CAP🧢 :
Evaluation of Persuasive and Creative Image Generation

Aysan Aghazadeh, Adriana Kovashka

Department of Computer Science, University of Pittsburgh

ICCV 2025

ICCV OCT 19-23, 2025 HONOLULU HAWAII

University of Pittsburgh

Introduction

Advertisement Generation Task

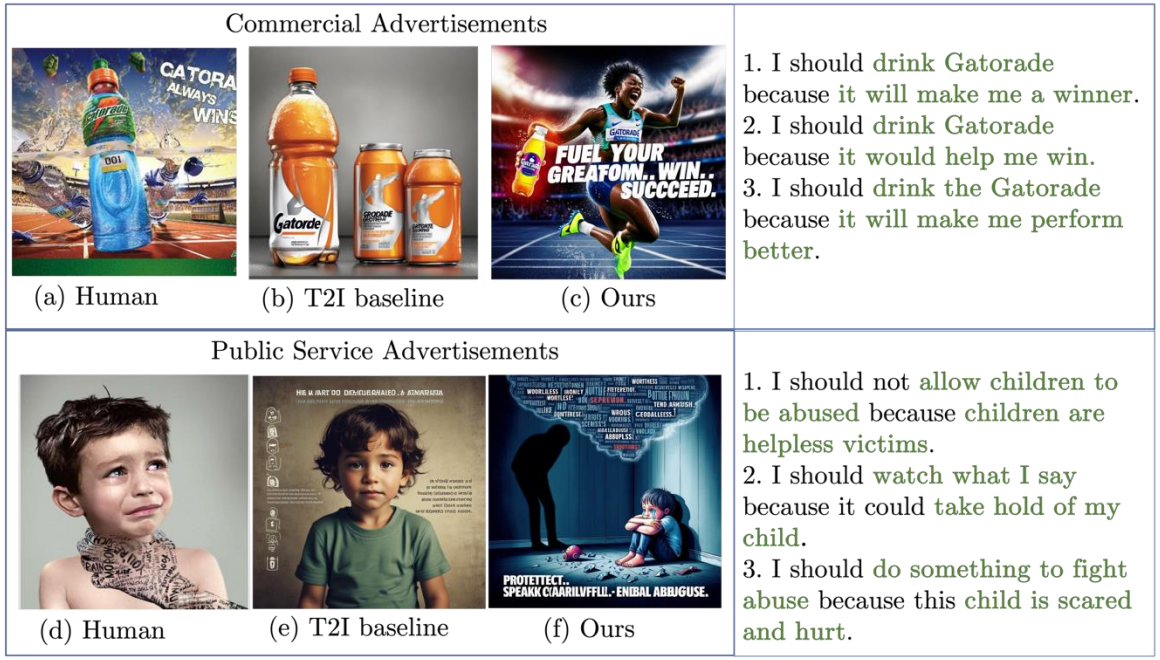
- **Definition**
  Given advertisement message, generating creative and persuasive image aligned with message
- **Advertisement message (Action-reason Statement)**
  Real advertisement image interpretation from PittAd dataset [1]
  - **Action:** The action the advertisement image should convince the audience to take
    Ex. I should drive a Subaru
  - **Reason:** The reason advertisement image use to convince the audience to take the action
    Ex. It is reliable

