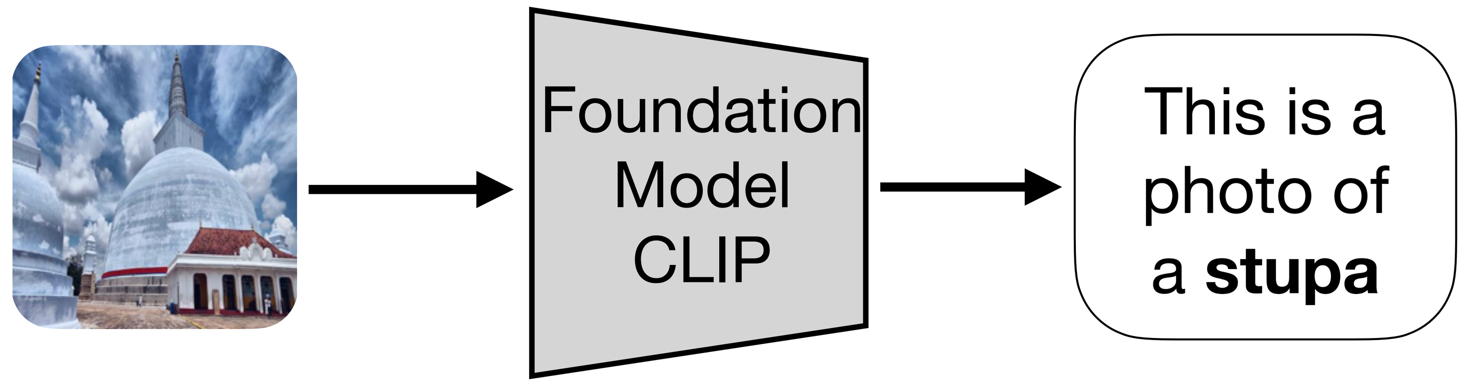
Trustworthy Foundation Models via Integrating Context

Chengzhi Mao, Junfeng Yang
Columbia University

Two Weaknesses of Today's Foundation Models

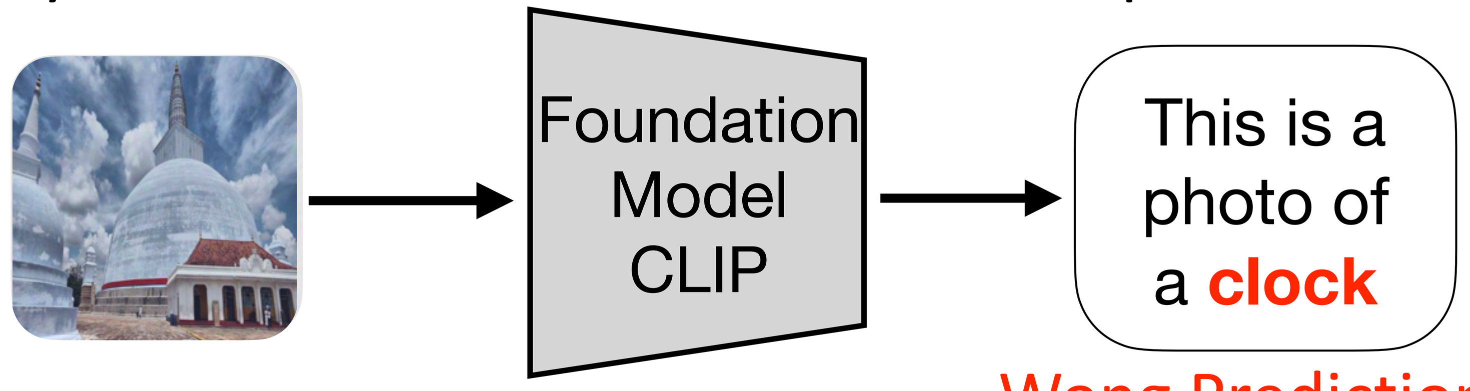
Not secure when handling open-world tasks (Weakness 1): Foundation models, like CLIP, are general purpose models. They can perform zero-shot recognition by retrieving language.

Normal Image



However, they are *vulnerable* under adversarial input.

