Revealing Visual Cognition with AI Simulator: Hierarchical Attention Entropy Derived from Artificial Neural Network

Jingyuan CAI, City University of Hong Kong

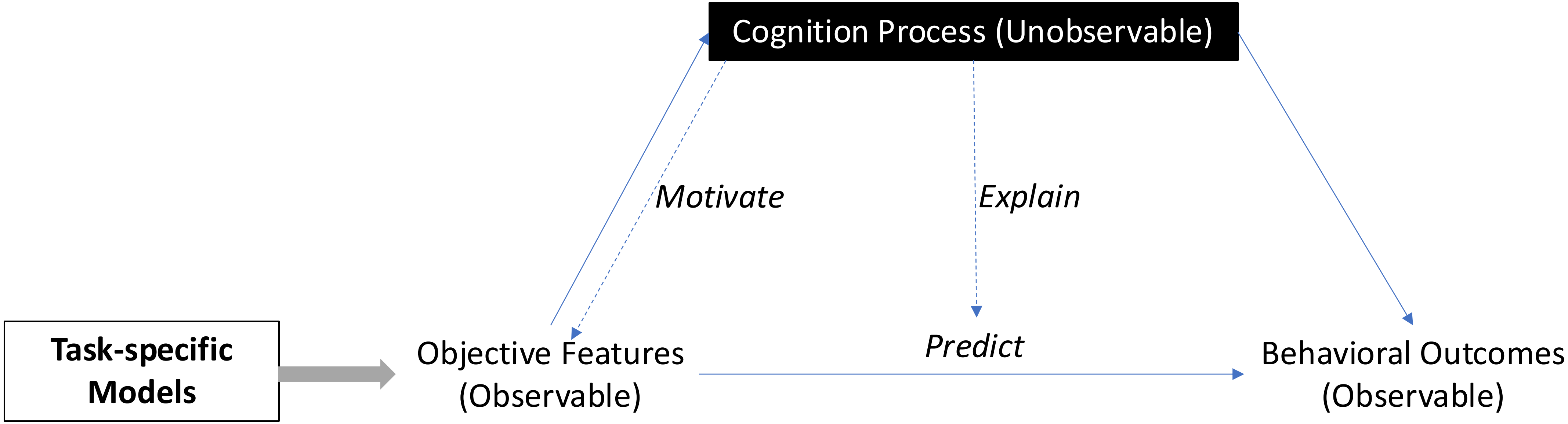
Chong Alex Wang, City University of Hong Kong

2025/04/25

ANN Models Exhibit Brain-like Visual Cognition

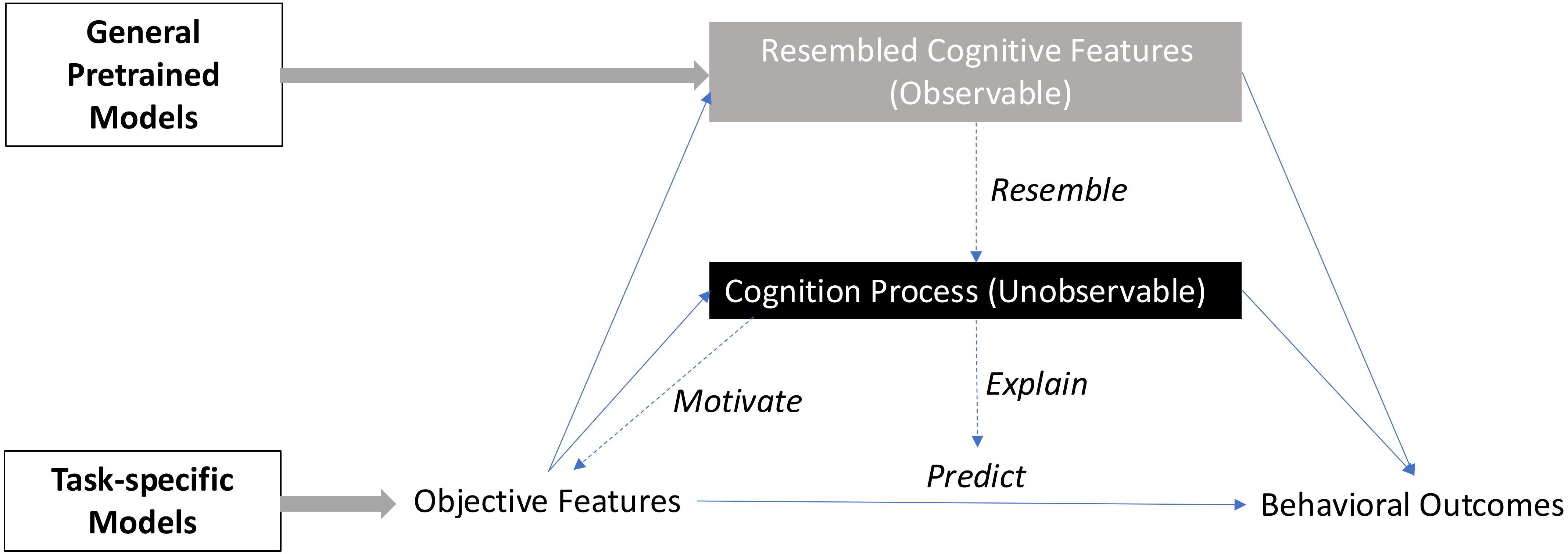
- Pre-trained with large volumes of human-generated data
- Layered information processing inspired by human hierarchical processing structure
- Recent neuroscience studies have revealed strong correlations between ANN layer-wise representations and hierarchical brain activations. (Yamins et al., 2014; Wenliang & Seitz, 2018; Caucheteux & King, 2022; Mischler et al., 2024)
- These findings suggest that ANN layers may contain rich, yet underexplored, insights into human cognitive processes.

Objective Feature-oriented Paradigm in ML-enabled Empirical Research


Cognition Process (Unobservable)
Motivate
Explain
Task-specific Models
Objective Features (Observable)
Predict
Behavioral Outcomes (Observable)

- Human cognition remains a black box in machine learning-enabled IS studies.

Turning Black Box into Grey Box



General Pretrained Models

Resembled Cognitive Features (Observable)

Resemble

Cognition Process (Unobservable)

Motivate

Explain

Task-specific Models

Objective Features

Predict

Behavioral Outcomes

Is it possible to leverage ANNs to resemble human visual cognitive processing, to extract cognitive features, and link them to user behavior?

