Benchmarking VLMs' Reasoning About Persuasive Atypical Images

Sina Malakouti[1]*, Aysan Aghazadeh[1]*, Ashmit Khandelwal[2], Adriana Kovashka[1]

[1]University of Pittsburgh [2]BITS Pilani

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

University of Pittsburgh

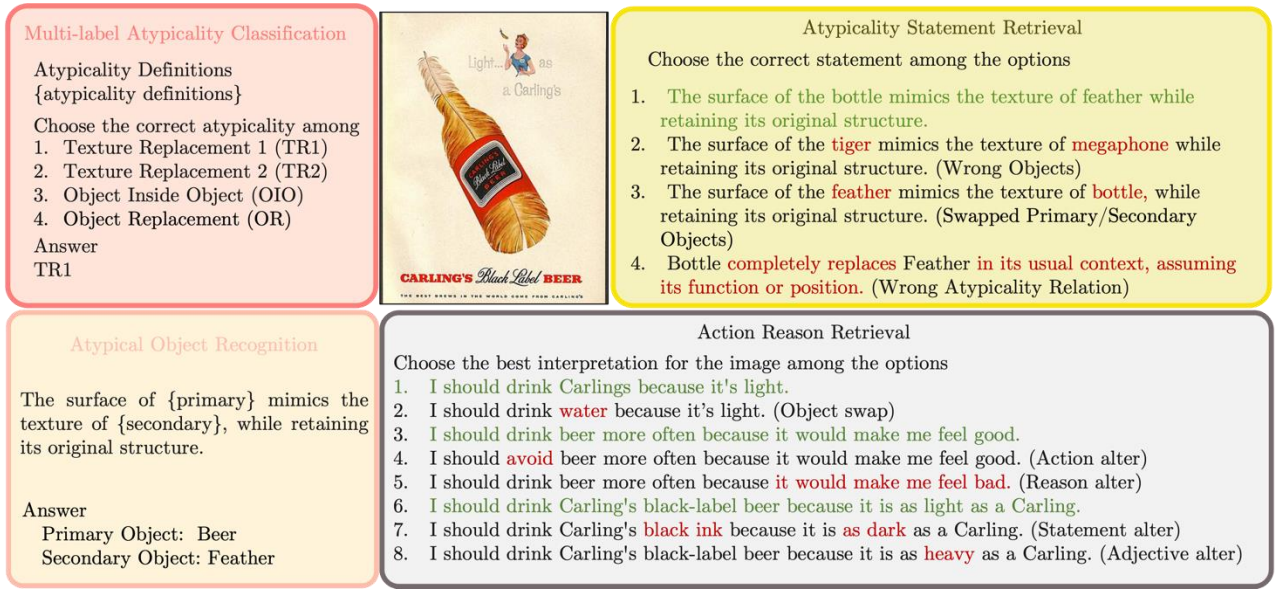
Motivation

- MLLMs and VLMs have shown promising reasoning capabilities and performance in different domain
- Existing benchmarks focus on common visual scenes and simple reasoning
  - Failing to challenge MLLM's visual reasoning capabilities
- We benchmark MLLMs/VLMs on **Atypicality Understanding** and **Advertisement Understanding**
  - Persuasive visual media (e.g., advertisement) uses creative visual rhetoric to capture attention and convey powerful messages

Overview

PersuasiveAdsVLM Benchmark

- Atypicality Understanding
  - Requires strong visual reasoning
  - Propose 3 novel tasks (classification, retrieval, and generative)
- Advertisement Understanding
  - Requires strong multi-step reasoning
  - Action-Reason Retrieval (Hussain, CVPR, 2017)
  - We generate semantically hard negatives to challenge model's reasoning capabilities



Overview

PersuasiveAdsVLM Benchmark

- Atypicality Understanding
  - Propose 3 novel tasks (Classification, retrieval, and generative)
- Advertisement Understanding
  - Action-Reason Retrieval following (Hussain, CVPR, 2017)
  - We generate semantically hard negatives to challenge model's reasoning capabilities

Hypothesize

- Understanding Atypicality helps to understand underlying message of an advertisement
- Propose a novel atypicality-aware verbalization

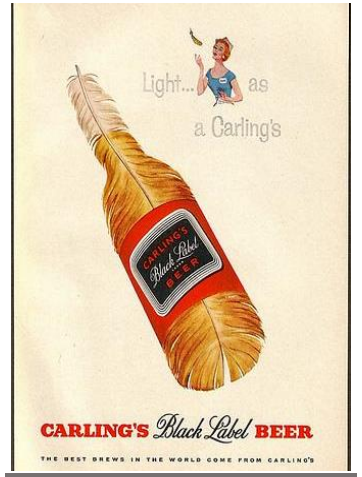
Atypicality Understanding and ARR Tasks

What is Atypicality?

