

Supplementary Materials of Paper Submission “Detecting Multimodal Situations with Insufficient Context and Abstaining from Baseless Predictions”

CCS CONCEPTS

• Computing methodologies → Natural language generation; Image representations; Natural language generation; Scene understanding.

ACM Reference Format:

. 2024. Supplementary Materials of Paper Submission “Detecting Multimodal Situations with Insufficient Context and Abstaining from Baseless Predictions”. In *Proceedings of (MM ’24)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 IMPORTANCE OF CONTEXT

Context provides critical information to explain situations, avoid misinterpretations, and leverage fine-grained knowledge for prediction. It is particularly important in visual language understanding. For example, the ambiguities in Figure 1 cannot be clarified without context. Lack of sufficient context can harm model learning and performance evaluation. However, ensuring adequate context exists in multimodal inputs with images and text is challenging and impractical for real-world scenarios, where additional context might not be available. Thus, the ability to abstain when needed context is missing is equally crucial.

2 ADDITIONAL IMPLEMENTATION DETAILS

2.1 Heuristics with Context-Aware Abstention

Since our method is data-centric and does not base its predictions on the output of the Vision Language Model (VLM), when deciding whether to abstain from an answer generated by a VLM, to account for the VLMs’ variance, we combine the VLM’s confidence with the prediction of the Context-Aware Abstention (CARA) detector according to a heuristic rule:

$$H = w(1 - C) + (1 - w)V \quad (1)$$

where V is the VLM’s confidence, C is CARA’s confidence, and $0 < w \leq 1$ is the weighting of CARA’s score. A high C indicates CARA predicts a need for context, so $1 - C$ represents CARA’s confidence in that the data point’s has sufficient context. We use the heuristic score H and a risk tolerance threshold to decide on abstaining or answering. This heuristic incorporates both CARA’s and VLM’s confidence scores via a weighted sum.

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted by ACM, provided that the copies are not made for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM ’24, October 28–November 01, 2024, Melbourne, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00
<https://doi.org/XXXXXXX.XXXXXXX>

Submission ID: 5104. 2024-04-20 03:52. Page 1 of 1–5.

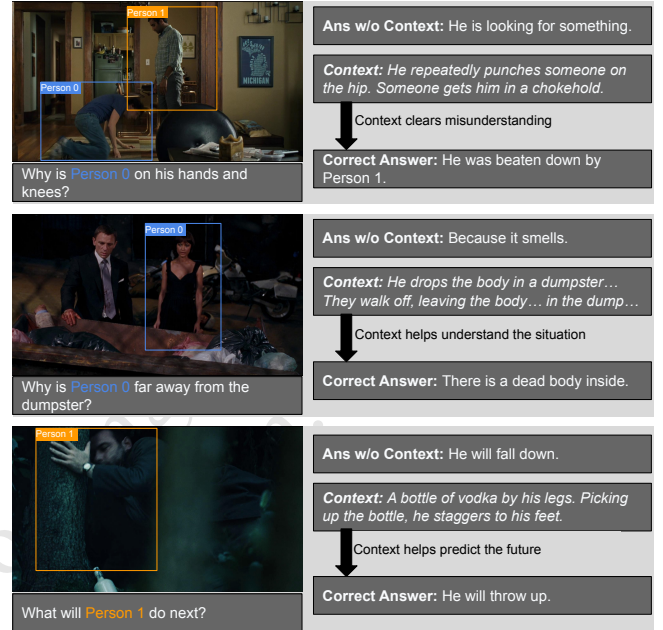



Figure 1: Three scenarios of how context can help understanding in an image-language reasoning task. In the first row, The fighting scene in the context suggests he is down because of the injury, but not what he seems to be doing in the image. In the second row, the context mentions the presence of a corpse invisible in the image, so the woman is more likely to stay away because of fear instead of distaste. In the third row, The appearance of a vodka bottle and his stumbling indicate he is drunk, which makes the correct answer more plausible.

3 ADDITIONAL ABLATION DETAILS

3.1 Context Modality

As introduced in Section 6.1.3 of the main paper, for the context selection module, we encode image context and text context using ViT [1] and Sentence-BERT[5], respectively. The two embeddings are combined and passed through a Multilayer Perceptron to obtain the final score. The context is inputted to VLM by appending the image/text context to the input sequence. Thus, we can control the modality of context the VLM can observe by appending the corresponding contexts to the inputs. Similarly, the modality the context selection module uses to select context can also vary by adding/removing the vision or language encoder. For instance, when using text to select text-only context, we append only the text context to the input sequence for the VLM, and we only use the embeddings from Sentence-BERT for the context selection module.

