医学技术研究院
Institute of Medical Technology

北京大学
未来技术学院
College of Future Technology
Peking University

Vision-to-Concept and Language Tokenizer


Hangzhou He


Lei Zhu


Xinliang Zhang


Shuang Zeng


Qian Chen


Yanye Lu

Medical Intelligence Laboratory (MILab)

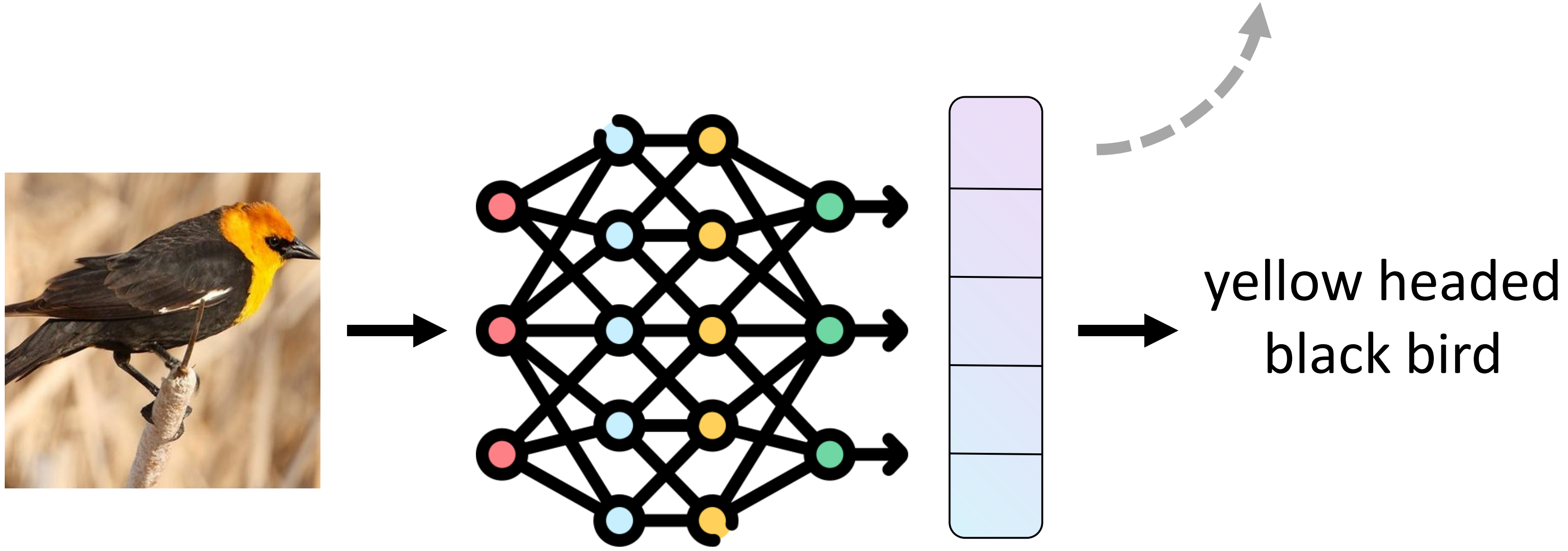
Imageomics-AAAI-25 – Philadelphia, USA

OUTLINE

1. Introduction
2. Vision-to-Concept Tokenizer (AAAI 2025)
3. Vision-to-Language Tokenizer (CVPR 2024)
4. Summary

Deep Neural Networks

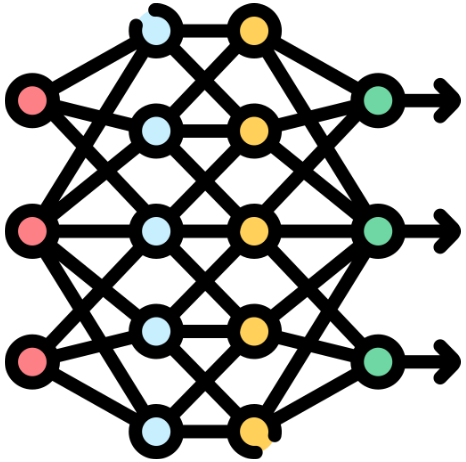
Deep learning has advanced the field of artificial intelligence with Deep Neural Networks (DNNs), which learn **feature representations** from data.



[1] Image borrowed from https://conceptlearning.github.io/

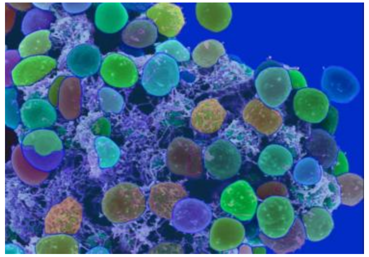
DNNs in Vision Tasks

Supervised training of DNNs with **a lot of labeled data** has proven highly effective for visual understanding and generation tasks.