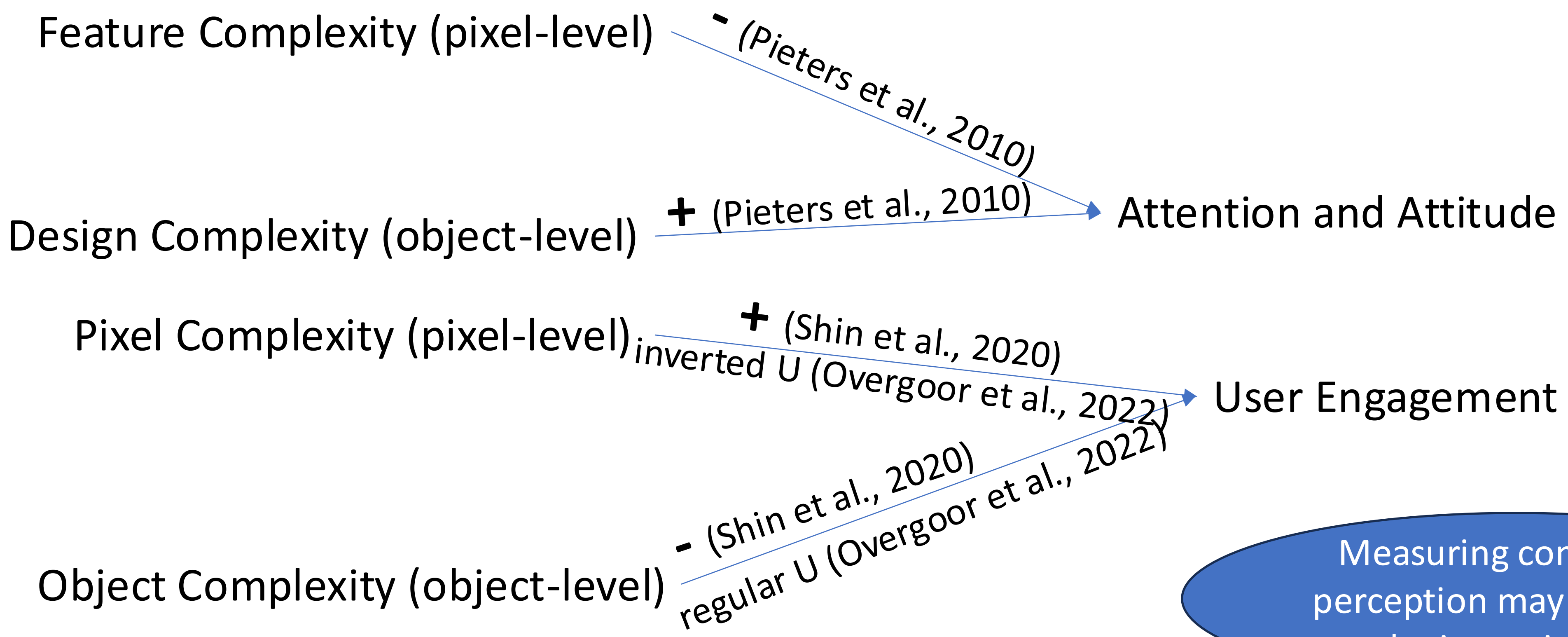
Validation Context: Visual Complexity Perception

- Visual complexity is “the amount of detail or intricacy of line in the picture.” (Snodgrass & Vanderwart, 1980)
- A key factor influencing user engagement (Shin et al., 2020; Overgoor et al., 2022), purchase intention, and consumer attitudes (Pieters et al., 2010; Wang et al., 2024).
- Current measurement:
  - Pixel-level complexity (pixel-level), object-level complexity (object-level) (Shin et al., 2020)
  - Feature complexity (pixel-level), design complexity (object-level) (Pieters et al., 2010)

Objective features derived from images

Validation Context: Visual Complexity Perception

Conflicting findings when linking visual complexity with outcome behavior



Feature Complexity (pixel-level)

- (Pieters et al., 2010)

Design Complexity (object-level)

+ (Pieters et al., 2010)

Attention and Attitude

Pixel Complexity (pixel-level)

+ (Shin et al., 2020)

inverted U (Overgoor et al., 2022)

Object Complexity (object-level)

- (Shin et al., 2020)

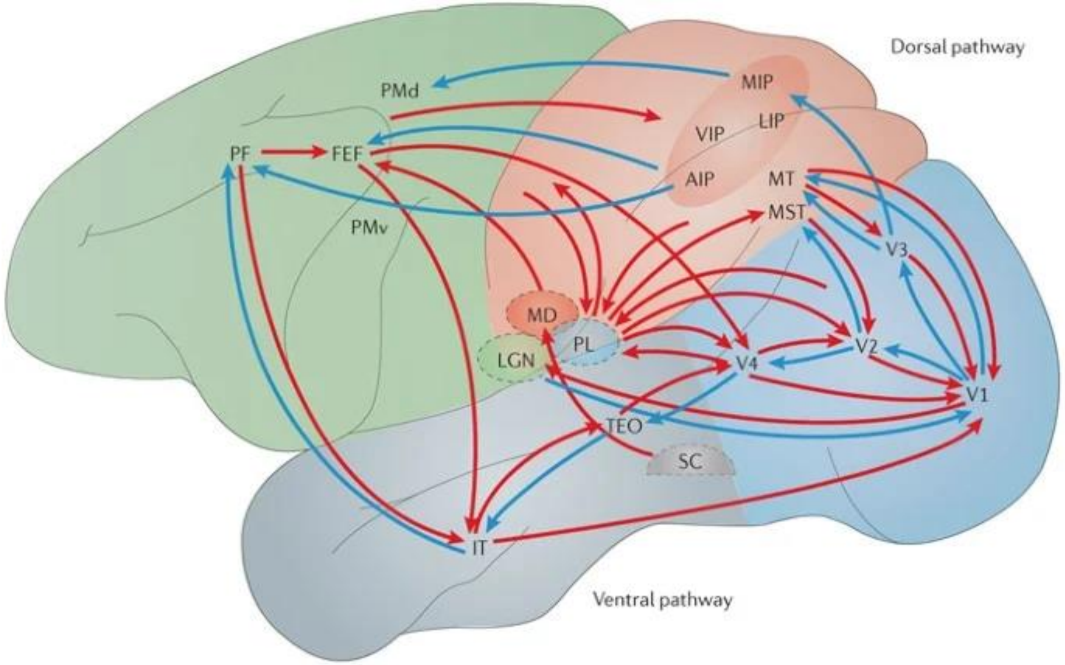
regular U (Overgoor et al., 2022)

User Engagement

Measuring complexity perception may reconcile the inconsistency

Visual Hierarchical Processing

- Primary visual sensory stage
  - “maps” the physical world onto brain tissue
- Object detection stage
  - recognize objects like faces, animals, etc.
- Information association stage
  - associate multisource information for semantic understanding



Nature Reviews | Neuroscience

Visual complexity perception can be separated into hierarchical stages

Can we map layered neural network to brain?