Implicit Prompt:
I should {action} because {reason}

I should drive a Subaru because it is reliable

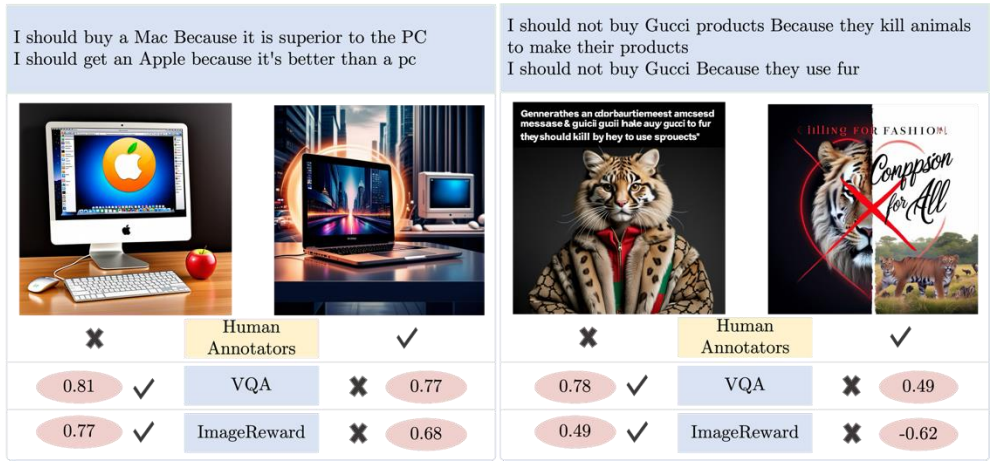
Advertisement Generation Task

- **Advertisement types**
  - **Commercial Advertisement (COM)**
    Advertisements promoting a product or a service
  - **Public Service Advertisement (PSA)**
    Advertisements with goal of changing/adding a behavior in society



Motivation

- Advertisement content criteria
  - Semantic Alignment: Conveying the message
  - Creativity: Being unique and relevant to the AR
  - Persuasion: Being convincing
- Existing T2I metrics
  - High performance in capturing visual mismatch
    - Objects, Object Attributes, Object Composition
  - Need for capturing the semantic mismatch
    - How well do these metrics capture semantic mismatch?
  - No metric for creativity/persuasion
    - How well do LLMs/MLLMs perform in evaluating creativity/persuasion?



Overview

- CAP
  - **Creativity**: C_{obj} is a metric for evaluating creativity in advertisement images with a focus on:
    - Uniqueness
    - Alignment
  - **Alignment**: AIM (Alignment of Image and Message) is an evaluation method for capturing both semantic and visual mismatch
  - **Persuasion**: $P_{comp+AIM}$ is the method for assessing how convincing the image is
    - Designed based on marketing criteria for persuasion
- T2I models
  - Existing models have significant performance in generating high-quality images from explicit descriptions.
    - How well do they perform when the prompt is implicit?
    - How creative these models are?

Related Works

Prompt-Image Alignment

- Image-Image Metrics: FID, IS
- Text-Image Metrics:
  - VLM-based methods - Low accuracy on complex prompts
    - CLIP-score[2]: Similarity of CLIP embeddings of image and text
  - Training LLMs/VLMs – Low accuracy on capturing semantic mismatch
    - VQA-score[3]: LMMs trained to answer the “Does the image show {prompt}?” question
    - Image-reward[4]: Reward model trained on RLHF data for image generation
  - Zero-shot LLMs/MLLMs
    - Davidsonian Scene Graph (DSG)[5]: Generates question given the prompt
      Depends on explicitness of the prompt

Persuasion and Creativity

- **Persuasion in Language**
  - Comparison of persuasion between the generated text and human written text [6]
  - Evaluation of models performing as a judge for persuasion of content [7]
  - Evaluation of persuasion of textual content and proposing methods to improve [8]
- **Non-computational Analysis of Persuasion and Creativity**
  - Analyze of effectiveness of different persuasion factors [9]
  - Introducing different persuasion factors and strategies in Ads [10]
  - Introducing different creativity factors [11]
  - Analyze of influence of creativity on persuasion [12]

