Appendix

A. Evaluating LLaVA on NegBench MCQs

In the main paper, we proposed a novel evaluation paradigm for negation understanding, aimed at simulating real-world scenarios as closely as possible. We then proceeded to evaluate joint embedding-based VLMs, particularly CLIP models, which are the dominant models for multimodal retrieval tasks, in addition to being popular for text-to-image generation, image captioning, and medical multimodal tasks. However, we recognize that there are other VLMs that can be useful in certain settings. In particular, instruction-tuned VLMs like LLaVA open up the path for conversational VLM chatbots. In this section, we evaluate LLaVA on the three natural image MCQ tasks in NegBench (COCO, VOC2007, and HardNeg-Syn). The results are in Figure 7.

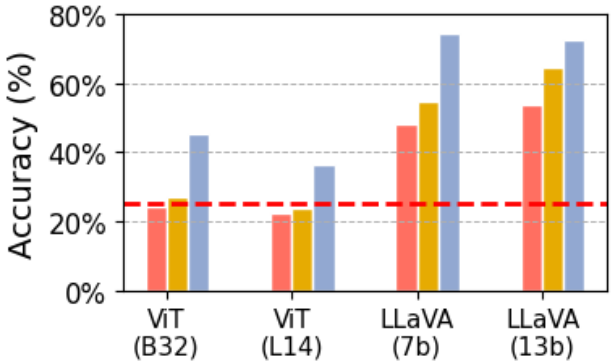

Figure 7. **Caption.**

LLaVA, an instruction-tuned VLM, demonstrates improvement. Figure 7 shows that LLaVA significantly outperforms CLIP models on the MCQ-Neg tasks. This is particularly notable because LLaVA uses a CLIP ViT-L/14 vision encoder, which we have shown in Figure 4 to struggle with negation. The key advantage of LLaVA might be in its use of the Vicuna LLM for text encoding. Unlike CLIP, which is pretrained on vision-language pairs that predominantly contain affirmative image captions, LLMs like Vicuna are trained on diverse textual corpora that include both affirmations and negations. This broader exposure allows LLaVA to better interpret negated statements. Additionally, LLaVA uses a learned projection layer to align vision and language representations, in contrast to CLIP's contrastive learning objective, which tends to ignore word order and subtle linguistic cues like negation [50]. We further explore these differences in Figure 8.

Limitations of LLaVA as a retrieval system. While LLaVA demonstrates improved negation understanding, it has significant limitations as a retrieval model compared to CLIP. CLIP learns a joint image-text embedding space, making it highly efficient for retrieval tasks by simply embedding both images and texts, and then computing cosine similarities. In contrast, LLaVA processes a single image-text pair at a time and generates text output, which makes image-to-text retrieval feasible only if all possible captions can fit into the model's context window. For MCQ-Neg, we applied this method by presenting the image alongside all possible captions and prompting LLaVA to select the correct one. However, this approach does not scale well with a large number of candidates and is not applicable for text-to-image retrieval, where fitting all dataset images into the context window is impractical. Therefore, advancing models like CLIP is crucial for real-world multimodal retrieval with negation. In the paper, we explored the data-centric reasons behind CLIP's failures in negation understanding and proposed synthetic data strategies to address them.

B. A Closer Look at VLM Negation Failures

To better understand the negation failures of VLMs, we further analyze the models' tendency to select specific template types when answering multiple-choice questions (MCQs) and provide further analysis into the embedding space of these models.

B.1. Template Selection Frequency

Figure 8 analyzes the frequency with which different models select specific template types (Affirmation, Negation, Hybrid) when answering multiple-choice questions, regardless of the correct answer. This analysis helps to reveal potential biases in model behavior and understand why models may struggle with negation. As shown in Figure 5 from the paper, most models perform poorly on Negation MCQs, reflecting a general struggle with negation understanding.

B.2. Template Selection Frequency

Figure 8 analyzes how often different models select templates of type Affirmation, Negation, or Hybrid when answering multiple-choice questions—regardless of whether the selected answer is correct. This helps reveal systematic biases in model decision-making.