- *Atypicality* is an unusual portrayal of objects
  - Often involves multiple objects engaged in an unusual relation
- We focus on 4 types of atypicality relations (Hussain, CVPR 2017)
  - Texture Replacement 1
  - Texture Replacement 2
  - Object Inside Object
  - Object Replacement

What is Atypicality?

- *Atypicality* is an unusual portrayal of objects
  - Often involves multiple objects engaged in an unusual relation
- We focus on 4 types of atypicality relations (Hussain, CVPR 2017)
  - **Texture Replacement 1**: Object's texture is borrowed from another object
  - Texture Replacement 2
  - Object Inside Object
  - Object Replacement



TR1

TR2

OIO

OR

What is Atypicality?

- *Atypicality* is an unusual portrayal of objects
  - Often involves multiple objects engaged in an unusual relation
- We focus on 4 types of atypicality relations (Hussain, CVPR 2017)
  - Texture Replacement 1
  - **Texture Replacement 2**: Texture created combining several small objects
  - Object Inside Object
  - Object Replacement

TR1

TR2

OIO

OR

What is Atypicality?

- *Atypicality* is an unusual portrayal of objects
  - Often involves multiple objects engaged in an unusual relation
- We focus on 4 types of atypicality relations (Hussain, CVPR 2017)
  - Texture Replacement 1
  - Texture Replacement 2
  - **Object Inside Object**: One object is inside another object in unexpected form
  - Object Replacement

TR1

TR2

OIO

OR

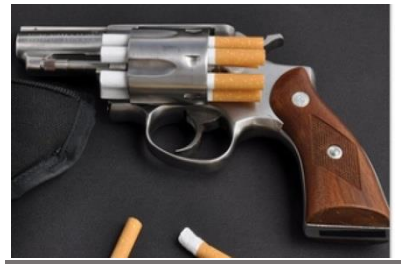
What is Atypicality?

- *Atypicality* is an unusual portrayal of objects
  - Often involves multiple objects engaged in an unusual relation
- We focus on 4 types of atypicality relations (Hussain, CVPR 2017)
  - Texture Replacement 1
  - Texture Replacement 2
  - Object Inside Object
  - **Object Replacement**: One object replaces another object in unexpected context

TR1

TR2

OIO



OR

What is Atypicality?

- *Atypicality* is an unusual portrayal of objects
  - Often involves multiple objects engaged in an unusual relation
- We focus on 4 types of atypicality relations (Hussain, CVPR 2017)
    - Texture Replacement 1
    - Texture Replacement 2
    - Object Inside Object
    - Object Replacement
- 3 Novel Atypicality Understanding tasks
  - Multi-label Atypicality Classification (MAC)
  - Atypicality Statement Retrieval (ASR)
  - Atypical Object Recognition (AOR)

Multi-label Atypicality Classification (MAC)

- Choosing the correct atypicality category for image
  - Texture Replacement 1, Texture Replacement 2, Object Inside Object, and Object Replacement
  - Not Atypicality (NA) to capture typical ads

Multi-label Atypicality Classification

Atypicality Definitions
{atypicality definitions}

Choose the correct atypicality among
1. Texture Replacement 1 (TR1)
2. Texture Replacement 2 (TR2)
3. Object Inside Object (OIO)
4. Object Replacement (OR)

Answer
TR1

Atypicality Statement Retrieval

Choose the correct statement among the options

1. The surface of the bottle mimics the texture of feather while retaining its original structure.
2. The surface of the tiger mimics the texture of megaphone while retaining its original structure. (Wrong Objects)
3. The surface of the feather mimics the texture of bottle, while retaining its original structure. (Swapped Primary/Secondary Objects)
4. Bottle completely replaces Feather in its usual context, assuming its function or position. (Wrong Atypicality Relation)

Atypical Object Recognition

The surface of {primary} mimics the texture of {secondary}, while retaining its original structure.

