Supplementary Materials (LogicAD: Explainable Anomaly Detection via VLM-based Text Feature Extraction)

Anonymous submission

Appendix

A.1 Guided CoT Prompts for all categories

Our Guided CoT prompts are all generated based on the general initial prompts: "`describe the image based on the object' size, location, color, length`", with only minor logical related modification, we can generate Guided Chain-of-Thought (Guided CoT), as follows:

- *breakfast_box*: "`what is on the left side of the image? and what is on the right side of the image? mandarin is equivalent to orange or tangerine, clementine; apple is equivalent to nectarine or peach),`"
- *juice_bottle*: "`what is color of the juice? what is the fruit? color of the juice should match the fruit (red, wine color for cherry, white for banana and yellow for orange), is the juice filled to around half of the neck in the bottle (only use the following word, around half the neck, full(for more than half the neck), largely empty, empty)? are there two stickers?, is the top sticker correct (square sticker with fruit, fruit match with juice, fruit is located in the middle of the label)? is the bottom sticker correct (100% juice, located at the bottom of the bottle, horizontally centred)? is the bottle with stickers symmetrically?`"
- *screw_bag*: "`Answer this question if there is only one object: is this washer or nut? Answer these questions if there are multiple objects: how many bolts are there? Describe the length of the shorter bolts, including the head, using the longer bolt as reference (only possible with 1/5, 2/5, 3/5, 4/5, and 1 of the longer bolt). All bolts longer than 3 times the diameter of the washer?`"
- *splicing_connector*: "`Answer this question if the image contains only one block of connectors: where is the vertical position of the cable (use the top, middle or bottom of the connectors for description)? Answer these questions if the image contains separate connector blocks: How many connectors are there? How many cables are there? Is the cable broken or not? Is the connector the same size?`"
- *pushpins*: "`how many pushpins are there in each compartment?`"

A.2 Sample Images from CountBench and UniformBench

Figure 1 shows nine representative images from the CountBench [1] dataset, where most objects display heterogeneous features, which means each object contains some unique features. For instance, one image contains five cars, each with distinct characteristics. In contrast, our custom dataset, UniformBench, primarily comprises homogeneous objects, as demonstrated in Figures 2 and 3. Our experimental results indicate that the accuracy of Autoregressive, multimodal Vision Language Models (AVLMs) can vary significantly with even minor changes in the arrangement of stones on a Go board. Furthermore, we observed that the performance of the GPT-4o model is noticeably superior when applied to web-based datasets, such as CountBench, which is a subset of LAION-400M [2]. However, its performance is considerably worse when tested on our custom dataset.

A.3 Prompts for Logic Reasoner

For the logic reasoner, LLM prompts are mainly used to generate formal image descriptions (Σ_0) and semantic analysis of the relevant predicates (Σ_{fa}, Σ_{na}, etc.).

- A task-independent instruction for the output format: "`Given the image description, output a formal specification as a set of`

[1] https://github.com/teaching-clip-to-count/teaching-clip-to-count.github.io/blob/main/CountBench.json