We observe that most CLIP-based models strongly over-select Negation templates, even when the correct answer is an Affirmation or Hybrid statement. This aligns with the results in Figure 5, where models struggle with Negation MCQs and tend to default to negated statements. This behavior supports our earlier claim of an *affirmation bias*: models trained with CLIP-like objectives tend to ignore function words like "not" and collapse positive and negative statements in their embedding space.

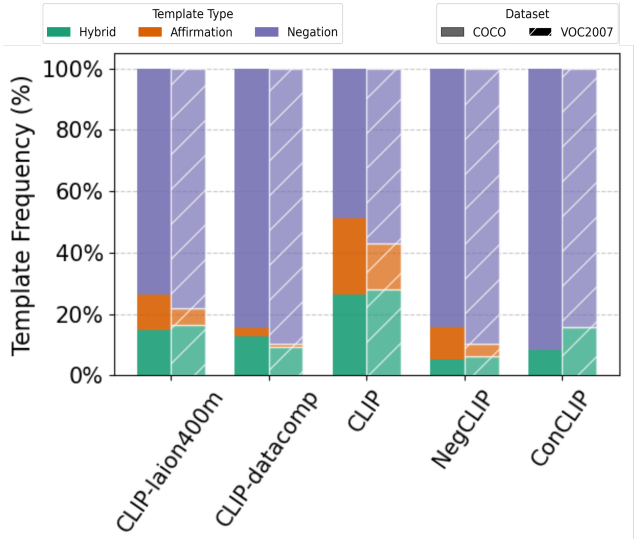

Figure 8. Template selection frequency for various models on COCO and VOC2007 datasets, broken down by template type (Affirmation, Negation, Hybrid).

Table 2. MCQ Total Accuracy (%) across different datasets for various models

Model	COCO	VOC2007	HardNeg-Syn
CLIP-OpenAI	16.27%	14.47%	18.24%
CLIP-Laion400M	24.26%	27.01%	44.60%
CLIP-datacomp	19.73%	19.72%	34.10%
NegCLIP	10.21%	8.51%	17.03%
ConCLIP	15.20%	20.43%	11.10%
CLIP-L14	22.44%	23.69%	36.51%
CLIP-H14	32.14%	38.26%	36.98%

B.3. Template Embedding Analysis

This subsection provides further details about the embedding analysis presented in Figure 6 of the main paper. We achieve this by:

1. Specifying templates used to generate the embeddings.
2. Expanding the embedding analysis to more models.

To generate the embeddings for the PCA projections, we used five categories of templates: Affirmation (single object), Negation (single object), Affirmation (two objects), Hybrid (one object affirmed, one negated), and Double Negation (two objects negated). Each category contains 24 templates, except for Affirmation (two objects) which has 23. The templates vary sentence structure and wording while maintaining the same core meaning.

- **Affirmation (single object):** 24 templates. Examples: "This image includes A", "A is present in this image", "This image shows A", "A is depicted in this image", "A appears in this image".
- **Negation (single object):** 24 templates. Examples: "This image does not include A", "A is not present in this image", "This image lacks A", "A is not depicted in this image", "A does not appear in this image".
- **Affirmation (two objects):** 23 templates. Examples: "This image includes A and B", "A and B are present in this image", "This image shows A and B", "A and B are depicted in this image", "A and B appear in this image".
- **Hybrid (one object affirmed, one negated):** 24 templates. Examples: "This image includes A but not B", "A is present in this image but not B", "This image shows A but not B", "This image features A but not B", "A appears in this image but not B".
- **Double Negation (two objects negated):** 24 templates. Examples: "This image includes neither A nor B", "Neither A nor B are present in this image", "This image shows neither A nor B", "Neither A nor B are depicted in this image", "Neither A nor B appear in this image".

While Figure 6 focused on CLIP, NegCLIP, and ConCLIP, Figure 9 presents an additional visualization with PCA projections for other CLIP models (varying in size and pretraining datasets). This broader analysis will provide a more comprehensive view of how different CLIP models handle negation in the embedding space.

C. Additional Insights and Context

D.1 How does this work fit into the broader landscape of negation and compositionality research?