- The self-attention mechanisms in Transformers correlate with human neural processing in vision and language comprehension (Lyu et al., 2024).
- As models perform better, their hierarchical feature extraction becomes more similar to processing hierarchies in the cortex recorded by EEG (Mischler et al., 2024) .
- Similarly, Transformer-based masked word prediction generates activation patterns correlated with brain responses recorded by fMRI and MEG (Caucheteux & King, 2022).
- Transformer self-attention correlates with human gaze durations during reading (Eberle et al., 2022).

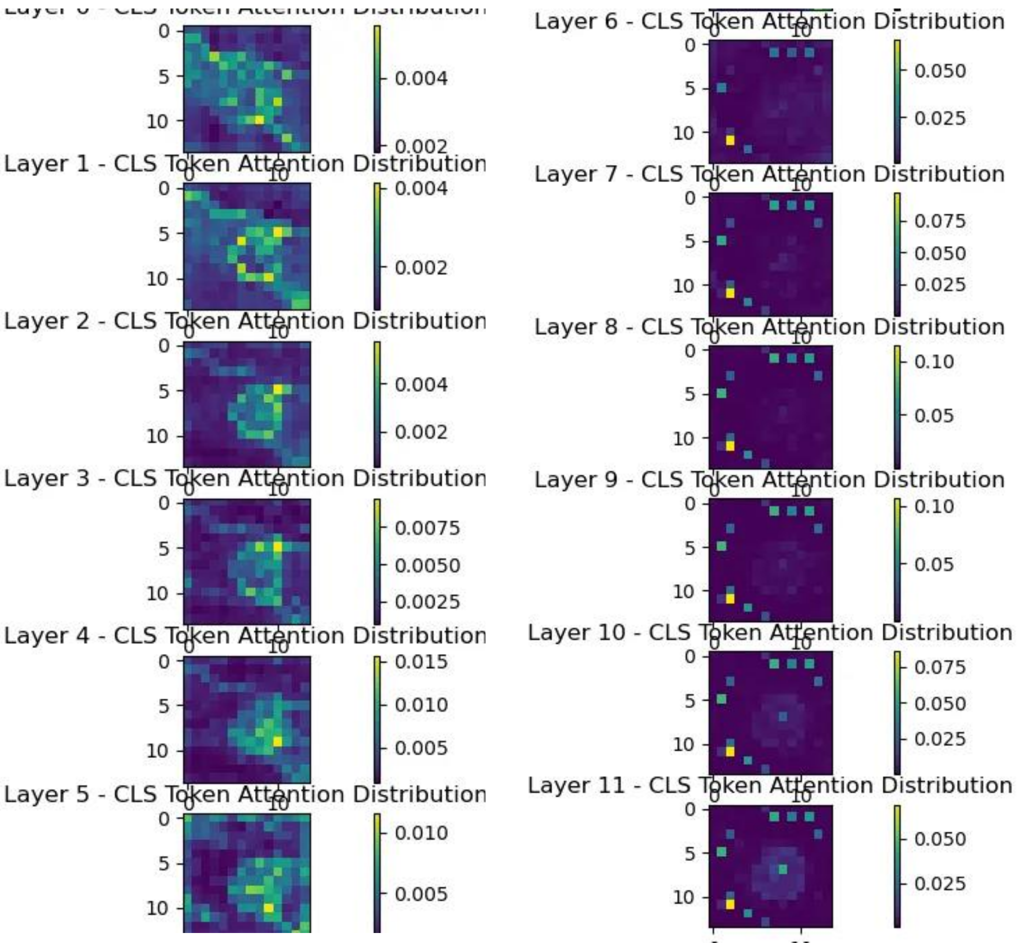
Can we map layered neural network to brain?

- Vision Transformers (ViT): As the receptive field size increases with network depth, more attention is paid to regions that are most semantically relevant (Dosovitskiy et al., 2020).

ANN layers may contain a wealth of information about the human visual cognitive process.

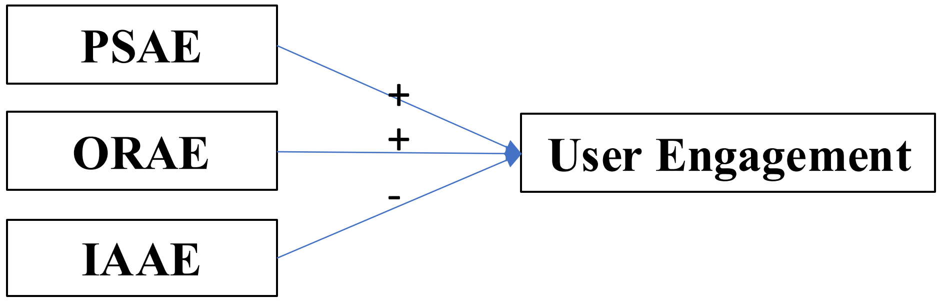
Hierarchical Attention Entropy

- We propose **Hierarchical Attention Entropy (HAE)** to approximate stagewise brain activation in visual complexity perception.
  - Primary Sensory Attention Entropy (PSAE)
  - Object Recognition Attention Entropy (ORAE)
  - Information Association Attention Entropy (IAAE)


