CAP 🧢 : Evaluation of Persuasive and Creative Image Generation

Aysan Aghazadeh, Adriana Kovashka
Department of Computer Science - University of Pittsburgh

ICCV HONOLULU HAWAII OCT 19-23, 2025

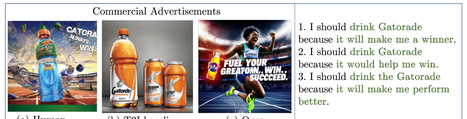
Overview

Task

Prompt → Ad Image

- Prompt: Action-Reason Message (AR)
  I should {action} because {reason}
- Ad Image: Creative and Persuasive

Commercial and Public Service Ads



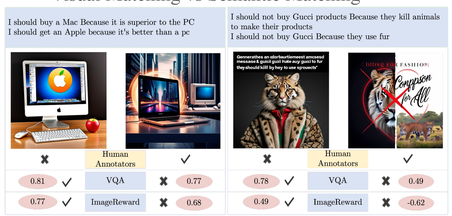
Challenge

- Prompt is implicit: Different images convey same AR message → Visual Matching is not enough
- Persuasion and Creativity are subjective

Alignment of Image and Message (AIM)

Motivation

- Visual Matching vs Semantic Matching

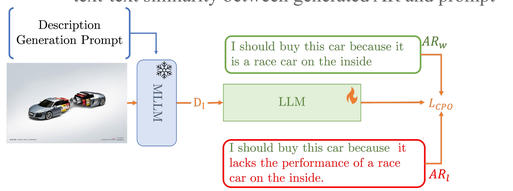


Related works

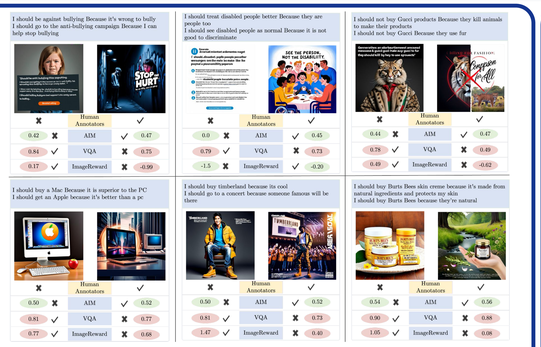
- Image-Image Similarity
  FID (Heusel, et al.), IS (Salimans, et al.)
- Image-Text Similarity:
  - CLIPScore (Hessel, et al.): sim(CLIP embeddings)
  - Training LLMs/VLMs
    VQA-score (Lin, et al.): P(Yes|Prompt, I) from LLM
    ImageReward (Xu, et al.) : Reward Model
  - Zero-shot LLMs: DSG (Cho, et al.)

Evaluation Method

- We fine-tune LLM to interpret the image **semantically** accurate
- **AR generation:** Image description generation→AR generation
- **Score:** Weighted average of action text similarity and reason text-text similarity between generated AR and prompt



$$AIM(I_{gen}, AR_m) = \frac{Sim(A_{gen}, A_m) + \alpha\, Sim(R_{gen}, R_m)}{1+\alpha}$$



Results

- **Baselines do not capture semantic mismatch**
- **+93%** higher agreement than baselines
  COM: 57%
  PSA: 140%
- Low 0-shot agreement (Krippendorff's Alpha)
  - Importance of training

Annotators	COM	PSA	All
H, ImageReward	0.12	0.06	0.11
H, VQAScore	0.38	0.24	0.31
H, CLIPScore	0.04	0.34	0.17
H, AIM (InternVL, LLAMA3) (0-shot)	0.17	0.26	0.18
H, AIM (InternVL, LLAMA3)	0.60	**0.82**	**0.68**
H, AIM (InternVL, QwenLM)	0.62	0.56	0.60
H, AIM (QwenVL, LLAMA3)	0.62	0.56	0.60
H, AIM (QwenVL, QwenLM)	**0.72**	0.56	0.65
H1, H2	0.86	0.85	0.86