Prior benchmarks such as CREPE and CC-Neg introduced limited forms of negation in vision-language tasks, focusing on compositionality or constrained template-based generation. More recently, SPEC [32] proposed fine-grained VQA tasks with a subset evaluating negation understanding. NaturalBench [19] presents a vision-centric QA protocol that reveals large performance gaps between humans and top-tier VLMs (e.g., GPT-4o, Qwen2-VL), often caused by answer biases such as a tendancy to say "Yes." over "No."

Our work complements and extends these efforts with several contributions:

- We introduce **NegBench**, a large-scale benchmark with 79K examples across retrieval and MCQ tasks, spanning images, video, and medical domains.
- We design **naturalistic negation prompts** using LLMs, covering a broad range of negation types and avoiding rigid linguistic templates.
- We generate **70M+ synthetic negation-enriched training samples**, supporting both contrastive and multiple-choice learning objectives.
- We conduct extensive experiments showing that our models **outperform prior negation-specific models (e.g., ConCLIP)** as well as SOTA VLMs (e.g., AIMv2) on negation tasks.

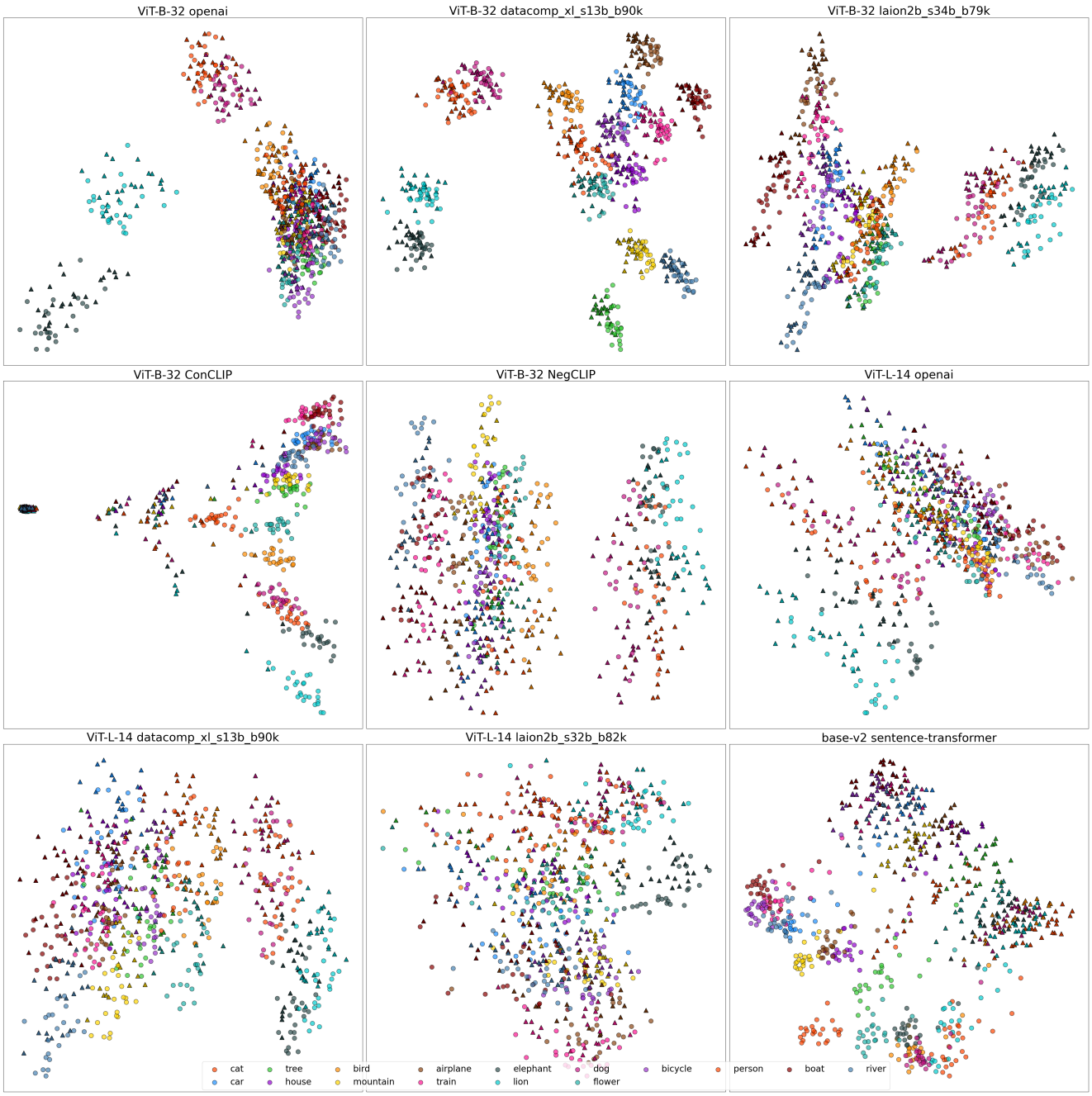

Figure 9. PCA projections of caption embeddings for various CLIP models and the Sentence Transformer. Each point represents a caption embedding. This figure complements Figure 6 by providing a broader view of embedding separation across different VLMs.

D.2 What is the significance of model scaling experiments and comparisons to recent architectures like AIMv2?

A common intuition is that larger models may better capture fine-grained distinctions such as negation. To evaluate this, we scale CLIP across ViT-B, L, and H variants, and additionally assess newer joint-embedding models such as SigLIP and AIMv2. Despite stronger performance on standard retrieval tasks, these models still struggle on MCQ-Neg and do not meaningfully close the gap—indicating that increased capacity alone does not resolve negation failures.

D.3 How are negative object queries constructed in retrieval and MCQ settings?

For datasets with dense annotations (COCO, VOC2007),

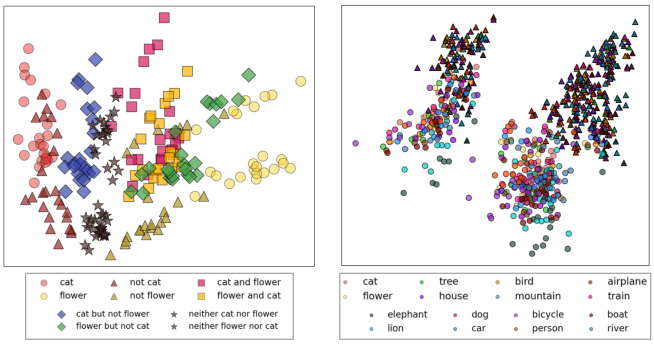

