



# Text-Guided Object Counting in Images

**GVHD: TS. Mai Tiến Dũng**

**Member:**

• Hứa Tấn Sang	- 22521239
• Cao Tiến Trung	- 22521553
• Nguyễn Anh Khoa	- 22520675

# Overview

---

01 Introduction

---

02 Problem Definition

---

03 Method & Dataset

---

04 Result & Conclusion

---

# Introduction

## 1. Giới thiệu

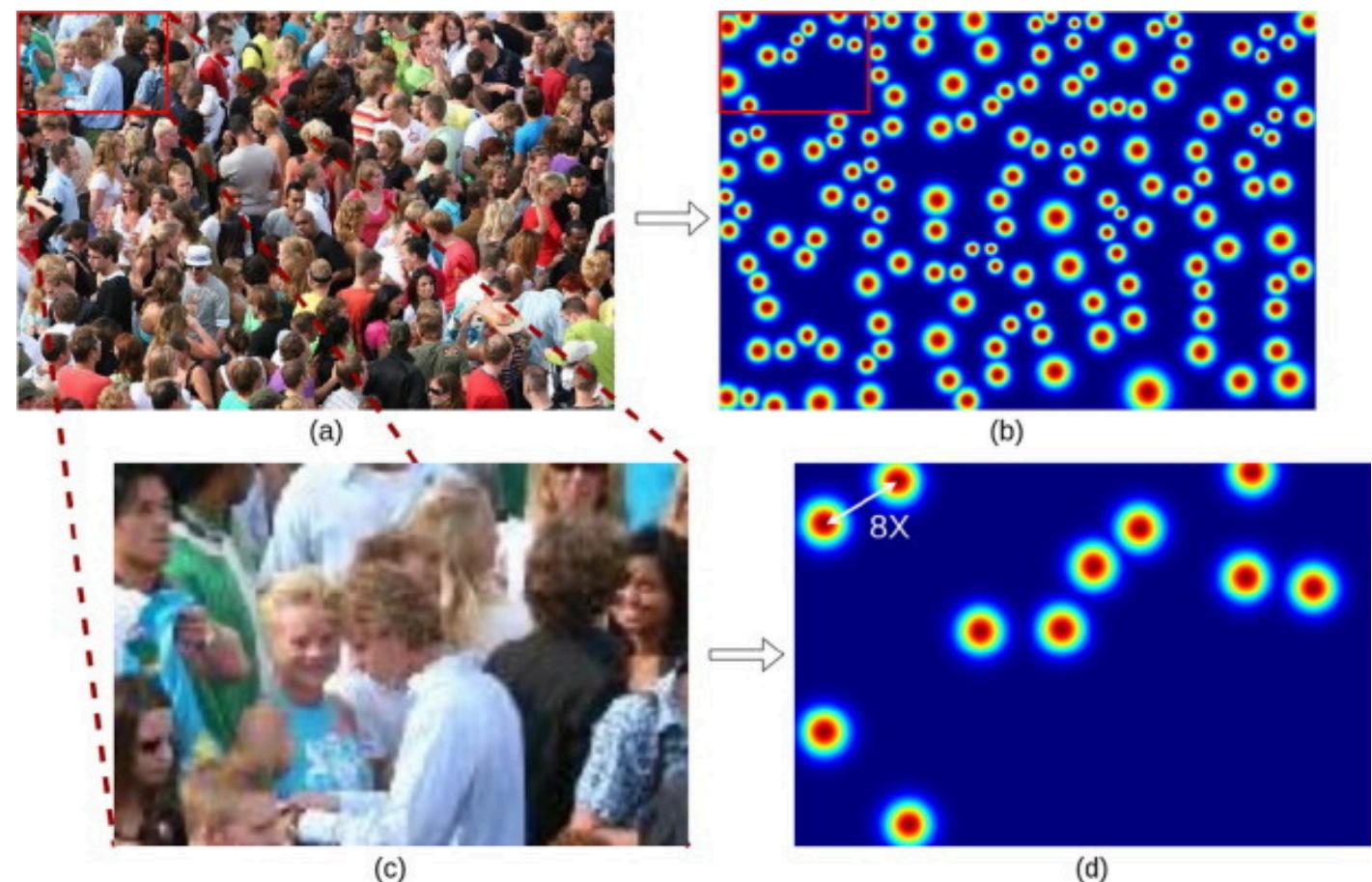
- Object Counting: Đếm số đối tượng thuộc các lớp cố định trong ảnh
- Text-Guided Counting: Sử dụng Prompt văn bản để chỉ định đối tượng cần đếm
- Cơ chế: Đếm các vật thể khớp với mô tả ngôn ngữ đầu vào
- Động lực: Ứng dụng cho quản lý kho hàng, Nông nghiệp,...

## 2. Ưu điểm

- Linh hoạt: Thay đổi Prompt → đếm mọi loại đối tượng
- Tổng quát hóa tốt: Không cần huấn luyện lại cho từng lớp
- Hiểu ngữ nghĩa: Đếm theo thuộc tính (vd: dâu tây, táo,...)