Answer
Primary Object: Beer
Secondary Object: Feather

Action Reason Retrieval

Choose the best interpretation for the image among the options

1. I should drink Carlings because it's light.
2. I should drink water because it's light. (Object swap)
3. I should drink beer more often because it would make me feel good.
4. I should avoid beer more often because it would make me feel good. (Action alter)
5. I should drink beer more often because it would make me feel bad. (Reason alter)
6. I should drink Carling's black-label beer because it is as light as a Carling.
7. I should drink Carling's black ink because it is as dark as a Carling. (Statement alter)
8. I should drink Carling's black-label beer because it is as heavy as a Carling. (Adjective alter)

Atypicality Statement Retrieval (ASR)

- Choosing the correct Atypicality statement
  - Each statement includes
    - Atypical relation between the objects
    - Objects involved in atypicality

Multi-label Atypicality Classification

Atypicality Definitions
{atypicality definitions}

Choose the correct atypicality among
1. Texture Replacement 1 (TR1)
2. Texture Replacement 2 (TR2)
3. Object Inside Object (OIO)
4. Object Replacement (OR)

Answer
TR1

Atypicality Statement Retrieval

Choose the correct statement among the options

1. The surface of the bottle mimics the texture of feather while retaining its original structure.
2. The surface of the tiger mimics the texture of megaphone while retaining its original structure. (Wrong Objects)
3. The surface of the feather mimics the texture of bottle, while retaining its original structure. (Swapped Primary/Secondary Objects)
4. Bottle completely replaces Feather in its usual context, assuming its function or position. (Wrong Atypicality Relation)

Atypical Object Recognition

The surface of {primary} mimics the texture of {secondary}, while retaining its original structure.

Answer
Primary Object: Beer
Secondary Object: Feather

Action Reason Retrieval

Choose the best interpretation for the image among the options
1. I should drink Carlings because it's light.
2. I should drink water because it's light. (Object swap)
3. I should drink beer more often because it would make me feel good.
4. I should avoid beer more often because it would make me feel good. (Action alter)
5. I should drink beer more often because it would make me feel bad. (Reason alter)
6. I should drink Carling's black-label beer because it is as light as a Carling.
7. I should drink Carling's black ink because it is as dark as a Carling. (Statement alter)
8. I should drink Carling's black-label beer because it is as heavy as a Carling. (Adjective alter)

Atypicality Statement Retrieval (ASR)

- Choosing the correct Atypicality statement
  - Each statement includes
    - Atypical relation between the objects
    - Objects involved in atypicality
  - We generate statements for each atypicality using pre-defined templates

$\mathcal{A}$	Definition $\mathcal{D}_{\mathcal{A}}$	Statement templates $\mathcal{S}_{\mathcal{A}}$
TR1	When the skin/texture of an object is replaced with another object to inherit an attribute of that.	The surface of *{primary object}* mimics the texture of *{secondary object}*, while retaining its original structure.
TR2	When something is made from lots of small things that are not usually part of it to inherit an attribute of the small objects.	*{primary object}* appears to be composed of numerous, smaller instances of *{secondary object}*, altering its texture.
OIO	When one thing is completely inside another thing where it is not common or natural.	*{primary object}* is visibly located within *{secondary object}*, in an unconventional manner.
OR	When one thing is used in a place or way where you usually find another thing to act as the original object.	*{primary object}* completely replaces *{secondary object}* in its usual context, assuming its function or position.

Atypicality Statement Retrieval (ASR)

- Choosing the correct Atypicality statement
  - Each statement includes
    - Atypical relation between the objects
    - Objects involved in atypicality
  - We generate statements for each atypicality using pre-defined templates
- Evaluation Setup
    - Replacing the objects with objects from other images
    - Replacing the atypicality with other categories
    - Swapping the objects in the statement

Atypicality Statement Retrieval

Choose the correct statement among the options

1. The surface of the bottle mimics the texture of feather while retaining its original structure.
2. The surface of the tiger mimics the texture of megaphone while retaining its original structure. (Wrong Objects)
3. The surface of the feather mimics the texture of bottle, while retaining its original structure. (Swapped Primary/Secondary Objects)
4. Bottle completely replaces Feather in its usual context, assuming its function or position. (Wrong Atypicality Relation)

Atypical Object Recognition (AOR)

- Complete the atypical statement by generating the missing objects
- Evaluation:
  - We use sentence similarity between the correct complete statement with the completed statement by model