[2] https://laion.ai/blog/laion-400-open-dataset/

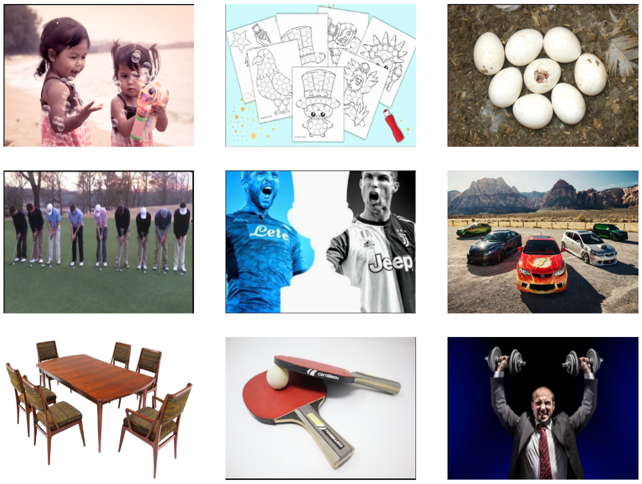


Figure 1: **Sample Images from CountBench.**

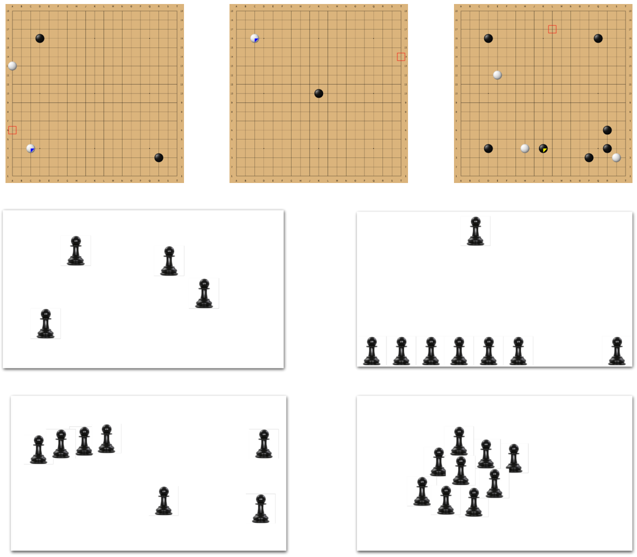


Figure 2: **Sample images from UniformBench include black and white stones on a Go board and pawns.** All the images, pawns and Go boards with stones were generated by us. These images were created using random patterns, as illustrated in the figure. The random generation of these patterns allows us to evaluate model performance on diverse yet controlled scenarios, providing a robust assessment of the models' capabilities in handling structured and uniform data.

```
(propositional) formulae following
some syntactical rules:
1. Each line consists of only one
piece of fact and an explanation of
the formula after a hashtag.
2. Connect words with underlines. Use
lowercase only.
3...."
```

- Two task-related examples, one for normal description and one for abnormal. Both can be fictitious.
  ```
  "TEXT:
  The juice has a red or wine color...
  The bottle is full...
  FORMULA:
  color(juice, red) OR color(juice,
  wine) # The juice has a red or...
  volume(full) # The bottle is full..."
  ```

- Prompts to generate (in-)equality in Σ_{na}:
  ```
  Are __ and __ synonymous or similar?
  Give a simple Yes or No answer.
  ```
  The placeholders __ are for objects to be distinguished, for example tangerine and mandarin. If LLM responds `No`, then $tangerine \neq mandarin$ is added to Σ_{na}.

- Prompts to identify functional predicates:
  ```
  When a formula $Inst$ means $Expl$,
  is the predicate $Pred$ functional
  regarding the last argument, i.e. two
  values of the last argument cannot be
  true simultaneously?
  ```
  Here `$Pred$` is a predicate that occurs in the normal specification, `$Inst$` is an instance of that predicate and `$Expl$` is the explanation of `$Inst$`. For example, we take $volume$ as `$Pred$`, $volume(full)$ as the instance and "The bottle is full" as the explanation. Based on the response it automatically decides whether

  $$\exists x. volume(x) \wedge (\forall x'. volume(x') \rightarrow x = x')$$

  will be added to Σ_{fa}.

Figure 3: **Sample Images from UniformBench, Beer Basket.**

MVTec LOCO AD Categories	Unmatch Cases	$\mathcal{M}_{FE}$ Accuracy	$\mathcal{M}_{LR}$ Accuracy
Breakfast Box	8.2%	45%	55%
Juice Bottle	2.3%	60%	40%
Pushpins	0%	-	-
Screw Bag	0%	-	-
Splicing Connector	4.8%	63%	37%
Mean	3.6%	56 %	44%

Table 1: **Ablation on the conflict between logic reasoner and format embedding.** The accuracy of using $\mathcal{M}_{FE}$ and $\mathcal{M}_{LR}$ are evaluated based **only** on the unmatched cases.

A.4 AVLMs hyperparameters and implementation details

AVLMs In our study, we utilize three advanced visual-language models (AVLMs): GPT-4o, LLaVA 1.6, and LLaVA 1.5. The GPT-4o model is accessed via the Azure platform (Azure GPT), specifically the model version *2024-05-13*. For the GPT-4o model, two key hyperparameters are adjusted: the temperature is set to 0.05, and the top_p is configured to 0.1. Regarding the LLaVA models, LLaVA 1.6 is employed with the checkpoint *llava-v1.6-vicuna-13b-hf*, while LLaVA 1.5 is utilized with the checkpoint *llava-1.5-7b-hf*. All experimental procedures involving the LLaVA 1.6 and LLaVA 1.5 models are conducted using one NVIDIA H100 GPU for inference.

GroundingDINO

The GroundingDINO model is implemented for region of interest (ROI) extraction. The prompt texts used with GroundingDINO are derived from Guided Chain of Thought (CoT) prompts, ensuring precise and contextually relevant extractions.

