Can Vision–Language Models Enable Direct Fish Counting in Aquaculture?

Zheng Miao Tien-Chieh Hung* Zhe Zhao

Department of Biological and Agricultural Engineering; Department of Computer Science, University of California, Davis

Abstract

Fish counting is a key task in aquaculture, yet existing vision-based methods are often tailored to specific facilities and lack scalability across environments. Vision-language models offer strong cross-domain generalization, making them promising for species-agnostic fish counting.

Introduction

- **Importance:** Fish counting for biomass estimation, feeding optimization, and farm management in aquaculture.
- **Limitations of existing methods:** Existing methods (Figure 1) are tailored to specific fish farm facilities, limited scalability and generalization across diverse aquaculture environments.
- **Our work:** We evaluate the performance of vision language models for zero shot fish counting in aquaculture.

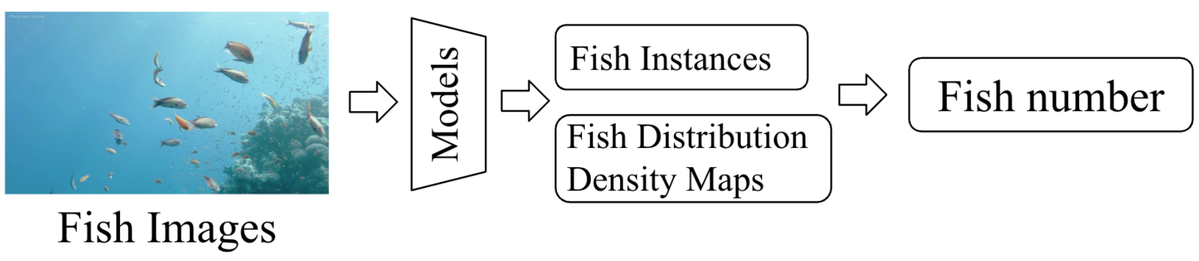

Figure 1. Pipeline of traditional machine learning models for fish counting.

Methodology

Zero-shot fish counting on the IOCFish5K dataset using two categories of vision–language models, without model retraining:

- **General-purpose vision–language models:** Qwen, GPT, and Gemini, using textual prompts for direct image-level object counting (Figure 2).
- **Density-based vision–language counting models:** CountGD, leveraging both textual and visual prompts to generate text-conditioned density maps (Figure 5).

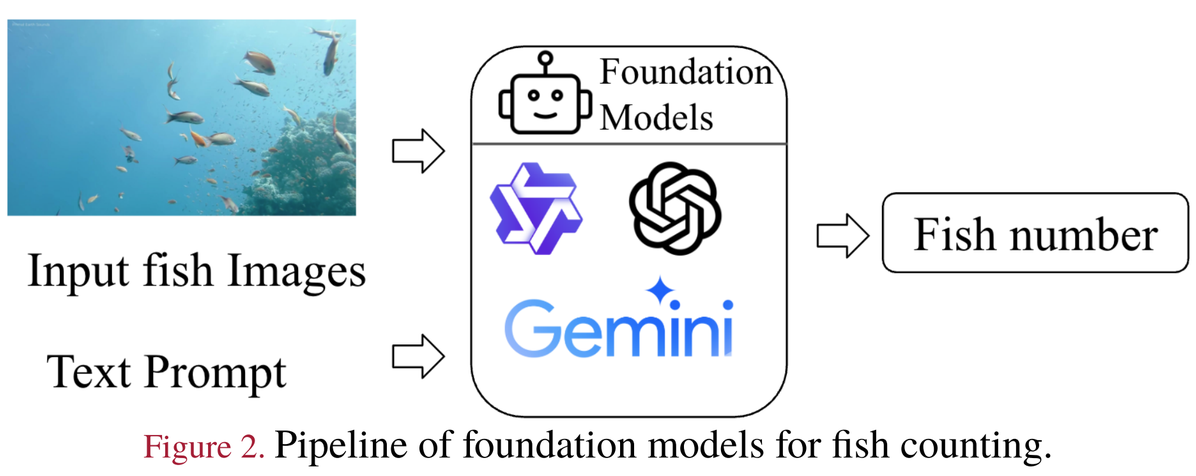

Figure 2. Pipeline of foundation models for fish counting.

Conclusion

This study evaluates the feasibility of applying vision language models to fish counting in aquaculture. While general purpose vision language models achieve competitive average accuracy without retraining, their predictions exhibit high error variance, limiting counting reliability. In contrast, vision language based density estimation provides more stable performance for zero-shot fish counting by better visual grounded ability.