Layer 0 - CLS Token Attention Distribution
0.004
0.002
Layer 1 - CLS Token Attention Distribution
0.004
0.002
Layer 2 - CLS Token Attention Distribution
0.004
0.002
Layer 3 - CLS Token Attention Distribution
0.0075
0.0050
0.0025
Layer 4 - CLS Token Attention Distribution
0.015
0.010
0.005
Layer 5 - CLS Token Attention Distribution
0.010
0.005
Layer 6 - CLS Token Attention Distribution
0.050
0.025
Layer 7 - CLS Token Attention Distribution
0.075
0.050
0.025
Layer 8 - CLS Token Attention Distribution
0.10
0.05
Layer 9 - CLS Token Attention Distribution
0.10
0.05
Layer 10 - CLS Token Attention Distribution
0.075
0.050
0.025
Layer 11 - CLS Token Attention Distribution
0.050
0.025
0
5
10

Hypothesis

- **Hypothesis 1**: Primary sensory attention entropy (PSAE) has a positive impact on user engagement.
- **Hypothesis 2**: Object recognition attention entropy (ORAE) has a positive impact on user engagement.
- **Hypothesis 3**: Information association attention entropy (IAAE) has a negative impact on user engagement.



Empirical Test

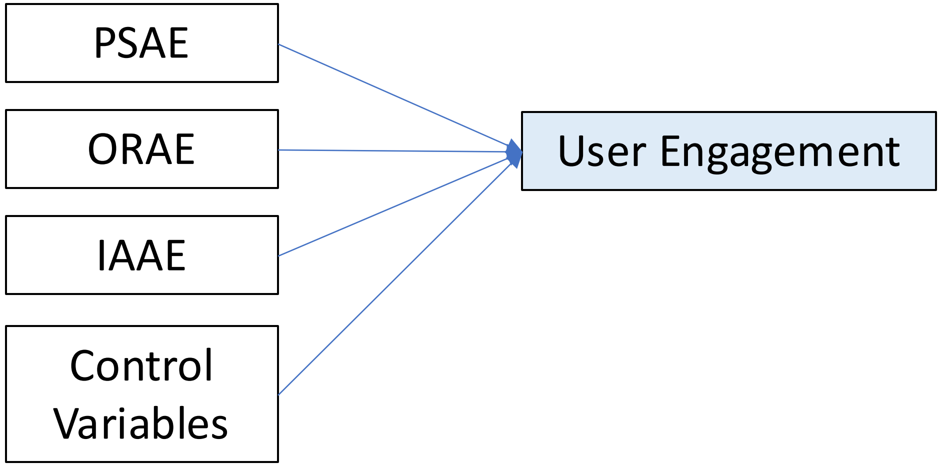
Dataset

- Instagram Dataset
- 16,007 posts by 8,985 influencers
- From July 2012 to May 2019

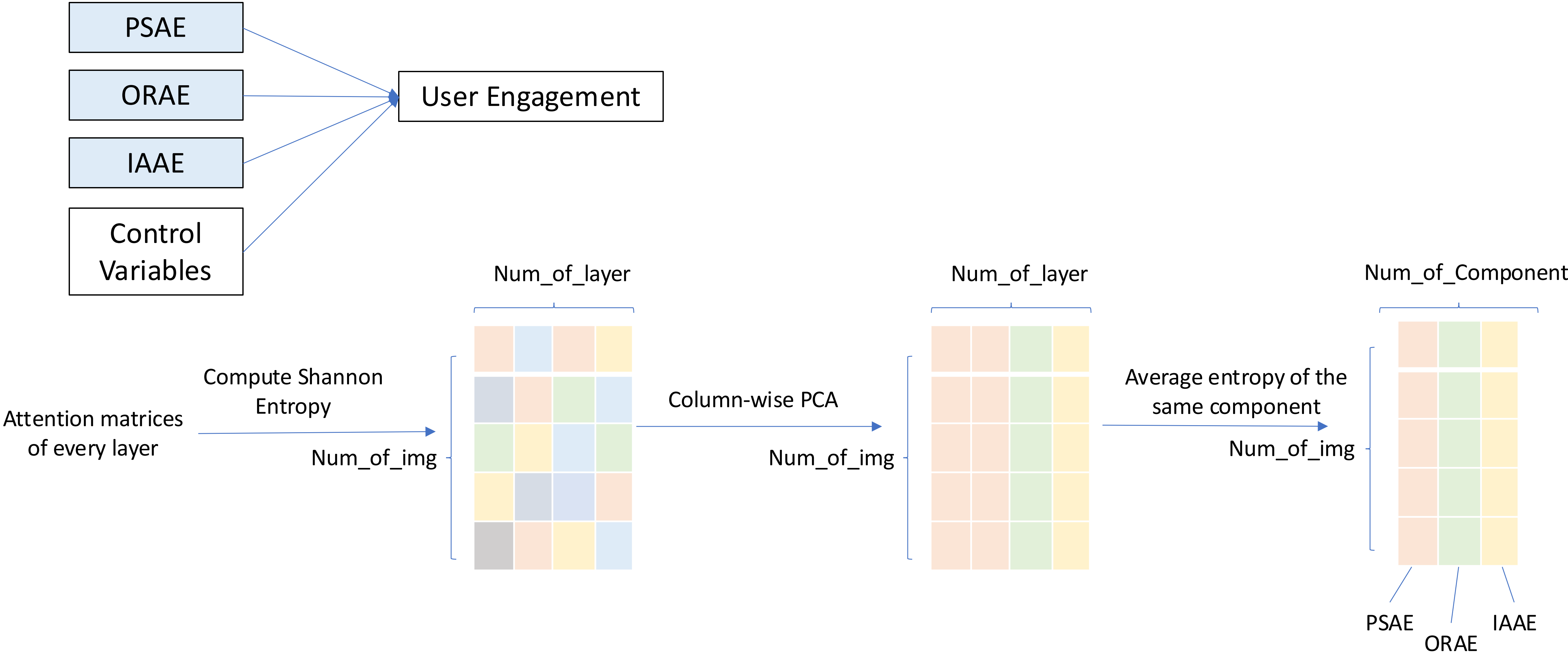
For each post:

- User engagement information (the number of likes and comments)
- Post information (images, text, publish time, sponsorship)
- Influencer information (name, category, number of followers, number of followees, number of posts).