## 3. Thách thức

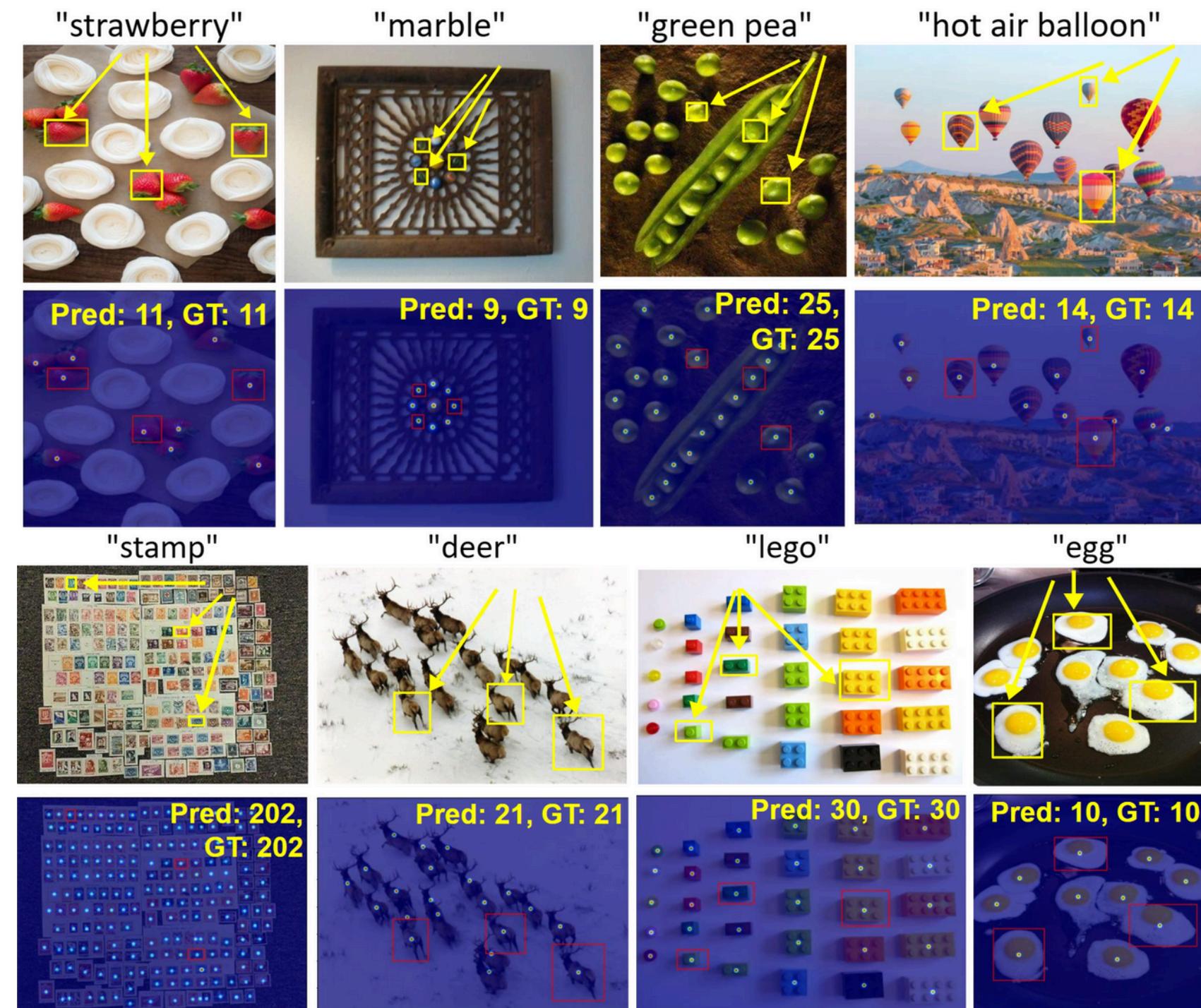
- Đối tượng khó: Nhỏ, dày đặc, chồng lấp
- Biến thiên ảnh: Góc nhìn, tỉ lệ, ánh sáng
- Prompt phức tạp: Thuộc tính tinh vi khó xử lý hơn đếm đơn giản



# Problem Definatoin

- |               |  |
|---------------|--|
| <b>Input</b>  | <ul style="list-style-type: none"><li>• Một ảnh RGB chứa một hoặc nhiều đối tượng khác nhau.</li><li>• Một text prompt là danh từ tiếng Anh chỉ tên một loại đối tượng cần đếm trong ảnh (ví dụ: “strawberry”, “apple”, “car”,...).</li><li>• NOTE:<ul style="list-style-type: none"><li>◦ Mỗi ảnh có thể chứa 0 hoặc nhiều đối tượng phù hợp với text prompt.</li><li>◦ Bài toán hướng tới tính mở lớp: đối tượng trong prompt là không cố định và có thể là lớp chưa thấy.</li></ul></li></ul> |
| <b>Output</b> | <ul style="list-style-type: none"><li>• Một giá trị số biểu thị số lượng đối tượng tương ứng với prompt xuất hiện trong ảnh.</li><li>• Density map hiển thị vị trí của đối tượng được đếm.</li></ul>   |