T2I Performance

Evaluation		Image		COM Ads			PSA Ads		
MLLM	LLM	T2I	$I_{input-text}$	AIM	C_{obj}	P_{c+A}	AIM	C_{obj}	P_{c+A}
InternVL	LLAMA3	SDXL	I_{AR}	0.50	2.03	0.62	0.32	1.33	0.48
		AuraFlow	I_{AR}	0.50	2.12	0.64	0.31	1.36	0.42
			I_{LLAMA3}	0.53	2.25	**0.70**	0.43	1.87	**0.60**
			I_{QwenLM}	**0.55**	**2.34**	0.59	**0.48**	**2.05**	0.54
		FLUX	I_{AR}	0.51	2.06	0.48	0.43	1.83	0.44
			I_{LLAMA3}	**0.54**	2.20	0.52	**0.47**	1.93	**0.53**
			I_{QwenLM}	0.53	**2.35**	**0.55**	**0.47**	**2.06**	0.46
	QwenLM	SDXL	I_{AR}	0.49	1.90	0.50	0.44	1.32	0.35
		AuraFlow	I_{AR}	0.47	1.91	**0.50**	0.30	1.28	0.41
			I_{LLAMA3}	0.50	**2.50**	**0.50**	**0.49**	1.83	0.44
			I_{QwenLM}	**0.51**	2.18	**0.50**	0.48	**2.04**	**0.45**
		FLUX	I_{AR}	0.50	1.94	0.49	0.47	1.86	0.45
			I_{LLAMA3}	**0.51**	2.04	0.49	**0.48**	1.93	0.45
			I_{QwenLM}	**0.51**	**2.18**	**0.56**	**0.48**	**2.03**	**0.46**
QwenVL	LLAMA3	SDXL	I_{AR}	0.52	2.06	0.52	0.45	1.88	0.38
		AuraFlow	I_{AR}	0.51	2.08	0.44	0.45	1.92	0.35
			I_{LLAMA3}	0.53	2.19	**0.54**	0.47	2.03	0.46
			I_{QwenLM}	**0.54**	**2.32**	**0.54**	**0.48**	**2.08**	**0.47**
		FLUX	I_{AR}	0.51	2.02	0.47	0.46	1.89	0.42
			I_{LLAMA3}	**0.53**	2.17	**0.49**	0.47	1.97	0.45
			I_{QwenLM}	**0.53**	**2.30**	0.47	**0.48**	**2.09**	**0.46**
	QwenLM	SDXL	I_{AR}	0.49	1.93	0.43	0.44	1.84	0.37
		AuraFlow	I_{AR}	0.48	1.95	0.43	0.44	1.87	0.36
			I_{LLAMA3}	0.50	2.18	0.46	0.46	1.97	0.43
			I_{QwenLM}	**0.52**	**2.20**	**0.47**	**0.48**	**2.05**	**0.44**
		FLUX	I_{AR}	0.48	1.94	0.44	0.46	1.92	0.40
			I_{LLAMA3}	0.49	2.03	**0.47**	0.47	1.97	0.43
			I_{QwenLM}	**0.52**	**2.20**	**0.47**	**0.48**	**2.06**	**0.44**

- Low AIM, C_{obj}, and P_{C+A} on implicit prompt
- Lower AIM, C_{obj}, and P_{C+A} on PSA ads (more implicit prompts)
- Higher scores when using explicit descriptions generated by LLMs

Creativity of Images

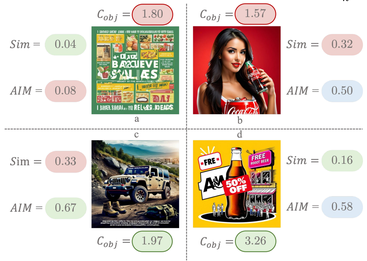
Creativity Criteria → Evaluation Method

- Uniqueness → Distant from visual elements in AR (Sim with elements ↓)
- Conveying the message → Higher semantic alignment of image and AR (AIM ↑)

$$C_{obj} = \frac{AIM(I_{gen}, AR_m)}{\frac{1}{n} \times \Sigma_{obj \in objects}\, Sim(I_{gen}, obj) + 0.01}$$

Results

- **LLMs do not evaluate creativity accurately** (No prior creativity metric)
- Agreement (Krippendorff's Alpha)
  0.5 (out of 1) improvement compared to LLMs
- Balancing criteria
  - Uniqueness of Image (Sim)
  - Conveying the AR message (AIM)



Annotators	COM	PSA	All
H, C_{LLM}	-0.03	0.15	0.04
H, C_{obj}	**0.57**	**0.53**	**0.54**
H1, H2	0.70	0.78	0.73

CAP

Contributions

Proposed **CAP evaluation:**

- **Creativity:** Balancing uniqueness and alignment
- **Alignment:** Evaluating what image convey instead of what it shows
- **Persuasion:** Incorporating marketing factors and AIM

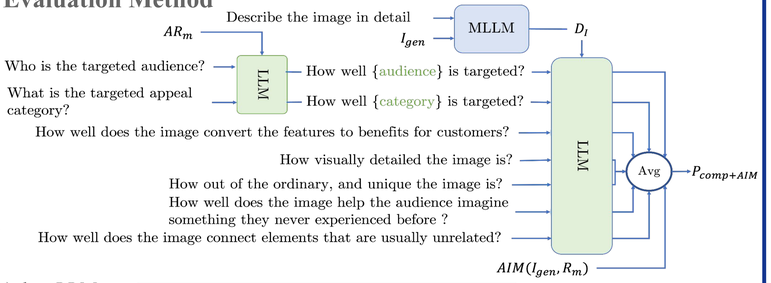
Highlighted the struggle of T2I models with **implicit prompt**

Persuasion of Images

Evaluation Overview

- We use effective persuasion factors analyzed in marketing research in computational way
- Reason in AR captures how to persuade the audience (Included in metric)

Evaluation Method



Results

- **LLMs do not evaluate persuasion accurately** (No prior persuasion metric)
- **+130%** higher agreement (Krippendorff's Alpha) than LLMs
  COM: 220%
  PSA: 180%
- Higher agreement when adding alignment of image and reason
  - Agreement increased by 23%
  - Higher increase in COM advertisements

Annotators	COM	PSA	All
H, P_{LLM}	0.27	0.26	0.27
H, P_{comp} (InternVL, LLAMA3)	0.83	0.54	0.65
H, $P_{comp+AIM}$ (InternVL, LLAMA3)	0.85	**0.75**	**0.80**
H, $P_{comp+AIM}$ (QwenVL, LLAMA3)	0.73	0.63	0.68
H, $P_{comp+AIM}$ (InternVL, QwenLM)	**0.89**	0.30	0.63
H, $P_{comp+AIM}$ (QwenVL, QwenLM)	**0.89**	0.74	0.74
H1, H2	0.80	0.56	0.70