Figure 10. PCA projections of caption embeddings for finetuned CLIP model on CC12M-NegCap. Each point represents a caption embedding.

we construct a co-occurrence matrix to identify object pairs that frequently appear together. We then generate negated prompts by selecting a plausible object that is *absent* from the current image but typically co-occurs with present objects. This ensures that the negation is realistic and visually grounded, rather than relying on unlikely or artificially constructed distractors.

D.4 What is the significance of the medical experiment, despite its simplicity?

The medical retrieval experiment uses a simple binary decision setup, which offers a clean, interpretable upper bound on model capability. Models are tasked with distinguishing statements like "has pneumonia" versus "does not have pneumonia." Despite the simplicity, we observe large performance drops under negation (up to 33%) for domain-specialized VLMs such as BioMedCLIP and CONCH. This reveals a persistent failure mode with real-world clinical implications, where affirming or negating a condition must be handled with precision to avoid dire consequences.

D. Dataset and Task Summary for NegBench

We provide a summary of the datasets and tasks used in NegBench, a framework designed to evaluate Visual Language Models (VLMs) on their understanding of negation across different modalities, including images, videos, and medical imaging. The benchmark includes both retrieval and multiple-choice question (MCQ) tasks, with two variations: templated and LLM-paraphrased. For synthetic data, we generate 10,000 images using Stable Diffusion, which serve as hard negatives for one another, enabling a more focused evaluation of negation comprehension in text-to-image retrieval tasks.

Each dataset contributes to either Retrieval-Neg or MCQ-Neg tasks, except for CheXpert, which has two distinct tasks (Affirmation Control and Negation Understanding) in both MCQ and binary classification formats. Additionally, we utilize original retrieval captions for COCO (5,000) and MSR-VTT (1,000), expanding the overall dataset size. VOC2007 does not include a Retrieval-Neg task as it lacks retrieval-style captions.