Multi-label Atypicality Classification

Atypicality Definitions
{atypicality definitions}

Choose the correct atypicality among
1. Texture Replacement 1 (TR1)
2. Texture Replacement 2 (TR2)
3. Object Inside Object (OIO)
4. Object Replacement (OR)

Answer
TR1

Atypicality Statement Retrieval

Choose the correct statement among the options

1. The surface of the bottle mimics the texture of feather while retaining its original structure.
2. The surface of the tiger mimics the texture of megaphone while retaining its original structure. (Wrong Objects)
3. The surface of the feather mimics the texture of bottle, while retaining its original structure. (Swapped Primary/Secondary Objects)
4. Bottle completely replaces Feather in its usual context, assuming its function or position. (Wrong Atypicality Relation)

Atypical Object Recognition

The surface of {primary} mimics the texture of {secondary}, while retaining its original structure.

Answer
Primary Object: Beer
Secondary Object: Feather

Action Reason Retrieval

Choose the best interpretation for the image among the options
1. I should drink Carlings because it's light.
2. I should drink water because it's light. (Object swap)
3. I should drink beer more often because it would make me feel good.
4. I should avoid beer more often because it would make me feel good. (Action alter)
5. I should drink beer more often because it would make me feel bad. (Reason alter)
6. I should drink Carling's black-label beer because it is as light as a Carling.
7. I should drink Carling's black ink because it is as dark as a Carling. (Statement alter)
8. I should drink Carling's black-label beer because it is as heavy as a Carling. (Adjective alter)

Action-Reason Retrieval (ARR)

- Choosing the correct statement to interpret an advertisement message
  - Action-Reason statement : *I should {**action**} because {**reason**}*
    - Action: What should I do?
    - Reason: Why should I do it?

Multi-label Atypicality Classification

Atypicality Definitions
{atypicality definitions}

Choose the correct atypicality among
1. Texture Replacement 1 (TR1)
2. Texture Replacement 2 (TR2)
3. Object Inside Object (OIO)
4. Object Replacement (OR)

Answer
TR1

Atypicality Statement Retrieval

Choose the correct statement among the options

1. The surface of the bottle mimics the texture of feather while retaining its original structure.
2. The surface of the tiger mimics the texture of megaphone while retaining its original structure. (Wrong Objects)
3. The surface of the feather mimics the texture of bottle, while retaining its original structure. (Swapped Primary/Secondary Objects)
4. Bottle completely replaces Feather in its usual context, assuming its function or position. (Wrong Atypicality Relation)

Atypical Object Recognition

The surface of {primary} mimics the texture of {secondary}, while retaining its original structure.

Answer
Primary Object: Beer
Secondary Object: Feather

Action Reason Retrieval

Choose the best interpretation for the image among the options
1. I should drink Carlings because it's light.
2. I should drink water because it's light. (Object swap)
3. I should drink beer more often because it would make me feel good.
4. I should avoid beer more often because it would make me feel good. (Action alter)
5. I should drink beer more often because it would make me feel bad. (Reason alter)
6. I should drink Carling's black-label beer because it is as light as a Carling.
7. I should drink Carling's black ink because it is as dark as a Carling. (Statement alter)
8. I should drink Carling's black-label beer because it is as heavy as a Carling. (Adjective alter)

Action-Reason Retrieval (ARR)

