TEXAS A&M UNIVERSITY

CSCE 689 – NLP For Science

Scientific VLMs: Miscellaneous

Hasnat Md Abdullah

February 20, 2025

Instructor: Dr. Yu Zhang (yuzhang@tamu.edu)

Agenda

- UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web
- BIOCLIP: A Vision Foundation Model for the Tree of Life
- MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Agenda

- UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web
- BIOCLIP: A Vision Foundation Model for the Tree of Life
- MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Urban Region Profiling

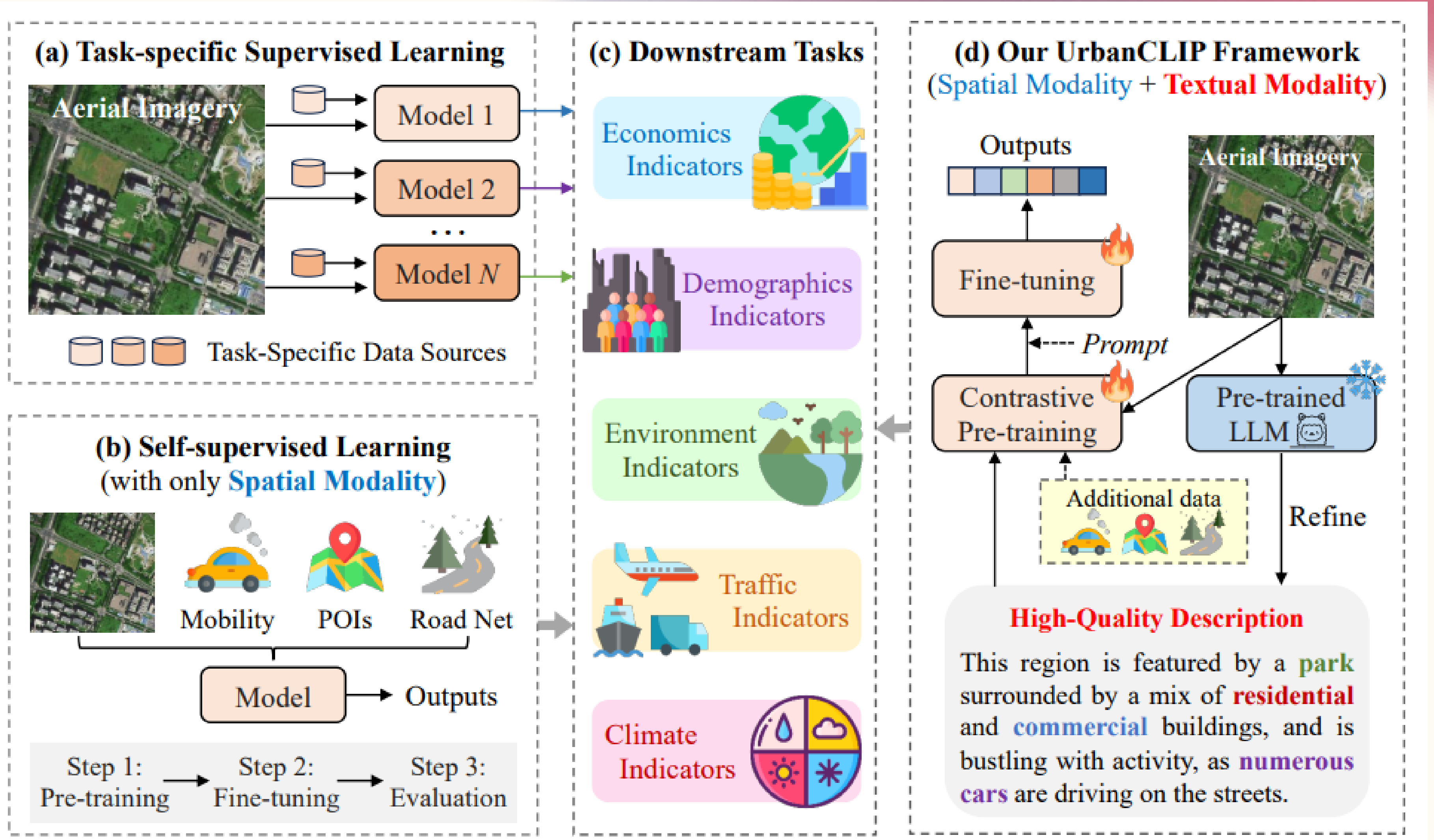
The process of representing and summarizing key features and attributes of urban areas.

Urban Indicators: Carbon Emission, Population, GDP, and Textual Description with more insights.

Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

: Regression / Predicting Scalar Values

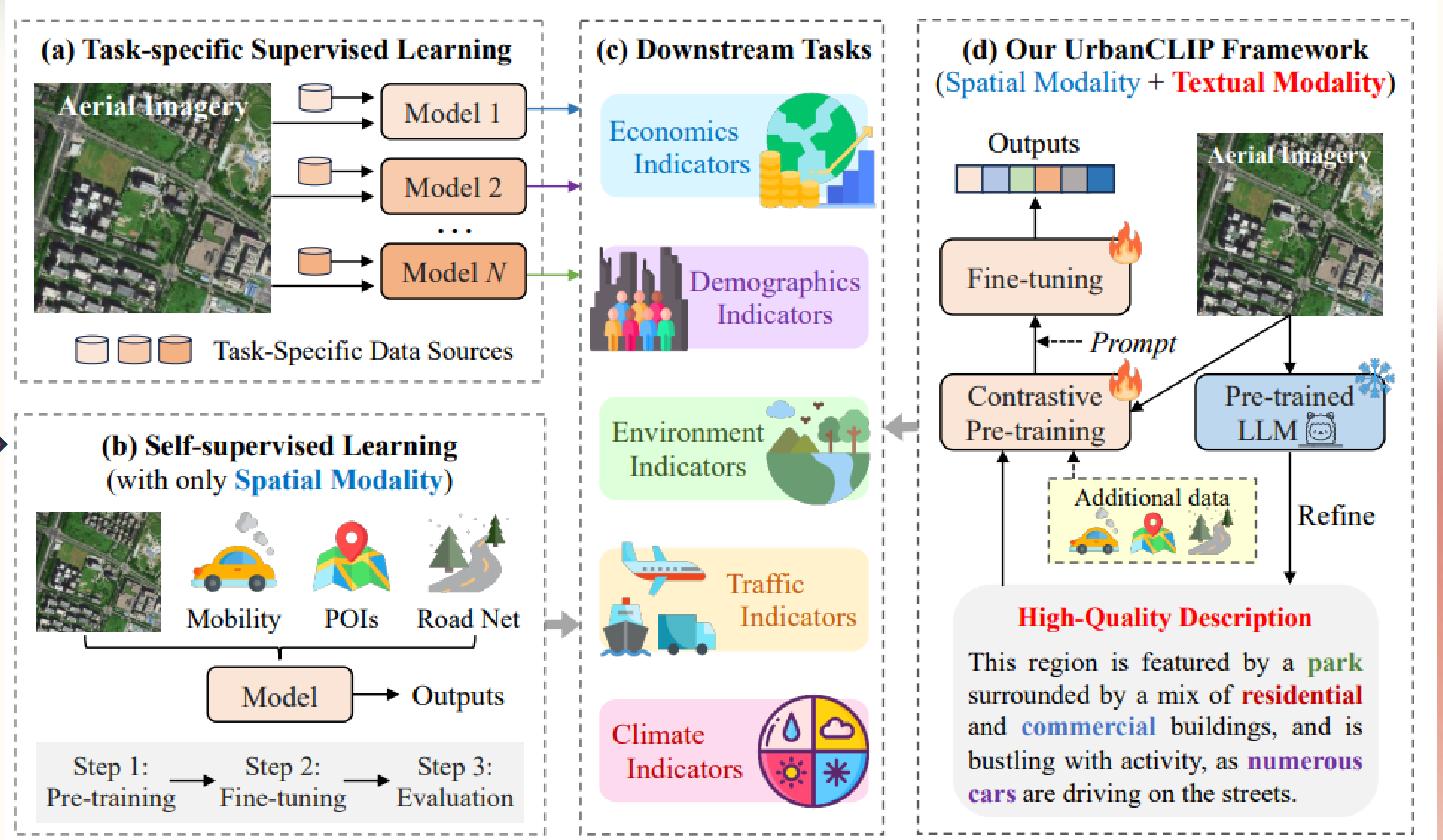
Urban Region Profiling



- Data Source: Satellite Imagery
- Requires considerable amount of labeled data, generalizability issue
- Tasks:
  - Poverty labels
  - Crop Yields
  - Population Land Cover
  - Commercial Activity

Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

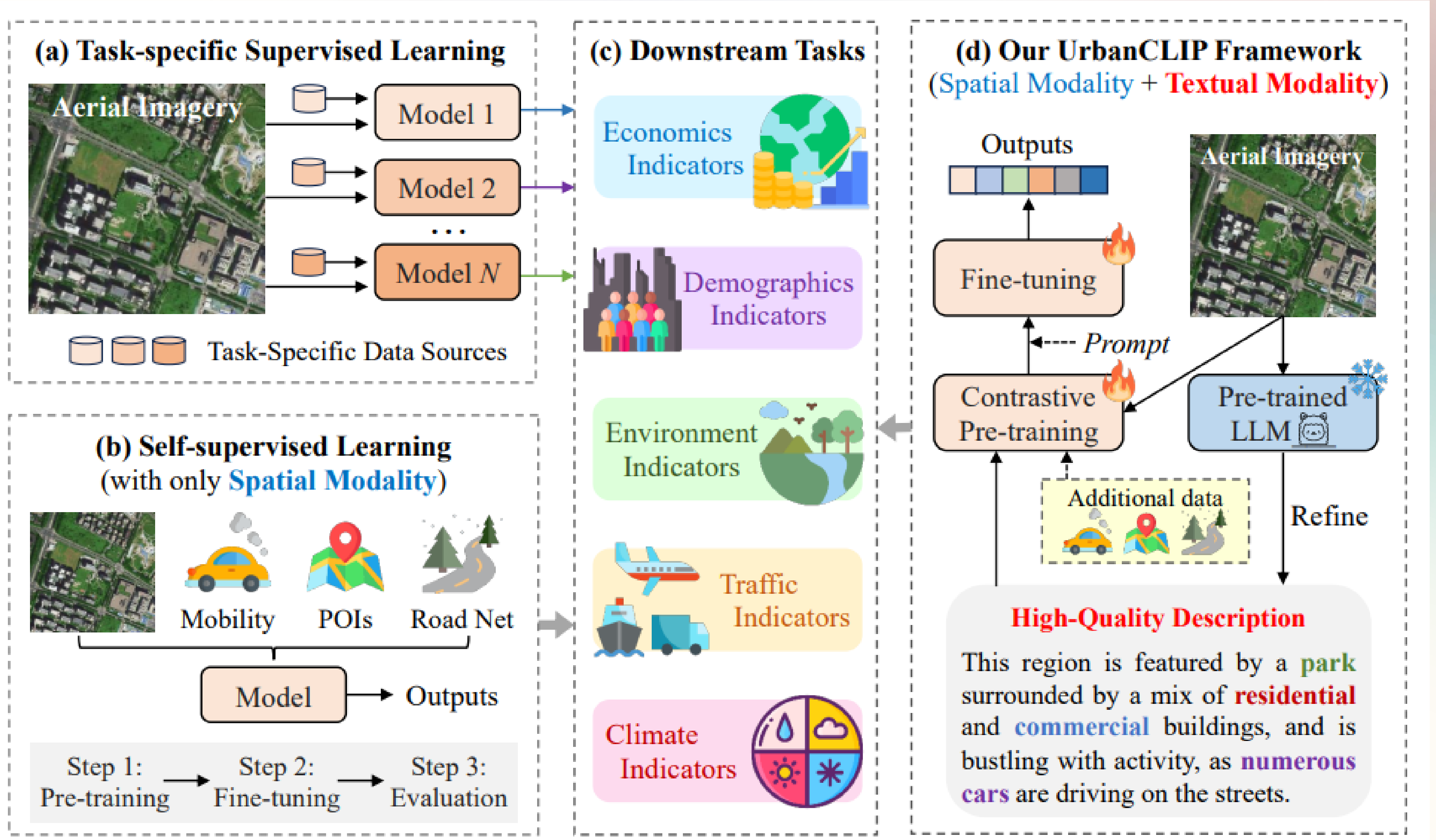
Urban Region Profiling



- Data Source: Satellite Imagery,
  - + Human inhabited areas & activities,
  - + Human Trajectory & Mobility
- Lacks Explainability in Natural Language

Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

Urban Region Profiling



- Data Source: Satellite Imagery,
  - Human inhabitated areas & activities,
  - Human Trajectory & Mobility
  - + Textual Modality
- RQs:
  - Can textual data complement Satellite Imagery? If so, in what ways?

Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

UrbanCLIP

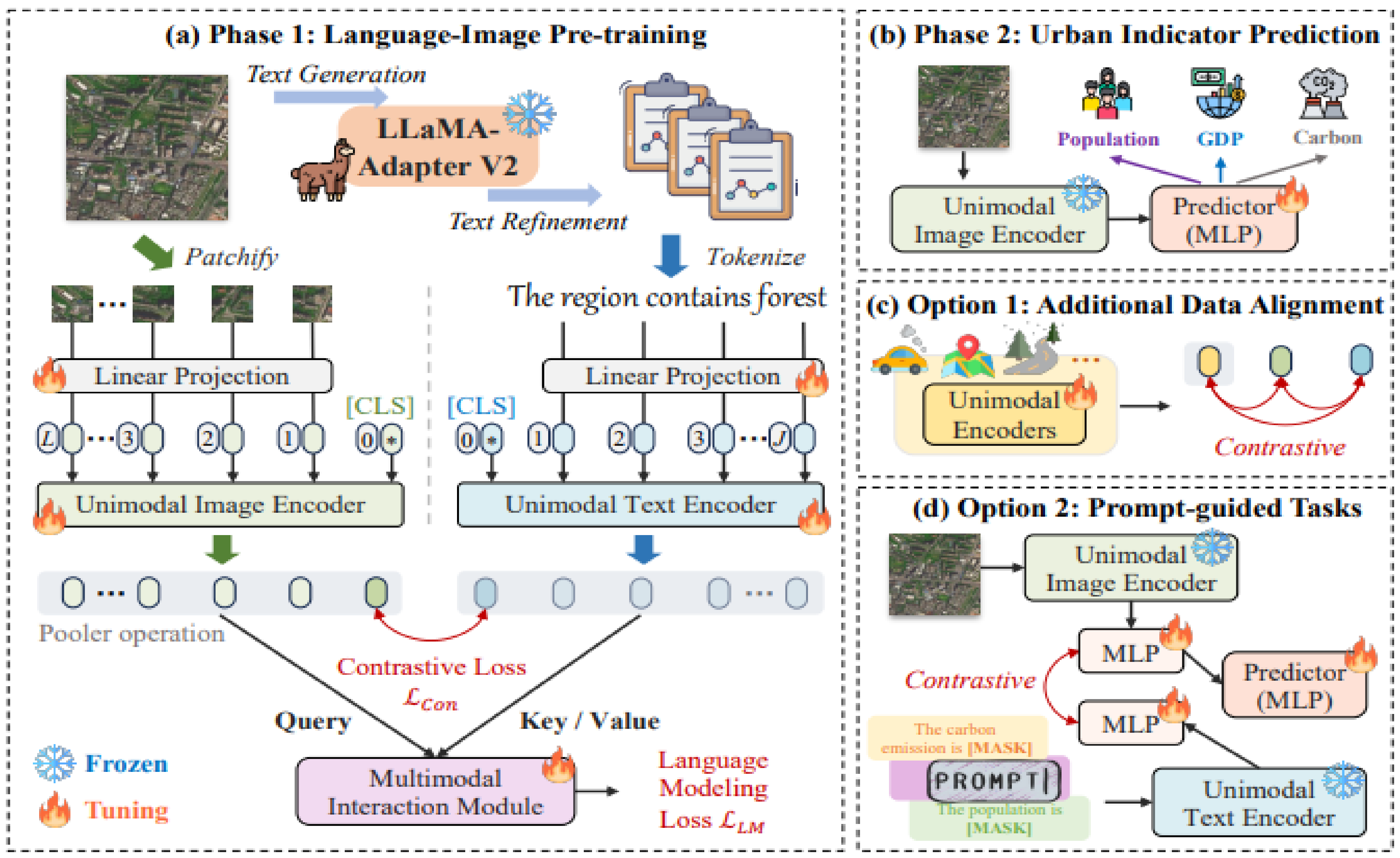


Figure 2: Overall framework of the proposed UrbanCLIP.