The total number of task variations across all datasets in NegBench is 18, and the total number of samples across all tasks and variations is 79,239. Table 3 summarizes the datasets, tasks, task versions, and sizes.

- **COCO**: 5,000 retrieval captions and 5,914 MCQ questions, resulting in 10,000 retrieval problems and 11,828 MCQ problems with templated and LLM-paraphrased variations.
- **VOC2007**: 5,032 MCQ questions, leading to 10,064 total samples. No retrieval task is provided due to the absence of retrieval-style captions.
- **MSR-VTT**: 1,000 retrieval captions and 1,000 MCQ questions, resulting in 2,000 samples per task, including both variations.
- **CheXpert**: Two MCQ tasks (4-choice) and two binary classification tasks. The 4-choice MCQ covers 690 samples for affirmation and 1,587 for negation, while the binary tasks each include 690 samples.
- **HardNeg-Syn**: 10,000 synthetic images, used to create 20,000 retrieval and 20,000 MCQ problems across templated and LLM-paraphrased versions.

Table 3. **Summary of datasets and tasks in NegBench.** Each task includes both templated and LLM-paraphrased versions, except for CheXpert tasks, which are templated only due to their straightforwardness (they directly evaluate diagnostic capabilities in the presence of negation words). The HardNeg-Syn dataset contains 10,000 synthetic images as hard negatives, offering a more targeted evaluation of negation understanding. The total number of task variations is 18, with a total of 79,239 samples across all tasks and variations.

Dataset	Task	Templated	LLM-Paraphrased	Task Size	Notes
COCO	Retrieval-Neg	✓	✓	10,000	Image retrieval with negated captions.
	MCQ-Neg	✓	✓	11,828	MCQ task with affirmative, negated, and hybrid options.
VOC2007	MCQ-Neg	✓	✓	10,064	MCQ task. No Retrieval-Neg for VOC2007.
MSR-VTT	Retrieval-Neg	✓	✓	2,000	Video retrieval task with negated captions.
	MCQ-Neg	✓	✓	2,000	Video-based MCQ task with temporal context.
CheXpert (4-choice)	Affirmation Control MCQ	✓	–	690	Medical image MCQ with 4 choices.
	Negation Understanding MCQ	✓	–	1,587	MCQ task with negation.
CheXpert (binary)	Affirmation Control	✓	–	690	Binary classification of medical images.
	Negation Understanding	✓	–	690	Binary classification, negated statements.
HardNeg-Syn	Retrieval-Neg	✓	✓	20,000	Synthetic image retrieval task.
	MCQ-Neg	✓	✓	20,000	MCQ task for synthetic images with 4 answer choices.

D.1. Details of HardNeg-Syn Construction

Object Label Selection

We gather a wide range of object text labels from existing datasets like ImageNet.

Scene Description

For each selected object label (`A`), LLaMA 3.1 generates:
A `{background_description}` and a related object `{B}`, crafting realistic scene contexts.

Image Generation

Using Stable Diffusion, we generate pairs of images:

> **Positive Image:** `{background_description}` with `{A}` next to `{B}`.
>
> **Negative Image:** `{background_description}` with `{A}`, excluding `{B}` in the negative prompt to ensure its absence.

Verification

We use OWL-ViT [27] to verify the presence and absence of `A` and `B`.

Caption Generation

Captions are generated using templates and paraphrased with LLaMA 3.1 for naturalness.

E. Visualizing the NegBench Evaluation Tasks

In Figures Figures 11 to 14, we visualize a few samples from the NegBench retrieval and MCQ tasks we introduced in the paper. We note that the datasets are diverse in terms of the nature of visual domain and real-world applicability.