Given Image:



Given Question:
What happened to the bus?

Q1: Is the question **ambiguous?**
(meaning there is no obvious answer to the question, and other people may likely have different answers.)

Ambiguous ☐
(There is **not** an obvious correct answer)

Unambiguous ☐
(There is an obvious correct answer)

Not sure ☐

Q2: If the given question is marked **ambiguous above, do you think it **lacks sufficient context information**?
(Do you need more specific context to determine an answer to the question?)**

Yes, lacking context ☐

No, not lacking context ☐

Figure 2: Interface layout for annotators in verifying the correctness of CARA’s detection results. We implemented this interface over the Amazon Turker platform to facilitate turkers to effectively understand the assignment and annotate the data. In practice, we also include plenty of annotated examples beforehand as the instruction or reference.

4 DATA COLLECTION

4.1 Context retrieval

The data points in VCR, VisualCOMET, and Visual SWAG are sourced from either ActivityNet [2], LSMDC [6], or YouTube. Since only LSMDC data points have consistent and ordered context information available, we initially remove all non-LSMDC sourced data points in the Data Filtering stage, as depicted in Figure 3,

In the Context Retrieval stage, we first sort the clips temporally. Then, we locate the source LSMDC clip for each QA data point. The script of the source clip serves as the text context of c_0 . The corresponding vision context is collected by finding the most relevant frame using a pre-trained CLIP [4] model, as mentioned in the main paper.

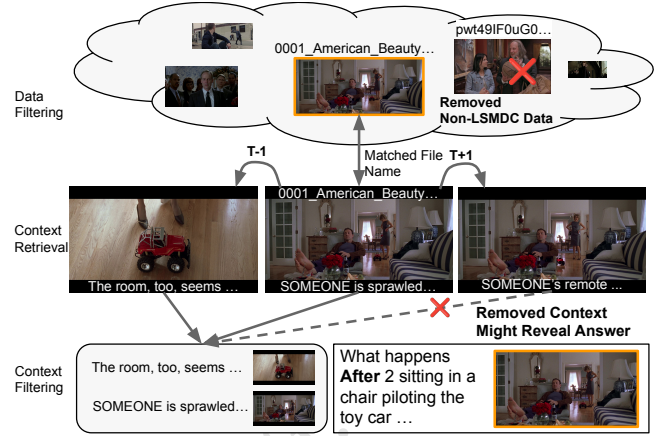


Figure 3: Dataset Construction Process: 1. Remove non-LSMDC data points. 2. Find the source clip for each image by matching the file names and save the corresponding captions as context. 3. Filter out context that can potentially reveal the correct answer.

The contexts at positive and negative indices are acquired with a similar procedure. For the context $c_{\pm n}$, we traverse n clips forward or backward and apply the procedure mentioned above. Collecting all the $c_{\pm n}$ will result in a context window size of $2n + 1$.

We set the maximum n to be 20. This means each data point will include a range from c_{-20} to c_{20} , totaling 41 context data sourced from LSMDC. We believe this adequately encompasses the necessary context for each question. Given that the average duration of LSMDC clips is 4.16 seconds, these 41 contexts collectively span approximately 2 minutes and 56 seconds of content.

Finally, in the Context Filtering stage, we remove the potentially cheating contexts for temporal questions by matching the keywords in the question. For example, Figure 3 shows contexts with positive indices are removed for questions asking about “After” to prevent the answer from leaking.

4.2 DATA QUALITY CONTROL FOR CASE

Building on the confidence-driven pseudo-labeling method (Section 5.2.1), we assembled a small data pool of 500 positive and 500 negative image-question pairs from the VCR validation set and Visual SWAG. With this curated data, we created the Context Ambiguity and Sufficiency Evaluation (CASE) Set, spanning both benchmarks to evaluate the efficacy of abstention methods in detecting samples with insufficient context. We evaluated these samples by Amazon Mechanical Turk workers to assess their ambiguity. We implemented the interface layout shown in Figure 2 and hired experienced annotators to manually verify the filtered samples. For each sample detected as positive (lacking sufficient context) by CARA, four experienced annotators re-verified it. The annotators were not informed of CARA’s prediction and answered two curated questions independently. Based on the annotation results, we calculated the voting percentage to determine if each question was considered ambiguous and lacking sufficient context. To ensure annotation consistency, we used Fleiss’ Kappa (κ) [3] to assess inter-annotator agreement. For determining if the question is ambiguous, κ is 0.81,

Table 1: Analysis of CARA abstained samples by humans, with percentages indicating "Abstained" samples where the model refrained from predicting, and "Ambiguous" and "Insufficient" denoting the proportions of abstained samples judged as such. Samples lacking context are considered ambiguous, but not vice versa. Majority of CARA-abstained samples are ambiguous, proving CARA works by removing ambiguous samples, not hard samples.