# Methodologies

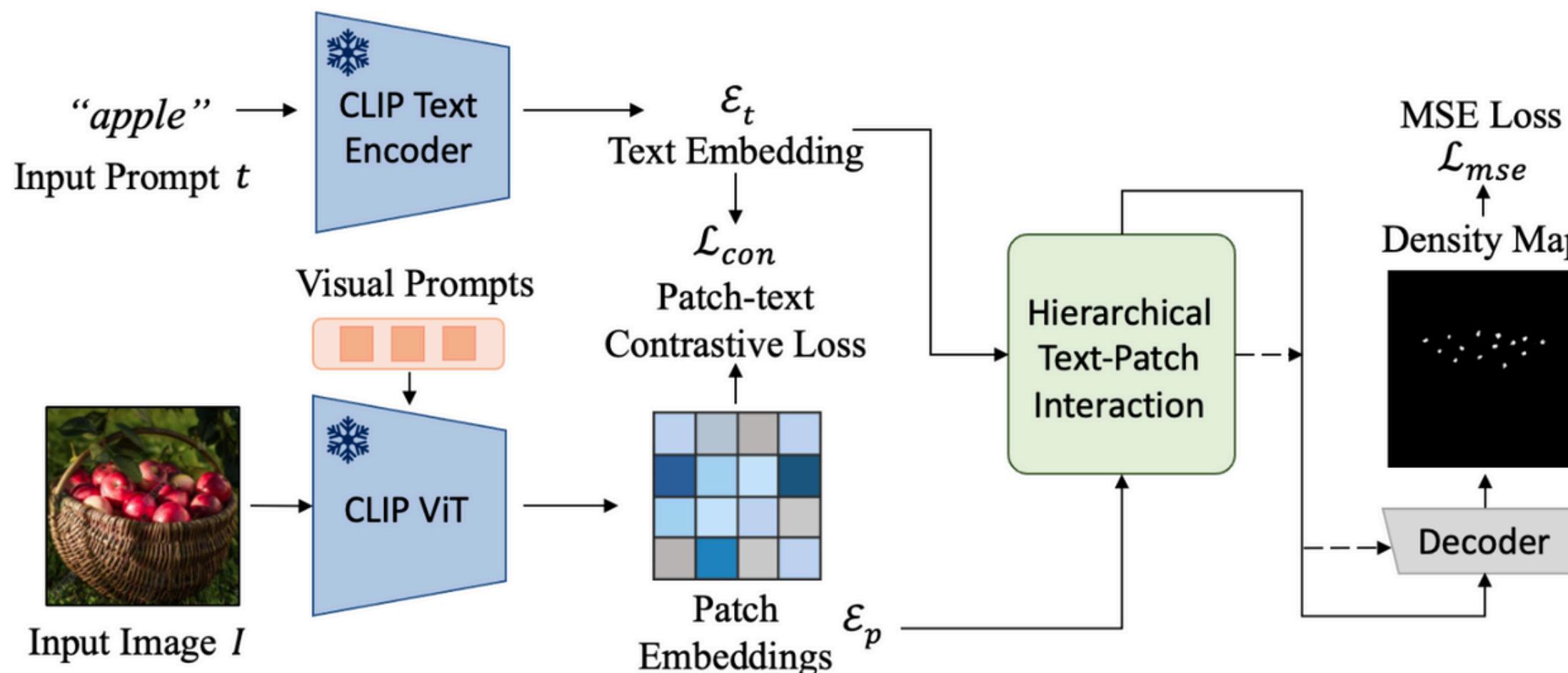


## 0. Data Augmentation

- **Tỉ lệ augmentation:**
  - 50% giữ nguyên
  - 30% augmentation thường
  - 20% Moisiac
- **Augmentation thường**
  - Áp dụng nhiều Gaussian
  - Thay đổi độ sáng và làm mờ ảnh (Color Jitter và Gaussian Blur)
  - Biến đổi hình học
  - Lật hình
- **Moisac Augmentation:**
  - Crop random 4 vùng từ ảnh gốc
  - Áp dụng Gaussian Blur và Color Jitter
  - Ghép 4 mảnh ảnh lại với nhau
  - Resize và crop cơ bản