Variables Operationalization



Variables Operationalization



PSAE

ORAE

IAAE

Control Variables

User Engagement

Attention matrices of every layer

Compute Shannon Entropy

Num_of_layer

Num_of_img

Column-wise PCA

Num_of_layer

Num_of_img

Average entropy of the same component

Num_of_Component

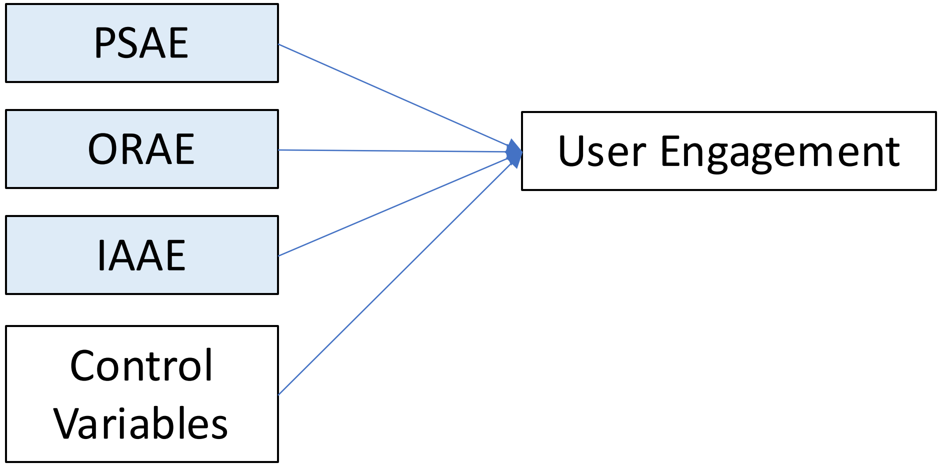
Num_of_img

PSAE

ORAE

IAAE

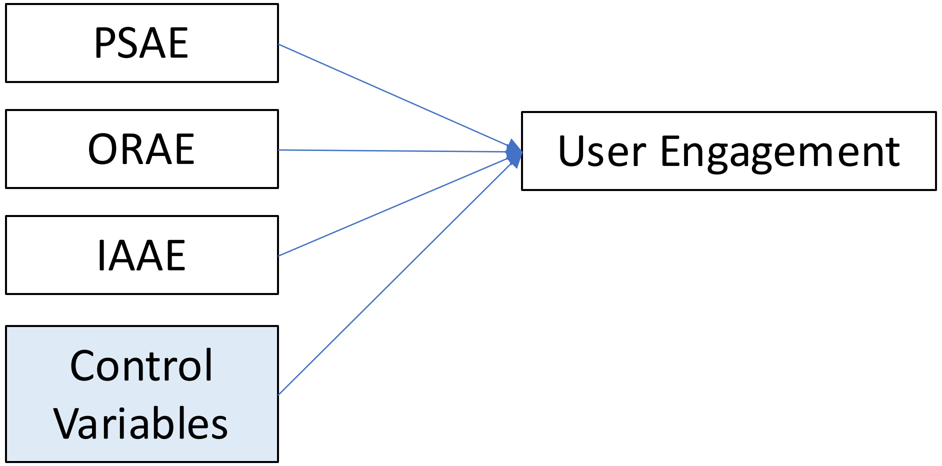
Variables Operationalization



Vision encoder of CLIP (Contrastive Language–Image Pre-training) neural network

Hierarchical Attention Entropy	Layers Included
Stage 1 (Primary Sensory Attention Entropy)	Layer 1, Layer 2, Layer 3, Layer 4, Layer 5
Stage 2 (Object Recognition Attention Entropy)	Layer 7, Layer 8
Stage 3 (Information Association Attention Entropy)	Layer 9, Layer 10, Layer 11