+ A Lot of data with manual labeling =

classification



segmentation

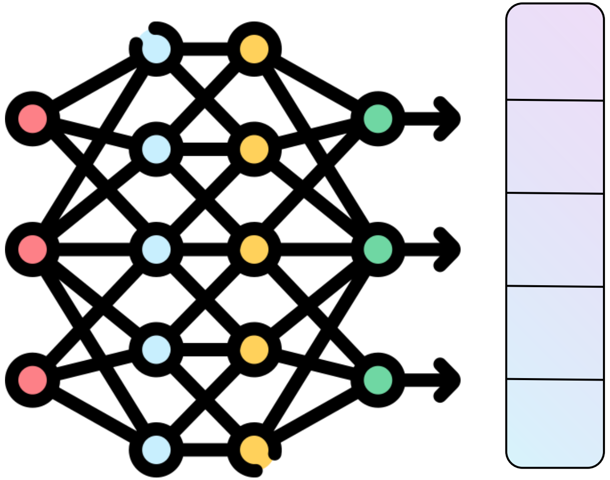
Image & video generation

[1] ImageNet: https://image-net.org/
[2] Segment Any Thing Model: https://segment-anything.com/
[3] Sora: https://openai.com/index/sora/

The Black-Box Problem

Can I teach the model to leverage concepts like *yellow headed* ?

Can I explain the decision-making process?



yellow headed black bird

Can I know what is encoded in that feature and learn from representation?

The Black-Box Problem

Can I teach the model to use concepts like *yellow headed* for classification?

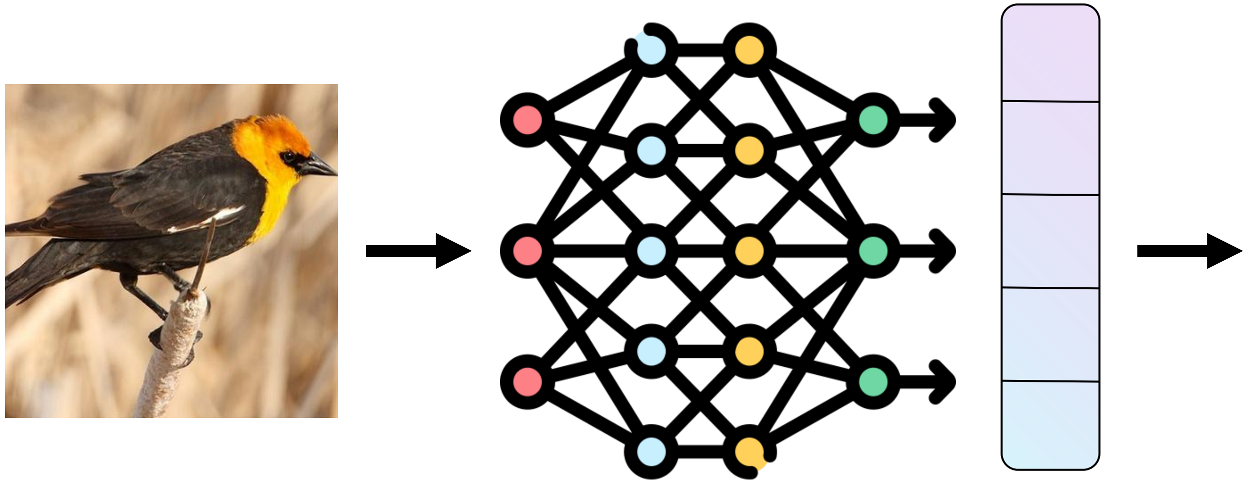
Can I explain the decision-

The answer is usually No!!!

Can I know what is encoded in that feature and learn from models?

Looking for Explanation Interface

Can we represent this feature in a **human-understandable** way?