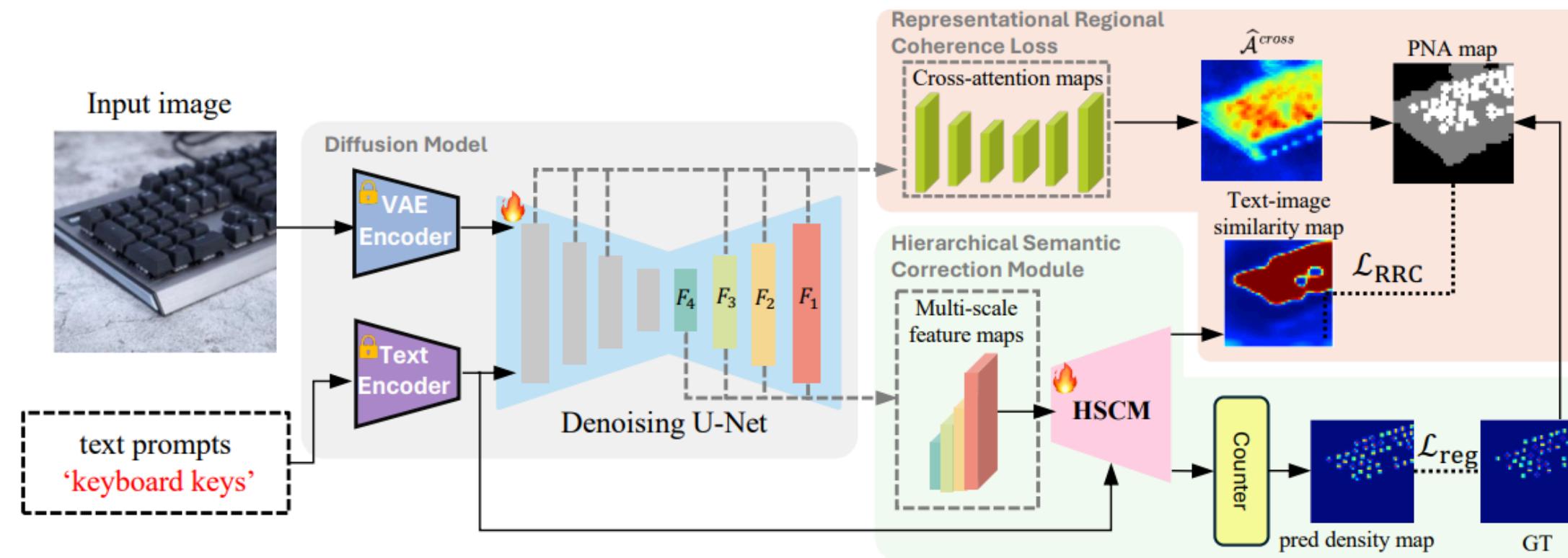
## 1. ClipCount

- CLIP-Count là một phương pháp đếm đối tượng (object counting) dựa trên mô hình CLIP, cho phép đếm số lượng bằng mô tả ngôn ngữ thay vì huấn luyện riêng biệt.
- Mô hình sử dụng CLIP để ánh xạ ‘text’ và ‘image’ vào chung một không gian vector để giúp mô hình có thể đếm được những hình ảnh chưa từng xuất hiện trong tập huấn luyện (Open-World Object Counting)