- *breakfast_box*: “`breakfast box`”
- *juice_bottle*:“"`juice bottle`”
- *screw_bag*: “`metal circle`”
- *splicing_connector*: “`fruit juice bottle`”
- *pushpins*: “`black square compartment`”

We select 0.3 as the `text_threshold` and 0.2 as `box_threshold` as the default threshold. As for the “pushpins” in MVTec LOCO AD, we have 0.1 for both thresholds.

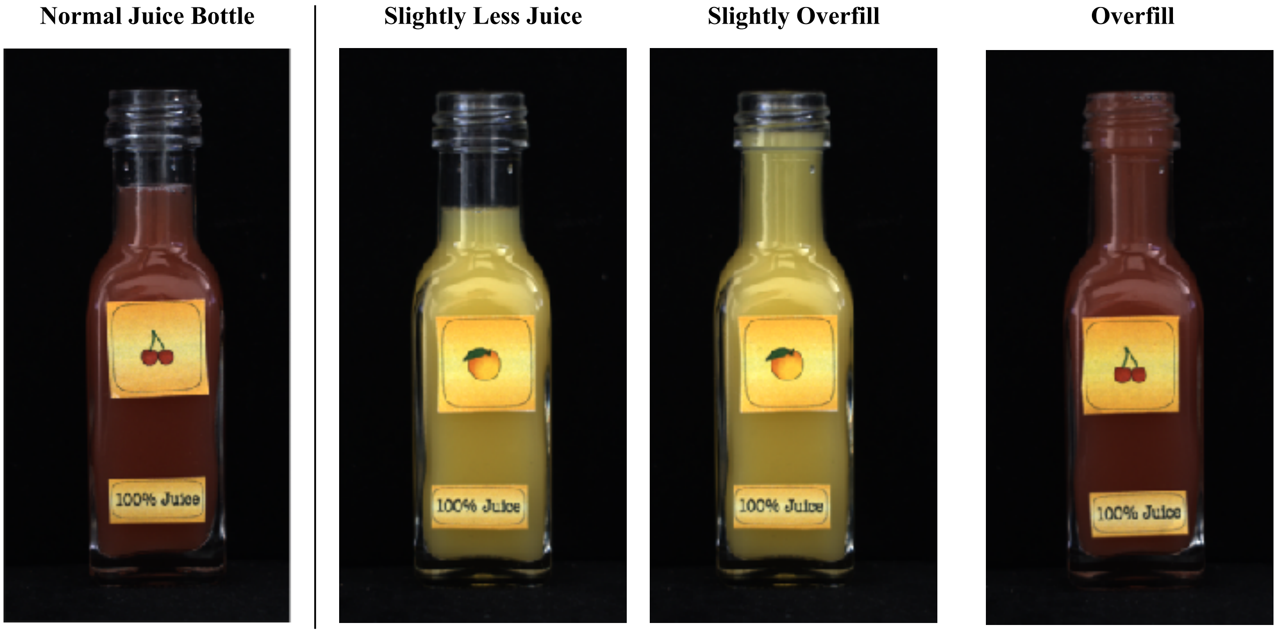
A.5 Failing and Unmatch Cases

Figure 4, Figure 5, Figure 4 and Figure 7 present some failing cases from the MVTec LOCO AD dataset. Since LogicAD has two directions, using *logic reasoner* and *format embedding*, unmatched cases can also occur. Figure 1 shows that for the MVTec LOCO AD dataset, only 3.6% of cases are unmatched, and $\mathcal{M}_{FE}$ (format embedding) performs only slightly better among the unmatched cases compared with $\mathcal{M}_{LR}$ (logic reasoner).

A.6 MVTec AD Evaluation

MVTec (AD)		LogicAD		WinCLIP	
	Category	AUROC	F_1-max	AUROC	F_1-max
Texture	Carpet	97.8	98.1	**99.8**	**98.3**
	Grid	95.3	**96.4**	**97.3**	94.7
	Leather	96.1	98.3	**98.9**	98.3
	Tile	**98.5**	97.1	97.8	96.4
	Wood	96.2	96.5	96.6	**95.8**
	Texture	96.9	**97.3**	**98.1**	96.7
Object	Capsule	**84.7**	**92.2**	77.3	91.5
	Screw	**89.1**	81.8	74.3	87.5
	Pill	**78.4**	**91.5**	78.1	91.2
	Hazelnut	**93.5**	**95.1**	92.2	89.7
	Transistor	**84.4**	**81.3**	81.1	62.6
	Toothbrush	**90.0**	**89.9**	87.1	88.1
	Zipper	**93.1**	**92.5**	84.3	89.8
	Bottle	79.5	81.5	**98.7**	**96.8**
	Cable	79.4	81.2	**85.9**	**85.1**
	Metal Nut	89.6	90.1	**92.2**	**93.2**
	Objects	86.2	87.7	85.1	88.7
	Average	**89.7**	90.9	88.9	**91.4**

Table 2: Anomaly detection (classification) performance comparison between LogicAD and WinCLIP.