- Encoder-Decoder Architecture
- Two Unimodal Encoders: Image (ViT) & Text (Decoder-only)
- Contrastive Loss between Vision and Text Modalities
- Decoder: Cross Attention between Image and Text Representation with Language Modeling Loss

Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

UrbanCLIP

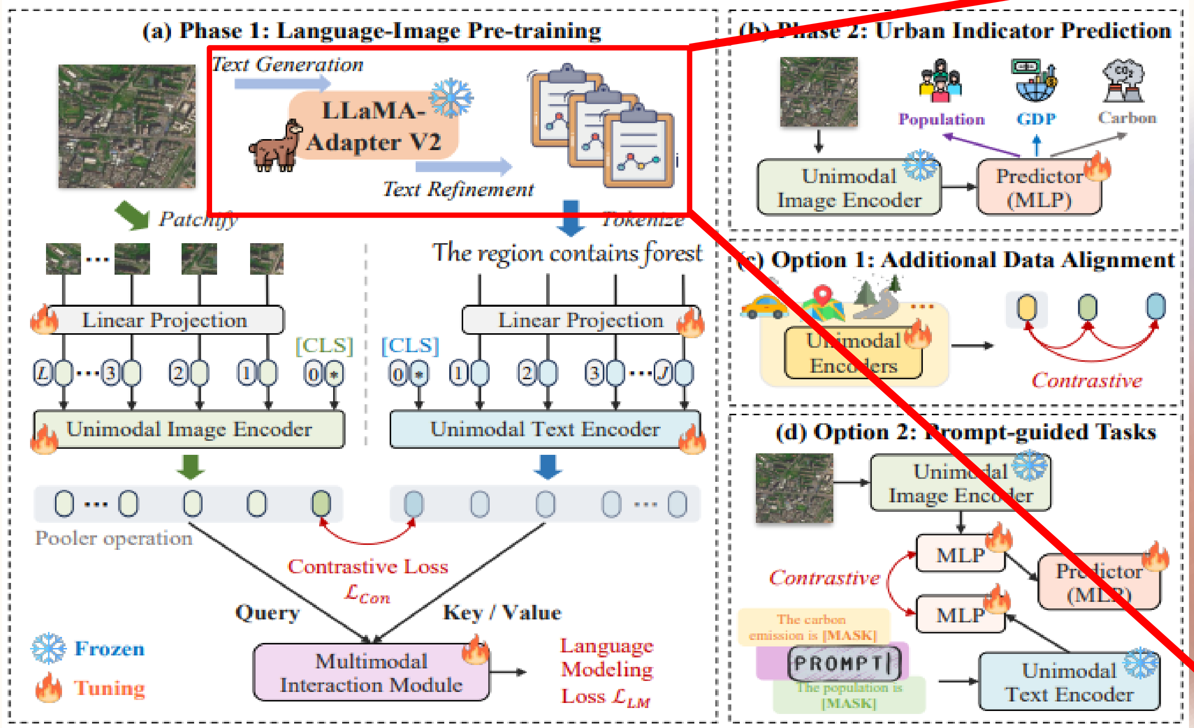


Figure 2: Overall framework of the proposed UrbanCLIP.

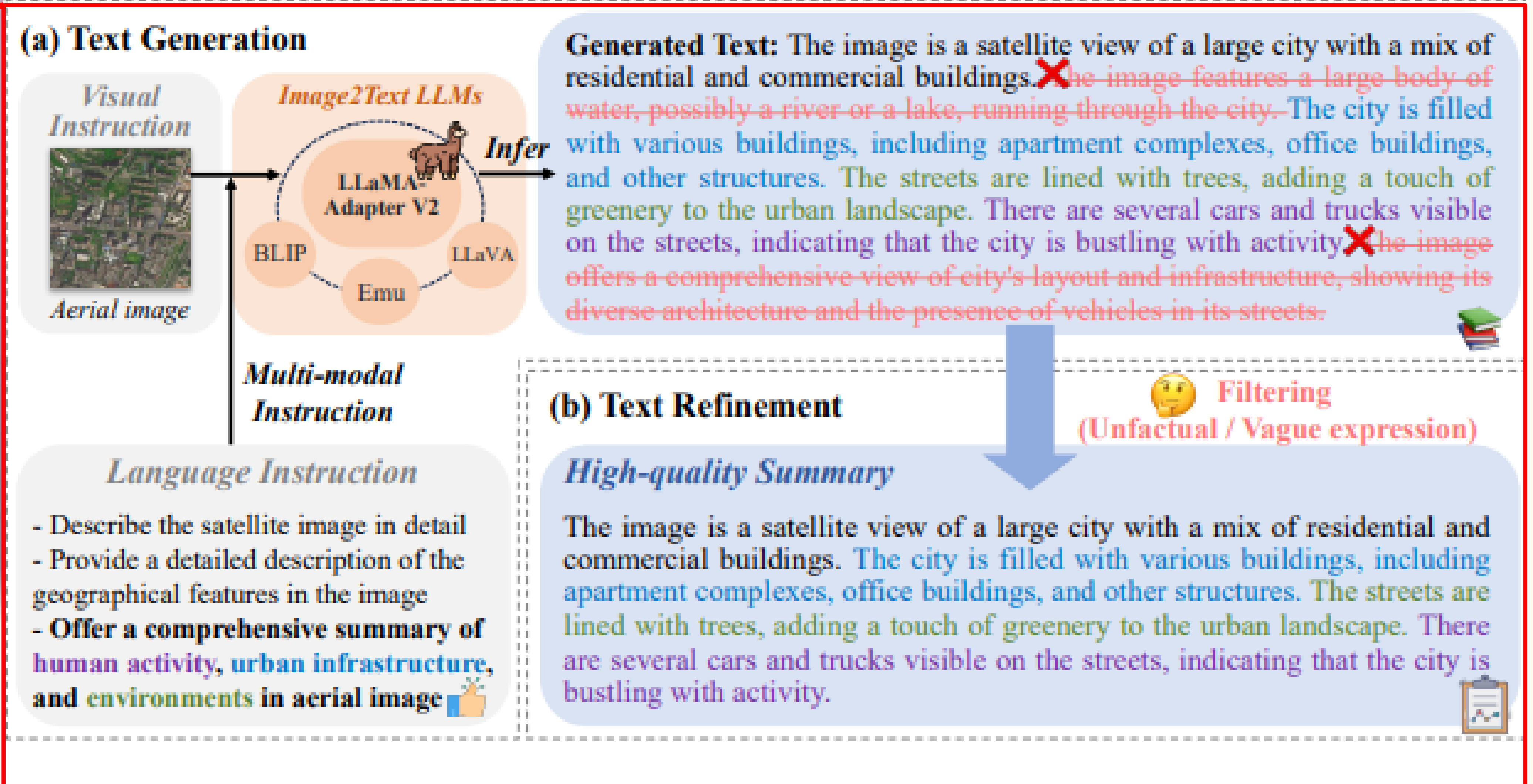


Figure 3: Text generation and refinement.

Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

UrbanCLIP

- **Cross Modal Representation Learning from satellite images and text descriptions**

Modality Alignment Task

$$\mathcal{L}_{Con} = \mathcal{L}_{Con}^{\text{Image}\rightarrow\text{Text}} + \mathcal{L}_{Con}^{\text{Text}\rightarrow\text{Image}}$$

$$= -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{e}_{pool}^{I}, \boldsymbol{e}^{T}\right)\right)}{\sum_{i=1}^{m} \exp\left(\text{sim}\left(\boldsymbol{e}_{pool}^{I}, \boldsymbol{e}_{i}^{T}\right)\right)} - \log \frac{\exp\left(\text{sim}\left(\boldsymbol{e}^{T}, \boldsymbol{e}_{pool}^{I}\right)\right)}{\sum_{i=1}^{m} \exp\left(\text{sim}\left(\boldsymbol{e}^{T}, \boldsymbol{e}_{pool_i}^{I}\right)\right)},$$

where sim$(\cdot)$ is inner product; $\mathcal{L}_{con}^{\text{Image}\rightarrow\text{Text}}$ and $\mathcal{L}_{con}^{\text{Text}\rightarrow\text{Image}}$ are image-to-text and text-to-image contrastive losses, repsectively.

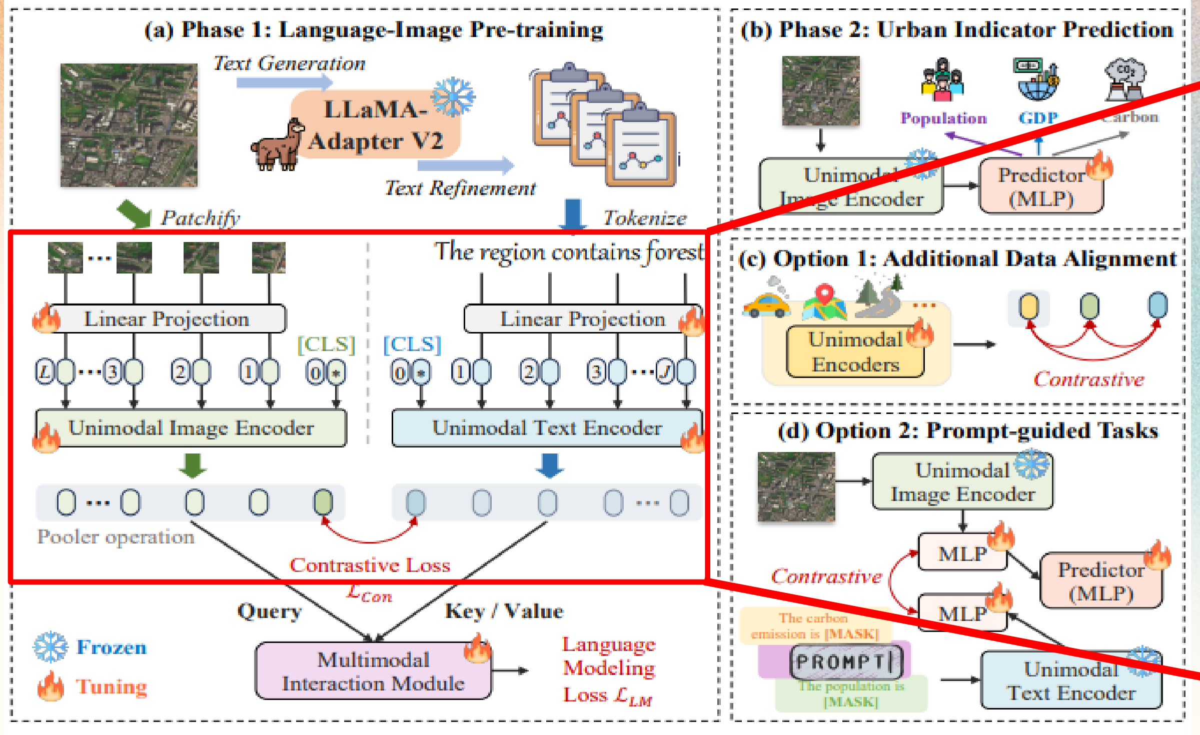
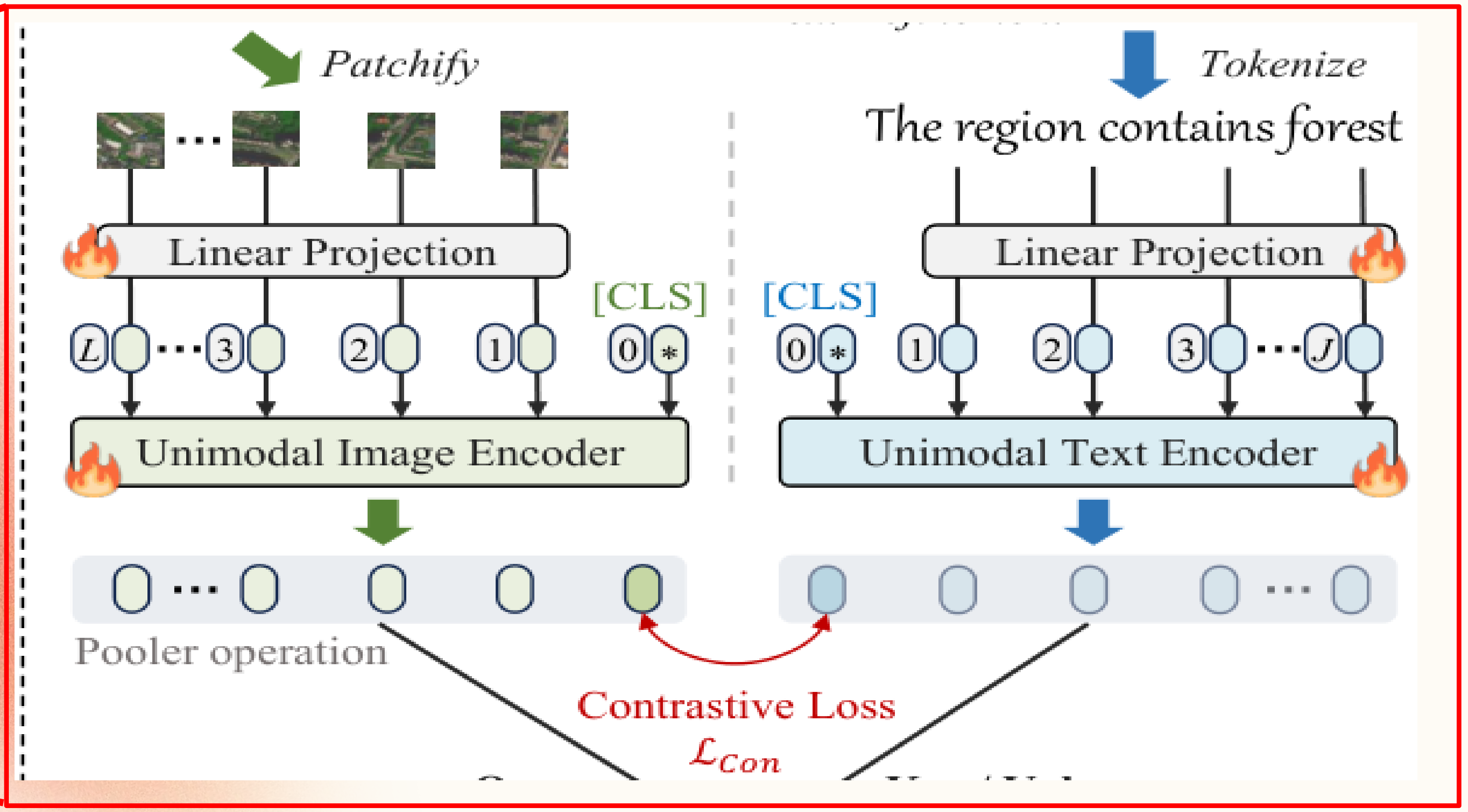


Figure 2: Overall framework of the proposed UrbanCLIP.



Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

UrbanCLIP

- **Cross Modal Representation Learning from satellite images and text descriptions**

Modality Interaction Task

learn to maximize the conditional likelihood of the paired text T:
$\mathcal{L}_{\mathrm{LM}} = -\sum_{l=1}^{L} \log P_\theta (T_l \mid T_{<l}, I)$.

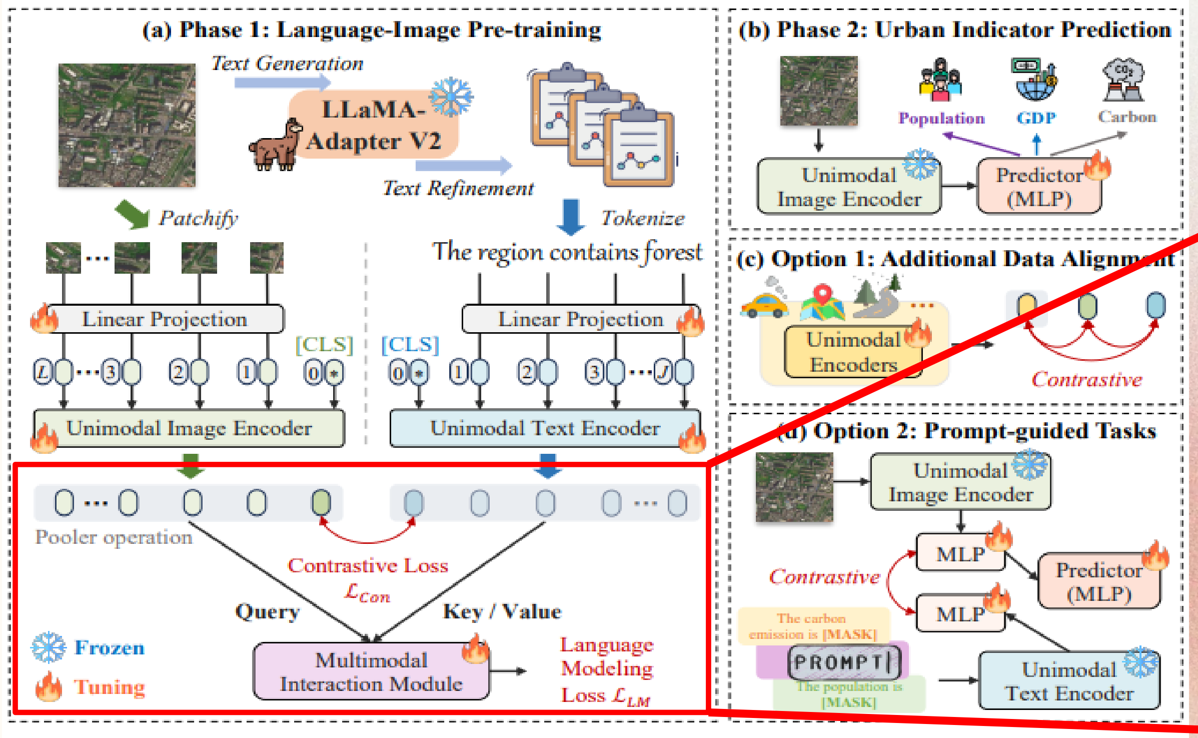
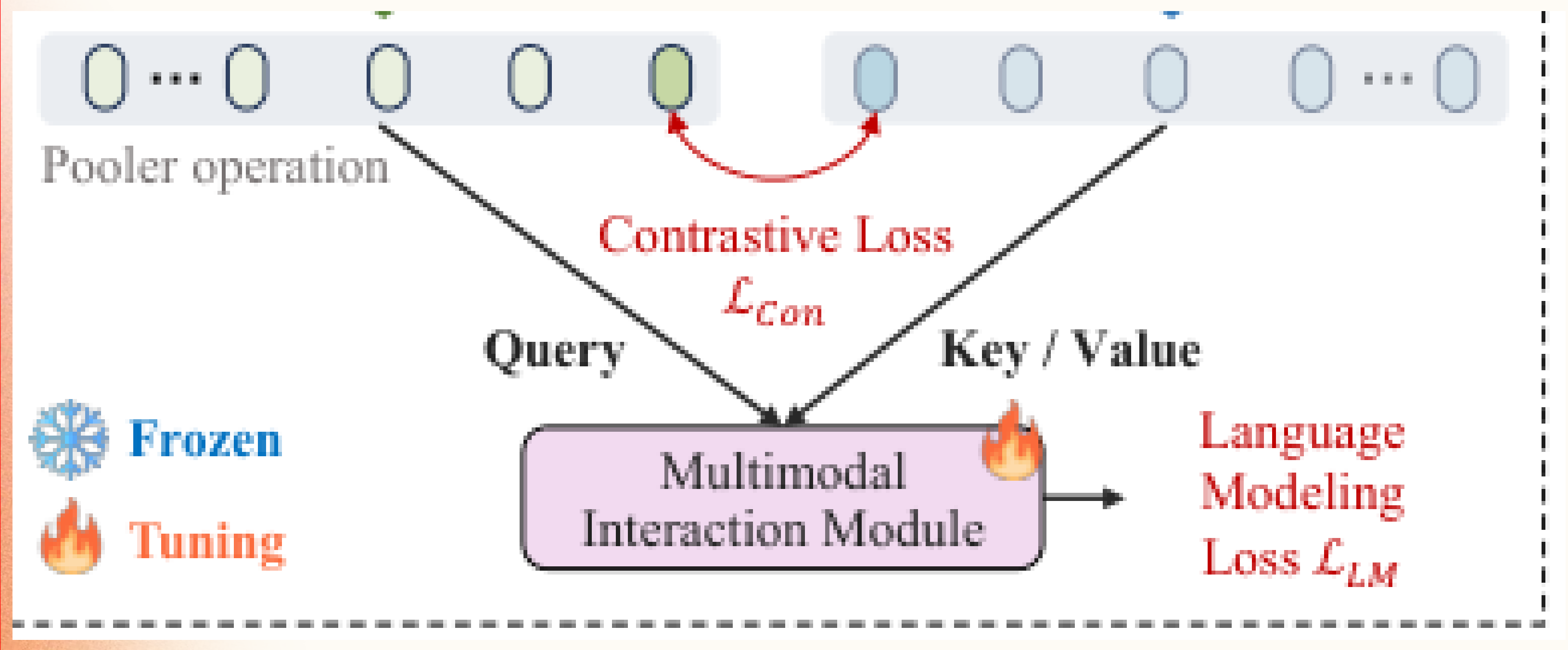


Figure 2: Overall framework of the proposed UrbanCLIP.



Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

UrbanCLIP

- **Urban Indicator Prediction**

Pre-training Objective

$$\mathcal{L}_{\text{Total}} = \lambda_{\text{Con}} \cdot \mathcal{L}_{\text{Con}} + \lambda_{\text{LM}} \cdot \mathcal{L}_{\text{LM}},$$

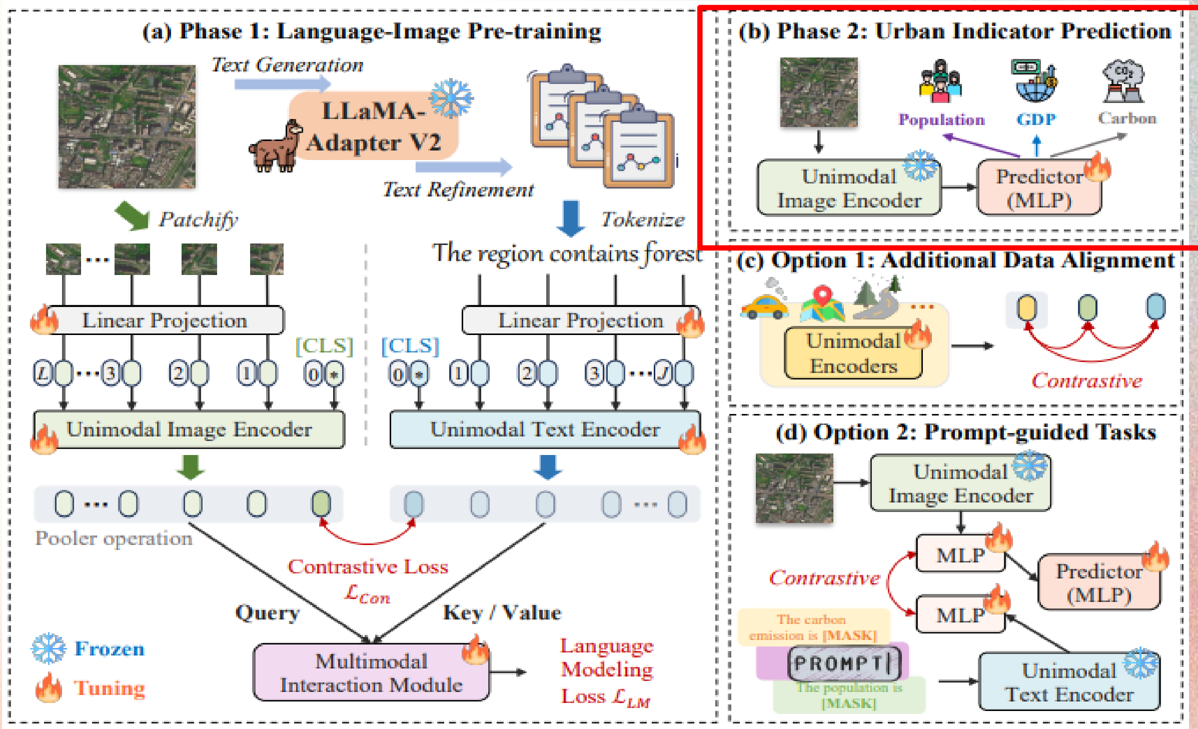
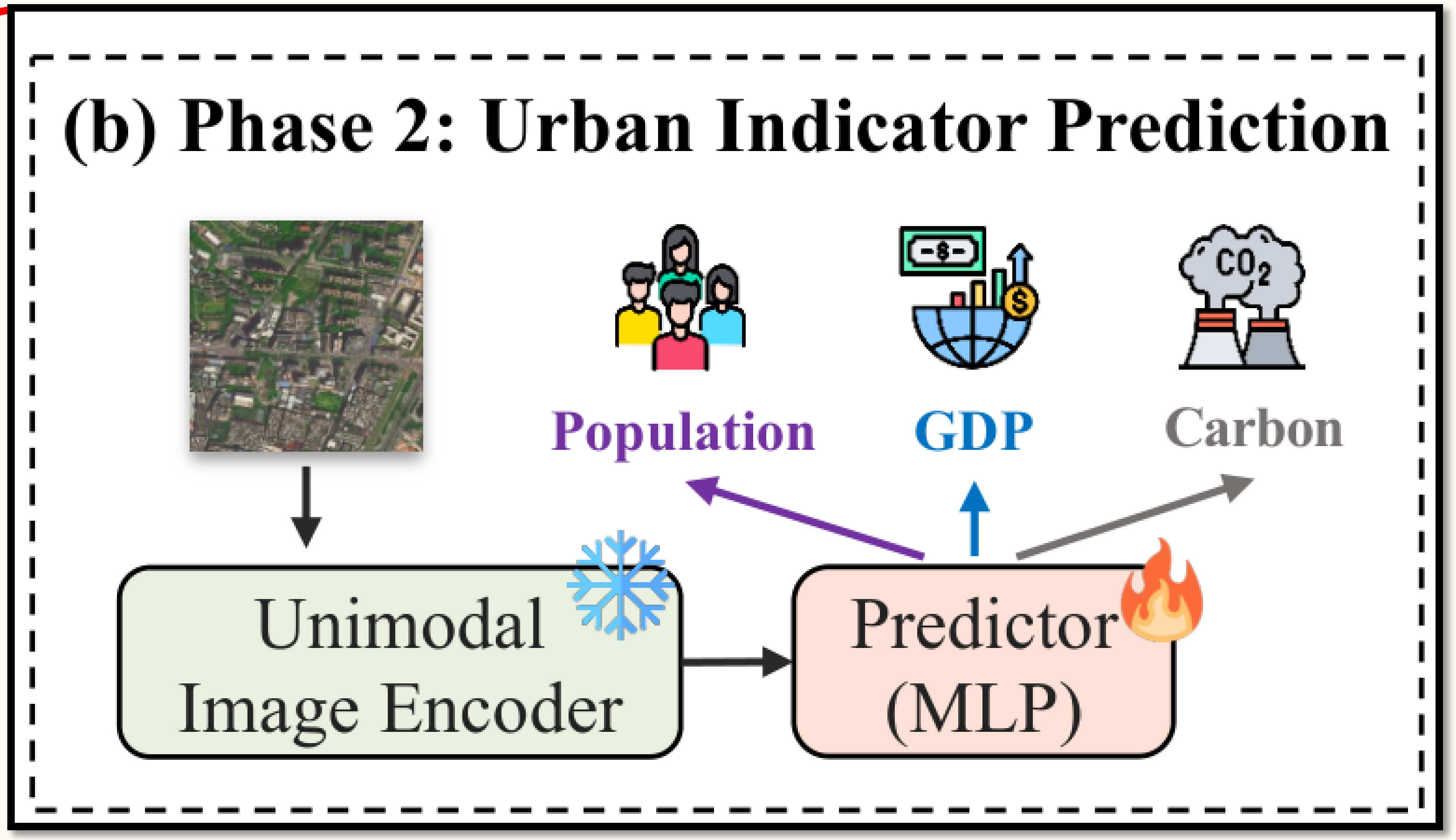


Figure 2: Overall framework of the proposed UrbanCLIP.