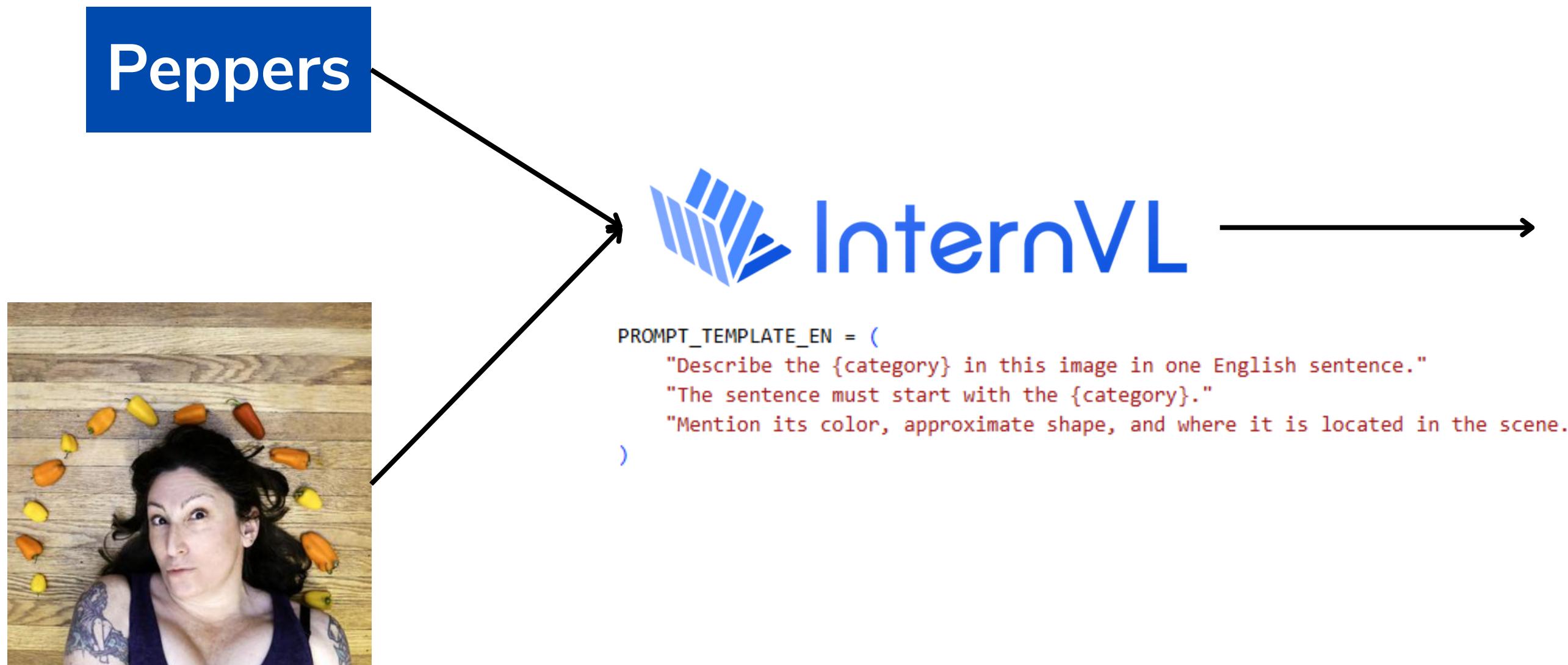
## 2. T2ICount

- T2ICount = Stable Diffusion (single-step) + HSCM + LRRC + Counter (density regression).
- Mô hình dùng HSCM sửa alignment theo kiểu coarse→fine, giúp đếm đúng lớp trong prompt, đặc biệt khi lớp đó là minority trong ảnh.
- Tận dụng pixel-level prior mạnh của diffusion: feature từ U-Net giàu chi tiết không gian hơn CLIP-only.
- Nhanh hơn diffusion chuẩn: chỉ chạy 1 denoising step nên nhẹ hơn nhiều so với multi-step diffusion nhưng vẫn giữ lợi thế representation.



# Methodologies

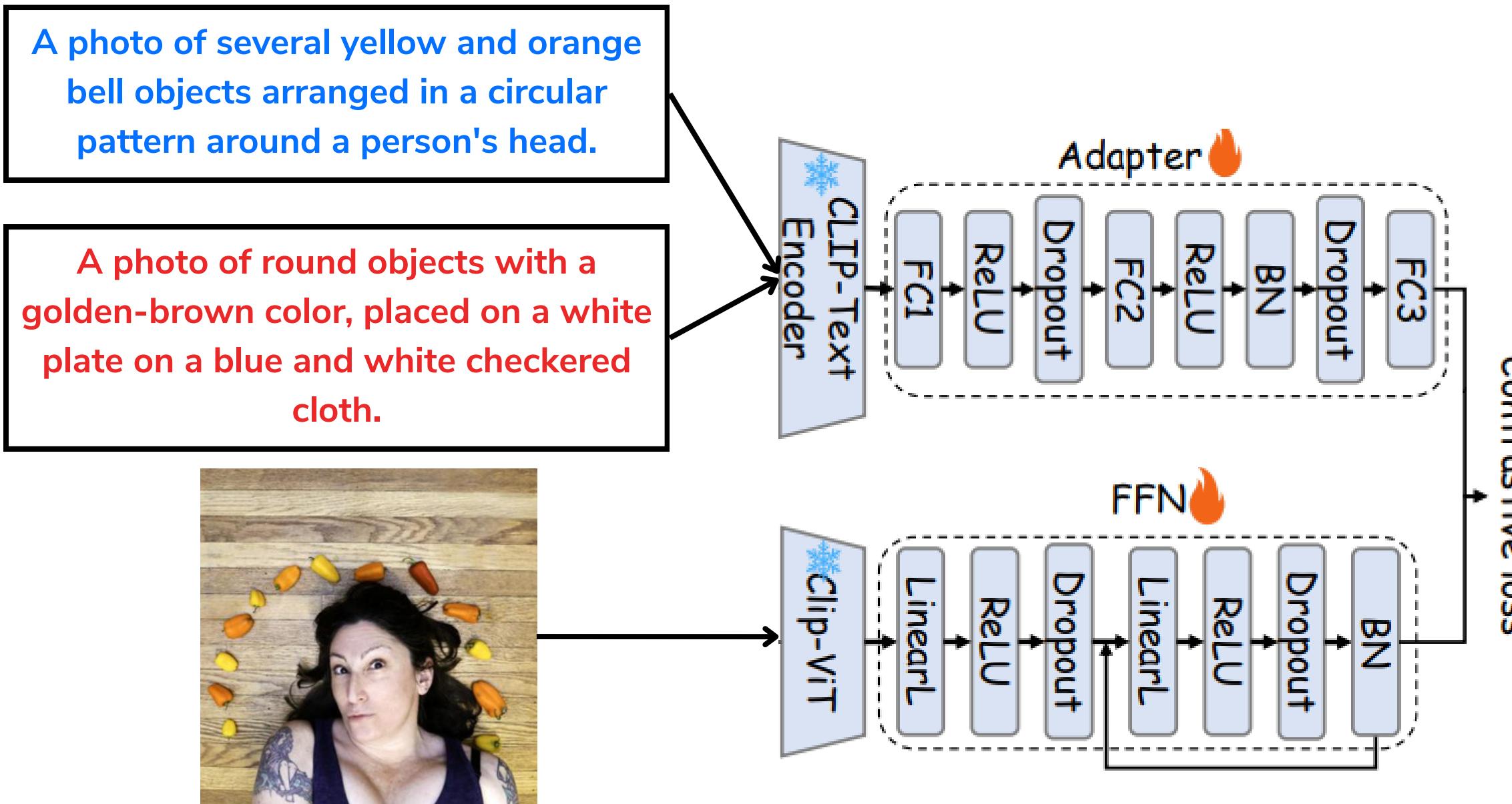
## 3. Proposed 1: InternVL+ ClipCount+ FFN&Adapter + Rank Loss



A photo of several yellow and orange bell objects arranged in a circular pattern around a person's head.