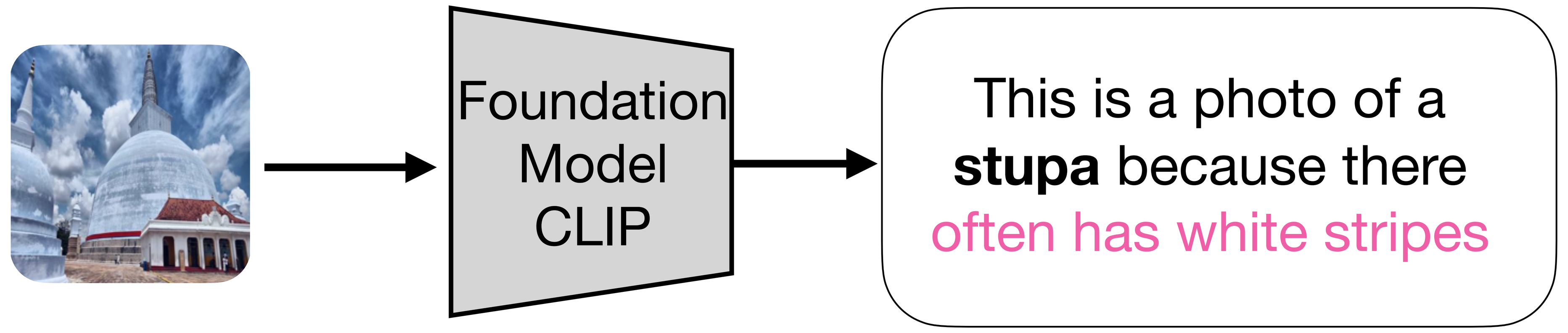
Adversarial Image



Wong Prediction!

Foundation models, such as LLAVA, mini-GPT4, and BLIP, rely on CLIP representation. CLIP will be a single point of failure because adversarial attacks that break CLIP will also fool those multi-modal LLM models. Secure CLIP vision encoder will be crucial.

Biased and hallucinated explanations (Weakness 2): Foundation models, such as CLIP, often *hallucinate* wrong rationales for their explanations.



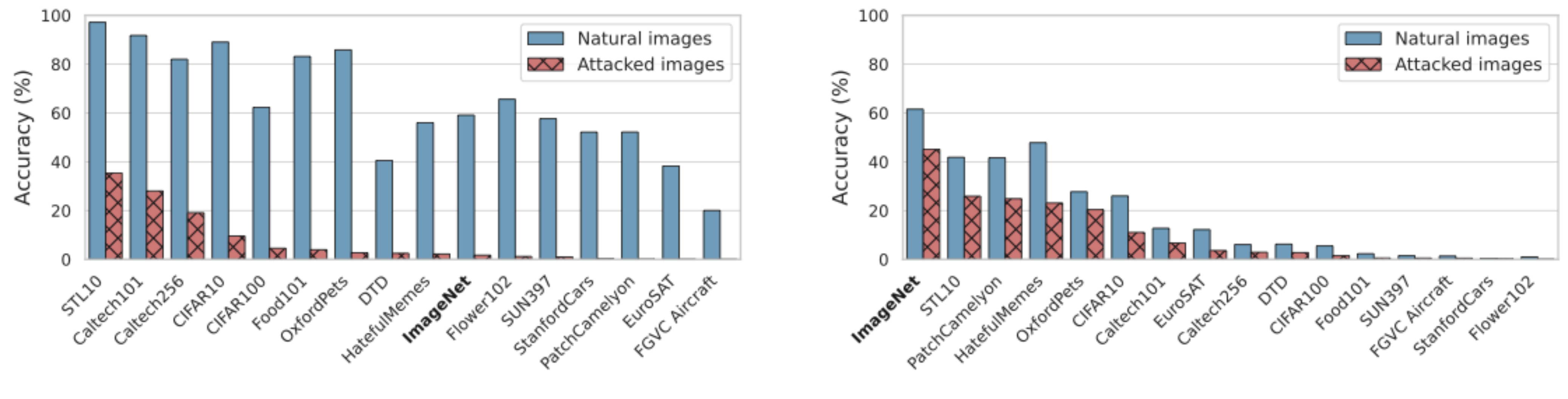
Wong Explanation!

This creates concerns when applying foundation models to applications where explanations are crucial, such as medical diagnosis.

Idea 1: Integrating Language Prior for Zero-Shot Adversarial Robustness

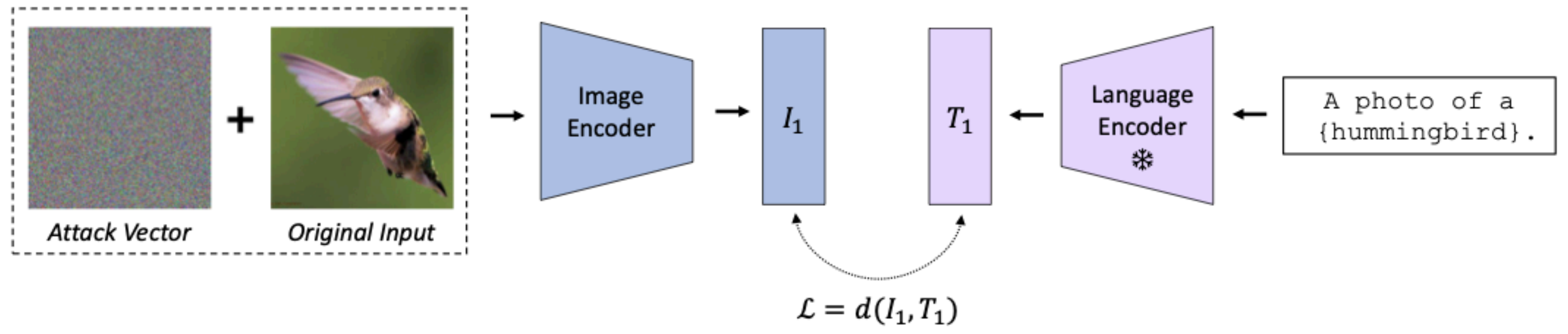
Training to secure foundation models: ICLR 2023

- Traditional robust training methods, like adversarial training, align vision representations with one hot label.
- Can only secure the task it has been adversarially trained on, but cannot generalize robustness to novel tasks.