Modality Alignment Task

Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

UrbanCLIP

- Multiple modality alignment and integration

Multimodality contrastive learning

- Satellite Images
- Text Description
- POI: parks, roads

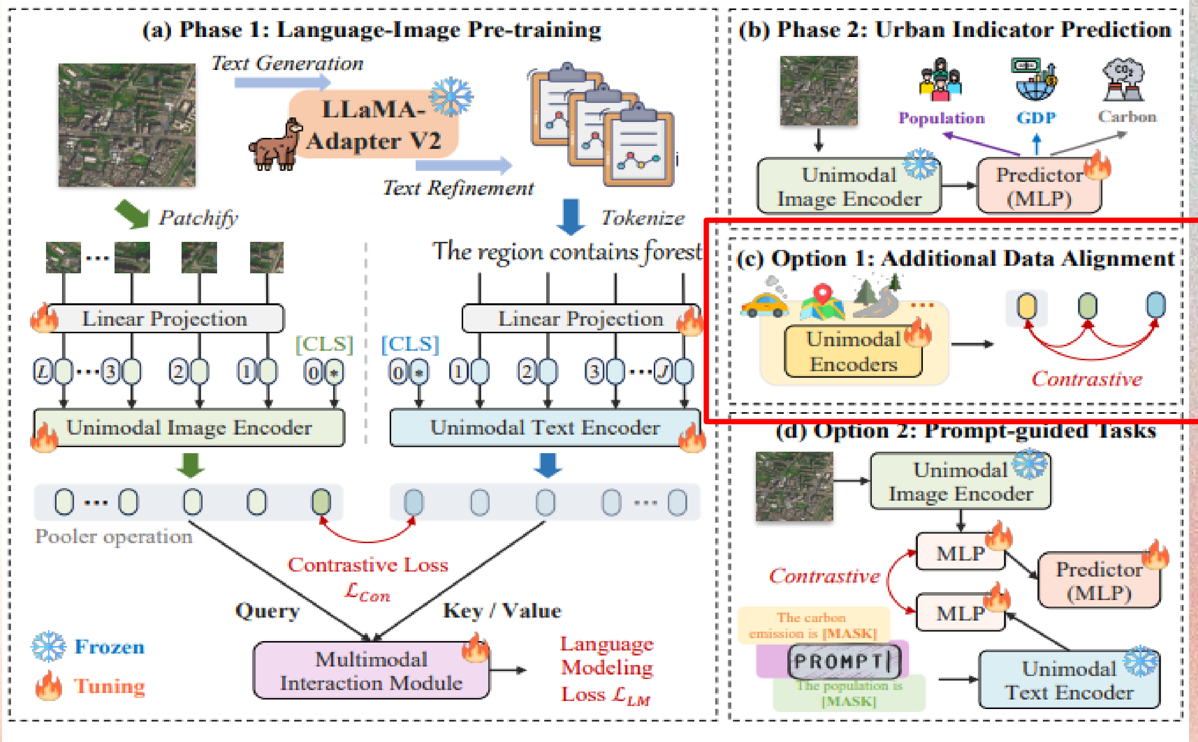
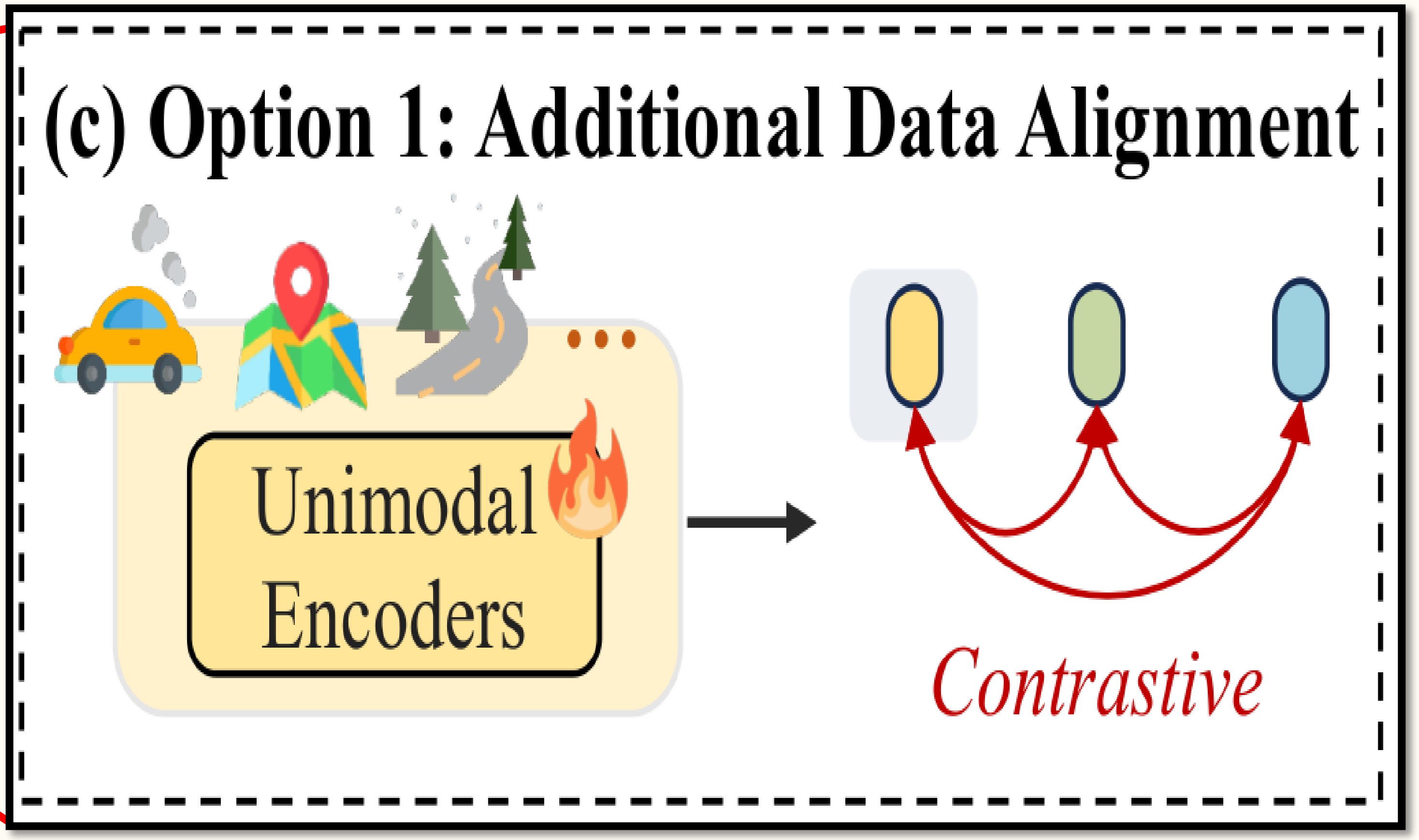


Figure 2: Overall framework of the proposed UrbanCLIP.



Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

UrbanCLIP

- **Multiple modality alignment and integration**

Instruction Tuning the pre-trained Unimodal Image-Text Encoders for regression tasks

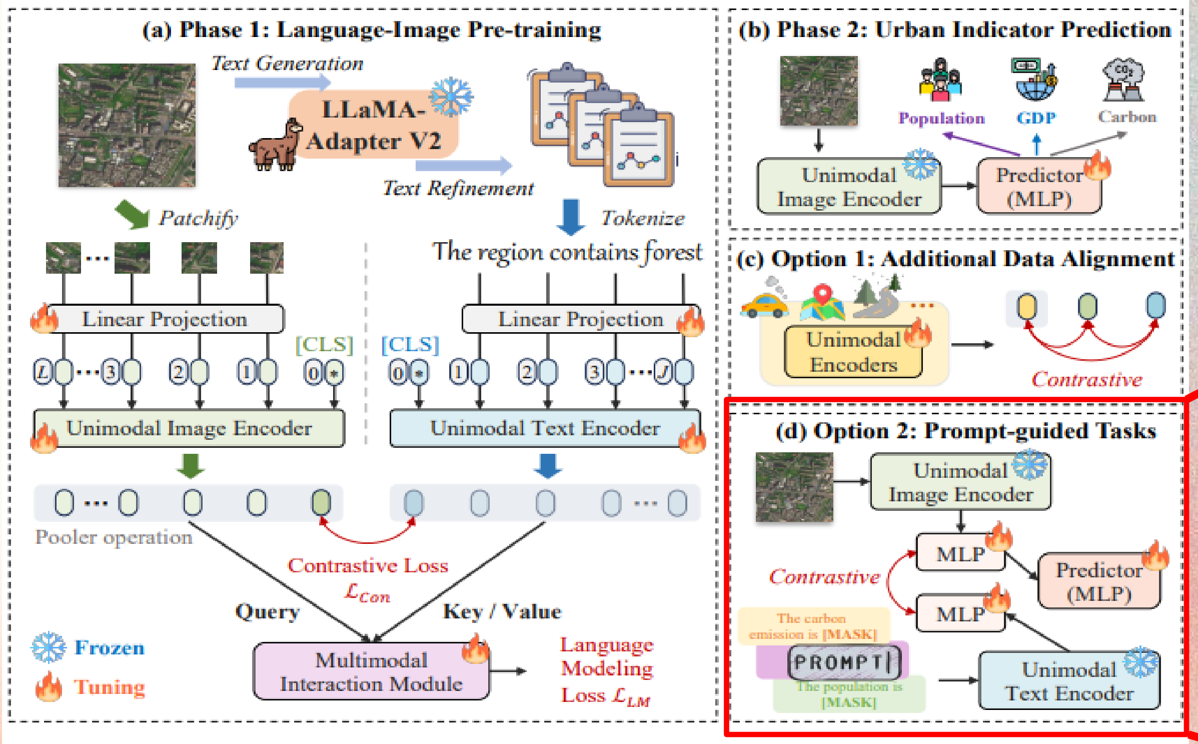
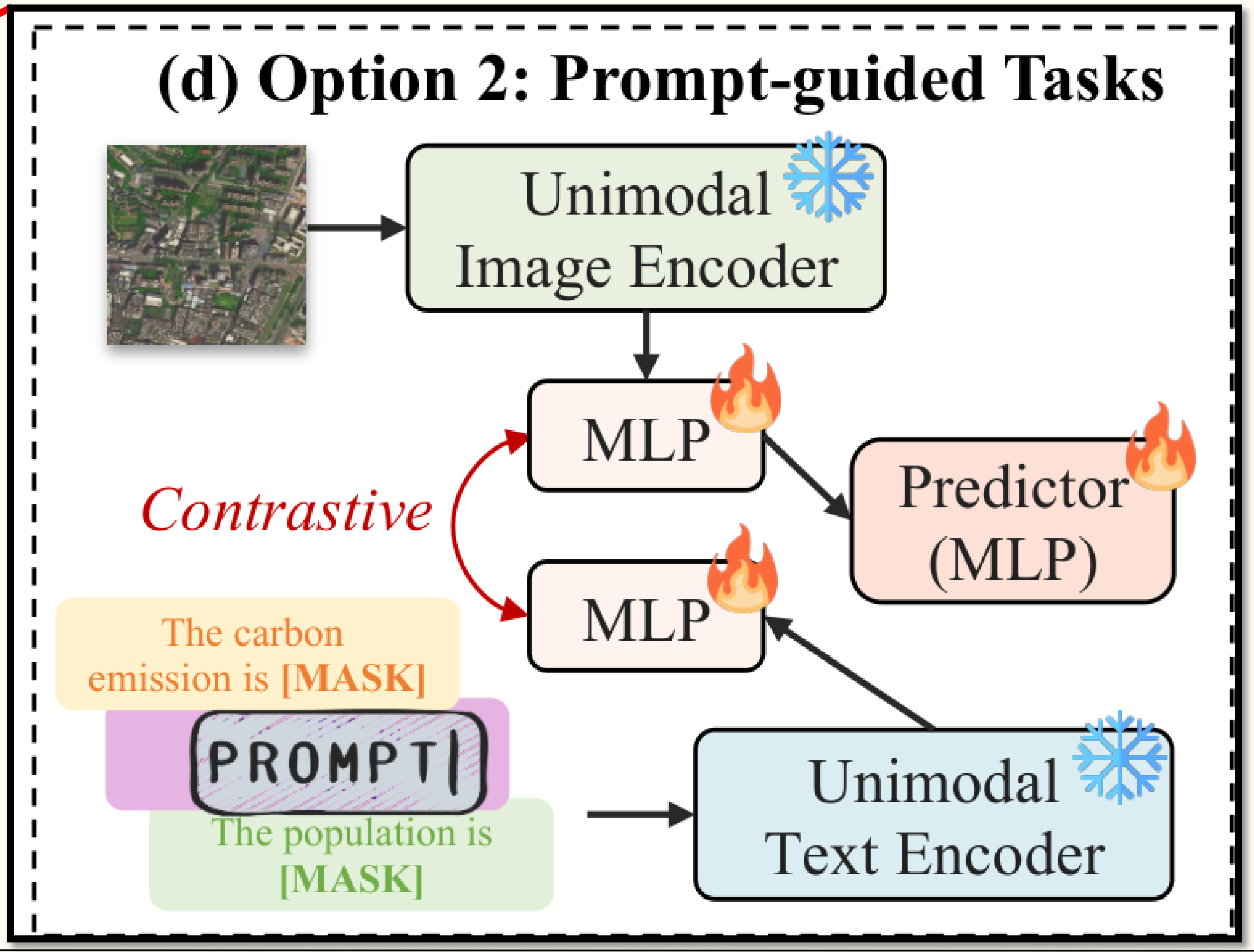


Figure 2: Overall framework of the proposed UrbanCLIP.



Source: UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web

Dataset & Metrics

- Satellite Imagery : Baidu Map API (256x256; 13 meters per pixel; 1 km^2)
- Textual Description : Generated by LLaMA-Adapter V2
- Urban Indicator : Population [WorldPop], GDP [2], Carbon Emission [ODIAC]
  - Cities: Beijing, Shanghai, zuangzhou, and Shenzhen

Table 1: Dataset statistics.

Dataset	Coverage		#Satellite Image	#Location Description
	Bottom-left	Top-right		
Beijing	39.75°N, 116.03°E	40.15°N, 116.79°E	4,592	20,642
Shanghai	30.98°N, 121.10°E	31.51°N, 121.80°E	5,244	23,455
Guangzhou	22.94°N, 113.10°E	23.40°N, 113.68°E	3,402	15,539
Shenzhen	22.45°N, 113.75°E	22.84°N, 114.62°E	4,324	18,113

Metrics

- **Prediction performance:**
  - Coefficient of determination (R^2 ↑),
  - Rooted mean squared error (RMSE ↓),
  - Mean absolute error (MAE ↓)

Source:
UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web
[1] WorldPop, open data for spatial demography. Scientific data 4, 1 (2017), 1–4
[2]Forecasting China's GDP at the pixel level using nighttime lights time series and population images. GIScience & Remote Sensing z4, 3 (2017)
[3] The Open-source Data Inventory for Anthropogenic CO 2, version 2016 (ODIAC2016), The Earth System Science Data.