	Abstention	VCR	VisualCOMET	Visual SWAG	VQA v2	GQA	OKVQA	A-OKVQA
Abstained	CARA	13.66	18.14	18.73	10.90	5.78	28.77	5.12
Ambiguous		88.00	98.00	78.00	69.00	70.00	64.00	69.00
Insufficient Context		82.00	98.00	74.00	47.00	42.00	46.00	53.00
Abstained	Selector MLP	24.83	25.90	24.08	21.07	17.05	34.08	25.08
Ambiguous		58.00	72.00	65.00	23.00	16.00	25.00	17.00
Insufficient Context		32.00	70.00	58.00	18.00	14.00	20.00	16.00

and for determining if the question lacks sufficient context, κ is 0.84.

5 ADDITIONAL EXPERIMENTS AND RESULTS

5.1 Abstention Results Verification

In Tables 4 and 5 of the main paper, we can observe that adding CARA on top of base VLMs can generally improve the performance across benchmarks. To further verify CARA’s effectiveness and ensure that CARA focuses on removing problematic ambiguous samples (including samples with insufficient context) instead of challenging but answerable ones, we conduct manual human verification to examine the filtered-out data by CARA. Specifically, we let human annotators verify 100 randomly sampled instances for each dataset where CARA predicts positive (i.e., need context). In Table 1, we show human verification results on different datasets. We label “ambiguous” for data points that have no obvious correct answer, as shown in the examples in Figure 6 of the supplementary materials. The ambiguity of these questions may vary. For example, the first question’s reference to laptops is ambiguous since there is more than one brand in the image, and some cannot be determined due to poor image quality. Among these, a significant portion of ambiguity is caused by insufficient context, which happens when the question is ambiguous. Still, such ambiguity can be alleviated when additional information about the scene (i.e., context) is provided. Examples of this type are shown in Figures 1, 2, and 6 of the main paper, as well as highlighted in Figure 6. We are surprised to find that CARA is able to identify other types of samples with ambiguities as well, such as those with ambiguous questions or poor image quality.

5.2 Qualitative Examples

5.2.1 Context Selection. In Figure 4 and Figure 5, our contextual model demonstrates superior performance over the non-contextual model across numerous instances. Take, for instance, the third example from the Visual SWAG dataset. Without context, the correct choice, A, appears arbitrary, leading to the model incorrectly selecting choice D. However, our contextual model effectively identifies and leverages the relevant context—“someone gets up and goes over to the cool box”—to correctly associate it with the answer “returns with four cans”.

5.2.2 Abstention. Figure 6 shows the prediction of CARA, with the abstained samples labeled with “Ambiguous” or “Insufficient

Context” by humans. We also provide BLIP2’s response to these questions. Compared to the non-abstained questions (bottom two), the abstained ones have significantly diverse answer references, indicating disagreement among annotators.

6 LIMITATION

Although CARA can be adapted to different problems and VLMs without needing to be retrained, the decision threshold and parameters for the heuristic rule in Equation (1) may require additional tuning to achieve optimal performances.

The context selection method defined in Section 5 of the main paper works only for segmented contexts, which in our case consists of short sentences and videos. However, when applying it in other scenarios, for example, when context is in the form of paragraphs, context needs to be broken into pieces to adapt our method. In addition, the loss function mentioned in Section 5.1 of the main paper requires the model to recompute the input m times given the context window size of m . This raises scalability issues for large context window sizes.