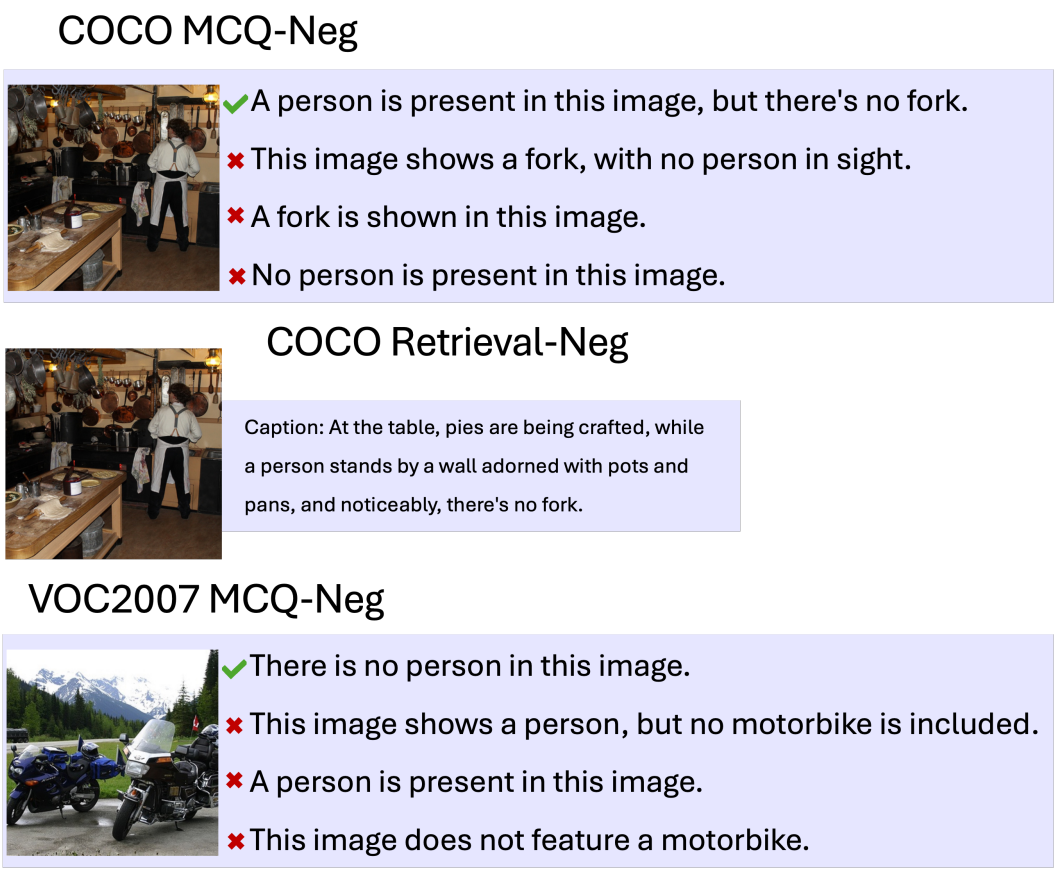

Figure 11. Examples of COCO and VOC2007 tasks, including Retrieval with negated captions and MCQ with negation.

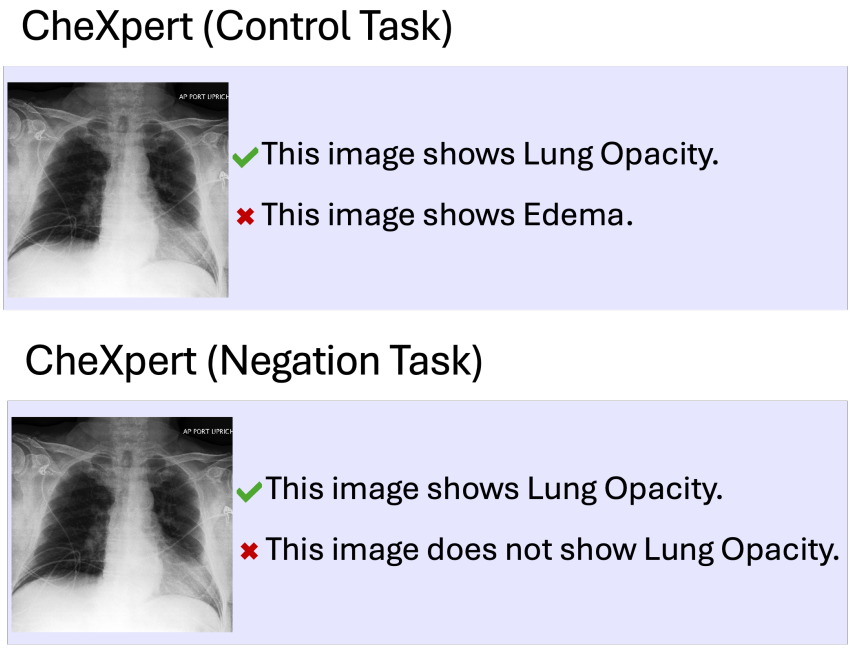

Figure 12. Examples of CheXpert MCQ tasks, including the Affirmation Control task and the Negation task.