Results

Table 1. Zero-shot fish counting on IOCFish5K (lower is better).

Model	Availability	Prompts	Training	MAE	RMSE
MCNN	Open	–	Fully-supervised	72.93	12.33
CSRNet	Open	–	Fully-supervised	38.12	8.86
IOCFormer	Open	–	Fully-supervised	15.91	5.84
Gemini-1.5-pro	Close	Text	Train-free	73.51	162.96
GPT-4o	Close	Text	Train-free	71.24	158.40
Qwen2.5VL-7B	Open	Text	Train-free	82.43	177.83
CountGD	Open	Text (+Visual)	Train-free	34.29	94.75

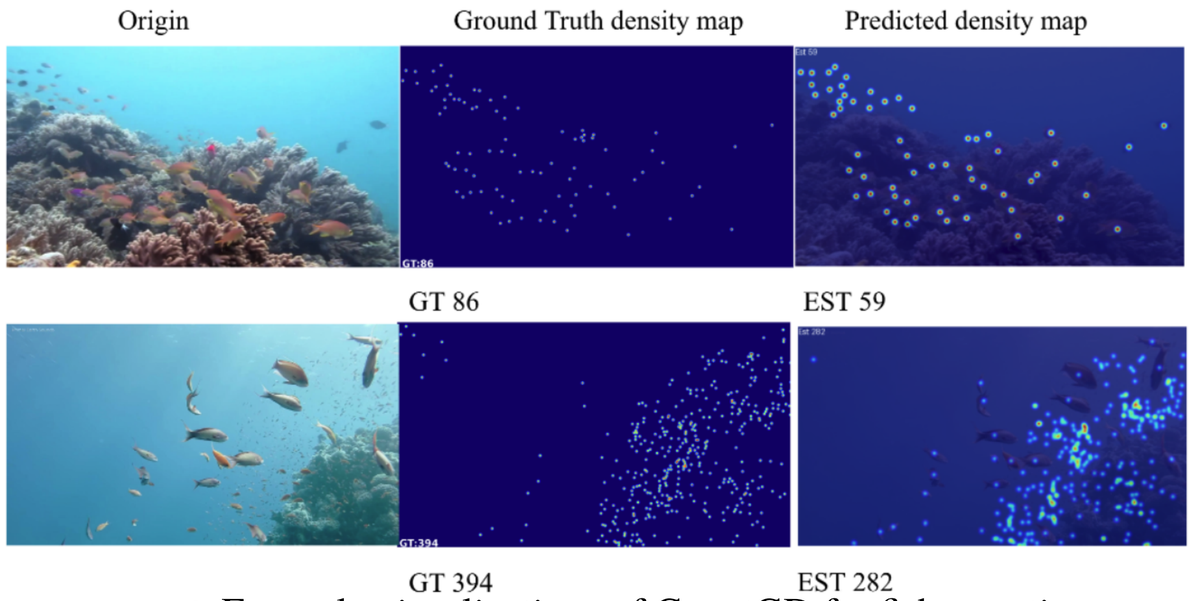

Figure 3. Example visualizations of CountGD for fish counting.