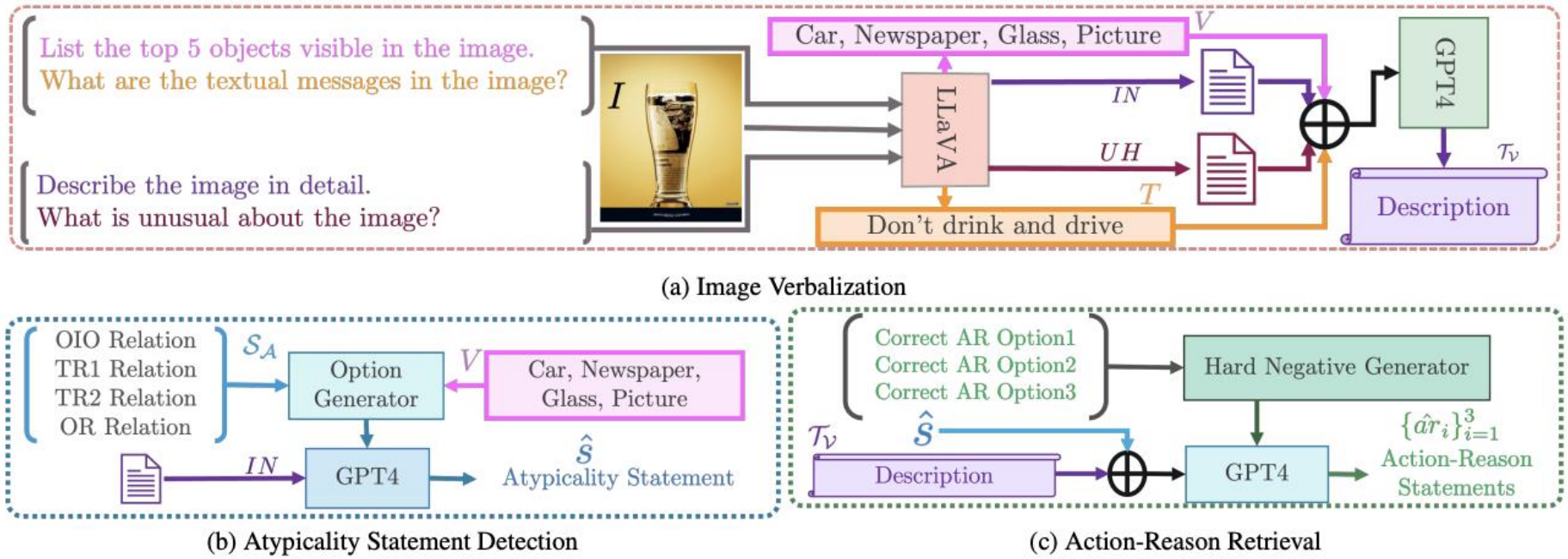
- Choosing the correct statement to interpret an advertisement message
  - Action-Reason statement : *I should {**action**} because {**reason**}*
    - Action: What should I do?
    - Reason: Why should I do it?
  - Evaluation Setup
    - Prior works mine negatives from other ads randomly or from similar topics, object detection is enough
    - we use LLM to generate statements that are **semantically** different from correct statement

Correct Option	I should drink beer more often Because it would make me feel good
Action Alter	I should abstain from beer because it would make me feel good.
Reason Alter	I should drink beer more often because it would make me feel bad.
Object Swap	I shouldn't drink water more often Because it would make me feel good
Statement Alter	I should drink beer more often because it enhances my physical fitness.
Adjective Alter	I should avoid beer more often because it would make me feel terrible.

Methodology (Action-Reason Retreival)

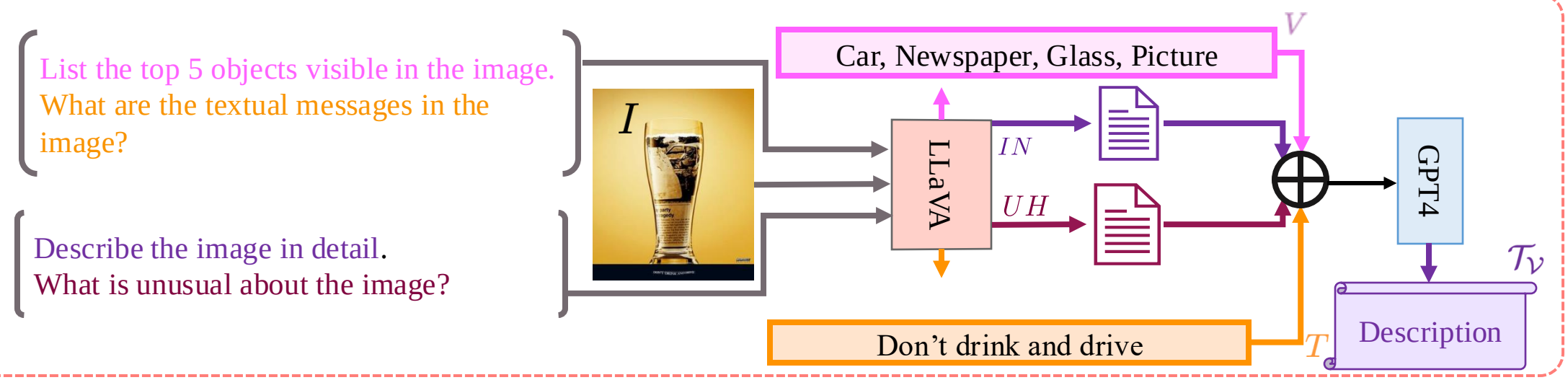
Proposed Approach

- Goal: retrieve all correct action-reason statements
- 3 Steps:
  - Image Verbalization (Atypicality-aware Verbalization)
  - Atypicality Statement Detection
  - Action-Reason Retrieval