REFERENCES

- [1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]
- [2] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Nieves. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.
- [3] Rosa Falotico and Piero Quatto. 2015. Fleiss’ kappa statistic without paradoxes. *Quality & Quantity* 49 (2015), 463–470. <https://api.semanticscholar.org/CorpusID:121849847>
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]
- [5] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. <https://arxiv.org/abs/1908.10084>
- [6] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2016. Movie Description. arXiv:1605.03705 [cs.CV]

	<p>The film makers</p> <p>A. pop their story in silhouette. B. look down at the destruction. C. glides out on a soft smooth surface. D. cuts up two creeping cars.</p>	<p>Context:</p> <p>-1: He sets the camera up to film the crash site in the distance. -2: SOMEONE takes the camera. -3: They are filming with the wrecked train in the distance behind them. -4: Filming a scene, SOMEONE shoots zombie SOMEONE as SOMEONE watches.</p>
	<p>The ambulance</p> <p>A. moves away from the van. B. is wheeled down the driveway. C. gets wheeled back onto a lower deck across the road. D. arrives at the bombed hotel.</p>	<p>Context:</p> <p>-1: Then ducks inside and guns down people. -2: SOMEONE sets charges which blow the door to SOMEONE's room. -3: He shoots a third agent. -4: SOMEONE uses a mirror to spot two agents outside SOMEONE's room, then opens fire.</p>
	<p>Someone</p> <p>A. returns with four cans. B. eyes the bearded general confidently then lowers his rock before releasing it. C. comes over and takes another slug. D. wears a strap over his eyes.</p>	<p>Context:</p> <p>-1: SOMEONE gets up and goes over to the cool box. -2: SOMEONE looks thoughtful. -3: SOMEONE drags on his cigarette. -4: SOMEONE stares at SOMEONE.</p>
	<p>Why is Person 0 wet?</p> <p>A. Someone large jumped into the pool and splashed water outside the pool onto Person 2. B. Person 0 doesn't have an umbrella and it's raining. C. They want to dry off. D. Person 0 has been rescued at sea.</p>	<p>Context:</p> <p>1: SOMEONE smiles. 0: SOMEONE continues to stare up at the giant statue as the crewman moves on. -1: A ship's crewman holding a clipboard steps up to SOMEONE.</p>
	<p>Why is Person 0 more defiant?</p> <p>A. Person 0 is the leader of the gang. B. Person 0 may be equipped to respond to the threat. C. Person 0 is disagreeing or not in full agreeance with a something said by Person 4 or Person 18. D. Person 0 is a bold person that doesn't follow the norm.</p>	<p>Context:</p> <p>1: Dayton residents crowd closer. 0: As he aims his gun at SOMEONE and SOMEONE. -1: SOMEONE gets out of his car</p>
	<p>What is Person 1 looking at?</p> <p>A. Person 1 is looking ahead of her in order to make a future transaction. B. She is looking at the dart board. C. She is looking at her nephew. D. She is looking at Person 1 who is looking back at her.</p>	<p>Context:</p> <p>1: SOMEONE eyes her quizzically. 0: SOMEONE gives an awkward nod and SOMEONE crosses to her drink at the bar. -1: Another of her throws bounces off the ceiling.</p>

Figure 4: Qualitative examples of Visual SWAG (example 1-3) and VCR (4-6) with/without context. Predictions made by context model are highlighted in **Green**. Predictions made by no context models are highlighted in **Red**. The selected context is highlighted in **Blue**. Correct choices are in Bold font for Visual SWAG and VCR examples.

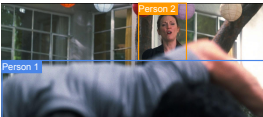


	<p>What happens after Person 2 is making a concerned face as she looks at Person 1 in the backyard?</p> <p>Prediction (with context): try to stop the fight Prediction(without context): walk away from Person 1</p>	<p>Context:</p> <p>-2: All four guys wrestle. -1: They struggle, tangled together. 0: They fall.</p>	<p>Ground Truths:</p> <p>Try to break up the fight, Call the police, Call an ambulance, Ask if they can help</p>
	<p>What is the intent of Person 1 is running as fast as he can on a road?</p> <p>Prediction (with context): get away from Person 2 Prediction(without context): get somewhere fast</p>	<p>Context:</p> <p>-1: SOMEONE leans out the passenger window and aims a gun at his back. 0: The man collapses. 1: SOMEONE gets out, approaches him, and aims again.</p>	<p>Ground Truths:</p> <p>Avoid getting shot, Get away from Person 2</p>
	<p>What happens after Person 2 stops what she is doing and watches in shock and horror as Person 1 grabs onto someone in a kitchen?</p> <p>Prediction (with context): try to help the woman Prediction(without context): finish her work</p>	<p>Context:</p> <p>-2: Then grabs SOMEONE. -1: Using her as a human shield, he backs out of the foyer and enters a kitchen. 0: Fighting his way outside.</p>	<p>Ground Truths:</p> <p>Scream for help, Try to help the woman being kidnapped</p>

Figure 5: Qualitative examples of VisualCOMET with/without context. Predictions made by context model are highlighted in **Green**. Predictions made by no context models are highlighted in **Red**. The selected context is highlighted in **Blue**.