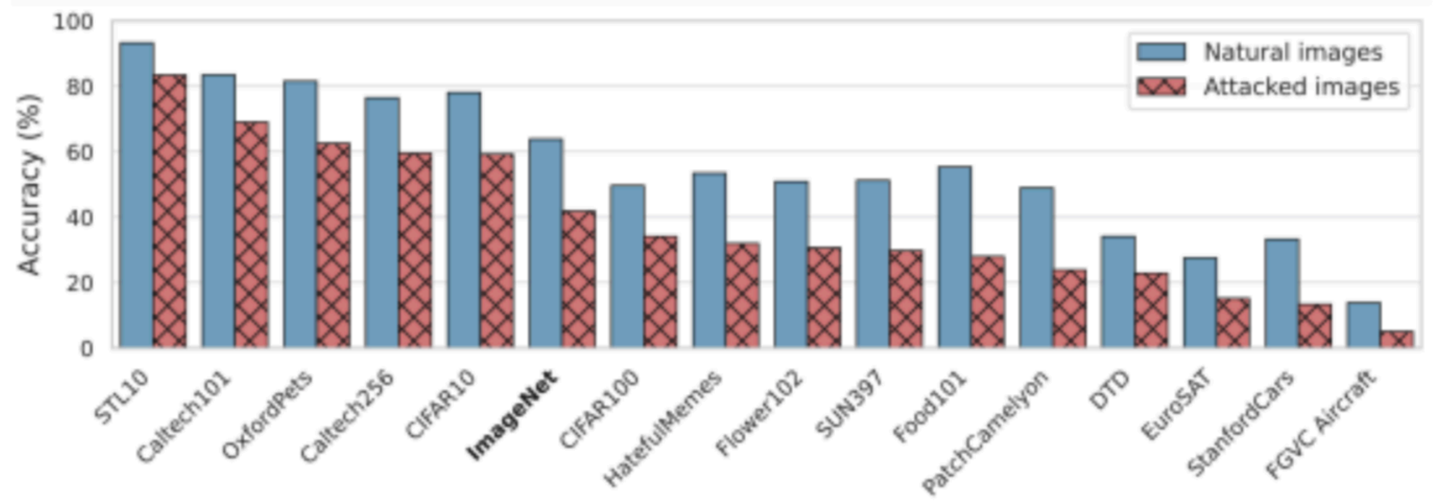


(a) CLIP (b) Adversarially Finetuned CLIP

Key Idea: <u>Align</u> vision representations to language representations during adversarial training. The inherent structure in language allows adversarial robustness transfer to zero-shot tasks.



Key Results:



Improving zero-shot adversarial robustness over 16 datasets by an average of 24%.

Idea 2: Integrating LLM and Web Knowledge to reduce Hallucinations on Explanations

CLIP retrieves incorrect rationales for explanations CVPR 2023

- Training data for foundation models can be biased.
- Do not incorporate all knowledge correctly and extensively.

Key Idea: <u>Align</u> vision representations to the correct rationales by incorporating knowledge from LLM reasoning and the Web.

Pipeline:

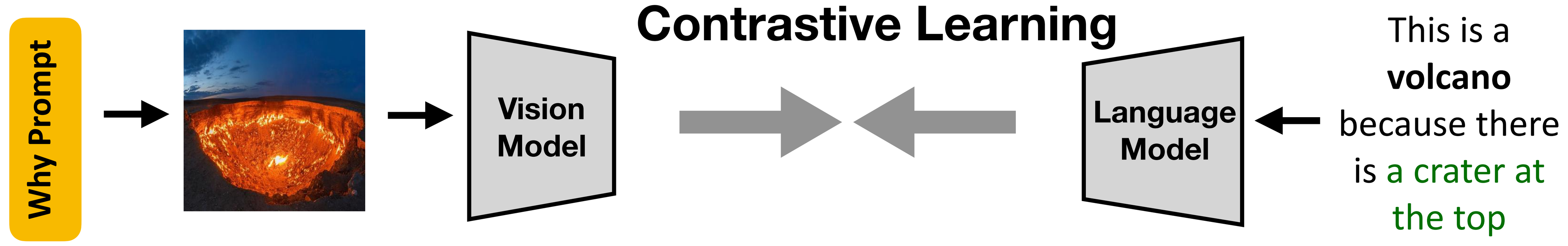
LLM Reasoning

What are the useful visual features for distinguishing a Volcano ? → LLM →
1. A crater at the top
2. A cone shaped mountain
3. Smoke Belowing from crater

→ Google Image Search → Images contain the described features

Web Knowledge

Align foundation model's vision representations with the rationales



Results: We can <u>correct the hallucinations</u> in foundation models and produce the right explanations, improving accuracy by 20%.