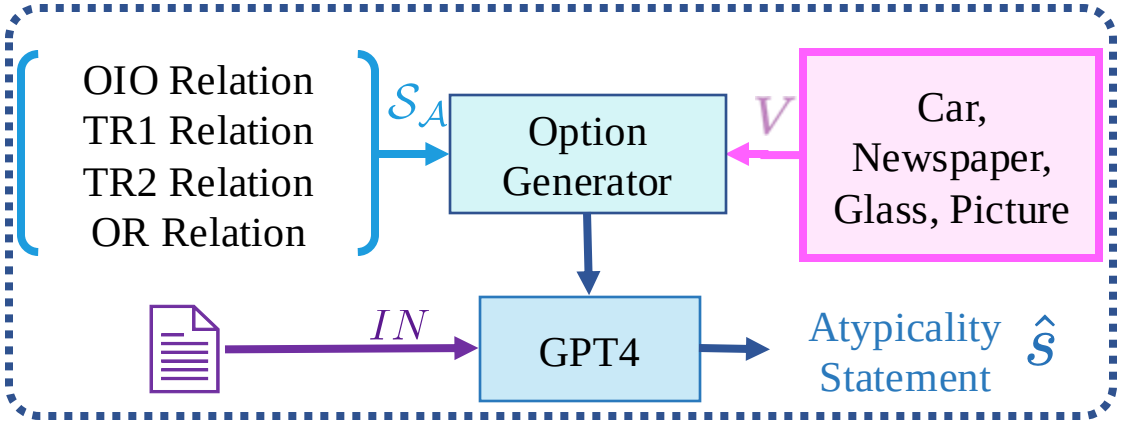
Step 1: Image Verbalization

- Verbalize image in atypicality-aware manner
  1. Basic information
     - Objects (V) and Textual elements (T)
  2. Atypicality-aware verbalization ($\mathcal{T}_{\mathcal{V}}$)
     - ImageNarrator (IN): Detail description of the image
     - UnusualHighlighter (UH): Extract unusualness
  3. Generating a coherent verbalization by combining all 4 components using an LLM



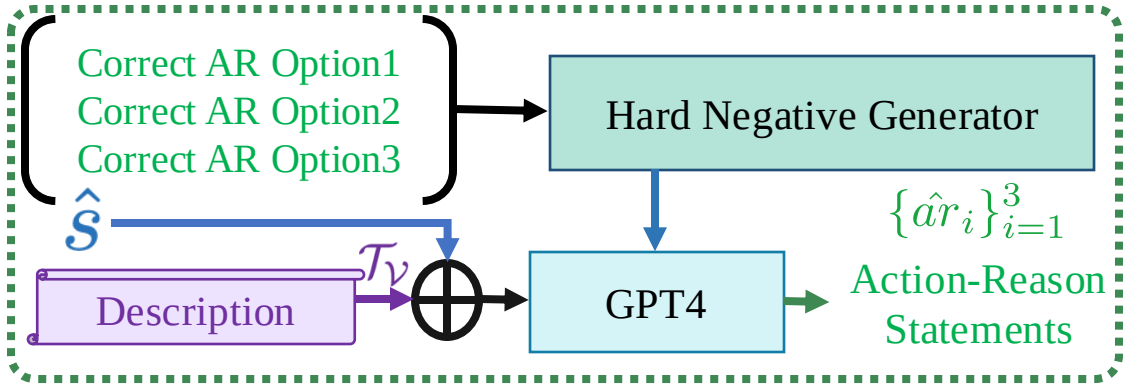
Step 2: Atypicality Detection

- Goal: Detect Atypicality Statement
- Steps
  1. Generate all possible *atypicality statements* given extracted objects and pre-defined templates ($\mathcal{S}_{\mathcal{A}}$)
  2. Detect the correct statement based on Image description (IN)



