platform)
- Avg. 35 ads per brand

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

[Instagram ads for example brands]

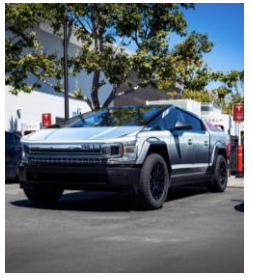
Brand A

A MIND-ALTERING …

Light work. Sabrina 4 …

Brand B

Dark mode Engaged …



Heard you wanted a …

…

Reference: Nike Official Instagram (@nike), Tesla Official Instagram (@teslamotors)

Experimental Settings

Datasets

- **Real-world Instagram ads** 
- 10 brands (Beauty / Fashion / Platform)
- Avg. 35 ads per brand

Baselines

- Basic Multimodal Baselines

  (i) Embedding from pretrained encoders + MLP rankers

  (ii) Zero-shot VLM, Few-shot VLM

- Social Media Popularity (SMP) Prediction

  (i) DEVL, (ii) ECSF, (iii) MMF

Experimental Settings

Performance Labels (Marketing Performance Metrics)

- **CTR** (Click-Through-Rate) ↑ : clicks / impressions
- **CPC** (Cost Per Click) ↓ : spend / clicks
- **CPM** (Cost Per Mille) ↓ : 1000 × spend / impressions

Evaluation Metrics

- **Evaluated with** NDCG@1/3/5

Experimental Results: Accuracy

- **ADvisor** achieves the best overall ranking performance.
  - Single-turn VLM predictors and pretrained multimodal encoders show limited effectiveness.

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**

Experimental Results: Ablation Study

- All key components of **ADvisor** are effective.

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

- **ADvisor-AB:** Augmentation using all brands
- **ADvisor-CB:** No cross-brand augmentation
- **ADvisor-PR:** Using only public data
- **ADvisor-RE:** No critique and refinement
- **ADvisor-RA:** VLM ranking

- The largest drop occurs in **ADvisor-RE**, showing the importance of self-critique and refinement.

Experimental Results: Case Study

- Generated **brand-specific criteria** are **meaningful**.

Fashion Brand A

Human Presence
High-performing advertisements feature a visible person with clear facial presence or eye contact ... capturing attention.

Headline Text Hook
On-image headlines enable immediate … comprehension and are positively associated with higher engagement.

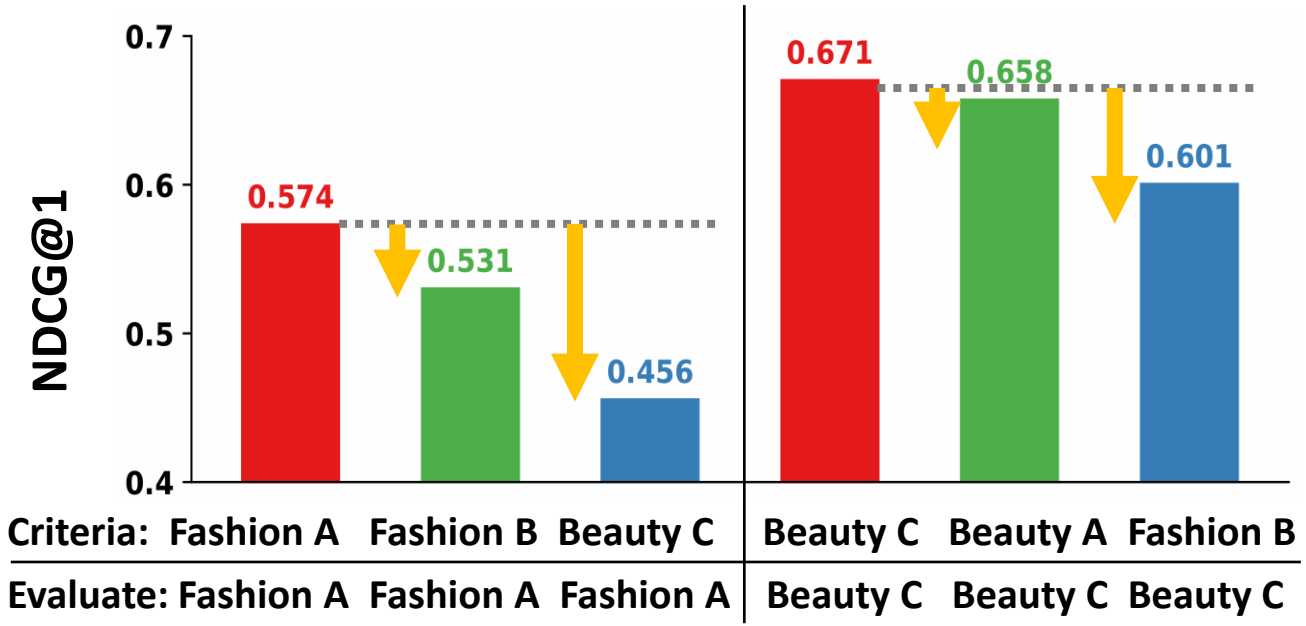
Logo Prominence
Clear and dominant display of the brand logo or signature mark … brand identity cues and increases engagement.