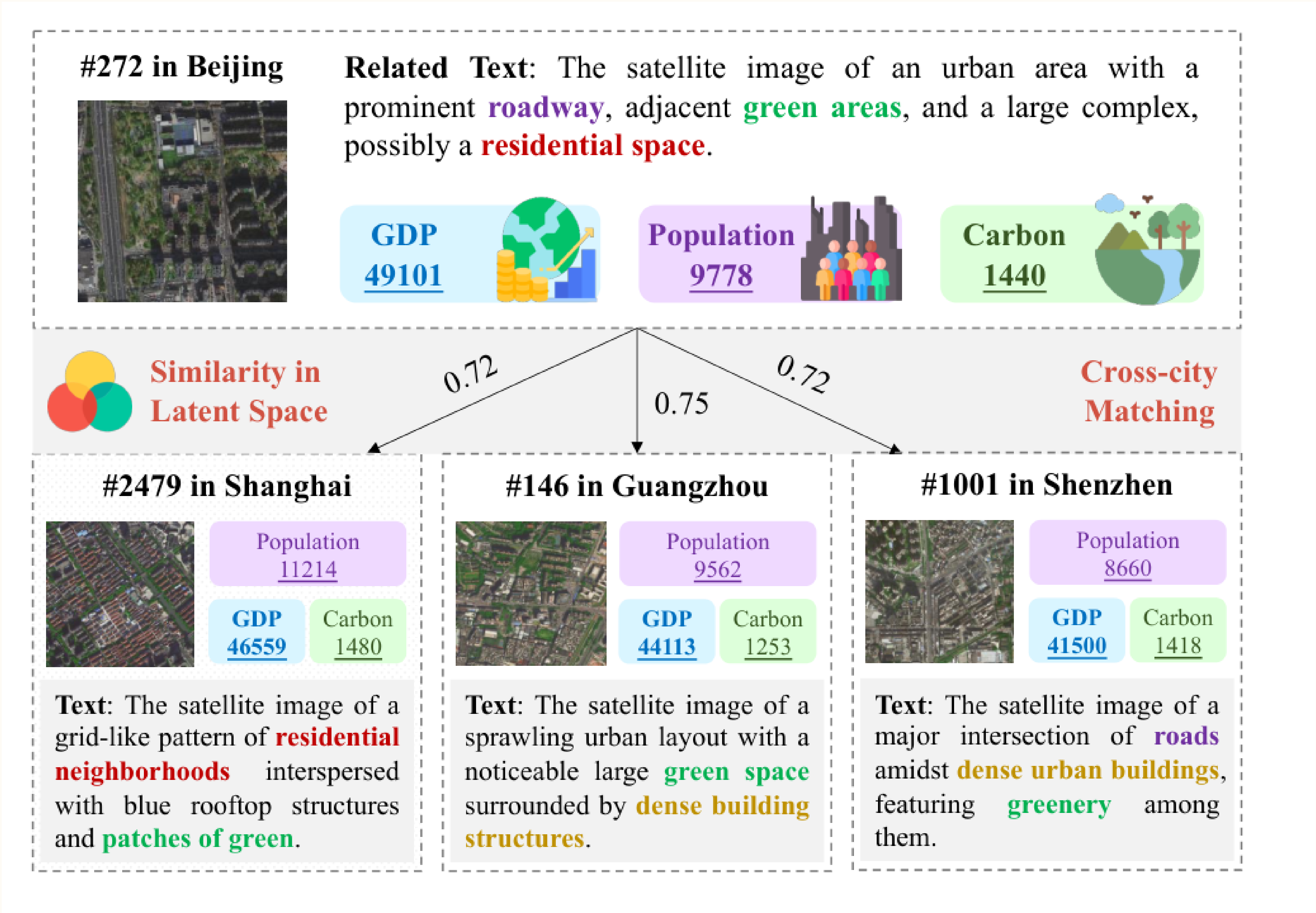
Table 2: Urban indicators prediction results in four datasets. The best results are in bold, and the second-best results are underlined. The last row indicates the relative improvement in percentage.

Dataset	Beijing									Shanghai								
Model	Carbon			Population			GDP			Carbon			Population			GDP		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Autoencoder	0.099	0.936	0.621	0.094	0.988	0.712	0.115	1.603	0.858	0.119	0.968	0.617	0.101	0.967	0.800	0.077	1.782	0.900
PCA	0.124	0.921	0.598	0.109	0.968	0.700	0.102	1.696	0.882	0.123	0.952	0.588	0.131	0.958	0.802	0.103	1.702	0.890
ResNet-18	0.393	0.599	0.411	0.202	0.858	0.680	0.203	1.280	0.758	0.451	0.512	0.460	0.233	0.852	0.692	0.217	1.297	0.777
Tile2Vec	0.599	0.512	0.468	0.204	0.813	0.635	0.182	1.356	0.792	0.572	0.462	0.390	0.249	0.801	0.620	0.169	1.380	0.806
READ	0.284	0.678	0.545	0.301	0.813	0.632	0.208	1.281	0.759	0.399	0.588	0.527	0.322	0.801	0.600	0.229	1.296	0.773
PG-SimCLR	0.613	0.489	0.360	0.362	0.799	0.599	0.317	1.114	0.688	0.597	0.442	0.356	0.410	0.790	0.584	0.319	1.181	0.725
UrbanCLIP	**0.662**	**0.327**	**0.302**	**0.407**	**0.788**	**0.589**	**0.319**	**1.102**	**0.684**	**0.652**	**0.331**	**0.300**	**0.429**	**0.778**	**0.578**	**0.320**	**1.119**	**0.702**
Improvement	8.11%	33.22%	16.00%	12.35%	1.39%	1.69%	0.73%	1.04%	0.62%	9.28%	25.12%	15.73%	4.59%	1.54%	1.06%	0.38%	5.28%	3.06%

Dataset	Guangzhou									Shenzhen								
Model	Carbon			Population			GDP			Carbon			Population			GDP		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Autoencoder	0.068	0.992	0.736	0.163	0.991	0.833	0.122	1.753	0.887	0.099	0.970	0.704	0.122	0.989	0.817	0.093	1.901	0.899
PCA	0.087	0.989	0.688	0.179	0.989	0.812	0.134	1.693	0.862	0.133	0.956	0.677	0.134	0.977	0.810	0.087	1.902	0.899
ResNet-18	0.388	0.500	0.513	0.244	0.883	0.711	0.215	1.290	0.791	0.409	0.556	0.503	0.250	0.880	0.701	0.165	1.398	0.844
Tile2Vec	0.482	0.499	0.501	0.269	0.855	0.683	0.173	1.346	0.799	0.466	0.501	0.486	0.289	0.841	0.649	0.123	1.500	0.881
READ	0.353	0.589	0.589	0.301	0.849	0.633	0.200	1.289	0.766	0.378	0.600	0.551	0.301	0.811	0.631	0.186	1.356	0.823
PG-SimCLR	0.503	0.401	0.401	0.370	0.823	0.603	0.309	1.109	0.702	0.523	0.412	0.417	0.386	0.791	0.610	0.290	1.172	0.741
UrbanCLIP	**0.587**	**0.390**	**0.389**	**0.388**	**0.801**	**0.602**	**0.309**	**1.109**	**0.700**	**0.597**	**0.373**	**0.387**	**0.391**	**0.791**	**0.602**	**0.293**	**1.153**	**0.734**
Improvement	16.77%	2.65%	3.02%	4.89%	2.70%	0.10%	0.10%	0.04%	0.37%	14.12%	9.58%	7.27%	1.48%	0.04%	1.39%	0.86%	1.65%	0.96%

Results

- UrbanCLIP consistently performed best
- Carbon > Population > GDP



Results

- Transferability and explainability of UrbanCLIP
- UrbanCLIP can capture similar spatial characteristics and distributions among comparable regions
  - Beijing (North)
  - Shanghai (East)
  - Guangzhou (South)
  - Shenzhen (South)

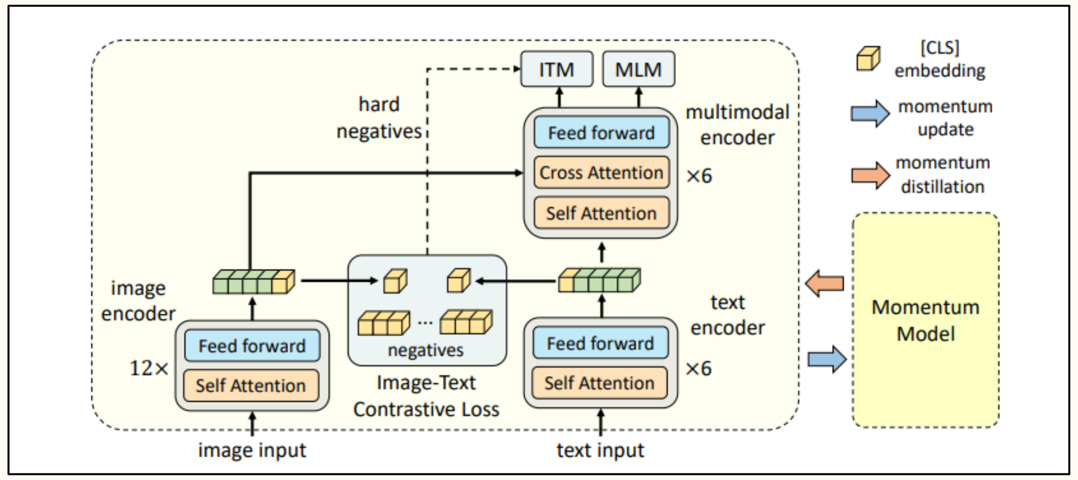
Take Aways

$$e^{T} = \mathrm{LayerNorm}\left(e_{E}^{T} + \text{M-MSA}\left(e_{E}^{T}\right)\right), \qquad (3)$$

Unidirectional Language Modeling

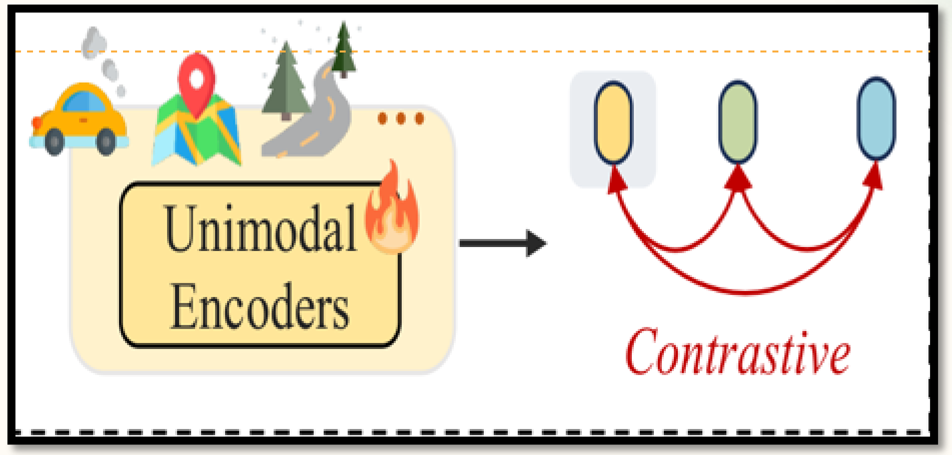
- Time Efficient
- Converts both Contrastive and Generative training in a single forward pass

Bidirectional: Vision-Language Learning [1]



Supports flexible infusion of multiple modalities

- Plug and play integration



Decoder Only Architecture for encoding Text

- Normally, BERT-style models with encoder only architectures are used
- Traditional bidirectional attention may encounter low-rank issues
- Limited Generative Capabilities

LIMITATIONS

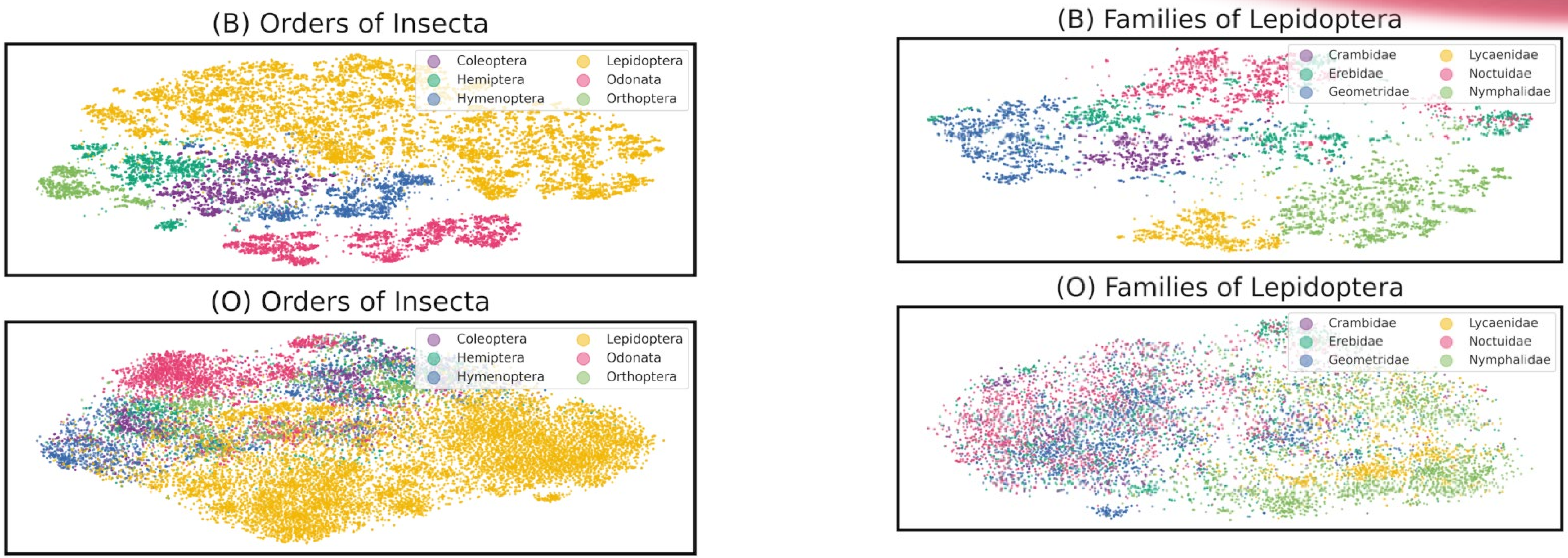
- Completely depends on Vision-Language Model for text description for satellite images
- Needs frequent updates to pre-trained model to use in real-life applications
- Text refinement technique is : **rule - based. Rules were not shared and it may induce** subjectivity and selective bias.

*Source: [1]*Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Agenda

- UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web
- BIOCLIP: A Vision Foundation Model for the Tree of Life
- MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

BioCLIP vs CLIP



TSNE- Visualization of image features, colored by taxonomic labels

---

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

Evolutionary Biology

General Tasks

- Species classification
- Individual identification
- Trait detection
- Understanding mechanisms of adaptation
- Abundance and population structure estimation
- Biodiversity monitoring and conservation

Challenges / Motivation

- Biologists **need significant ML expertise** to label data and train models .
- Existing biological datasets **lack** the necessary **scale, diversity, or fine-grained taxonomic labels** to train effective models.
- Current general vision models (like CLIP and OpenCLIP) **fail to provide fine-grained distinctions** needed for biological research
- **Need for Generalization** – A useful model must **extend beyond the taxa** it was trained on to cover the entire **tree of life** effectively.

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

BioCLIP: Vision Foundation Model for Tree of Life

Challenges / Motivation

- Biologists **need significant ML expertise** to label data and train models .
- Existing biological datasets **lack** the necessary **scale, diversity, or fine-grained taxonomic labels** to train effective models.
- Current general vision models (like CLIP and OpenCLIP) **fail to provide fine-grained distinctions** needed for biological research
- **Need for Generalization** – A useful model must **extend beyond the taxa** it was trained on to cover the entire **tree of life** effectively.

Objective of BioCLIP

- Generalize to taxa not seen during training.
- Learn fine-grained representations of biological images.
- Perform well in low-data regimes (zero-shot or few-shot learning).

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

BioCLIP: Overview

Objective of BioCLIP

- Generalize to taxa not seen during training.
- Learn fine-grained representations of biological images.
- Perform well in low-data regimes (zero-shot or few-shot learning).