Step 3: Action-Reason Retrieval

- Goal: Given a set of action-reason statements, identify the correct options
  - Atypicality-aware Verbalization ($\mathcal{T}_{\mathcal{V}}$) and Predicted atypicality statement ($\hat{s}$)



Results

Atypicality Understanding tasks

- MAC and ASR
  - VLMs lack reasoning capability on atypicality understanding
  - V+T is not informative enough:
    - V+T only lists the visual and textual elements
  - VLMs are effective for verbalization of the image.
- MAC
  - UH verbalization is better for classifying the atypicality
    - Directly describes the unusualness in the image
- ASR
  - IN verbalization is better for ASR task
    - It includes both information about atypicality and objects

Classifier	Method	*MAC* Precision ✓	*MAC* Precision ×	*MAC* Recall ✓	*MAC* Recall ×	*MAC* F1-score ✓	*MAC* F1-score ×	*ASR* Acc
LLaVA	I	27.75	27.75	42.38	52.71	21.24	26.03	18.83
	IN	25.12	31.40	42.44	53.04	25.06	31.32	20.90
	UH	44.35	30.44	42.04	52.44	24.16	29.98	17.90
InstructBLIP	I	34.81	27.60	41.43	50.73	17.72	20.18	19.76
Vicuna	$V+T$	36.70	30.64	41.73	45.78	32.52	31.66	14.30
	IN	37.71	32.04	**43.70**	**45.91**	**34.51**	**32.09**	**23.29**
	UH	**39.41**	**33.33**	36.05	42.88	27.35	30.36	14.74
GPT 3.5	$V+T$	41.46	35.36	23.21	21.54	28.18	24.95	50.00
	IN	46.28	42.50	25.13	14.75	**28.49**	19.64	**50.55**
	UH	**49.10**	**43.34**	**27.38**	**30.92**	27.06	**28.24**	50.05
GPT 4	$V+T$	40.38	35.95	22.56	6.69	22.66	10.99	52.44
	IN	**54.78**	**53.40**	27.19	13.64	30.58	20.91	**57.70**
	UH	53.49	51.01	**29.15**	**28.89**	**34.62**	**33.05**	56.89

Atypicality Understanding Tasks

- AOR
  - Scores
    - >0.7 strong semantic overlap
    - (0.5, 0.7) moderate semantic overlap
    - <0.5 weak semantic overlap
  - MLLMs and VLMs struggle finding the objects
    - 65% of responses have weak semantic overlap
    - Maximum average similarity: 0.59

Model	***Avg.* similarity** ($\hat{s}$ to s^+) **score**	***% of scores*** $> \mathbf{0.7}$	$> \mathbf{0.6}$	$> \mathbf{0.5}$
BLIP2 |	0.45	8.77	19.78	35.43
InstructBLIP |	0.46	9.54	21.24	40.76
MiniGPT4 |	0.51	15.24	32.28	51.71
LLaVA |	**0.59**	**31.41**	**51.35**	**65.16**
GPT-4V	0.67	46.94	61.63	77.14

Action-Reason Retrieval

- MLLMs/VLMs underperform LLMs
  - LLMs outperforms MLLMs/VLMs
- Atypicality-aware verbalization improves the performance
- Atypicality-aware verbalization outperforms basic verbalization (V+T)
  - Atypicality statement in prompt improves the VLMs' performance

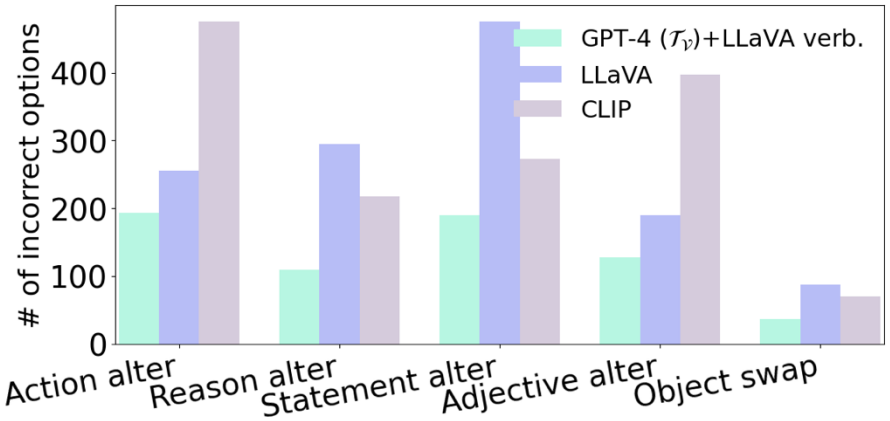
Model	I	$I + s^+$	$I + \hat{s}$	$I + \bar{s}$
LLaVA	26.00	35.18	**54.28**	28.16
InstructBLIP	20.44	23.25	**23.40**	19.69
GPT-4V	86.87	**89.35**	87.24	86.96
GPT-4 ($\mathcal{T}_\mathcal{V}$)	**96.77**	91.42	96.76	90.20

