Reproducing and Extending Findings from LVLM-LP

COMP8539 Advanced Topics in Computer Vision

August 31, 2025

Paper: https://link.springer.com/chapter/10.1007/978-3-031-73195-2_8
ArXiv: https://arxiv.org/abs/2403.09037
Repository: https://github.com/Qinyu-Allen-Zhao/LVLM-LP

Overview

Your task is to reproduce the core findings of the paper and explore possible extensions of the proposed method. The assignment consists of three parts as follows.

Task 1: Reproduce Core Experiments (4 points)

Reproduce the main experiments on linear probing presented in the paper. There are six sub-tasks (identify unanswerable visual questions, defense against jailbreak attack, identify deceptive questions, estimate uncertainty in math solving, mitigate hallucination, and classify images). You need to reproduce results on **2–3 datasets of your choice**.

For each selected dataset, you are required to:

(i) Prepare the dataset (download, format, and load it correctly).

(ii) Prepare the model (load the specified or chosen vision-language model).

(iii) Run inference to obtain the **logits of the first token** for each input.

(iv) Perform linear probing using the extracted logits to evaluate classification performance.

Note on compute: If you face OOM issues running models from the repository, you may use smaller models such as InternVL3-1B, InternVL3-2B, or Qwen2.5-VL-3B-Instruct.

Task 2: Reproduce the Decoding Strategy (2 points)

The paper introduces a decoding strategy that utilizes the first token. While the original implementation modifies the `transformers` library, you may implement the decoding strategy in your own way (e.g., by writing wrapper functions or adding hook functions).

- Carefully follow the guidance in the repository README.
- Implement the decoding strategy and **show example outputs** that reflect the method's effectiveness (e.g., cases where hallucinations are reduced or unanswerable questions are detected).
- You may choose any reasonable approach, as long as it aligns with the paper's goal.
- As in Task 1, smaller models may be used if resources are limited.

Task 3: Extension Task (4 points)

This part is open-ended and rewards creativity and technical depth. You can select one or multiple directions from the list below and implement an initial prototype with discussion.

You are encouraged to think beyond the given list — these are only suggestions. Innovative and meaningful ideas outside this list will be credited.

Your mark will be based on your understanding and innovation of your work.

Note: Consider your available compute resources when choosing a direction. Some topics are feasible with lightweight models, while others require stronger hardware.

1. **Beyond Classification:** Apply the first-token probing idea to other vision tasks such as *object detection* or *semantic segmentation*.
   Resources: Feasible with smaller models on limited hardware.
   Recommended reading:
   LLaVA-Grounding: `https://arxiv.org/abs/2312.02949`
   LISA: `https://arxiv.org/abs/2308.00692`

2. **Video Understanding:** Investigate whether the method (e.g., probing the first token) can be extended to video classification tasks, given that recent vision-language models support video inputs.
   Resources: A bit more resource-intensive; better suited for stronger GPUs.
   Recommended reading:
   Qwen-VL 2.5: `https://arxiv.org/abs/2502.13923`
   LLaVA-OneVision: `https://arxiv.org/abs/2408.03326`

3. **Image Generation:** Explore whether the first token's information can guide or condition image generation, leveraging the generative capabilities of models like GPT-4o or CogVLM.
   Resources: Proof-of-concept possible on smaller models; full experiments require higher compute.
   Recommended reading:
   Show-o2: `https://arxiv.org/abs/2506.15564`
   Bagel: `https://arxiv.org/abs/2505.14683`

4. **Integration with Reinforcement Learning (e.g., GRPO):** Since the first token reflects the correctness of the response, can it serve as a reward signal in reinforcement learning?
   Resources: Highly challenging; typically requires substantial compute and is optional.
   Recommended reading:
   DeepSeek-R1: `https://arxiv.org/abs/2501.12948` (You can find a few good implementations of GRPO on GitHub.)
   Reasoning Models Know When They're Right: `https://arxiv.org/abs/2504.05419v1`

Assessment Instructions

Same as Assignment 1, you will present your findings in your lab. This will take place in Week 9. Make sure you present your findings very clearly. Presentation is worth 3 marks, and peer marking is worth 2 marks.