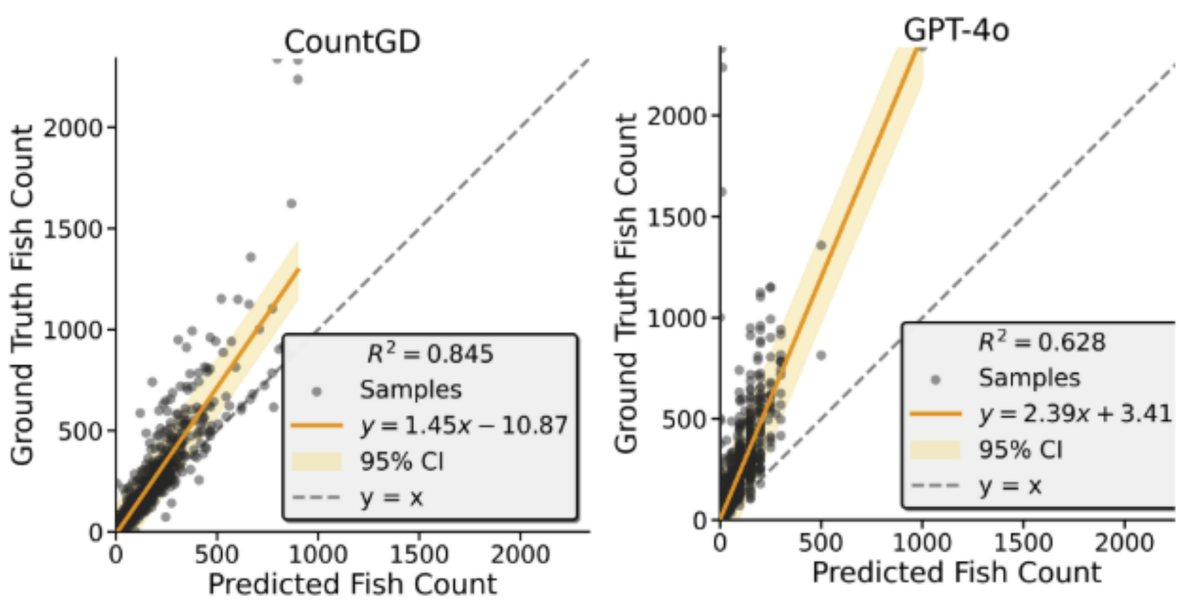

Figure 4. Regression analysis of predicted counts vs. ground truth.

Discussion

This instability of VLMs for fish counting in aquaculture is likely caused by hallucination arising from weak visual grounding and reasoning. Promising directions to mitigate the instability of VLMs include:

- **Prompt engineering:** Designing structured prompts to reduce ambiguous or implausible counting outputs;
- **In-context learning:** Incorporating visual and semantic context to strengthen visual grounding;
- **Model adaptation:** Adapting VLMs to counting tasks for more scenarios understanding.

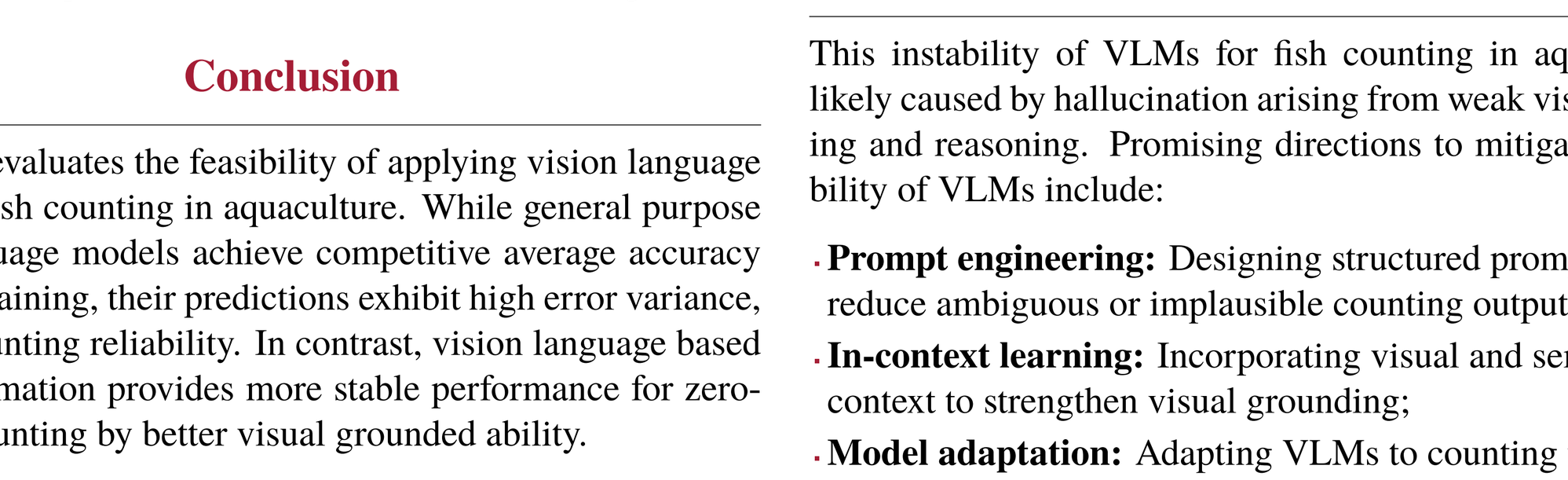

Figure 5. Pipeline of density-based vision–language counting for fish counting, e.g. Countgd.

Contact: zhmiao@ucdavis.edu