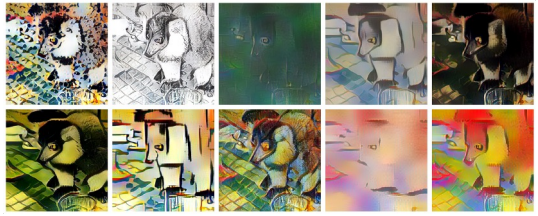
HIDDEN IN PLAIN SIGHT: Evaluating Abstract Shape Recognition in Vision-Language Models

Arshia Hemmat[1], Adam Davies*[2], Tom A. Lamb*[1], Jianhao Yuan*[1], Philip Torr[1], Ashkan Khakzar[1], Francesco Pinto[1]
*Equal Contribution. [1]University of Oxford. [2]University of Illinois at Urbana-Champaign

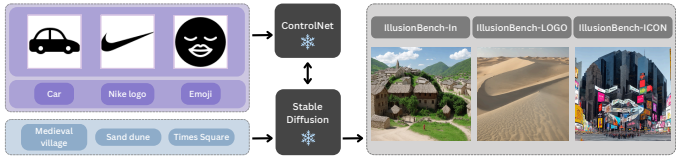
Problem

- Do image classifiers rely on shape or texture?
- Limitations of existing benchmarks:
  - Lack of coherent, naturalistic, complex visual scenes.
  - Missing shape information, poor fine-grained details.



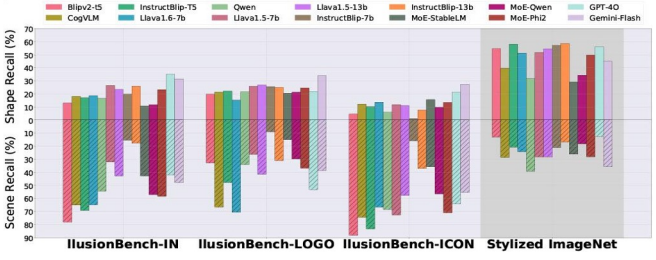
Approach

To address these issues, we introduce `IllusionBench`:



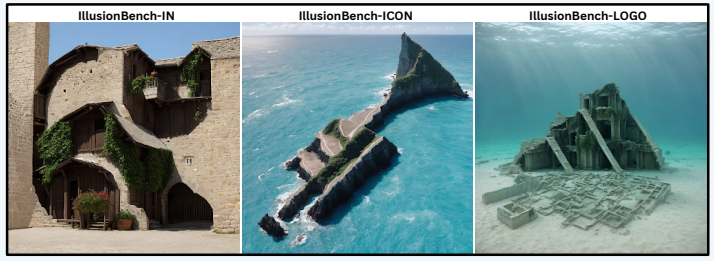
- Use ControlNet to condition Stable Diffusion on…
  - Conditioning images (binary masks of target shape).
  - Scene description (e.g., “medieval village” or “sand dune”).

Zero-shot Shape Recognition



- SOTA VLMs biased towards scene/texture.
- Shape recognition gap between open and closed-source models.

Can SOTA vision-language models recognise these shapes while ignoring scene/texture?

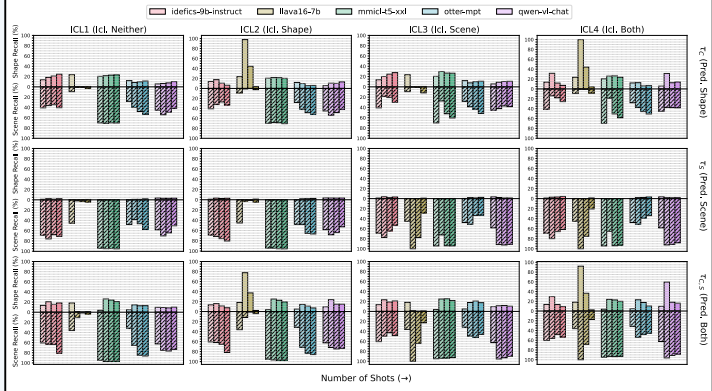


Zero-shot? ❌
Few-shot? ❌
Fine-tuned? 🤔

PAPER

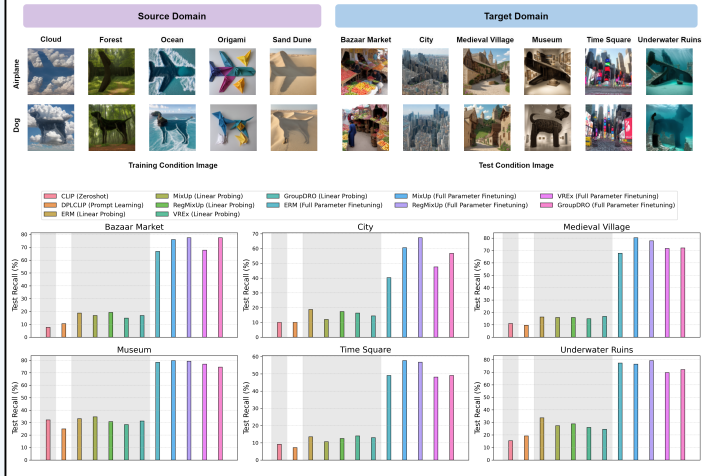
WEBSITE

Few-shot Shape Recognition



- ICL does not consistently solve the problem.
- VLMs still biased towards scene/texture.

Domain Generalisation



- CLIP models fail zero-shot or via probing.
- Can be fine-tuned to learn domain-generalisable features.