Contextual Relevance
Showing clear use cases or seasonal cues, such as … lifestyle scenarios, helps viewers understand applicability.

Fashion Brand B

Experimental Results: Case Study

- Generated **criteria** are **brand-specific,** and **effective**.
  - Using *similar brand's criteria* leads up to **7.5%** performance drop.
  - Using *different category's criteria* leads up to **20.6%** performance drop.



Experimental Results: Online A/B Testing

- A/B Test Settings
  - Platform: Instagram
  - Compared **top-2 ads selected by ADvisor** & **top-2 ads selected by human experts**
  - Duration: 7 days (Nov. 26 – Dec. 2, 2025)
  - Target audience: female users aged 25+
  - Additional Performance Metric
    - **ROAS (**Return on Ad Spend) ↑ : 1000 × revenue / cost

- **ADvisor** demonstrates real-world effectiveness against human experts.

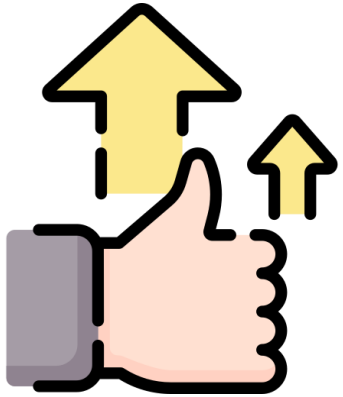
Method	CTR ↑	CPC ↓	ROAS ↑
Human Marketers	8.37%	428	1,070%
ADVISOR	**10.14%**	**231**	**1,219%**
Improvement (%)	21.15%	46.03%	13.93%

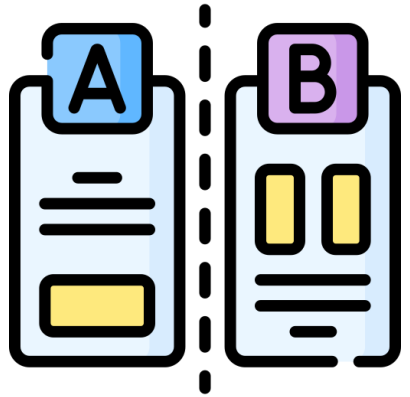
Outline

1. Introduction.
2. Proposed method.
3. Experiments.
4. Conclusion.

Conclusion

- We introduce a novel, practical task, **brand-specific advertisement ranking**.
- We propose **ADvisor** for **brand-specific advertisement ranking.**
- **ADvisor** consists of brand-aware criteria generation, reflection-based scoring, and brand-specific ranking**.**
- **ADvisor** is effective on **both offline** real-world ad ranking & **online** A/B testing.





KAIST AI
Kim Jaechul Graduate School

MADUP
For your innovative growth

ACL 2026
SAN DIEGO JULY 2-7

Thank You!

Q & A

<u>Kyungho Kim*</u>

Yeonje Choi*

Gyurim Hwang

Sejin Chung

Hongseok Lee

Myeong Ho Song

Yeongho Kim

Sunwoo Kim

Jongha Lee

Juyeon Kim

Kijung Shin