# Methodologies

## 3. Proposed 1: InternVL+ ClipCount+ FFN&Adapter + Rank Loss



$$d_i^+ = \|\mathbf{z}_i - \mathbf{t}_i^+\|_2$$

$$d_i^- = \|\mathbf{z}_i - \mathbf{t}_i^-\|_2$$

$$\text{gap} = \overline{d^-} - \overline{d^+}$$

# Methodologies

## 3. Proposed 1: InternVL+ ClipCount+ FFN&Adapter + Rank Loss

**Contrastive Loss**

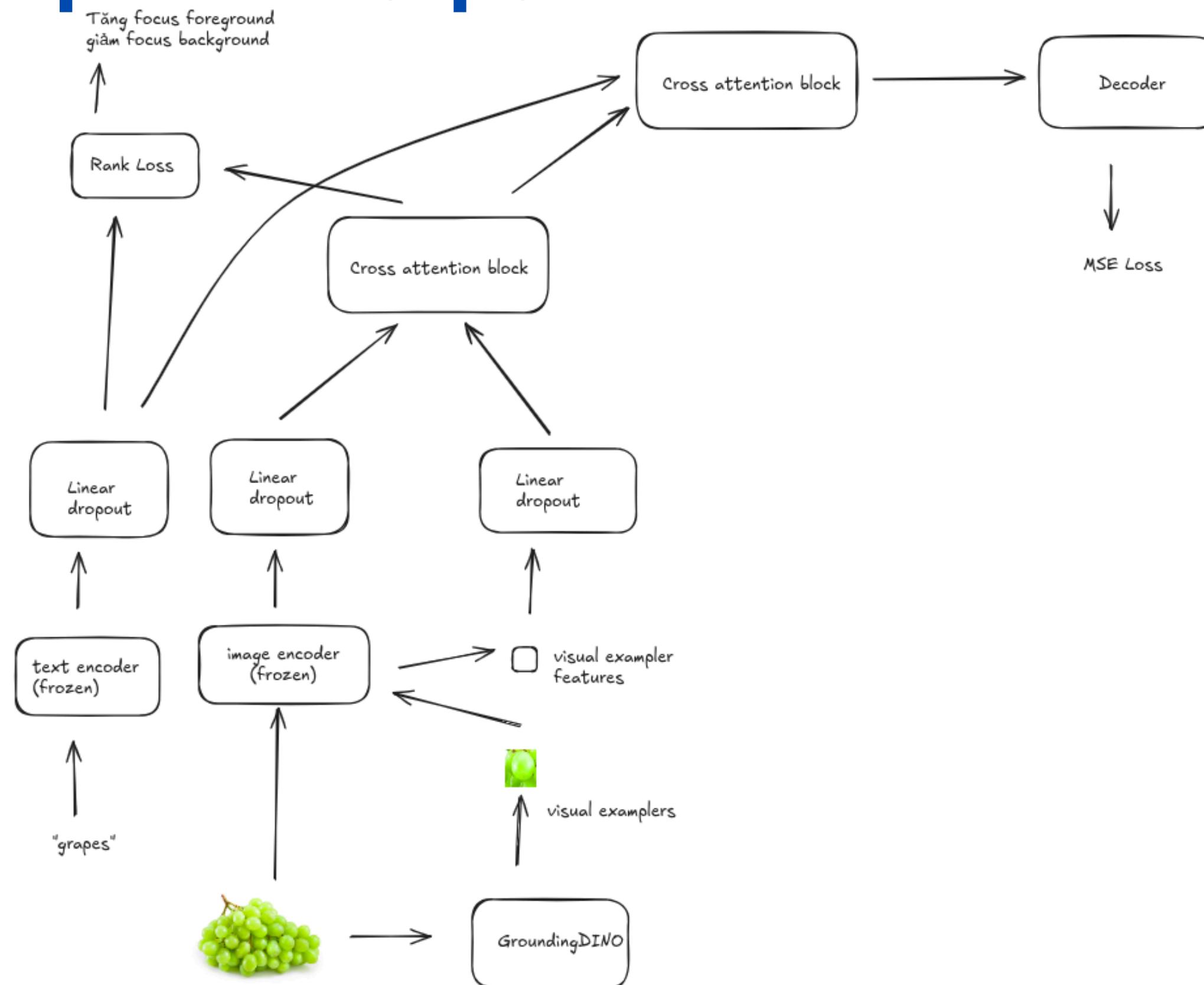
$$\mathcal{L}_{con} = -\log \frac{\sum_{i \in P} \exp\left(\frac{s(F_l^i, F_t)}{\tau}\right)}{\sum_{i \in P} \exp\left(\frac{s(F_l^i, F_t)}{\tau}\right) + \sum_{k \in B} \exp\left(\frac{s(F_l^k, F_t)}{\tau}\right)}$$

**Rank Loss**

$$d_{pq} = \frac{s(F_l^p, F_t) - s(F_l^q, F_t)}{\tau_1} \quad L_{rank} = \frac{1}{|P||B|} \sum_{p \in P} \sum_{q \in B} \max(0, \lambda - d_{pq})$$