Contributions

- TREEOFLIFE-10M Dataset; 454k Taxa; ML-ready
- BIOCLIP Model; Contrastive loss + taxonomic hierarchy
- Comprehensive Benchmarking; classification, rare species classification
- BioCLIP learns hierarchical representation

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

BioCLIP: TreeOfLife 10M dataset

Diversity Gap

- Existing largest ML-ready Biology Dataset: iNat21 [1]; 2.7M images – 10k Species
- IUCN [2] report (2022): 2M Species {Bird & Reptile having 10K species each}

Data Source + Curation

- iNat21 (training split)
- Encyclopedia of Life (EOL) – 6.6M images, adding 440K taxa
  - 1M+ [insect species]
  - 10K+ [birds species]
  - 10K+ [reptiles species]
- BIOSCAN-1M [3]: 1M lab images of insects from 494 families
- Taxonomic inconsistencies across sources addressed by unifying labels using **ITIS [4], EOL[2], and iNaturalist [5]**.

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life. ***[1]*** *Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In CVPR (2015)* ***[2]*** International Union for Conservation of Nature **[3]** A step towards worldwide biodiversity assessment: The BIOSCAN-1M insect dataset. [NerulPS 2024] **[4]** Integrated Taxonomic Information System **[5]** The iNaturalist species classification and detection dataset. (CVPR 2018)

BioCLIP: TreeOfLife 10M dataset

Dataset	Description	Images	Unique Classes
iNat21	Citizen scientist labeled image dataset from iNaturalist for fine-grained classification.	2.7M	10,000
BIOSCAN-1M	Expert labeled image dataset of insects for classification.	1.1M	7,831
EOL	A new dataset with citizen scientist images sourced from Encyclopedia of Life and taxonomic labels standardized by us.	6.6M	448,910
TREEOFLIFE-10M	**Largest-to-date ML-ready dataset of biology images with taxonomic labels.**	**10.4M**	**454,103**

Table 1. Training data sources used in TREEOFLIFE-10M. We integrate and canonicalize taxonomic labels across the sources (§2.2).

- **10+ million images, 454K+ unique taxonomic names.**
- **Phyla [1] Coverage:** Includes insects, birds, reptiles, fungi, plants, and other taxa (visualized in **treemap** →).

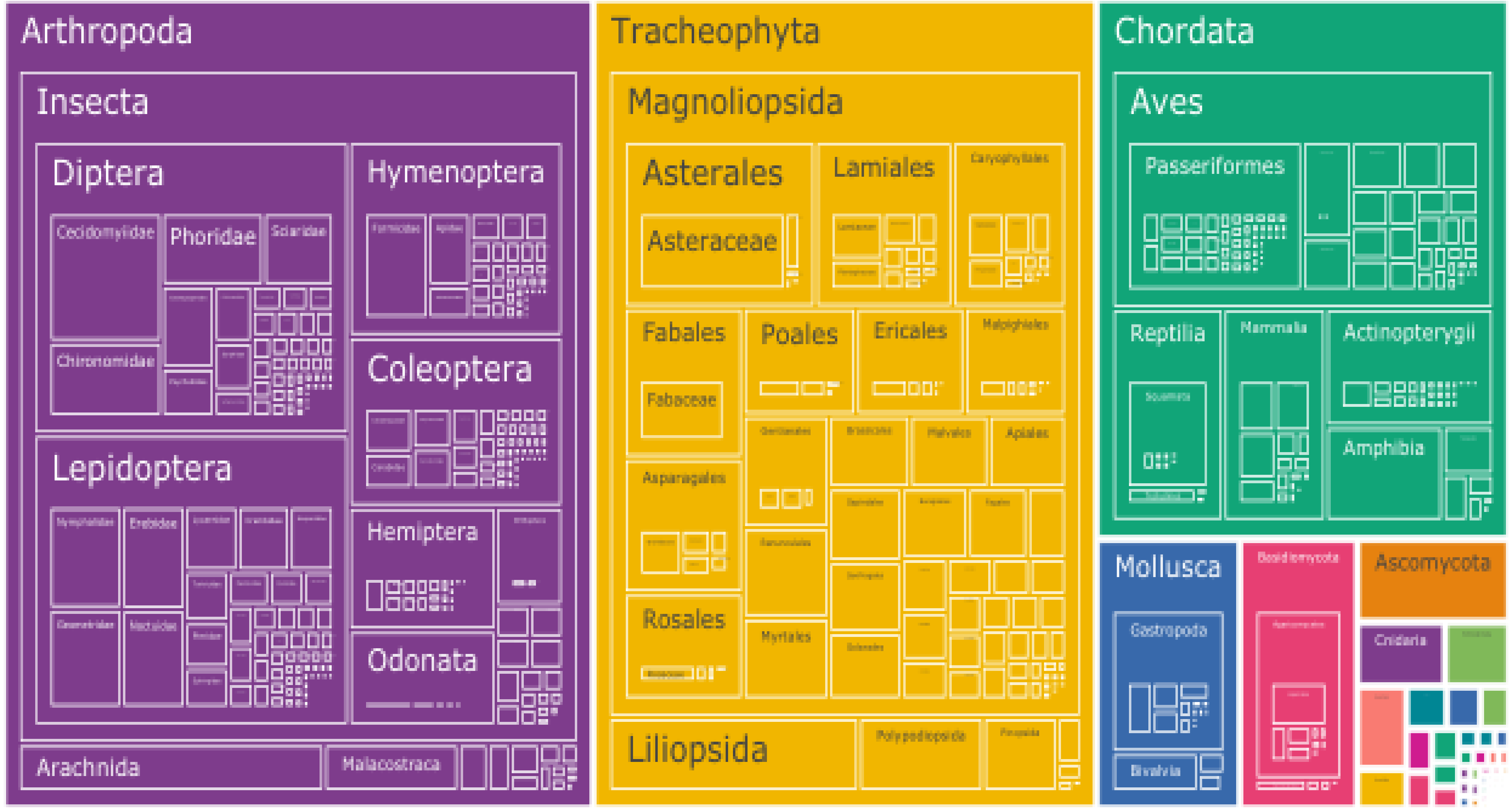


Figure 2. Treemap of the 108 phyla in TREEOFLIFE-10M. Different colors are different phyla; nested boxes represent classes, orders, and families. Box size, not number of inner boxes, represents relative number of samples.

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

[1] Major group of animals or plants that share fundamental characteristics

BioCLIP: Modeling

- Trains 2 Unimodal embeddings models [(Vision | Text) Encoders]
- Objectives
  - Maximize feature similarity of (Image, Text)$^{+}$ pairs*
  - Minimize feature similarity of (Image, Text)$^{-}$ pairs*

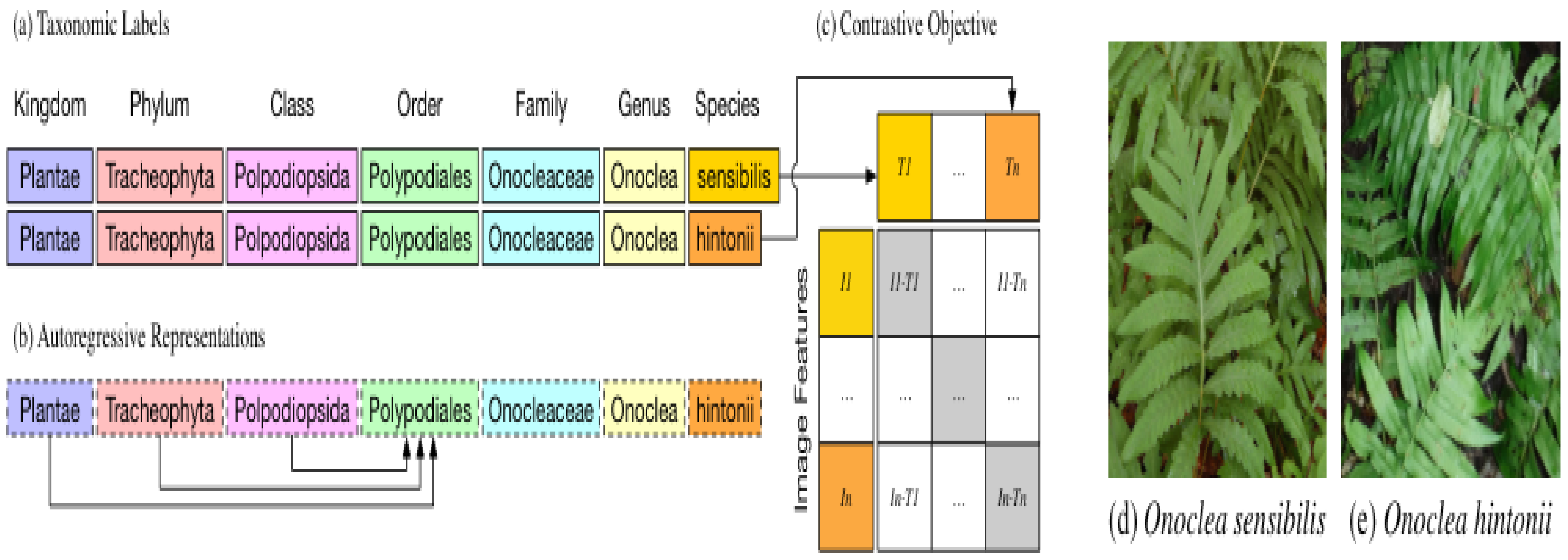


Figure 1. (a) Two taxa, or taxonomic labels, for two different plants, *Onoclea sensibilis* (d) and *Onoclea hintonii* (e). These taxa are identical except for the species. (b) The autoregressive text encoder naturally encodes the hierarchical structure of the taxonomy. See how the Order token(s) (Polypodiales) can incorporate information from the Kingdom, Phylum and Class tokens, but nothing later in the hierarchy. This helps align the visual representations to this same hierarchical structure (see §4.6). (c) These hierarchical representations of taxonomic labels are fed into the standard contrastive pre-training objective and are matched with image representations (d) and (e).

How to make taxonomic structure?

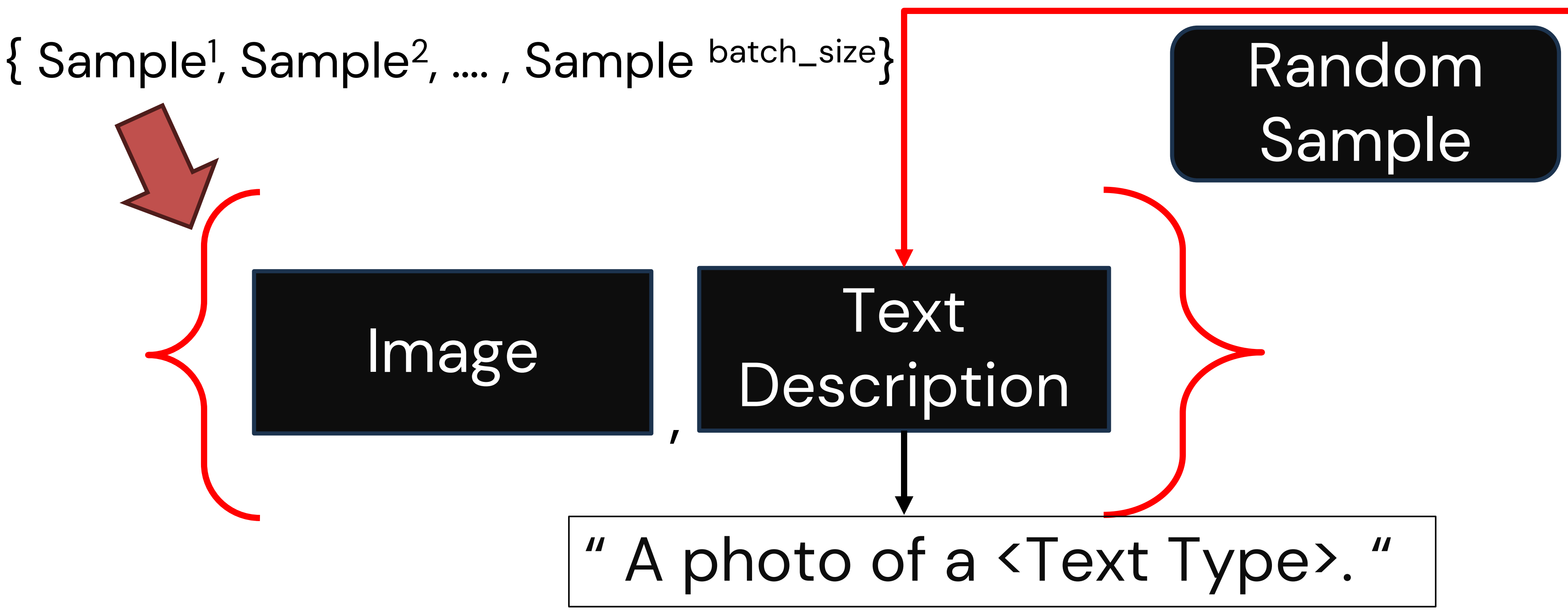
Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

** (+) means pairs are from training data, (-) means pairs are from all other possible pairs in a batch.*

BioCLIP: Modeling

Training Strategy

- Mixed text-type training strategy



Text Type	Example
Common	black-billed magpie
Scientific	*Pica hudsonia*
Taxonomic	*Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia*
Scientific + Common	*Pica hudsonia* with common name black-billed magpie
Taxonomic + Common	*Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia* with common name black-billed magpie

Table 3. Text types considered in the training of BIOCLIP.

- Taxonomic: seven-level biology taxonomy
- Scientific name: genus and species
- Common name: Regular English Word

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

.

BioCLIP:Training Strategy

Pre-training BioCLIP	Image Encoder: ViT-B/16 [+ OpenAI CLIP weights]
	Text Encoder: 77-token Causal autoregressive transformer
	Data: TREEOFLIFE-10M
	Epochs: 100; Batch size: 32768 samples
	Hardware: 8x – NVIDIA A100-80GB GPU

Zero-Shot Learning

- Same settings as CLIP

Few-shot Learning

- Randomly sample k examples for each class
- Obtain k image embeddings from pre-trained models
- Centroid of each class: Average Feature Vector of K embeddings
- Apply mean subtraction + L2 normalization to (centroid | test feature vector)
- Choose Class with nearest centroid to test vector

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

BioCLIP: Evaluation Dataset

	Name	Description	Examples	Classes	Labels
Animals	Birds 525	Scraped dataset of bird images from web search. [68]	89,885	525	Taxonomic
	Plankton	Expert-labeled in situ images of plankton [35].	4,080	102	Mixed
	Insects	Expert and volunteer-labeled in-the-wild citizen science images of insects [74].	4,680	117	Scientific
	Insects 2	Mixed common and scientific name classification for insect pests [91].	4,080	102	Mixed
Plants & Fungi	PlantNet	Citizen science species-labeled plant images, some drawings [27].	1,000	25	Scientific
	Fungi	Expert-labeled images of Danish fungi [66].	1,000	25	Scientific
	PlantVillage	Museum-style leaf specimens labeled with common names [25].	1,520	38	Common
	Medicinal Leaf	Species classification of leaves from mature, healthy medicinal plants [71].	1,040	26	Scientific
	PlantDoc	17 diseases for 13 plant species [76].	1,080	27	Common
	RARE SPECIES	Subset of species in the IUCN Red List categories: Near Threatened through Extinct in the Wild (iucnredlist.org).	12,000	400	Taxonomic

Table 2. Datasets used for evaluation. All tasks are classification evaluated with Top-1 accuracy.