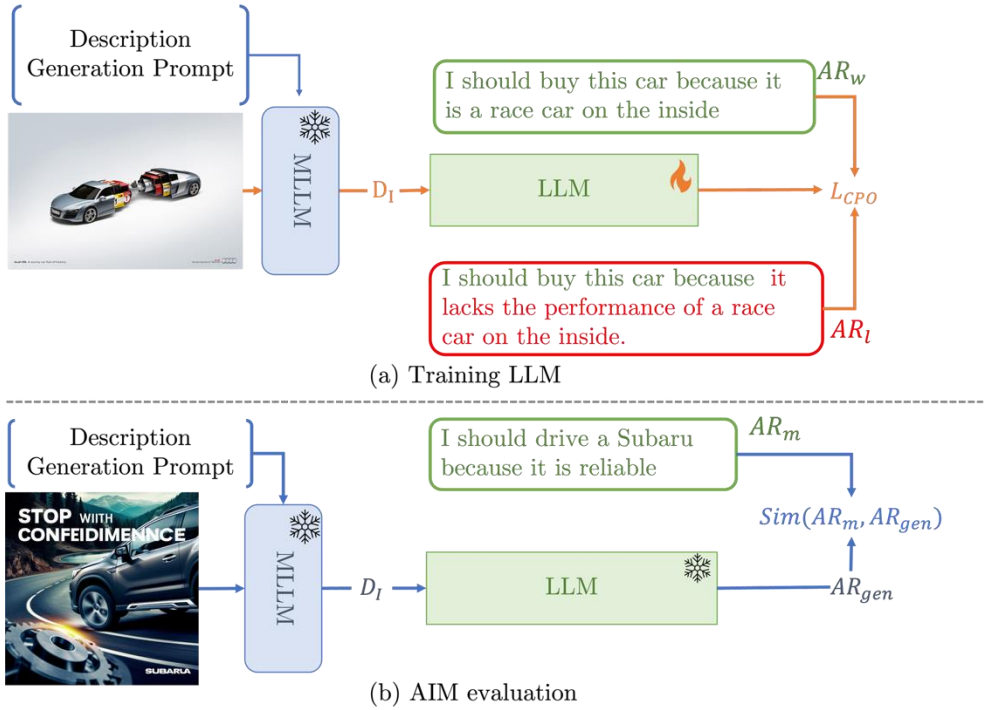
CAP Framework
Creativity, Alignment, and Persuasion Metrics

University of Pittsburgh

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Metrics - Alignment

- Existing alignment metrics:
  - High performance in capturing visual mismatch
  - Fail in capturing the semantic mismatch
- Previous results:
  - LLMs perform better in Ad understanding
- Forcing LLM to generate AR statements semantically correct
  - Using Contrastive Preference Optimization
    - Accepted: Correct Action-reason
    - Rejected: Semantically challenging negatives
- Inference:
  - Describe image
  - Generate action-reason
  - Return weighted text-text similarity score