# Methodologies

## 4. Proposed 2: ClipCount + Rank Loss + Exemplars



# Dataset

## 1. FSC-147

- **Giới thiệu:** Bộ dữ liệu lớn nhất và phổ biến nhất cho bài toán Few-shot Object Counting.
- **Quy mô:** 6,135 ảnh với 147 danh mục vật thể đa dạng (nhà bếp, động vật, xe cộ...).
- **Cơ chế Few-shot:** Cung cấp 3 hộp mẫu (exemplar boxes) mỗi ảnh để định nghĩa đối tượng cần đếm.
- **Đặc điểm Open-world:** Tập Train và Test không trùng lặp lớp đối tượng, yêu cầu mô hình có khả năng học đặc trưng thay vì ghi nhớ lớp.
- **Cấu trúc bộ dữ liệu:**
  - train: 89 unique classes, 0 missing images in class map
  - dev: 29 unique classes, 0 missing images in class map
  - test: 29 unique classes, 0 missing images in class map
- **Phân giao giữa các set**
  - train  $\cap$  dev: 0 classes
  - train  $\cap$  test: 0 classes
  - dev  $\cap$  test: 0 classes

### Learning To Count Everything

Viresh Ranjan<sup>1</sup> Udbhav Sharma<sup>1</sup> Thu Nguyen<sup>2</sup> Minh Hoai<sup>1,2</sup>

<sup>1</sup>Stony Brook University, USA

<sup>2</sup>VinAI Research, Hanoi, Vietnam

#### Abstract

Existing works on visual counting primarily focus on one specific category at a time, such as people, animals, and cells. In this paper, we are interested in counting everything, that is to count objects from any category given only a few annotated instances from that category. To this end, we pose counting as a few-shot regression task. To tackle this task, we present a novel method that takes a query image together with a few exemplar objects from the query image and predicts a density map for the presence of all objects of interest in the query image. We also present a novel adaptation strategy to adapt our network to any novel visual category at test time, using only a few exemplar objects from the novel category. We also introduce a dataset of 147 object categories containing over 6000 images that are suitable for the few-shot counting task. The images are annotated with two types of annotation, dots and bounding boxes, and they can be used for developing few-shot counting models. Experiments on this dataset shows that our method outperforms several state-of-the-art object detectors and few-shot counting approaches. Our code and dataset can be found at <https://github.com/cvlab-stonybrook/LearningToCountEverything>.

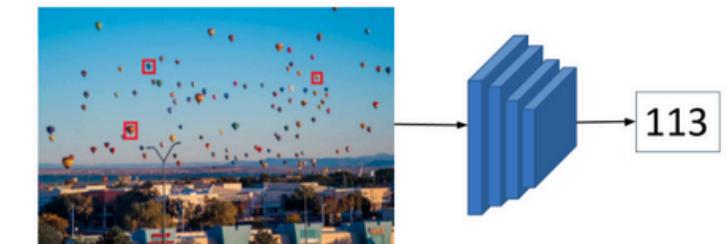


Figure 1: Few-shot counting—the objective of our work. Given an image from a novel class and a few exemplar objects from the same image delineated by bounding boxes, the objective is to count the total number of objects of the novel class in the image.

annotations for millions of objects on several thousands of training images, and obtaining this type of annotation is a costly and laborious process. As a result, it is difficult to scale these contemporary counting approaches to handle a large number of visual categories. Second, there are not any large enough unconstrained counting datasets with many visual categories for the development of a general counting method. Most of the popular counting datasets [14–16, 43, 49, 55] consist of a single object category.

In this work, we address both of the above challenges. To handle the first challenge, we take a detour from the existing

## 2. FSC-147-S

- **Giới thiệu:** Là một tập con của FSC-147
- **Quy mô:** 230 ảnh, gồm 84 class
- **Công dụng:** Để test mô hình sau khi đã huấn luyện

# Dataset

### T2ICount: Enhancing Cross-modal Understanding for Zero-Shot Counting

Yifei Qian<sup>1\*</sup>, Zhongliang Guo<sup>2\*</sup>, Bowen Deng<sup>1</sup>, Chun Tong Lei<sup>3</sup>, Shuai Zhao<sup>4</sup>, Chun Pong Lau<sup>3</sup>,  
Xiaopeng Hong<sup>5</sup>, Michael P. Pound<sup>1†</sup>

<sup>1</sup>University of Nottingham <sup>2</sup>University of St Andrews <sup>3</sup>City University of Hong Kong

<sup>4</sup>Nanyang Technology University <sup>5</sup>Harbin Institute of Technology

{yifei.qian, bowen.deng, michael.pound}@nottingham.ac.uk, zg34@st-andrews.ac.uk, {ctlei2, cplau27}@cityu.edu.hk,  
shuai.zhao@ntu.edu.sg, hongxiaopeng@hit.edu.cn