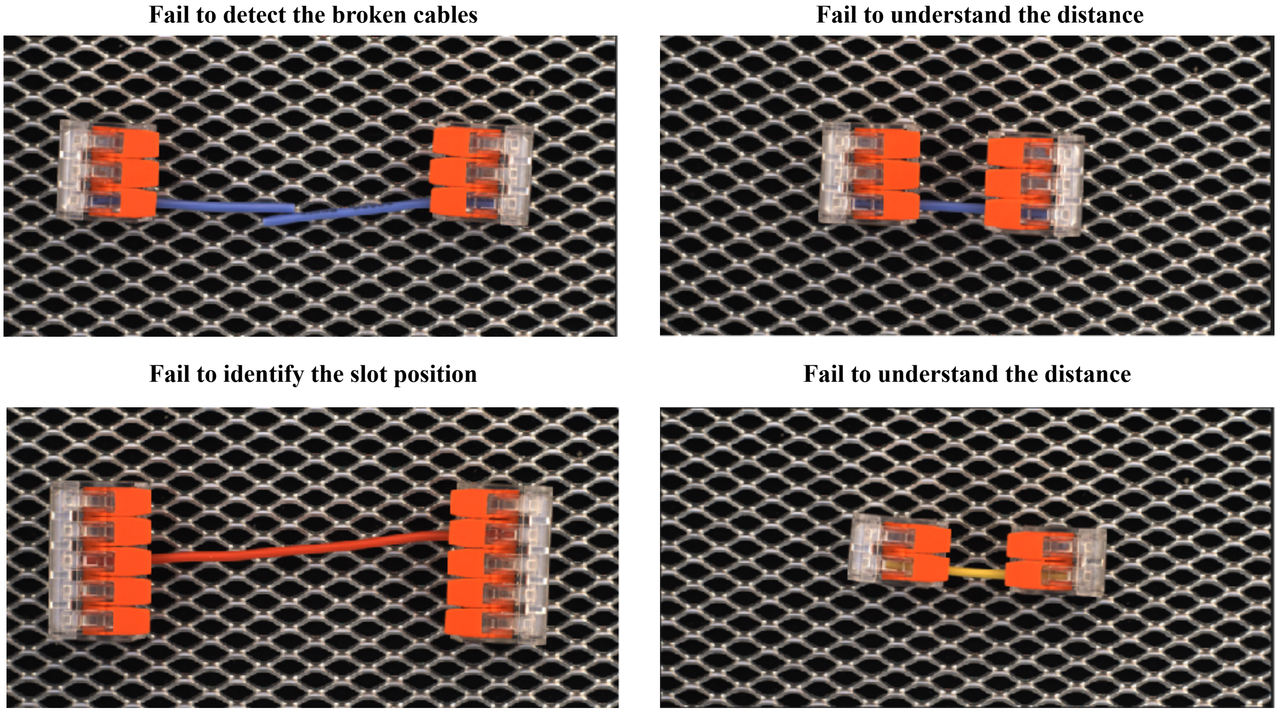
Normal Juice Bottle

Slightly Less Juice

Slightly Overfill

Overfill

Figure 4: **Failing cases: Juice Bottle in MVTec LOCO AD.** AVLMs still encounter challenges in accurately determining the fill level of juice bottles. For example, bottles with slightly less juice and those with normal fill levels are nearly indistinguishable for AVLMs, even when utilizing the Guided CoT, such as, "`is the juice filled to around half of the neck in the bottle?`". Note that these failing cases present significant challenges even for human observers.