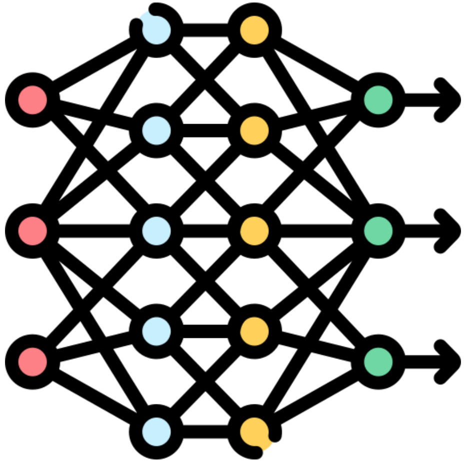


yellow headed black bird

Language as a Bidirectional Interface

integrate knowledge

learn from representations

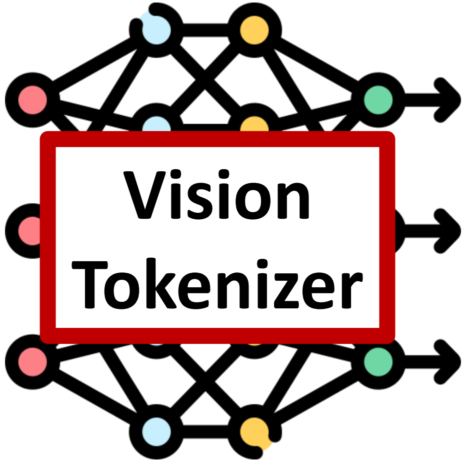


- *yellow headed*
- *black beak*
- *black wing*
- *black undertail*

yellow headed black bird

explain the decision

Building Vision Tokenizers



- *yellow headed*
- *black wing*
- *black undertail*
- *black beak*

yellow headed
black bird

Vision Encoder

Global:
CLIP, BioCLIP, ⋯
Local:
VQ-GAN

Codebook

visual concepts,
bio/medical terminology
tokens of language models,
⋯

Concept Codebook Generation

Concept Vocabulary

$\{adj_1, adj_2, \cdots, n_1, n_2, \cdots\}$

$\{adj_1\text{–}n_1, \cdots, adj_i\text{–}n_j, \cdots\}$

$\{pp_1\text{–}\ adj_1\text{–}n_1, pp_2\text{–}adj_1\text{–}n_1, \cdots\}$

Construct vocabulary based on **word frequency** from *Web Corpus*

atomic: white, fur, happy, ⋯

bigram: white fur, high tree, ⋯

trigram: with white fur, ⋯

Home > Blog >

All Our N-gram are Belong to You

August 3, 2006 · Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Google Research

[1] vocabulary from https://norvig.com/ngrams/

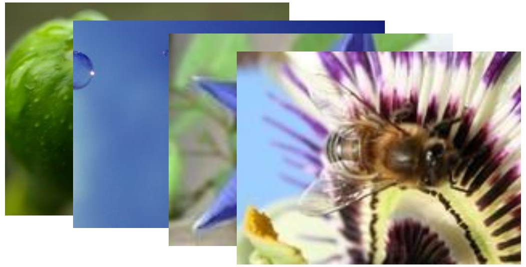
Concept Codebook Generation

Concept Vocabulary

$\{adj_1, adj_2, \cdots, n_1, n_2, \cdots\}$

$\{adj_1\text{–}n_1, \cdots, adj_i\text{–}n_j, \cdots\}$

$\{pp_1\text{–}\ adj_1\text{–}n_1, pp_2\text{–}adj_1\text{–}n_1, \cdots\}$



Vision-Language Models (VLMs)

class names

unlabeled images

select **class-related images** from large-scale unlabeled images using vision-language models