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

Metrics – Creativity

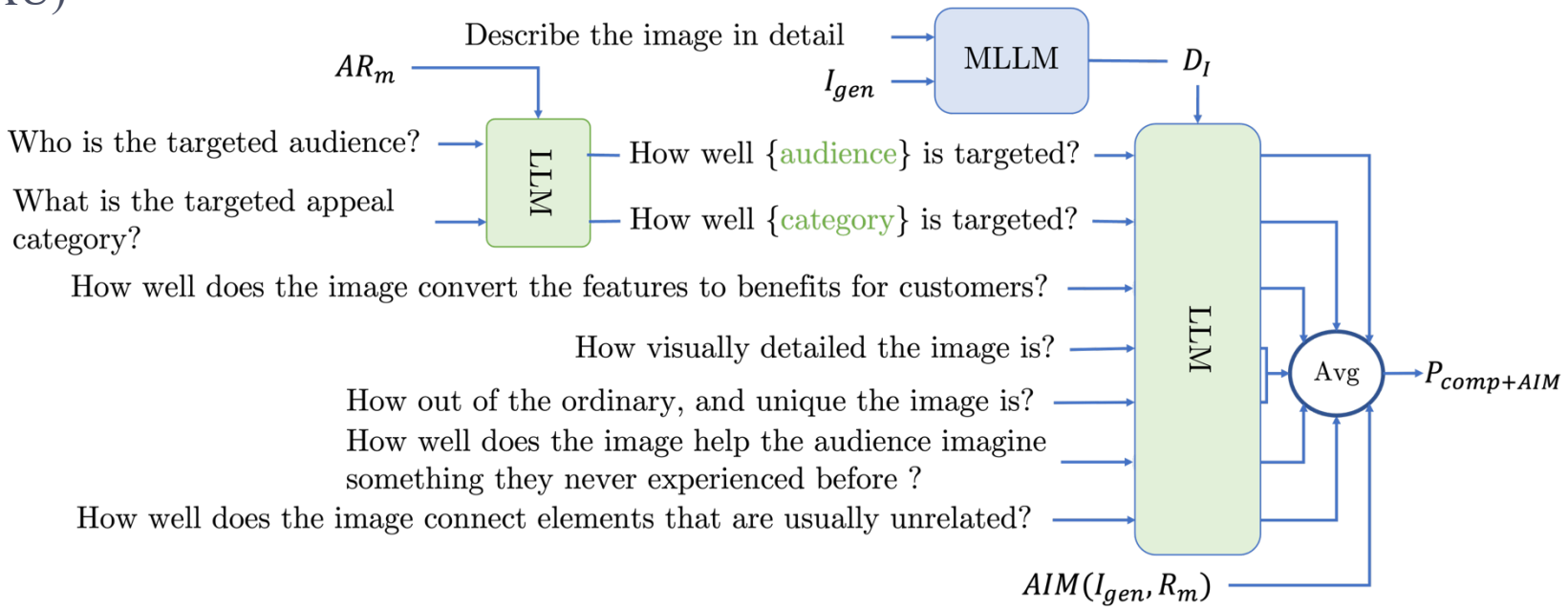
- Creativity in advertisements:
  - Uniqueness:
    - Be distant from the basic representation of the visual element in the prompt
  - Alignment:
    - While being unique, it must be relevant to the advertisement message
- Evaluation:
  - Extract the list of visual elements from the prompt
    - Example: I should drink this beer because it is light
      - [beer]
  - Being distant from basic representation:
    - Text-image similarity between a visual element and an image↓
    - Alignment ↑

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

Metrics - Persuasion

- Persuasion and its factors are subjective.
  - Combining the factors can make it less subjective.
- Persuasion factors in marketing:
  - Targeting correct audience(AU)
  - Appeal Category (AP)
    - Ethos
    - Pathos
    - Logos
  - Features to benefit (B)
  - Elaboration (E)
  - Originality (O)
  - Imagination (I)
  - Synthesis (S)

- The reason in the image should be the same as the reason in the action-reason statement.



Results

Results – Alignment Agreement

- Existing image-text alignment metrics struggle when the text is implicit.
- LLMs in 0-shot experiments struggle in accurately representing the message in the image.
- AIM using InternVL and LLAMA3-instruct is the most accurate in public service advertisements.
- AIM with QWenVL and QWen(LM) is the most accurate in commercial advertisements.

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

I should be against bullying Because it's wrong to bully
I should go to the anti-bullying campaign Because I can help stop bullying

✖	Human Annotators	✓
0.42 ✖	AIM	✓ 0.47
0.84 ✓	VQA	✖ 0.75
0.17 ✓	ImageReward	✖ -0.99

I should treat disabled people better Because they are people too
I should see disabled people as normal Because it is not good to discriminate

✖	Human Annotators	✓
0.0 ✖	AIM	✓ 0.45
0.79 ✓	VQA	✖ 0.73
-1.5 ✖	ImageReward	✓ -0.20

I should not buy Gucci products Because they kill animals to make their products
I should not buy Gucci Because they use fur

✖	Human Annotators	✓
0.44 ✖	AIM	✓ 0.47
0.78 ✓	VQA	✖ 0.49
0.49 ✓	ImageReward	✖ -0.62

I should buy a Mac Because it is superior to the PC
I should get an Apple because it's better than a pc

✖	Human Annotators	✓
0.50 ✖	AIM	✓ 0.52
0.81 ✓	VQA	✖ 0.77
0.77 ✓	ImageReward	✖ 0.68