Variables Operationalization

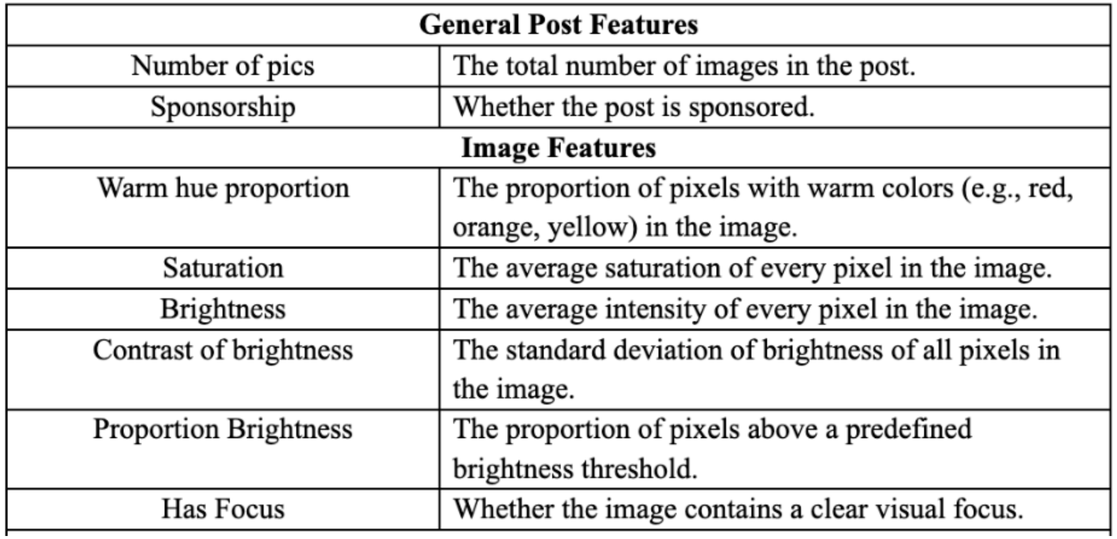

PSAE

ORAE

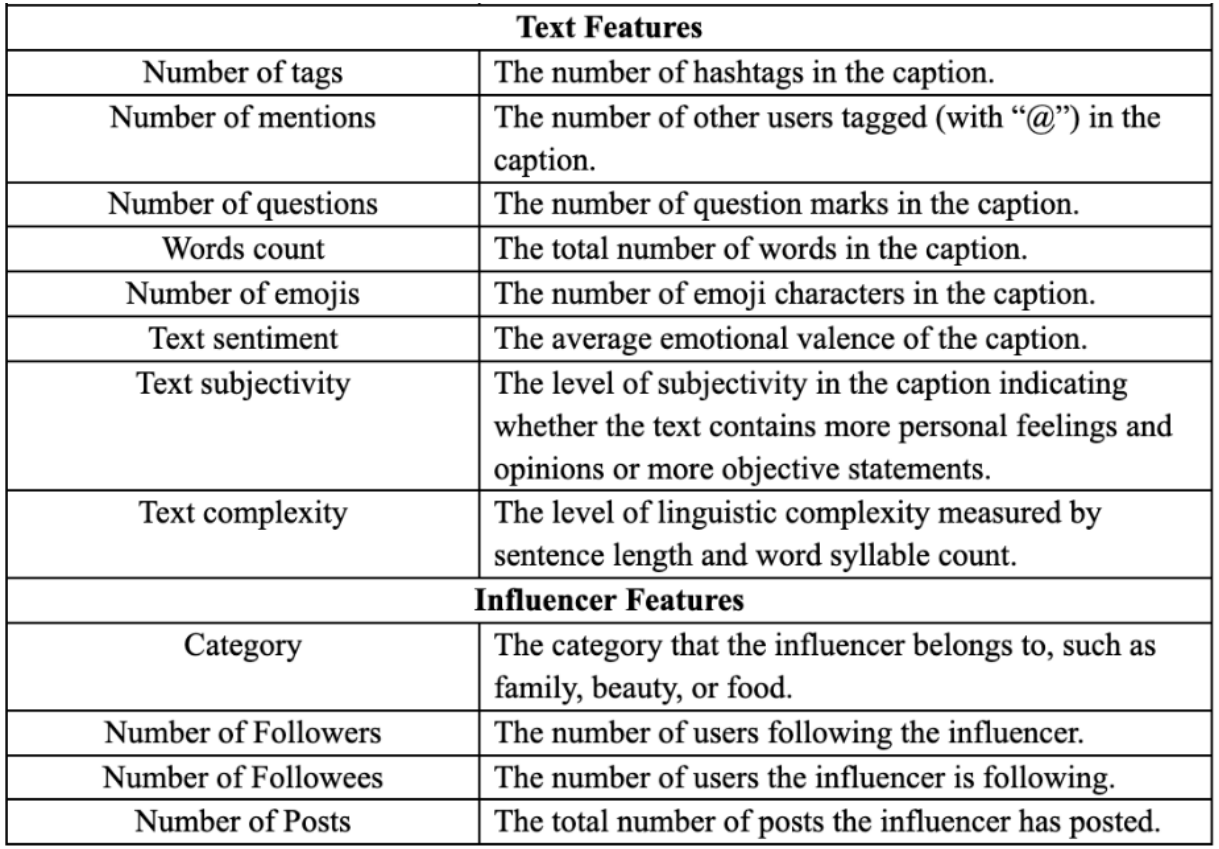
IAAE

Control Variables

User Engagement



General Post Features	
Number of pics	The total number of images in the post.
Sponsorship	Whether the post is sponsored.
Image Features	
Warm hue proportion	The proportion of pixels with warm colors (e.g., red, orange, yellow) in the image.
Saturation	The average saturation of every pixel in the image.
Brightness	The average intensity of every pixel in the image.
Contrast of brightness	The standard deviation of brightness of all pixels in the image.
Proportion Brightness	The proportion of pixels above a predefined brightness threshold.
Has Focus	Whether the image contains a clear visual focus.



Text Features	
Number of tags	The number of hashtags in the caption.
Number of mentions	The number of other users tagged (with "@") in the caption.
Number of questions	The number of question marks in the caption.
Words count	The total number of words in the caption.
Number of emojis	The number of emoji characters in the caption.
Text sentiment	The average emotional valence of the caption.
Text subjectivity	The level of subjectivity in the caption indicating whether the text contains more personal feelings and opinions or more objective statements.
Text complexity	The level of linguistic complexity measured by sentence length and word syllable count.
Influencer Features	
Category	The category that the influencer belongs to, such as family, beauty, or food.
Number of Followers	The number of users following the influencer.
Number of Followees	The number of users the influencer is following.
Number of Posts	The total number of posts the influencer has posted.

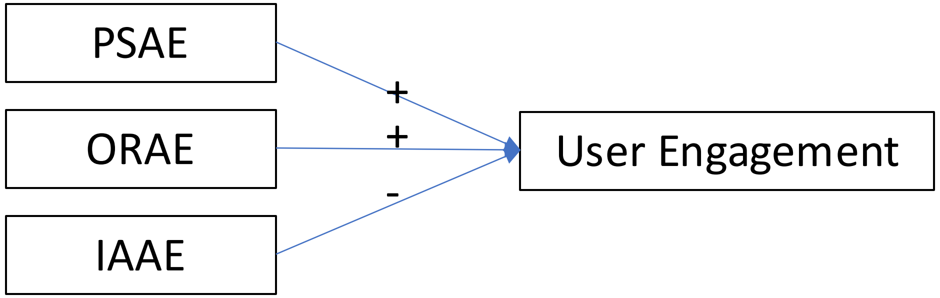
Estimation Model

$$\log\big(E(Likes_i)\big) = \alpha_{10} + \alpha_{11}PSAE_i + \alpha_{12}ORAE_i + \alpha_{13}IAAE_i + \alpha_{14}Controls_i$$
$$+Month_i + Quarter_i + DayOfWeek_i + IsHoliday_i + \varepsilon_{1i}$$

$$\log\big(E(Comments_i)\big) = \alpha_{20} + \alpha_{21}PSAE_i + \alpha_{22}ORAE_i + \alpha_{23}IAAE_i + \alpha_{24}Controls_i$$
$$+Month_i + Quarter_i + DayOfWeek_i + IsHoliday_i + \varepsilon_{2i}$$