Classifier	Verb.	*Precision@k* k=1	k=2	k=3	*Top-k Acc* k=1	k=2	Avg
CLIP |	I	**61.04**	33.86	22.66	23.72	44.61	37.18
	$I + T$	46.15	24.36	16.24	15.15	31.25	26.63
LLaVA	I	59.67	**38.27**	**26.06**	**32.92**	**48.14**	**41.01**
	$I + \mathcal{T}_\mathcal{V}$	59.45	37.37	25.14	27.49	47.07	39.30
InstructBLIP	I	15.05	10.03	7.80	13.04	13.04	11.79
InternVL-V1	I	52.22	32.79	22.17	22.51	40.66	30.07
Vicuna	$V + T$	64.13	40.71	27.57	21.49	**43.41**	39.46
	$\mathcal{T}_\mathcal{V}$ (Ours)	67.38	44.01	29.94	23.20	41.95	41.30
	$\mathcal{T}_\mathcal{V} + \hat{s}_{IN}$ (Ours)	68.32	**44.52**	30.25	22.95	43.24	41.86
	$\mathcal{T}_\mathcal{V}$ (GPT-4 Verb.) (Ours)	**68.49**	**44.52**	**30.37**	**24.06**	43.24	**42.14**
GPT-4	$V + T$	93.73	84.42	70.50	71.50	89.87	82.00
	$\mathcal{T}_\mathcal{V}$ (Ours)	93.99	86.35	72.96	74.94	91.16	83.88
	$\mathcal{T}_\mathcal{V} + \hat{s}_{IN}$ (Ours)	**95.54**	**87.55**	**74.62**	**88.42**	**93.40**	**87.91**

Semantically Hard Negatives

- Comparison of easy and hard negatives
  - Performance drops by 75.8 for CLIP from easy negatives to semantically hard negative
  - Robustness of LLMs on hard negatives
  - **MLLMs/VLMs reliance on visual differences**
- Least challenging type of negatives for all the models is **Object swap**
- Most challenging types of negative for CLIP is **Action Alter**
- Most challenging types of negative for LLAVA and GPT4 is **Statement Alter**

Neg. Strategy	Model	*Multi* *Precision@k* k=1	k=2	k=3	*Single* Acc
12 Neg.	CLIP (I)	98.79	**97.58**	**92.20**	**96.77**
	CLIP $(I+T)$	97.58	**97.58**	87.10	90.32
	LLaVA (I)	93.47	74.08	56.33	94.31
	GPT4 $(\mathcal{T}_\mathcal{V})$	**99.60**	96.98	91.13	93.52
18 Hard Neg.	CLIP (I)	64.52	34.48	22.98	20.97
	CLIP (I+T)	47.18	25.40	16.94	15.73
	LLaVA (I)	59.67	38.27	26.06	26.80
	GPT4 $(\mathcal{T}_\mathcal{V})$	**96.77**	**87.30**	**74.60**	**96.77**



Findings

- Current MLLMs can’t detect atypical objects directly
  - Due to unconventional structure
  - Unseen during training
- MLLMs show some promise extract valuable information about atypical aspects
- MLLMs lack strong reasoning capabilities even compared to LLM counter part
- Atypicality is essential in understanding and designing effective ads

Conclusion

- We introduce 3 novel task
  - Multi-label Atypicality Classification (MAC)
  - Atypicality Statement Retrieval (ASR)
  - Atypical Object Recognition (AOR)
- We show that current MLLMs and VLMs lack reasoning capabilities on these tasks
- We propose an atypicality-aware verbalization method
  - Results show that LLMs with informative verbalization have higher performance than MLLMs/VLMs
  - Results show that atypicality improves the performance of the models
- We expand the PittAds dataset introducing semantically challenging negative options resulted in drop of the performance of VLMs by 75.8

Thank you