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

Model	Birds 525	Plankton	Insects	Insects 2	PlantNet	Fungi	PlantVillage	Med. Leaf	PlantDoc	Rare Species	Mean (Δ)	
	Animals				**Plants & Fungi**							
Random Guessing	0.2	1.2	1.0	1.0	4.0	4.0	2.6	4.0	3.7	0.3	2.2	
Zero-Shot Classification												
CLIP	49.9	3.2	9.1	9.8	58.5	10.2	5.4	15.9	26.1	31.8	21.9	–
OpenCLIP	54.7	2.2	6.5	9.6	50.2	5.7	8.0	12.4	25.8	29.8	20.4	−1.5
BIOCLIP	**72.1**	**6.1**	**34.8**	**20.4**	**91.4**	40.7	**24.4**	**38.6**	**28.4**	**38.0**	**39.4**	+17.5
– iNat21 Only	56.1	2.6	30.7	11.5	88.2	**43.0**	18.4	25.6	20.5	21.3	31.7	+9.8
One-Shot Classification												
CLIP	43.7	25.1	21.6	13.7	42.1	17.2	49.7	70.1	24.8	28.5	33.6	–
OpenCLIP	53.7	32.3	23.2	14.3	45.1	18.4	53.6	71.2	26.8	29.2	36.7	+3.1
Supervised-IN21K	60.2	22.9	14.7	14.4	46.7	16.9	**62.3**	58.6	27.7	28.0	35.2	+1.6
DINO	40.5	**37.0**	23.5	16.4	30.7	20.0	60.0	79.2	23.7	31.0	36.2	+2.6
BIOCLIP	71.8	30.6	**57.4**	**20.4**	64.5	**40.3**	58.8	**84.3**	**30.7**	**44.9**	**50.3**	+16.7
– iNat21 Only	**74.8**	29.6	53.9	19.7	**67.4**	35.5	55.2	75.1	27.8	36.9	47.5	+13.9
Five-Shot Classification												
CLIP	73.5	41.2	39.9	24.6	65.2	27.9	71.8	89.7	35.2	46.0	51.5	–
OpenCLIP	81.9	52.5	42.6	25.0	68.0	30.6	77.8	91.3	42.0	47.4	55.9	+4.4
Supervised-IN21K	83.9	39.2	32.0	25.4	70.9	30.9	**82.4**	82.3	44.7	47.3	53.9	+2.4
DINO	70.8	**56.9**	46.3	28.6	50.3	34.1	82.1	94.9	40.3	50.1	55.4	+3.9
BIOCLIP	90.0	49.3	**77.8**	**33.6**	**85.6**	**62.3**	80.9	**95.9**	**47.5**	**65.7**	**68.8**	+17.3
– iNat21 Only	**90.1**	48.2	73.7	32.1	84.7	55.6	77.2	93.5	41.0	55.6	65.1	+13.6

Table 4. Zero-, one- and five-shot classification top-1 accuracy for different models. **Bold** indicates best accuracy. All models use the same ViT-B/16 architecture. "iNat21 Only" follows the same procedure as BIOCLIP but uses iNat21 instead of TREEOFLIFE-10M. Δ denotes the difference in mean accuracy with CLIP. Supervised-IN21K [78] and DINO [15] are vision-only models and cannot do zero-shot classification.

Results

BIOCLIP's strong zero-shot performance on the diverse tasks and classes in TREEOFLIFE-10M.

Text Type
Common
Scientific
Taxonomic
Scientific + Common
Taxonomic + Common

Dataset	Train↓Test→	Com	Sci	Tax	Sci+Com	Tax+Com
ToL-1M	Com	24.9	9.5	10.8	22.3	21.0
	Sci	11.0	22.3	4.5	21.5	8.0
	Tax	11.8	10.1	26.6	16.0	24.8
	Sci+Com	24.5	12.9	12.6	28.0	24.9
	Tax+Com	20.5	8.0	19.7	24.0	30.4
	Mixture	**26.1**	**24.9**	**26.7**	**29.5**	**30.9**
iNat21-2.7M	Mixture	20.4	14.7	15.6	20.9	21.3
ToL-10M	Mixture	31.6	30.1	34.1	37.0	38.0

Table 5. Zero-shot accuracy on species not seen during training (RARE SPECIES task). Different rows and columns indicate different text types during training and test time, respectively. **Blue** indicates best accuracy and Orange indicates second-best. Using the taxonomic name over the scientific name always improves accuracy (22.3→26.6 and 28.0→30.4). The final rows use the full iNat21 dataset and TREEOFLIFE-10M for reference.

Results

Using **mixed text types** for training yields consistently strong performance across all text types during testing.

Results

Objective	Mean 1-Shot	Mean 5-shot
Cross-entropy	16.5	26.2
Hier. cross-entropy	19.3	30.5
CLIP	**44.7**	**63.8**

Table 6. One- and five-shot classification top-1 accuracy for different pre-training objectives on TREEOFLIFE-1M. Results are macro-averaged over all the test sets in Tab. 4.

CLIP objective massively outperforms both baselines

- Objectives
  - Maximize feature similarity of (Image, Text)$^+$ pairs*
  - Minimize feature similarity of (Image, Text)$^-$ pairs*

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

** (+) means pairs are from training data, (-) means pairs are from all other possible pairs in a batch.*

Model	Animals				Plants & Fungi							
	Birds 525	Plankton	Insects	Insects 2	PlantNet	Fungi	PlantVillage	Med. Leaf	PlantDoc	Rare Species	Mean (Δ)	
Random Guessing	0.2	1.2	1.0	1.0	4.0	4.0	2.6	4.0	3.7	0.3	2.2	
Zero-Shot Classification												
CLIP	49.9	3.2	9.1	9.8	58.5	10.2	5.4	15.9	26.1	31.8	21.9	–
OpenCLIP	54.7	2.2	6.5	9.6	50.2	5.7	8.0	12.4	25.8	29.8	20.4	−1.5
BioCLIP	**72.1**	**6.1**	**34.8**	**20.4**	**91.4**	40.7	**24.4**	**38.6**	**28.4**	**38.0**	**39.4**	+17.5
– iNat21 Only	56.1	2.6	30.7	11.5	88.2	**43.0**	18.4	25.6	20.5	21.3	31.7	+9.8
One-Shot Classification												
CLIP	43.7	25.1	21.6	13.7	42.1	17.2	49.7	70.1	24.8	28.5	33.6	–
OpenCLIP	53.7	32.3	23.2	14.3	45.1	18.4	53.6	71.2	26.8	29.2	36.7	+3.1
Supervised-IN21K	60.2	22.9	14.7	14.4	46.7	16.9	**62.3**	58.6	27.7	28.0	35.2	+1.6
DINO	40.5	**37.0**	23.5	16.4	30.7	20.0	60.0	79.2	23.7	31.0	36.2	+2.6
BioCLIP	71.8	30.6	**57.4**	**20.4**	64.5	**40.3**	58.8	**84.3**	**30.7**	**44.9**	**50.3**	+16.7
– iNat21 Only	**74.8**	29.6	53.9	19.7	**67.4**	35.5	55.2	75.1	27.8	36.9	47.5	+13.9
Five-Shot Classification												
CLIP	73.5	41.2	39.9	24.6	65.2	27.9	71.8	89.7	35.2	46.0	51.5	–
OpenCLIP	81.9	52.5	42.6	25.0	68.0	30.6	77.8	91.3	42.0	47.4	55.9	+4.4
Supervised-IN21K	83.9	39.2	32.0	25.4	70.9	30.9	**82.4**	82.3	44.7	47.3	53.9	+2.4
DINO	70.8	**56.9**	46.3	28.6	50.3	34.1	82.1	94.9	40.3	50.1	55.4	+3.9
BioCLIP	90.0	49.3	**77.8**	**33.6**	**85.6**	**62.3**	80.9	**95.9**	**47.5**	**65.7**	**68.8**	+17.3
– iNat21 Only	**90.1**	48.2	73.7	32.1	84.7	55.6	77.2	93.5	41.0	55.6	65.1	+13.6

Results

- Can BIOCLIP Classify More Than Species?
  - Task: plant diagnosis with the PlantVillage and PlantDoc (has diseased plant image)
  - BIOCLIP has learned useful visual representations that are useful even with only one labeled example

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life

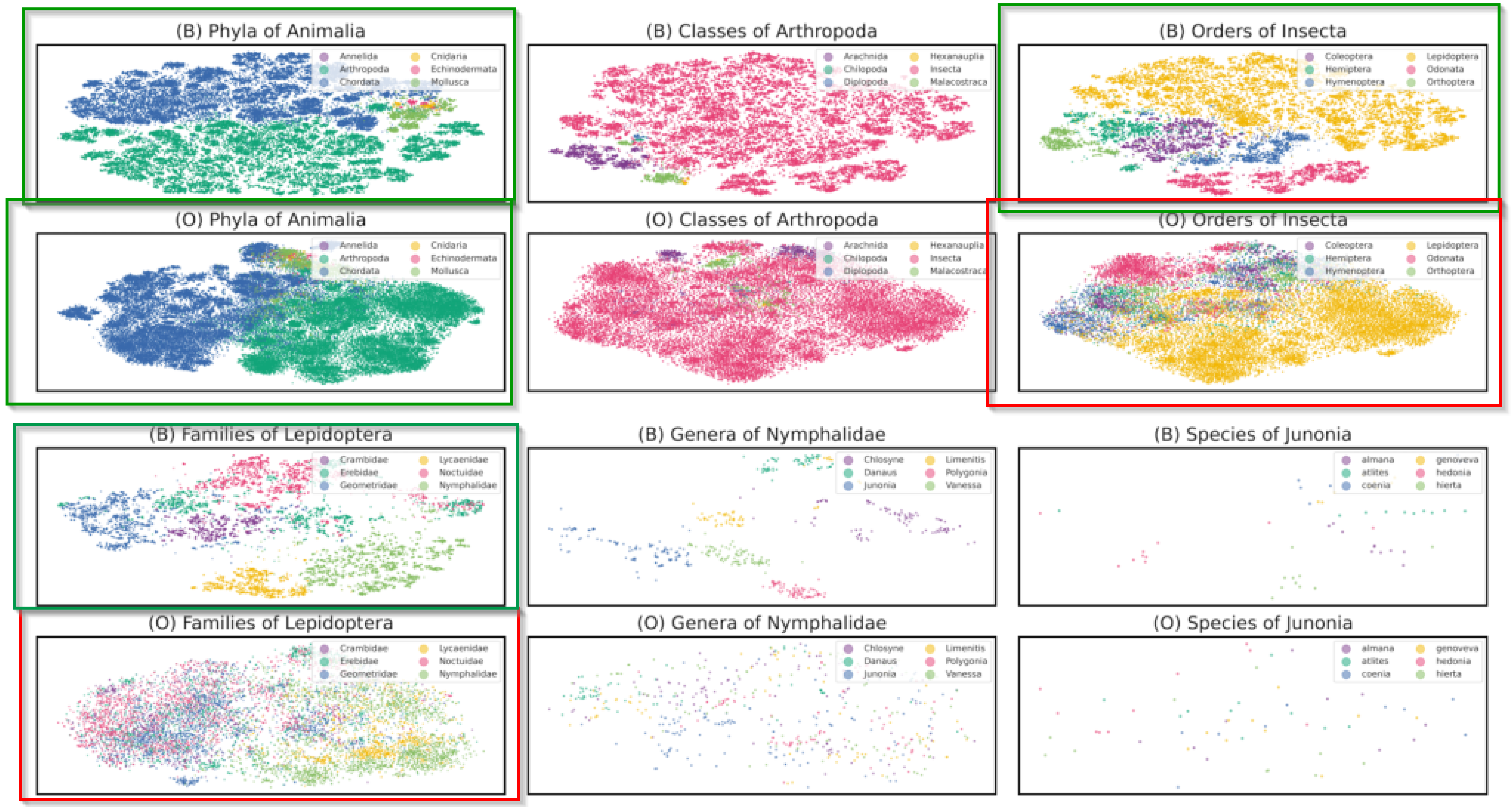


Figure 3. T-SNE visualization of image features, colored by taxonomic labels. BIOCLIP (B) is visualized in the first and third row and OpenAI's CLIP (O) is visualized in the second and fourth rows. BIOCLIP's features better preserve the hierarchical structure: while both BIOCLIP and CLIP cleanly separate the phyla in the Animalia Kingdom (top left), only BIOCLIP successfully separates the orders in the Insecta Class (top right) and the families in the Lepidoptera Order (bottom left).

Results

- Does BIOCLIP Learn the Hierarchy?
- Task: plant diagnosis with the PlantVillage and PlantDoc (has diseased plant image)
  - BIOCLIP has learned useful visual representations that are useful even with only one labeled example

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life.

Take Aways

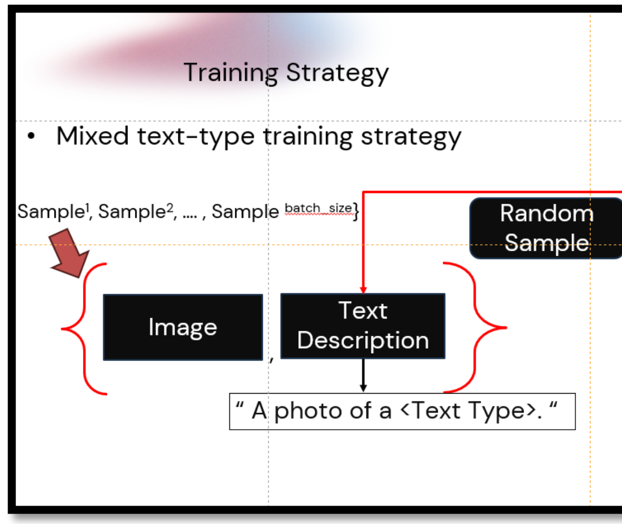
$$e^{T} = \mathrm{LayerNorm}\left(e_{E}^{T} + \text{M-MSA}\left(e_{E}^{T}\right)\right), \quad (3)$$

BioCLIP can achieve superior zero - shot generalization

Mix - Type training strategy can enhance classification task

One - shot Learning was enough to achieve Task transferability of BioCLIP

– Classifier trained model could detect plant disease



Text Type	Example
Common	black-billed magpie
Scientific	*Pica hudsonia*
Taxonomic	*Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia*
Scientific + Common	*Pica hudsonia* with common name black-billed magpie
Taxonomic + Common	*Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia* with common name black-billed magpie

Table 3. Text types considered in the training of BIOCLIP.

- Taxonomic: seven-level biology taxonomy
- Scientific name: genus and species
- Common name: Regular English Word

General Tasks

- Species classification
- Individual identification
- Trait detection
- Understanding mechanisms of adaptation
- Abundance and population structure estimation
- Biodiversity monitoring and conservation

LIMITATIONS

- **Computational Constraints** (mix type analysis used only 10% subset of TreeOfLife-10M dataset)
- **Domain Specificity:** Applicability to broader tasks remains unknown

Source: BIOCLIP: A Vision Foundation Model for the Tree of Life.

Agenda

- UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web
- BIOCLIP: A Vision Foundation Model for the Tree of Life
- MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Artificial General Intelligence (AGI)

AI system that reaches "at least 90th percentile of skilled adults" [1]

How to create benchmarks for measuring Expert AGI?