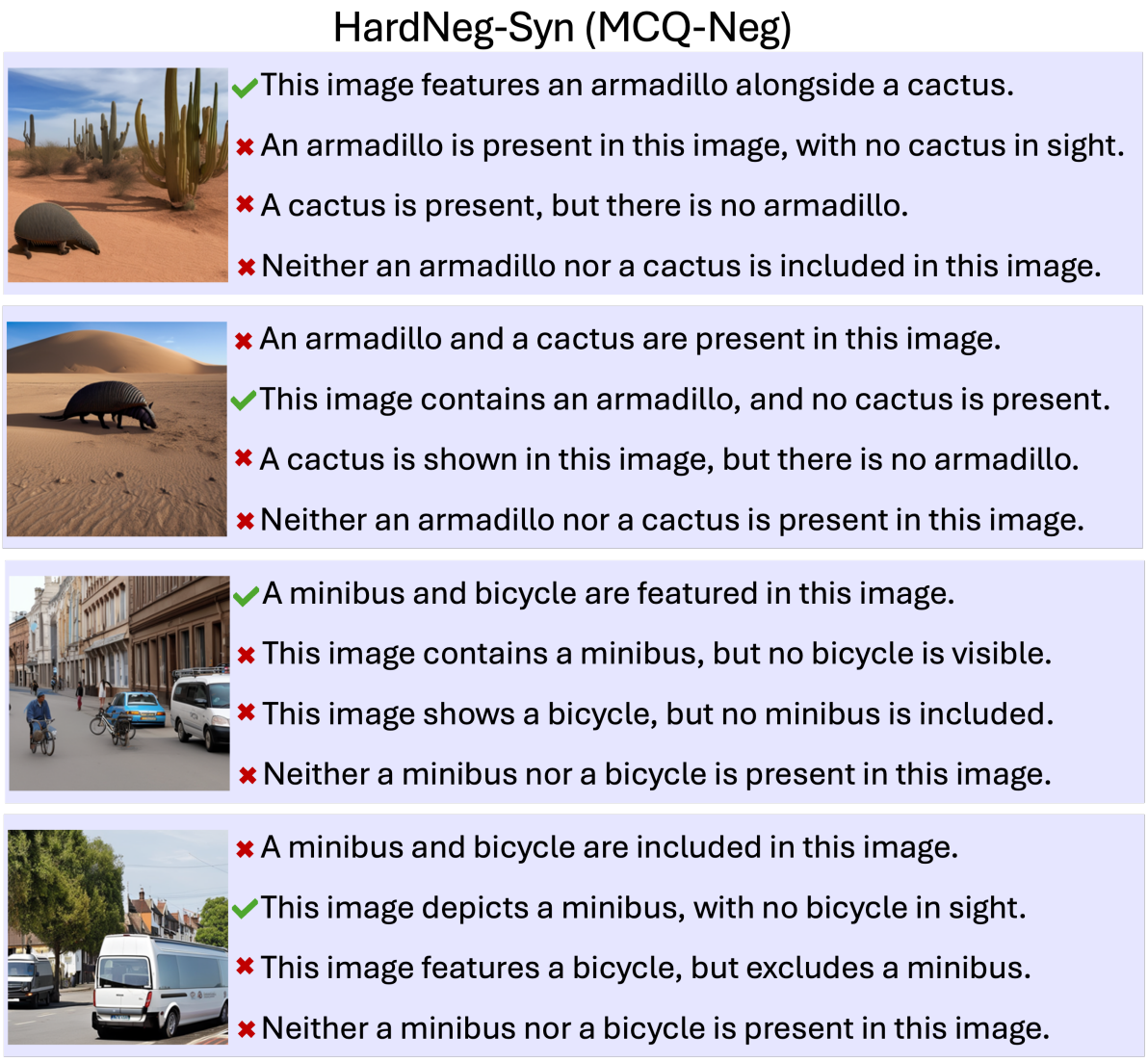

Figure 13. Examples of HardNeg-Syn (MCQ-Neg) tasks. Images in this dataset are constructed in pairs, with each pair differing by a single object (the cactus in the first pair), making the dataset particularly suitable for studying negation understanding.