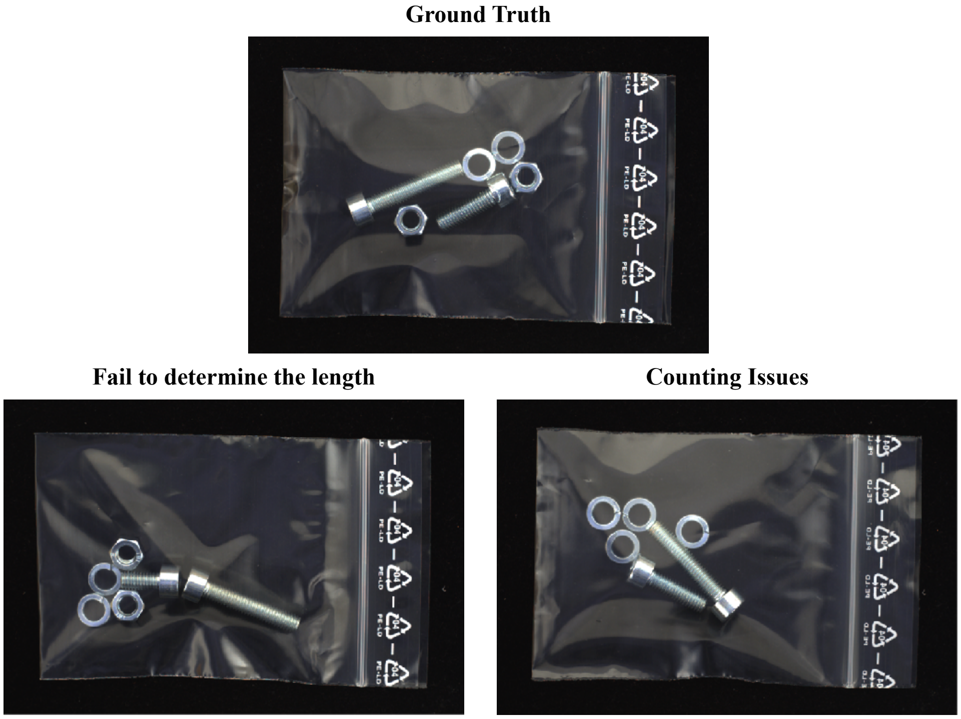
Fail to detect the broken cables

Fail to understand the distance

Fail to identify the slot position

Fail to understand the distance

Figure 5: **Failing cases, splicing connector in MVTec LOCO AD.** We observed that most of the "broken" cases can be detected by our model, but there are still a few failing cases. Additionally, our model sometimes fails to fully understand the concepts of "long" and "short" distances and "slot position."

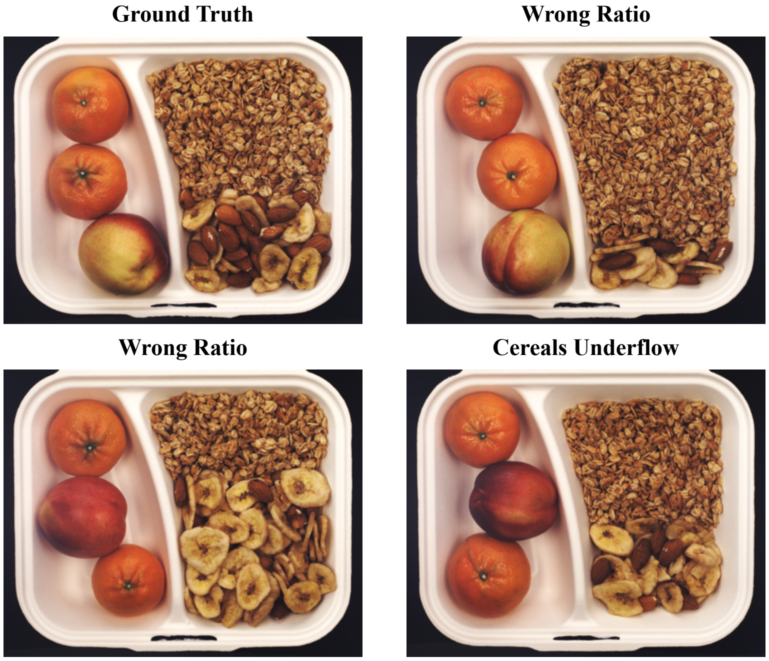


Ground Truth

Fail to determine the length

Counting Issues

Figure 6: **Failing cases, screw bag in MVTec LOCO AD.** We observed that even with Guided CoT, "`Describe the length of the shorter bolts, including the head, using the longer bolt as reference (only possible with 1/5, 2/5, 3/5, 4/5, and 1 of the longer bolt)`.", the failing cases can still occur, same as "juice bottle", these failing cases are also challenging for human observers.



Ground Truth

Wrong Ratio

Wrong Ratio

Cereals Underflow

Figure 7: **Failing cases, breakfast box in MVTec LOCO AD.** We observed that with simple prompts, such as, "`What is on the left side of the image?`" and "`What is on the right side of the image?`", our model occasionally fails to comprehend the concepts of "ratio" and "underflow."