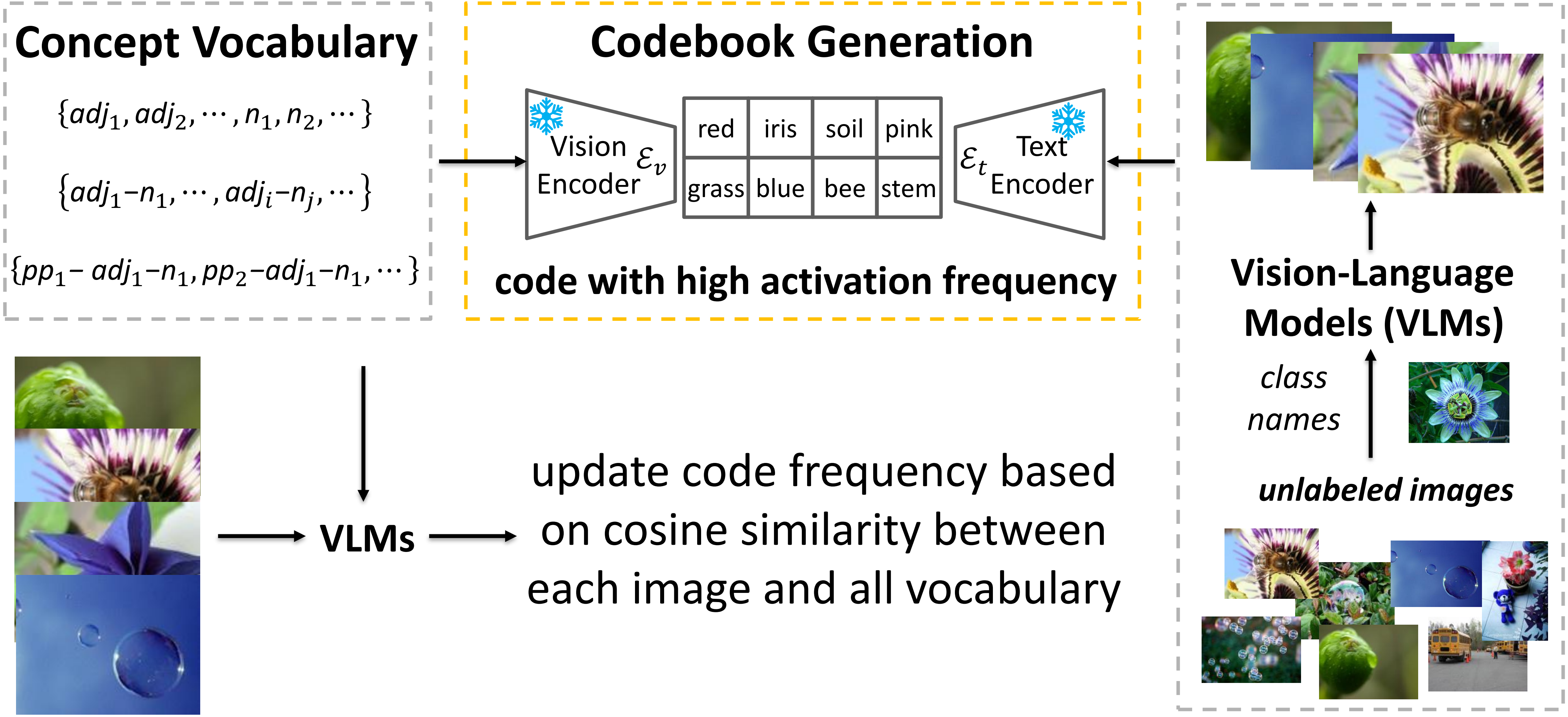
target class

Passionflower

+

images found by pretrained VLMs

Concept Codebook Generation



Vision-to-Concept Tokenizer

images from same class

↓

VLM Vision Encoder

↓

top-k closest concepts

↑

VLM Text Encoder

↑

codebook

red	iris	soil	pink
grass	blue	bee	stem

Vision-to-Concept Tokenizer

Discovering Concepts from Images

images from same class

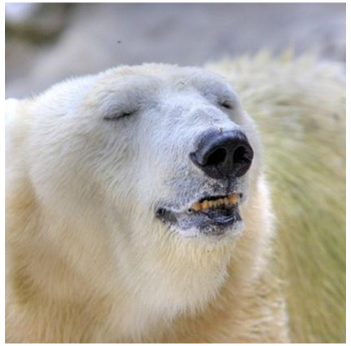
VLM Vision Encoder

top-5 closest concepts

VLM Text Encoder

codebook

red	iris	soil	pink
grass	blue	bee	stem

Class Name	Top-5 Concepts	Class Name	Top-5 Concepts
Acadian flycatcher 	• green upperpart • green breast • green • white breast • long bill	**Brambling** 	• black head • brown back • common bird • orange breast • has a black tail
American redstart 	• black head • orange wing • orange breast • gray underpart • black wing	**Polar bear** 	• white bear • white enclosure • white animal • cold zoo • white fur

Discovering Concepts from Images

images from same class

↓

VLM Vision Encoder

↓

top-5 closest concepts

↑

VLM Text Encoder

↑

codebook

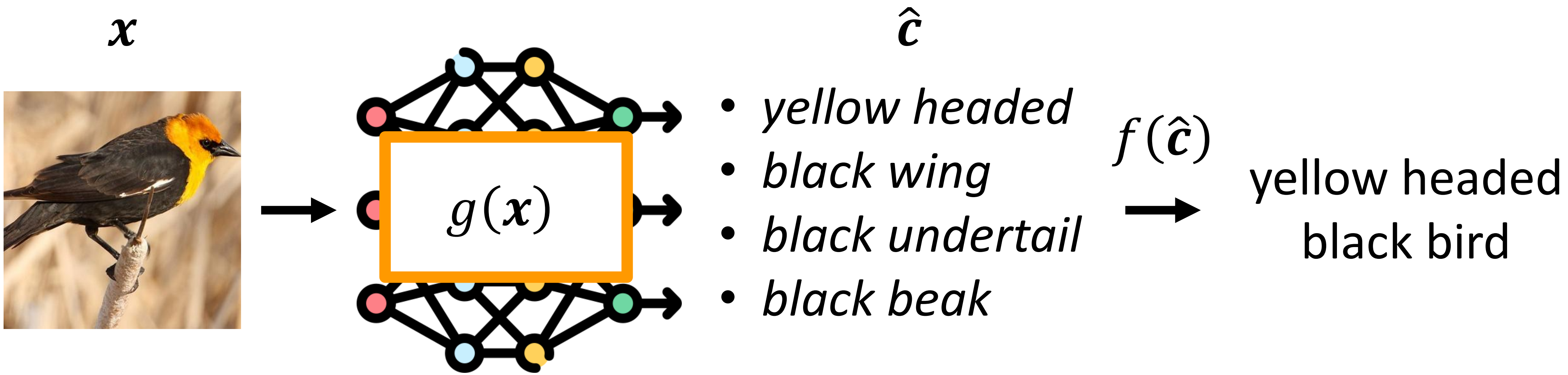
red	iris	soil	pink
grass	blue	bee	stem

Class Name	Top-5 Concepts	Class Name	Top-5 Concepts
Hammer 	• is a toolkit • brown handle • black handle • part of toolkit • metal	**School bus** 	• yellow bus • yellow vehicle • is a yellow bus • stop sign • ready student
Hot pot 	• hot bowl • hot dishes • red soup • hot soup • black pot	**Carousel** 	• single rider • carnival • happy spin • happy rider • young rider