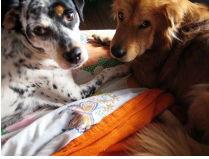











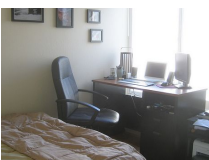

	<p>Question: What brand of laptops are these?</p> <p>VLM Prediction: Laptop</p> <p>CARA Decision: Abstain</p> <p>Answer References: Apple, Windows</p> <p>Ambiguous</p>		<p>Question: Where is the cat coming from?</p> <p>VLM Prediction: A building</p> <p>CARA Decision: Abstain</p> <p>Answer References: Outdoors, Doorway, Home</p> <p>Ambiguous Insufficient Context</p>
	<p>Question: Why are the dogs so close to each other?</p> <p>VLM Prediction: They are sleeping</p> <p>CARA Decision: Abstain</p> <p>Answer References: They're friends, Playing</p> <p>Ambiguous Insufficient Context</p>		<p>Question: Why is the hydrant so close to the wall?</p> <p>VLM Prediction: It is in the way</p> <p>CARA Decision: Abstain</p> <p>Answer References: Safety, Built after</p> <p>Ambiguous Insufficient Context</p>
	<p>Question: How is the man feeling?</p> <p>VLM Prediction: Happy</p> <p>CARA Decision: Abstain</p> <p>Answer References: Serious, Confused, Sad</p> <p>Ambiguous Insufficient Context</p>		<p>Question: What is the cat looking at?</p> <p>VLM Prediction: Laptop</p> <p>CARA Decision: Abstain</p> <p>Answer References: Couch, Person, Sofa</p> <p>Ambiguous</p>
	<p>Question: Why is this person taking a photo of a bottle?</p> <p>VLM Prediction: Because it is empty</p> <p>Answer References: Most likely instagram, Unknown, Interested, to share with someone</p> <p>Ambiguous Insufficient Context</p>		<p>Question: Why is she sitting in a refrigerator?</p> <p>VLM Prediction: She is drinking</p> <p>CARA Decision: Abstain</p> <p>Answer References: To rest, To sit, For fun, Joke</p> <p>Ambiguous Insufficient Context</p>
	<p>Question: What is in the bottle?</p> <p>VLM Prediction: Wine</p> <p>CARA Decision: Abstain</p> <p>Answer References: Honey, Coffee, Soap</p> <p>Ambiguous</p>		<p>Question: Why is a doll sitting at the table?</p> <p>VLM Prediction: It is a gift</p> <p>CARA Decision: Abstain</p> <p>Answer References: Toy, Game, Joke, Tea party</p> <p>Ambiguous Insufficient Context</p>
	<p>Question: What is in the cup?</p> <p>VLM Prediction: Blueberries</p> <p>CARA Decision: Abstain</p> <p>Answer References: Ice cream, Butter, Food</p> <p>Ambiguous</p>		<p>Question: Why are there two bags on the bed?</p> <p>VLM Prediction: They are backpacks</p> <p>CARA Decision: Abstain</p> <p>Answer References: Two travelers, For storage, Temporary storage, Roommates</p> <p>Ambiguous Insufficient Context</p>
	<p>Question: why is the cat wearing such a silly hat?</p> <p>VLM Prediction: It is funny</p> <p>CARA Decision: Abstain</p> <p>Answer References: Mean owners, Halloween, Dress up</p> <p>Ambiguous Insufficient Context</p>		<p>Question: What is likely to happen to the guy wearing the green hat?</p> <p>VLM Prediction: Skateboarding</p> <p>CARA Decision: Abstain</p> <p>Answer References: Fall, Land, Hurt</p> <p>Ambiguous</p>
	<p>Question: How many chairs are in the photo?</p> <p>VLM Prediction: One</p> <p>CARA Decision: Not Abstain</p> <p>Answer References: One</p>		<p>Question: What color is the man's coat?</p> <p>VLM Prediction: Brown</p> <p>CARA Decision: Not Abstain</p> <p>Answer References: Brown</p>

Figure 6: Additional qualitative examples answered by BLIP2. The labels "Ambiguous" and "Insufficient Context" under samples abstained by CARA are determined by human annotators.