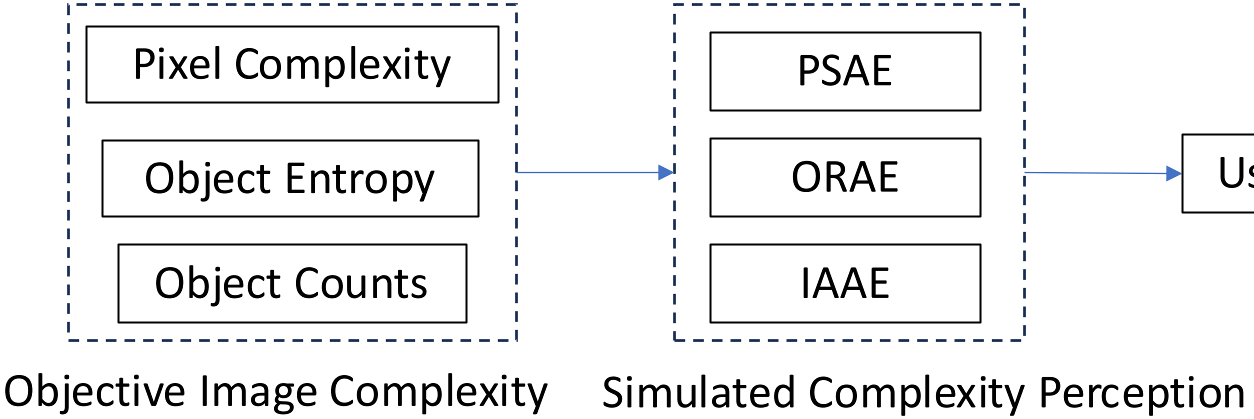
Main Results

Variables	Likes		Comments	
	Estimates	S.E.	Estimates	S.E.
PSAE	0.9768***	0.103	1.0552***	0.105
ORAE	0.9602***	0.071	0.8193***	0.072
IAAE	-0.9230***	0.070	-0.7291***	0.071



Potential capacity difference in processing low-level and high-level visual features

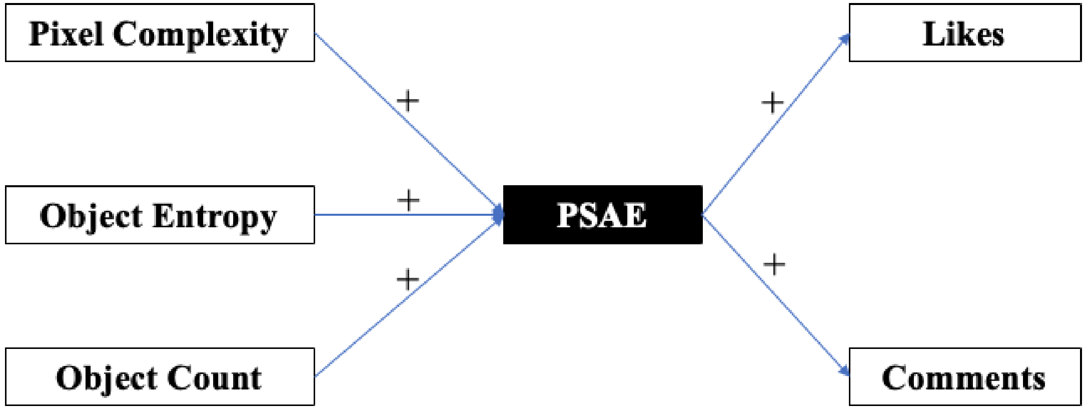
Mediation Analysis



3 (pixel complexity/object entropy/object count) *3 (PSAE, ORAE, and IAAE)*2 (likes/comments) bootstrap-based mediation analysis (1,000 iterations)

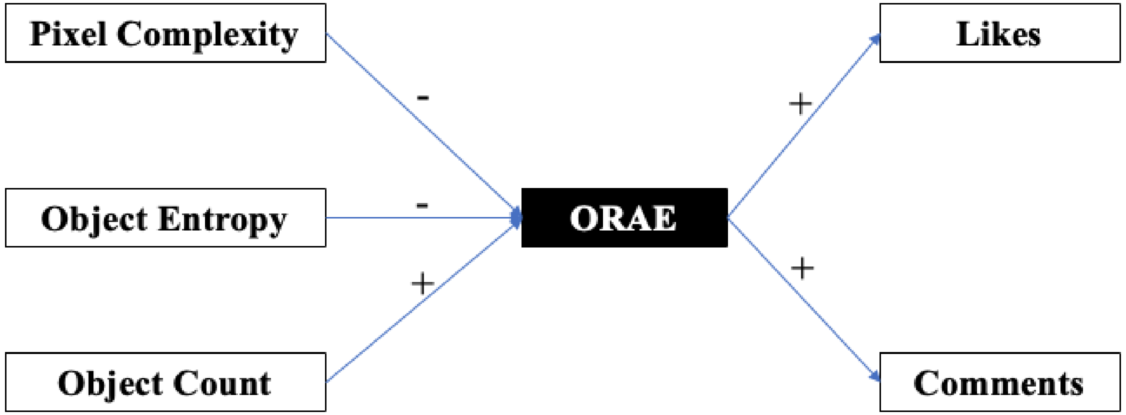
Mediation Analysis

IV	Mediator	DV	Indirect Effect	95% CI	Significant
Pixel Complexity	PSAE	Likes	0.1509	[0.0500, 0.2515]	TRUE
		Comments	0.152491	[-0.0008, 0.3142]	FALSE
Object Entropy		Likes	0.0146	[0.0080, 0.0216]	TRUE
		Comments	0.01299	[0.0022, 0.0233]	TRUE
Object Count		Likes	0.0035	[0.0004, 0.0067]	TRUE
		Comments	0.0055	[0.0008, 0.0100]	TRUE