Concept Bottleneck Models (CBMs)

CBMs decompose a DNN into two functions:

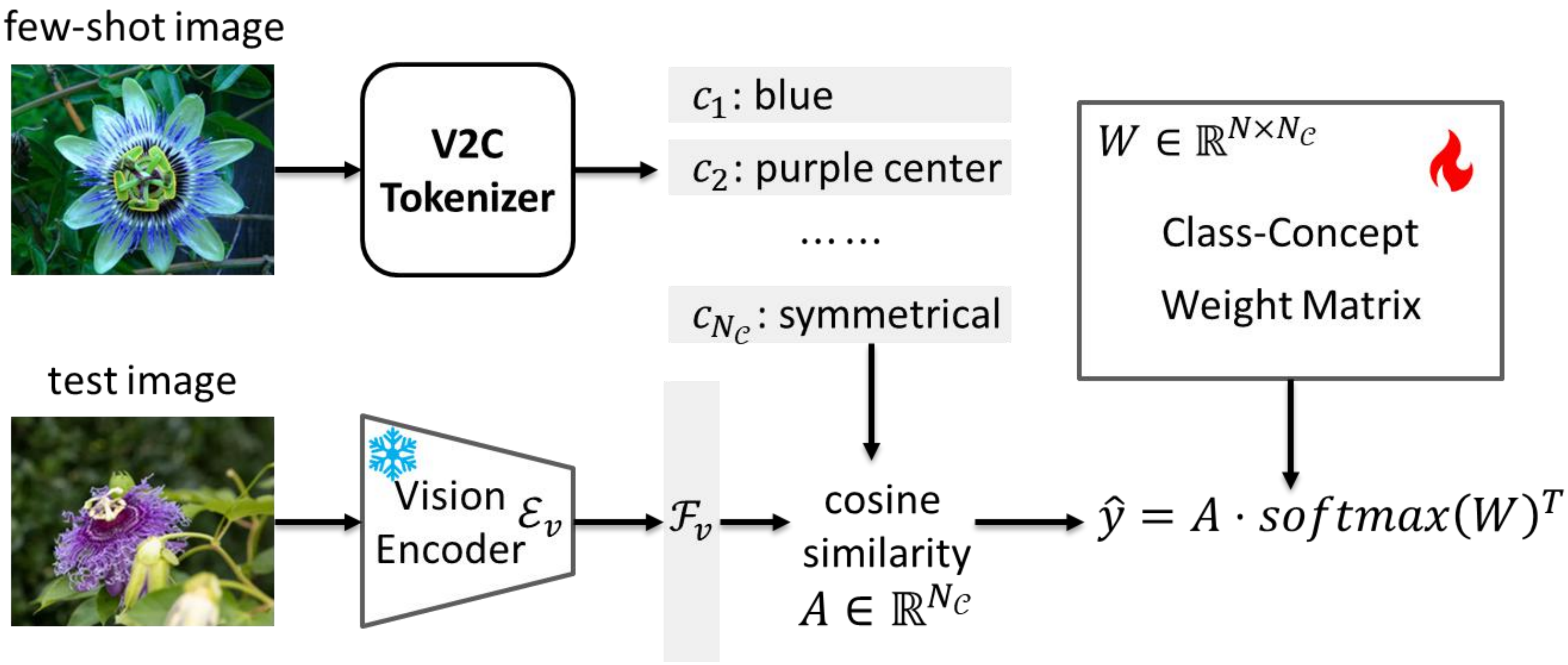
1. A *concept encoder* $g(\boldsymbol{x}) = \hat{\boldsymbol{c}}$ predicting **concepts from the input features**
2. A *label predictor* $f(\hat{\boldsymbol{c}}) = \hat{y}$ predicting **task labels from the concepts**



[1] Concept Bottleneck Models (ICML2020)

Building CBMs with V2C Tokenizer

With V2C Tokenizer, we can build CBMs without concept labels!