I should buy timberland because its cool
I should go to a concert because someone famous will be there

✖	Human Annotators	✓
0.50 ✖	AIM	✓ 0.52
0.81 ✓	VQA	✖ 0.73
1.47 ✓	ImageReward	✖ 0.40

I should buy Burts Bees skin creme because it's made from natural ingredients and protects my skin
I should buy Burts Bees because they're natural

✖	Human Annotators	✓
0.54 ✖	AIM	✓ 0.56
0.90 ✓	VQA	✖ 0.88
1.05 ✓	ImageReward	✖ 0.08

Results - Creativity

- LLM fails in evaluating creativity.
- Agreement among human annotators shows that creativity is more subjective and alignment
- Our proposed metric shows good agreement with human annotators.

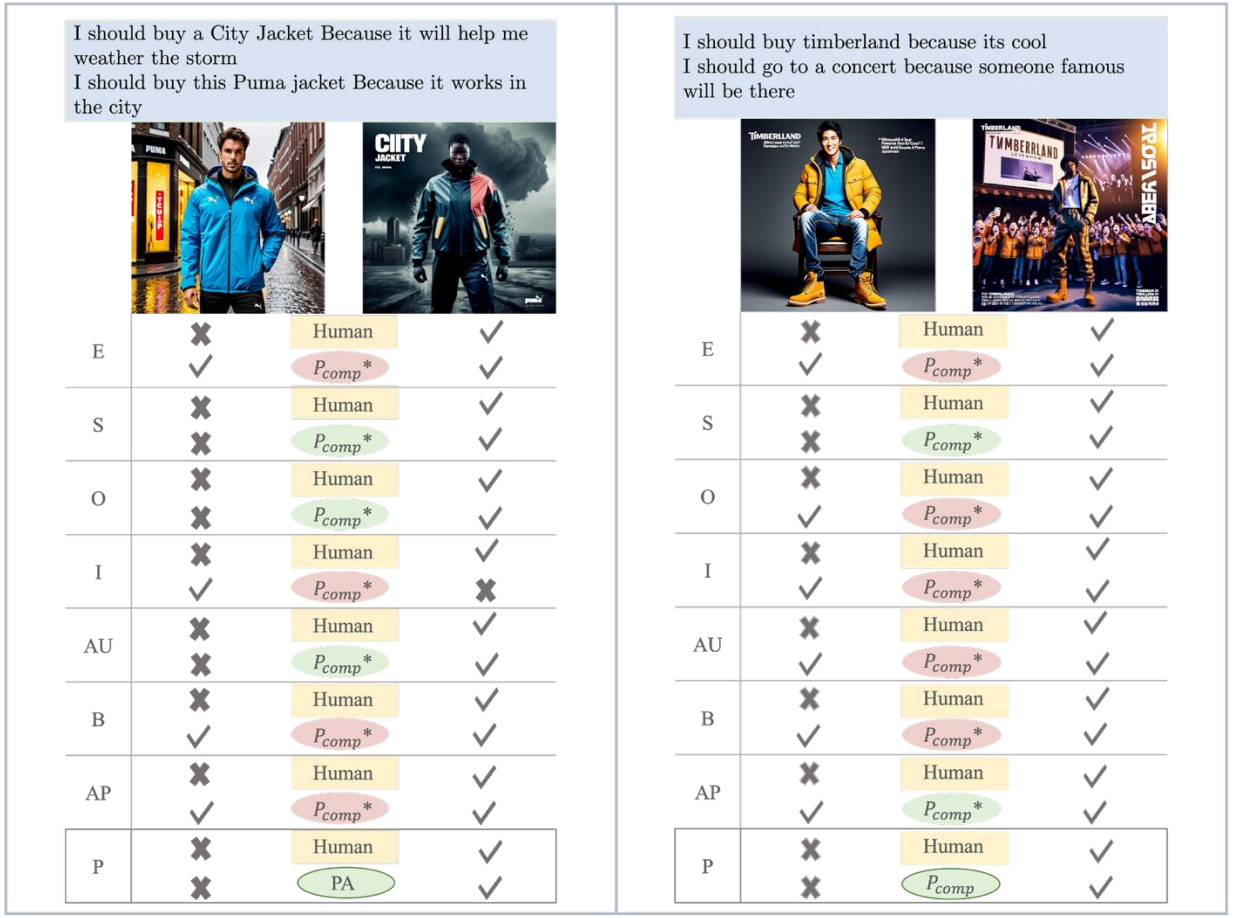
Annotators	COM	PSA	All
H, C_{LLM}	-0.03	0.15	0.04
H, C_{obj}	**0.57**	**0.53**	**0.54**
H1, H2	0.70	0.78	0.73