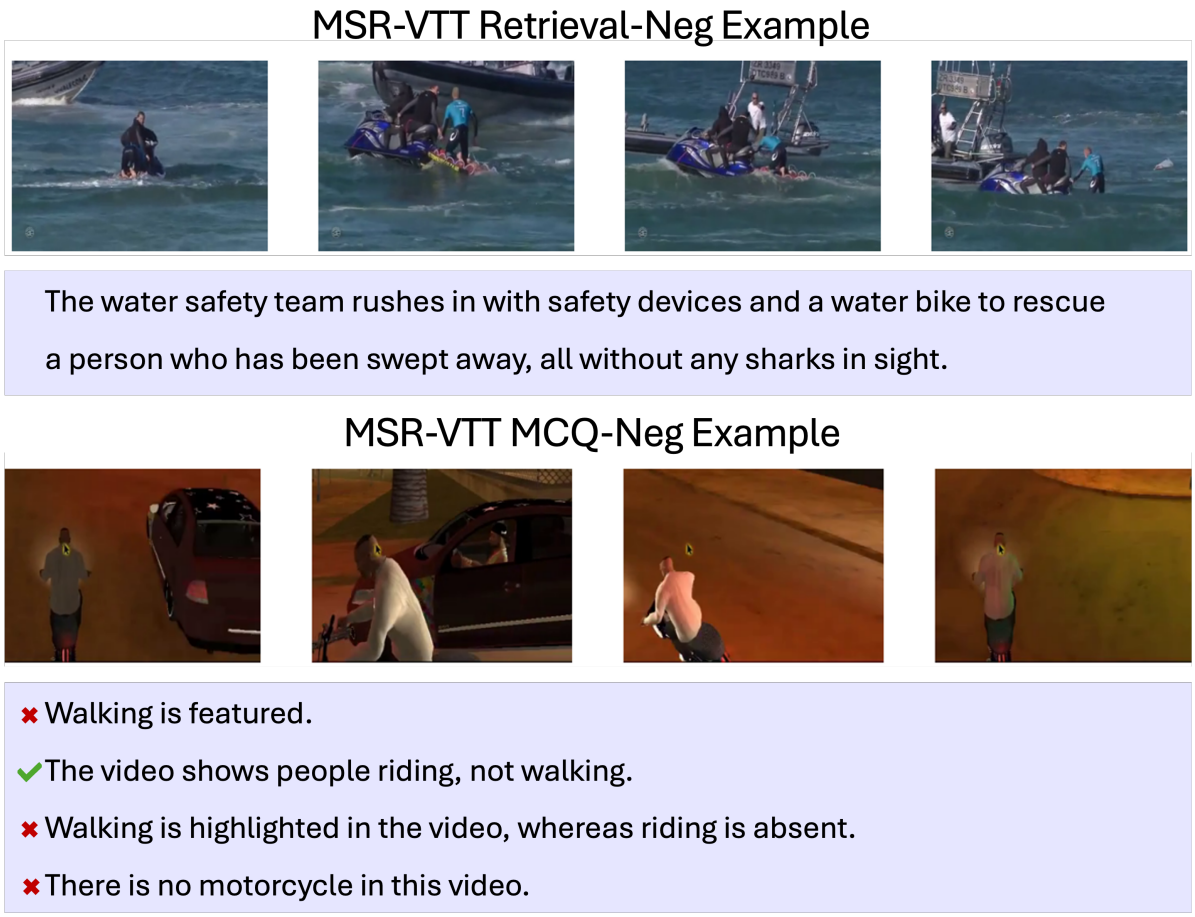

Figure 14. Examples of MSR-VTT tasks, including Retrieval-Neg (with negated captions about a complex water rescue scene) and MCQ-Neg (with answer choices about the presence or absence of actions like walking).