Building CBMs with V2C Tokenizer

Average classification accuracy (%) on 10 datasets

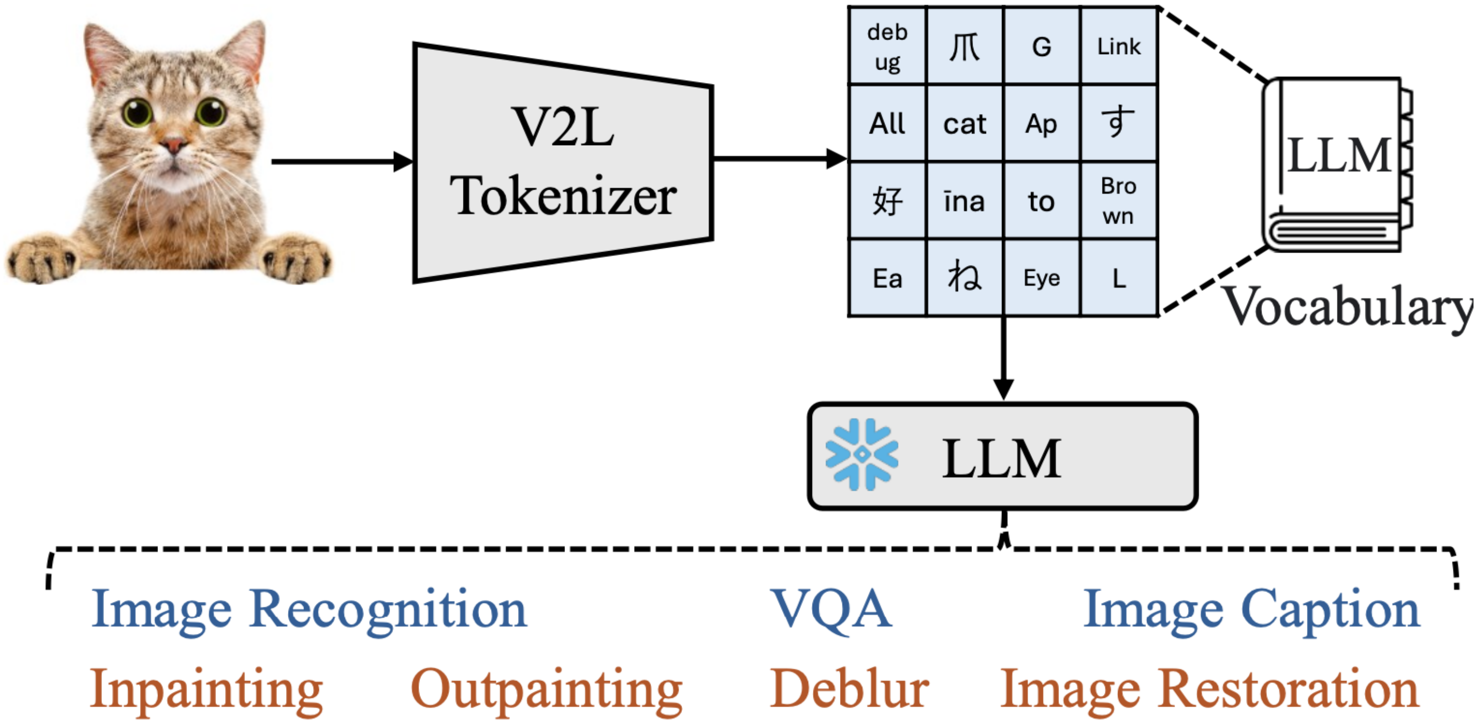
Method	1-shot	2-shots	4-shots	8-shots	16-shots	All
ViT-L/14	51.8	65.3	72.3	77.1	81.6	86.9
CBM	57.8	64.0	71.1	75.8	79.7	85.6

Scaling with the number of unlabeled images

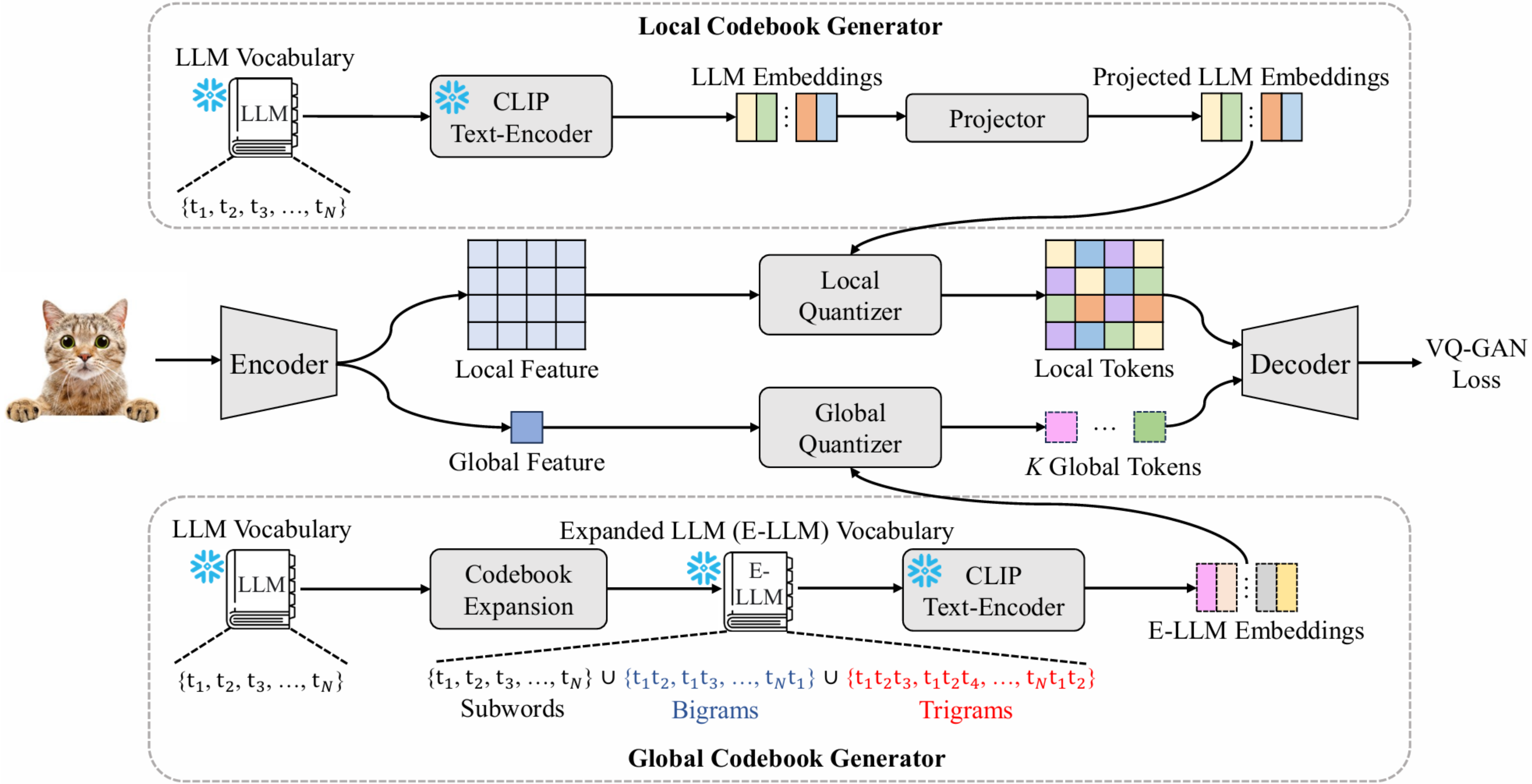
Tasks	1k	40k	80k	120k	160k	200k
bird	80.3	81.4	81.6	81.9	82.2	**83.0**
texture	73.1	76.3	76.8	77.4	77.6	**78.0**