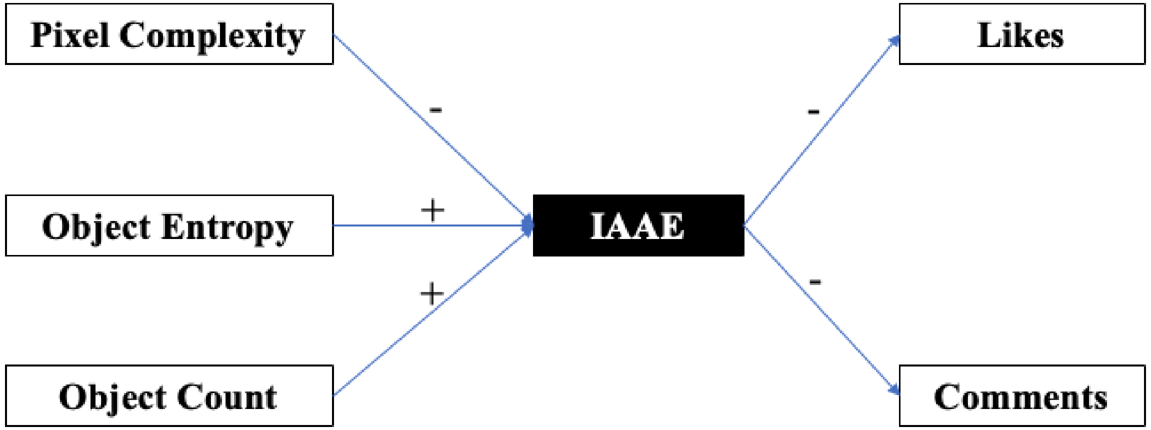
Mediation Analysis

IV	Mediator	DV	Indirect Effect	95% CI	Significant
Pixel Complexity	ORAE	Likes	-0.4233	[-0.6321, -0.2182]	TRUE
		Comments	-0.5565	[-0.8230, -0.2791]	TRUE
Object Entropy		Likes	-0.0016	[-0.0028, -0.0006]	TRUE
		Comments	-0.0013	[-0.0027, -0.0002]	TRUE
Object Count		Likes	0.0019	[0.0007, 0.0034]	TRUE
		Comments	0.0014	[0.0001, 0.0031]	TRUE

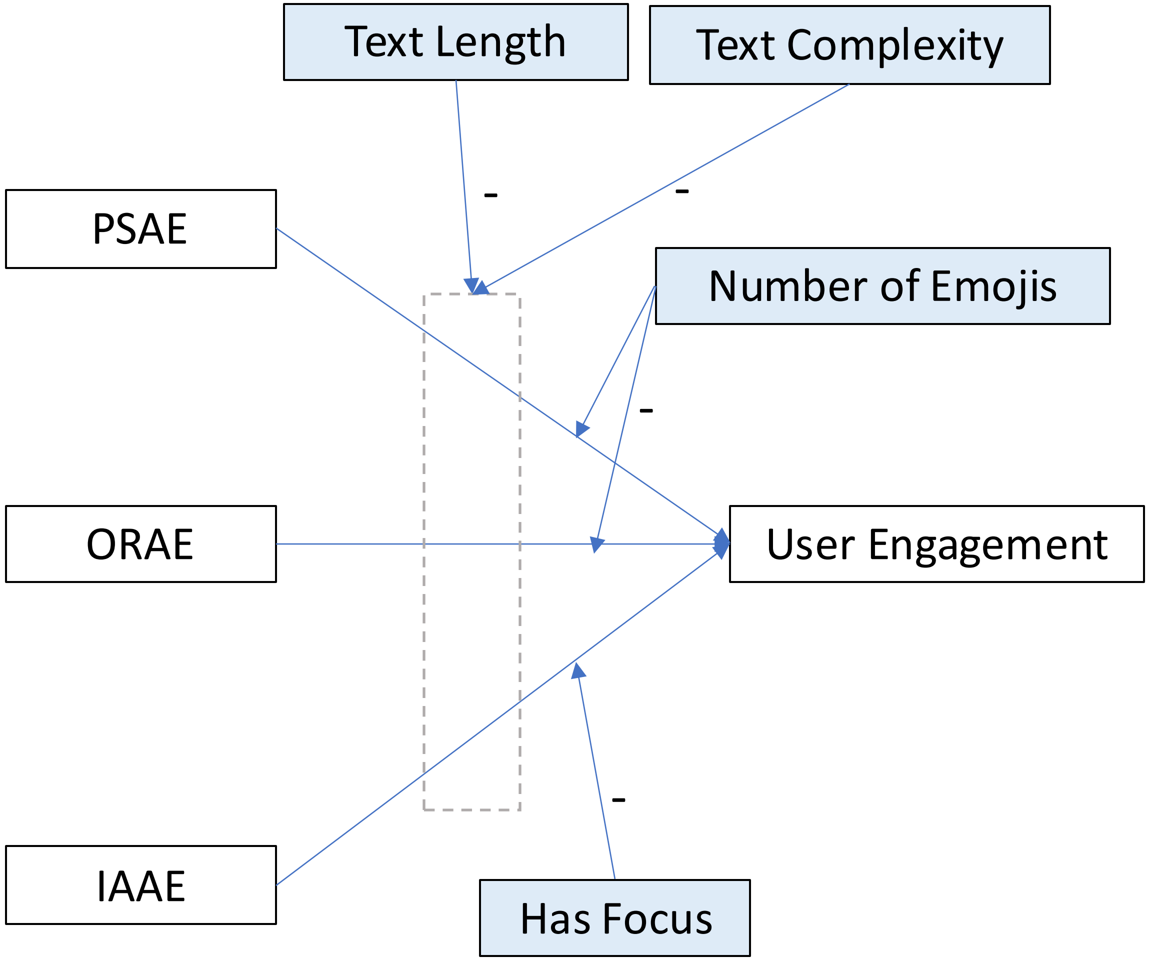


Mediation Analysis

IV	Mediator	DV	Indirect Effect	95% CI	Significant
Pixel Complexity	IAAE	Likes	0.1098	[0.0629, 0.1593]	TRUE
		Comments	0.0766	[0.0085, 0.1435]	TRUE
Object Entropy		Likes	-0.0087	[-0.0126, -0.0047]	TRUE
		Comments	-0.0091	[-0.0152, -0.0035]	TRUE
Object Count		Likes	-0.0060	[-0.0087, -0.0035]	TRUE
		Comments	-0.0050	[-0.0085, -0.0015]	TRUE



Additional Analysis



下面的不需要

Additional Analysis

- Results are stable to patch sizes of the model and sample dataset.
- We also tested the vision encoder of BLIP2, and the results generally hold a similar trend.

Implications

- Methodological implication:
  - A novel approach for data-intensive theory building: ANN attention entropy offers a computational proxy for latent cognitive processes.
- Theoretical implication:
  - Extend the visual complexity literature: a layered visual complexity framework distinguishes between different levels of complexity perception.
- Practical implication:
  - Creators should consider potential variations in viewers’ cognitive capacities at different visual levels.

Limitations

- Our analysis focuses solely on Instagram
- Our alignment between ANN and the human brain is currently limited to stage-level activation.
- We cannot guarantee the results to be stable in every multimodal transformer model, as their correlations with brains may differ due to different pre-training methods.

Thanks for listening!