#### Abstract

*Zero-shot object counting aims to count instances of arbitrary object categories specified by text descriptions. Existing methods typically rely on vision-language models like CLIP, but often exhibit limited sensitivity to text prompts. We present T2ICount, a diffusion-based framework that leverages rich prior knowledge and fine-grained visual understanding from pretrained diffusion models. While one-step denoising ensures efficiency, it leads to weakened text sensitivity. To address this challenge, we propose a Hierarchical Semantic Correction Module that progressively refines text-image feature alignment, and a Representational Regional Coherence Loss that provides reliable supervision signals by leveraging the cross-attention maps extracted from the denoising U-Net. Furthermore, we observe that current benchmarks mainly focus on majority objects in images, potentially masking models' text sensitivity. To address this, we contribute a challenging re-annotated subset of FSC147 for better evaluation of text-guided counting ability. Extensive experiments demonstrate that our method achieves superior performance across different benchmarks. Code is available at <https://github.com/cha15yq/T2ICount>.*

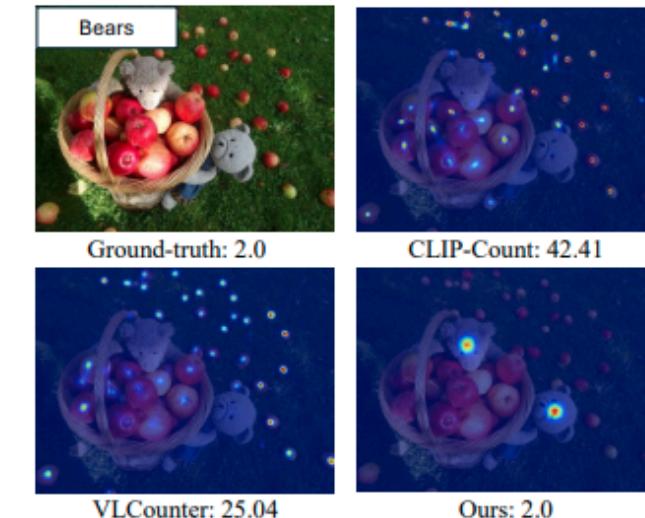


Figure 1. Visualizations of density maps predicted by official pretrained models of two recently proposed text-guided zero-shot object counting methods, CLIP-Count [7] and VLCounter [8], which demonstrate poor text sensitivity compared to the proposed T2ICount.

# Metrics

Độ đo	MAE	RMSE
Công thức	$MAE = \frac{\sum_1^n  y_i - \bar{y}_i }{n}$	$RMSE = \sqrt{\frac{\sum_1^n  y_i - \bar{y}_i ^2}{n}}$
Ý nghĩa	<ul style="list-style-type: none"><li>Cho biết sai số trung bình của mô hình với mỗi bức ảnh</li></ul>	<ul style="list-style-type: none"><li>RMSE tạo ra trọng số lớn đối với những kết quả có sai lệch lớn</li><li>Giúp quan sát được độ ổn định của mô hình</li></ul>

# Result

		FSC-147				FSC-147-S	
		Validate		Test		Test	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
T2iCount		<b>13.82</b>	<b>58.7</b>	<b>12.76</b>	<b>97.94</b>	<b>5.23</b>	<b>8.17</b>
ClipCount		19.5506	65.5509	17.86	103.7802	48.96	108.25
ClipCount using description		20.4243	69.8081	18.0831	102.5722	49.01	108.32
Proposed 1		<b>18.2918</b>	<b>62.698</b>	<b>17.4</b>	<b>104.107</b>	43.106	<b>106.7149</b>
Proposed 2		21.8035	67.3189	28.1863	126.6654	<b>38.8137</b>	113.7645

# Conclusion

- Mô hình T2ICount học ở mức pixel-level trong quá trình denoise + feature fusion đa tầng, nên localization tốt hơn → giảm lỗi khi vật thể nhỏ/đông/che khuất. Do đó cho ra kết quả tốt hơn nhiều so với các mô hình ClipCount-based vốn chủ yếu học ở mức global-level. Tuy nhiên, mô hình nặng và phức tạp khiến mô hình train lâu và lâu hội tụ.
- Mô hình ClipCount chủ yếu chỉ đếm majority class => ClipCount không thực sự “đếm theo text”, phần text-image alignment vẫn còn yếu
- Mô hình Proposed 1 có cải tiến so với ClipCount nhưng text-image alignment còn yếu.

# Conclusion

## Về Clip Count + Rank Loss + Exampler:

- Ảnh object được crop từ GroundingDINO có chất lượng chưa tốt; khi resize lên kích thước lớn hơn, ảnh bị mờ, làm suy giảm thông tin thị giác.
- Chưa áp dụng các module tăng cường đặc trưng (feature enhancement), nên chưa khai thác được representation ảnh hiệu quả như các phương pháp trong các nghiên cứu liên quan.
- GroundingDINO đôi khi trả về crop chứa nhiều object cùng lúc, khiến mô hình gặp khó khăn trong việc xác định và tập trung vào object cần đếm.

# Reference

Ruixiang Jiang, et al. CLIP-Count: Towards Text-Guided Zero-Shot Object Counting. arXiv arXiv:2305.07304

Yifei Qian, et al. T2ICount: Enhancing Cross-modal Understanding for Zero-Shot Counting.

arXiv:2502.20625

Shiwei Zhang, et al. Enhancing Zero-shot Object Counting via Text-guided Local Ranking and Number-evoked Global Attention,



# Thank You

Questions for more information <3