Vision-to-Language Tokenizer

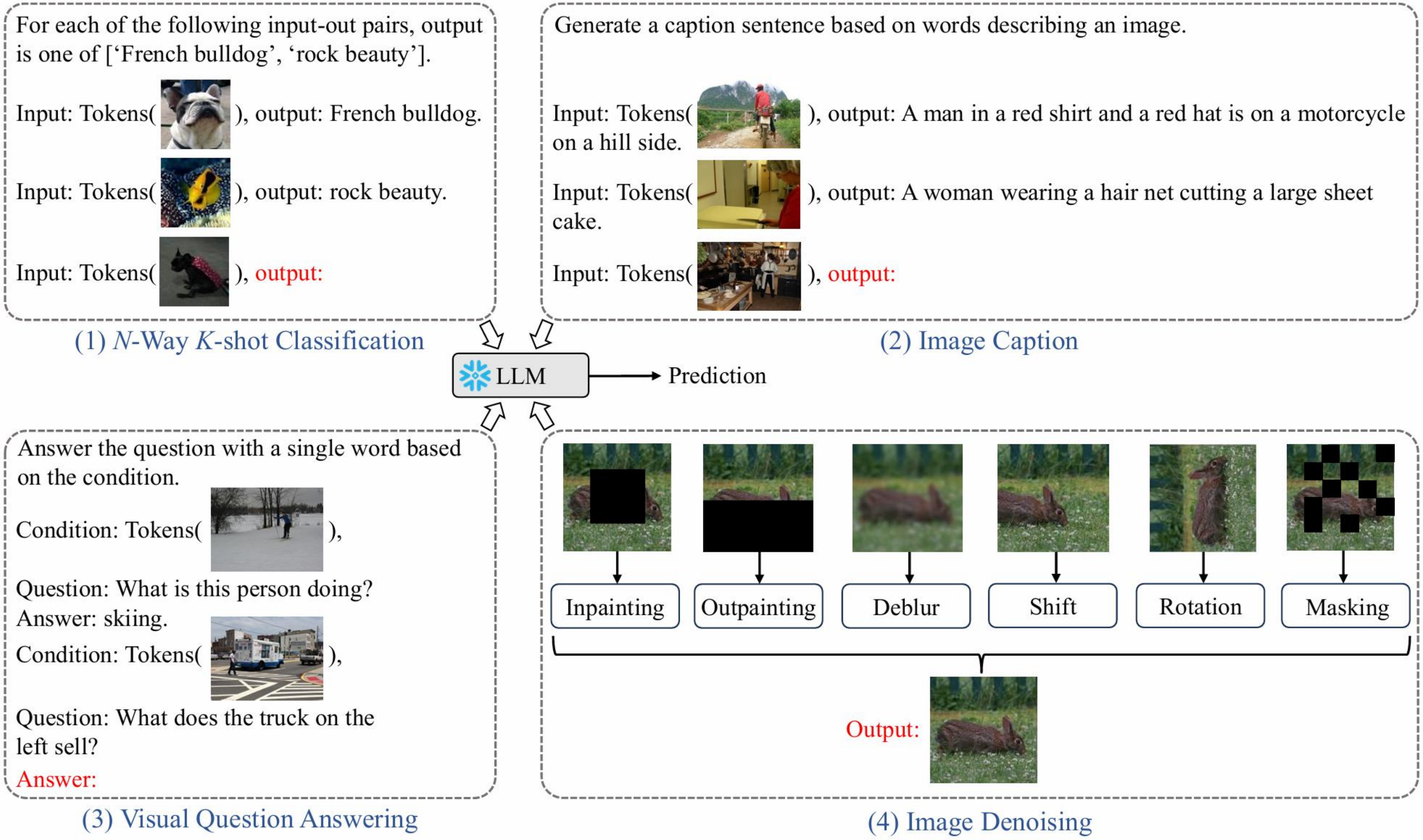
A Frozen Large Language Models (LLMs) can use the linguistic representation of images for directly visual understanding and generation!