- Uniqueness and alignment:
  - Uniqueness alone can score the creativity of irrelevant images high.

Results - Persuasion

- Agreement on each component is low.
  - Synthesis, Imagination, and AP:
    - Agreement among annotators is low.
- Agreement on all:
  - $P_{comp+AIM^*}$: image with average score of all components is the winner.
  - H, the image that wins the most over different components is the winner.
  - High agreement on all components combined.

Annotators	E	S	O	I	AU	B	AP	All
H, $P_{comp+AIM}$*	-0.15	-0.03	0.24	0.06	0.21	0.05	0.25	0.78
H1, H2	0.74	0.40	0.74	0.40	0.53	0.54	0.34	0.89



Results - Persuasion

- Agreement on each component is low.
  - Synthesis, Imagination, and AP:
    - Agreement among annotators is low.
- Agreement on all:
  - $P_{comp+AIM^*}$: image with average score of all components is the winner.
  - H, the image that wins the most over different components is the winner.
  - High agreement on all components combined.
- LLM struggles in evaluating persuasion.
- Combination of components is helpful.
  - Adding the correct reason increases the agreement.

Annotators	COM	PSA	All
H, P_{LLM}	0.27	0.26	0.27
H, P_{comp} (InternVL, LLAMA-Instruct)	0.83	0.54	0.65
H, $P_{comp+AIM}$ (InternVL, LLAMA-Instruct)	0.85	**0.75**	**0.80**
H, $P_{comp+AIM}$ (QwenVL, LLAMA-Instruct)	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

Annotators	E	S	O	I	AU	B	AP	All
H, $P_{comp+AIM}$*	-0.15	-0.03	0.24	0.06	0.21	0.05	0.25	0.78
H1, H2	0.74	0.40	0.74	0.40	0.53	0.54	0.34	0.89

Results – T2I Models Performance

- I_{AR} are the images generated by the T2I model when prompted with action-reason statements.
  - Implicit messages
- $I_{\{LLM\}}$ are the images generated by the T2I models when prompted with descriptions generated by LLMs.
  - Explicit descriptions
- T2I models struggle in generating creative and persuasive images when the prompt is implicit.

I should buy a Mac Because it is superior to the PC
I should get an Apple because it's better than a pc

I_{AR}

$I_{\{LLM\}}$

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}	0.53	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**

Conclusion

- Introduced CAP framework, evaluation metrics for:
  - **Creativity**, balancing the alignment and uniqueness criteria
  - **Alignment**, capturing the semantic mismatch between the image and prompt
  - **Persuasion**, reducing the subjectiveness of different factors by combining them and adding AIM
- Highlighted the struggle of T2I models in generating images given implicit prompts
- Highlighted the struggle of T2I models in generating creative and persuasive images

Acknowledgements

- This work was partly supported by NSF Grant No. 2006885 and partly by the University of Pittsburgh Center for Research Computing and Data, RRID:SCR_022735, through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681.
- We gratefully acknowledge the support of our annotators.

ICCV OCT 19-23, 2025 HONOLULU HAWAII

University of Pittsburgh

Any Questions?

Contact: aya34@pitt.edu

Visit: https://aysanaghazadeh.github.io/CAP/