Global and Local Codebook Generator



Visual Understanding and Generation



Visual Understanding and Generation

➢ Few-shot Classification

Method	#Tokens	Task Induction: / N-way K-shot: / #Repetitions:	2-1 0	✓ 2-1 0	✓ 2-3 0	✓ 2-5 0	✓ 2-1 1	✓ 2-1 3	✓ 2-1 5	Avg	5-1 0	✓ 5-1 0	✓ 5-3 0	✓ 5-5 0	✓ 5-1 1	✓ 5-1 3	✓ 5-1 5	Avg
Frozen [47]	-	-	1.7	33.7	66.0	66.0	63.0	65.0	63.7	51.3	0.9	14.5	34.7	33.8	33.8	33.3	32.8	26.3
LQAE [25]	256	GPT-3.5	1.5	35.2	68.2	69.8	68.5	68.7	65.9	54.0	1.0	15.7	35.9	36.5	31.9	36.4	45.9	29.0
SPAE [54]	5	GPT-3.5	5.3	77.2	84.4	86.0	79.4	77.2	77.1	69.5	-	-	-	-	-	-	-	-
SPAE [54]	5	PaLM-2 (340B)	32.2	84.0	88.5	88.4	85.1	83.6	82.4	77.7	23.6	64.2	68.0	69.9	63.4	62.0	60.2	58.8
Ours	5	LLaMA-2 (7B)	34.2	73.1	89.0	93.4	79.6	80.6	79.1	75.6	36.2	54.6	88.6	91.1	70.7	72.8	74.4	69.8
Ours	5	LLaMA-2 (13B)	44.4	77.9	91.9	94.4	81.5	82.8	82.0	79.3	**45.4**	69.6	89.9	91.3	75.8	75.7	77.2	75.0
Ours	5	LLaMA-2 (70B)	41.7	87.1	94.8	96.1	88.9	89.2	89.1	83.9	**45.4**	**81.5**	92.3	93.0	85.7	86.1	86.3	81.5
SPAE [54]	21	PaLM-2 (340B)	27.9	84.8	92.5	92.6	84.8	85.2	85.4	79.0	20.2	65.1	73.7	74.3	66.4	67.0	66.3	61.9
Ours	21	LLaMA-2 (7B)	36.5	76.3	91.2	95.3	84.0	84.4	83.7	78.8	37.1	44.8	91.8	94.0	73.9	82.2	85.3	72.7
Ours	21	LLaMA-2 (13B)	**48.7**	73.1	92.4	95.7	80.9	83.8	82.0	79.5	42.1	62.7	93.0	94.5	72.8	79.6	82.0	75.2
Ours	21	LLaMA-2 (70B)	46.5	**89.1**	**96.9**	**97.8**	**91.4**	**92.7**	**92.9**	**86.7**	45.0	79.7	**94.9**	**95.6**	**89.3**	**90.7**	**90.2**	**83.5**

➢ Caption & VQA

A dog is sitting in front of a computer.
A group of people in a kitchen.

A picture of a sign that says stop.
A bathroom with a bathtub and shower.

Q1: What food item is shown?
Pizza Burger
Q2: What country did this food originate from?
Italy Japan
Q3: What is the leafy substance?
Basil Lettuce

➢ Reconstruction & Generation

Input VQ-GAN LQAE SPAE **Ours**

Input VQ-GAN LQAE SPAE **Ours**

Input VQ-GAN LQAE SPAE **Ours**

Summary

1. **Language** as a **Bidirectional Explainable Interface** for vision tasks.

2. **V2C** and **V2L Tokenizer** to get linguistic representations of images.

3. **Efficient** to build and **Interpretable** and for use.

pretrained VLMs
unlabeled images
frozen LLMs

general & fine-grained
visual concepts

Lab Website | **Slides & Website** | **V2C Tokenizer** | **V2L Tokenizer**

MILab
Medical Intelligence Lab