- College Level Exams → MMLU [2], AGIEval [3] ;only text-based
- Existing Multimodal Benchmarks (ScienceQA [4]) focus on commonsense/daily knowledge, not expert-level knowledge– reasoning

---

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

[1] Operationalizing progress on the path to agi. Arxiv [2023]

[2] Measuring massive multitask language understanding. ICLR [2020]

[3] Agieval: A human-centric benchmark for evaluating foundation models. NAACL [2023]

[4] Learn to explain: Multimodal reasoning via thought chains for science question answering. ANIPS [2022]

MMMU: A Massive Multi - discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

- Designed For: College Level (multi-discipline) (multimodal understanding) and (reasoning)
- Problem Source: Exams, Quizzes, Text books
  - 6 Common discipline: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering.
- Covers:
  - 11.5K multimodal questions ← 30 diverse subjects 183 subfields.
  - Expert level reasoning : applying "Fourier Transform" or "Equi-
  - librium Theory" to derive the solution

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

MMMU: A Massive Multi - discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

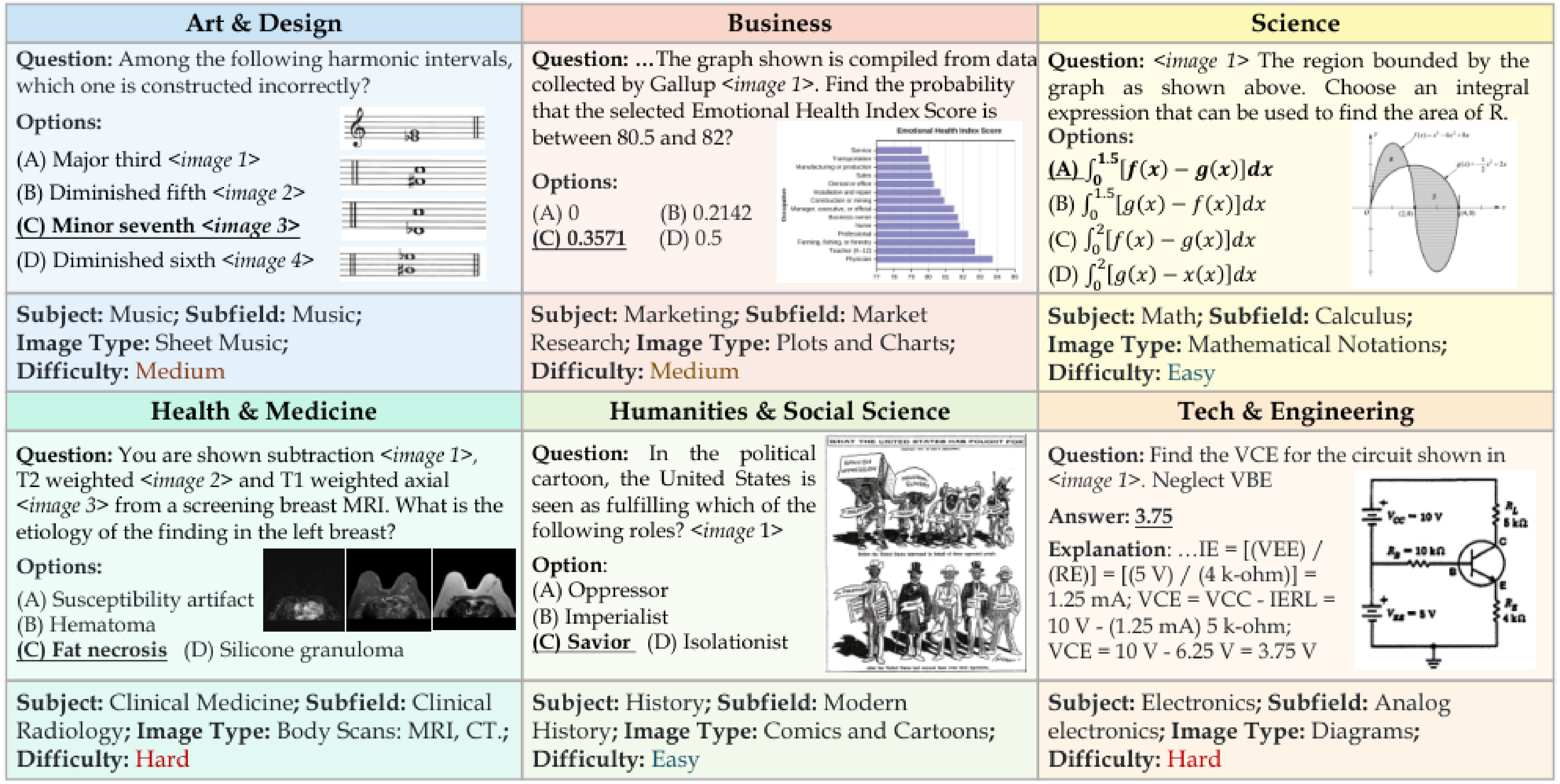


Figure 2. Sampled MMMU examples from each discipline. The questions and images need expert-level knowledge to understand and reason.

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

MMMU: A Massive Multi - discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

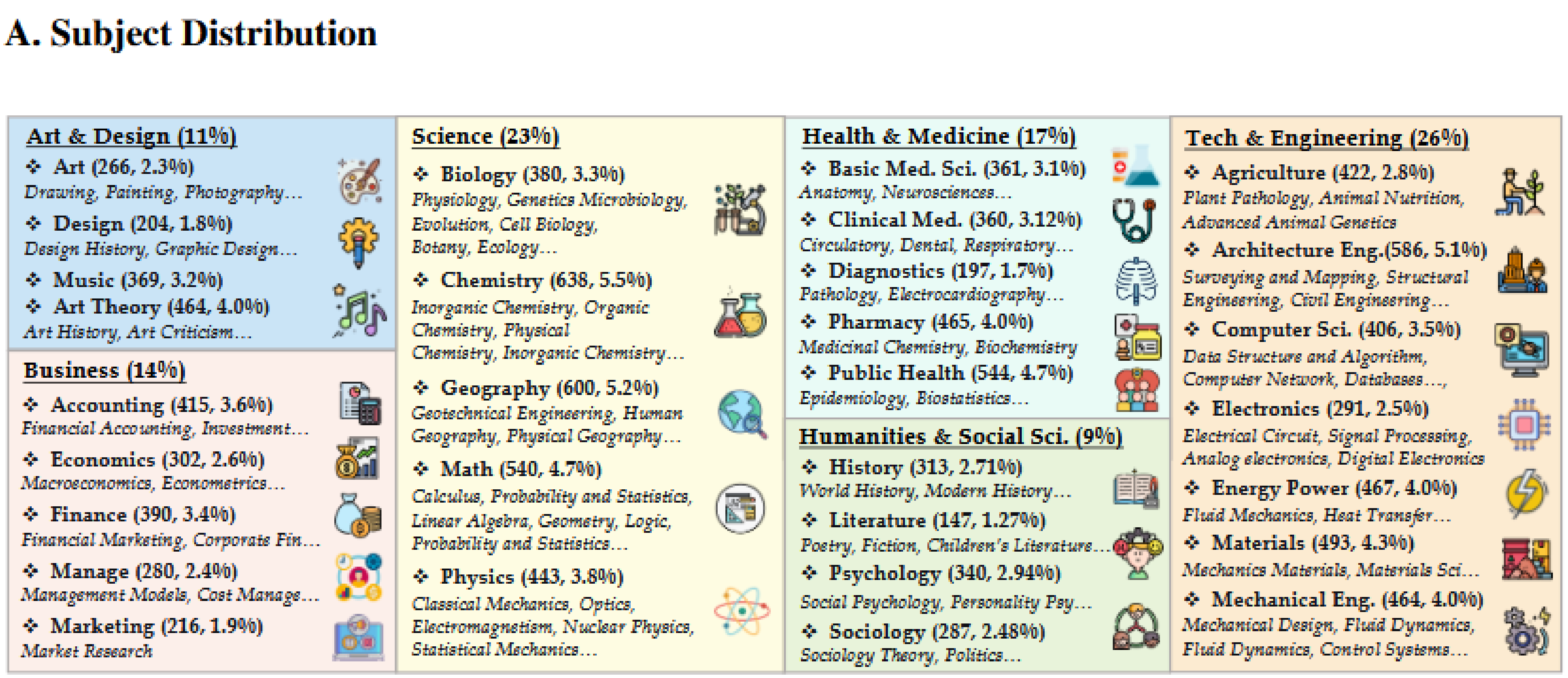


Figure 7. MMMU contains 11.5K multimodal questions covering six broad disciplines, 30 subjects, and 183 subfields.

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

MMMU: A Massive Multi - discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Statistics	Number
Total Questions	11550
Total Disciplines/Subjects/Subfields	6/30/183
Image Types	30
Dev:Validation:Test	150:900:10500
Difficulties (Easy: Medium: Hard)	28%:45%:27%
Multiple-choice Questions	10861 (94.03%)
Open Questions	689 (5.97%)
Questions with an Explanation	2035 (17.62%)
Image in the Question	11264 (97.52%)
* Images at the beginning	2006 (17.81%)
* Images in the middle	4159 (36.92%)
* Images at the end	5679 (50.42%)
Image in Options	389 (3.37%)
Example with Multiple Images	854 (7.39%)
Average question length	59.33
Average option length	9.17
Average explanation length	107.92

Table 1. Key statistics of the MMMU benchmark.

Curation + Quality Control

- Subjects like law and linguistics were excluded due to the lack of multimodal content.
- 50 university students and co-authors:
  - sourced from: textbooks, online resources, and, custom creation
- Mitigation strategy for potential data contamination:
  - Avoid readily available answers
  - Compliance with copyright and licensing regulations.
- Followed a standardized protocol to maintain consistency.
- Quality Control:
  - Duplicate Detection
  - Format and Typo Checking
  - Difficulty Categorization

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

MMMU: A Massive Multi - discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI : EVALUATION

Baselines

- Large Mulitmodal Models (LMMs)
- Text-only LLMs
- Human Experts
  - 90 College Senior Students
  - 30 Subjects , 900 validation questions (3 student/subject)
  - Allowed to consult books but **NO INTERNET**
- Metrics: micro-averaged accuracy; rule-based evaluation pipeline

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Results

	Validation Overall (900)	Test Overall (10,500)	Art & Design (1,163)	Business (1,428)	Science (2,426)	Health & Medicine (1,752)	Human. & Social Sci. (947)	Tech & Eng. (2,784)
Random Choice	22.1	23.9	24.1	24.9	21.6	25.3	22.8	24.8
Frequent Choice	26.8	25.8	26.7	28.4	24.0	24.4	25.2	26.5
Expert (Worst)	76.2	-	-	-	-	-	-	-
Expert (Medium)	82.6	-	-	-	-	-	-	-
Expert (Best)	88.6	-	-	-	-	-	-	-
Large Multimodal Models (LMMs): Text + Image as Input								
OpenFlamingo2-9B [4]	28.7	26.3	31.7	23.5	26.3	26.3	27.9	25.1
Kosmos2 [63]	24.4	26.6	28.8	23.7	26.6	27.2	26.3	26.8
Adept Fuyu-8B [6]	27.9	27.4	29.9	27.0	25.6	27.0	32.5	26.4
MiniGPT4-Vicuna-13B [94]	26.8	27.6	30.2	27.0	26.2	26.9	30.9	27.2
LLaMA-Adapter2-7B [88]	29.8	27.7	35.2	25.4	25.6	30.0	29.1	25.7
CogVLM [77]	32.1	30.1	38.0	25.6	25.1	31.2	41.5	28.9
Qwen-VL-7B-Chat [5]	35.9	32.9	47.7	29.8	25.6	33.6	45.3	30.2
InstructBLIP-T5-XXL [16]	35.7	33.8	48.5	30.6	27.6	33.6	49.8	29.4
BLIP-2 FLAN-T5-XXL [35]	35.4	34.0	49.2	28.6	27.3	33.7	51.5	30.4
InternLM-XComposer2-VL* [17]	43.0	38.2	56.8	32.8	30.1	39.8	60.7	31.8
Yi-VL-34B* [84]	45.9	41.6	56.1	33.3	32.9	45.9	66.5	36.0
LLaVA-1.6-34B* [46]	51.1	44.7	58.6	39.9	36.0	51.2	70.2	36.3
InternVL-Chat-V1.2* [11]	51.6	46.2	**62.5**	37.6	**37.9**	49.7	70.1	**40.8**
VILA1.5* [39]	**51.9**	**46.9**	62.1	**40.6**	37.7	**51.7**	**74.0**	39.5
Qwen-VL-MAX* [65]	51.4	46.8	64.2	39.8	36.3	52.5	70.4	40.7
SenseChat-Vision-0423-Preview* [68]	54.6	50.3	62.7	44.1	42.3	55.7	74.7	**43.5**
GPT-4V(ision) (Playground) [60]	56.8	**55.7**	**65.3**	**64.3**	**48.4**	**63.5**	**76.3**	41.7
Claude 3 Opus* [72]	59.4	-	-	-	-	-	-	-
Gemini 1.5 Pro* [23]	62.2	-	-	-	-	-	-	-
GPT-4o* [61]	**69.1**	-	-	-	-	-	-	-
Large Language Models (LLMs): Only Text as Input								
Llama2 7B [75]	30.1	28.7	30.7	27.2	26.7	27.7	32.6	29.8
FLAN-T5-XXL [14]	32.1	31.2	36.8	**28.9**	26.7	32.8	44.8	28.3
+ OCR	34.7	**31.9**	36.2	28.8	26.2	32.6	**50.5**	**29.7**
+ LLaVA Caption	**34.8**	**31.9**	**38.4**	27.8	**27.0**	**33.2**	49.9	28.7
Vicuna-13B [12]	33.3	31.0	35.1	**30.1**	24.7	31.4	44.8	30.1
+ OCR	**35.4**	31.9	37.1	28.6	**26.5**	32.0	49.3	30.0
+ LLaVA Caption	33.9	**32.7**	**42.0**	26.8	26.2	**33.4**	**49.4**	**31.4**
GPT-4 Text [59]	34.9	33.8	32.9	28.5	30.6	41.3	53.0	28.4

Table 2. Selected results of different models on the MMMU **validation** and **test set**. Besides reporting the performance of LMMs, we additionally add text-only LLM baselines. The best-performing model in each category is **in-bold**, and the second best is underlined. *: results provided by the authors. Due to the page limit, we show other models' results in Appendix Table 4. The live-updating leaderboard is available at: https://mmmu-benchmark.github.io/#leaderboard

InternVL-Chat-V1.2* [11]	51.6	46.2	**62.5**	37.6	**37.9**	49.7	70.1	**40.8**
VILA1.5* [39]	**51.9**	**46.9**	62.1	**40.6**	37.7	**51.7**	**74.0**	39.5
Gemini Nano2* [22]	32.6	-	-	-	-	-	-	-
Marco-VL*	41.2	40.4	56.5	31.0	31.0	46.9	66.5	33.8
Reka Edge* [62]	42.8	-	-	-	-	-	-	-
Qwen-VL-PLUS* [64]	45.2	40.8	59.9	34.5	32.8	43.7	65.5	32.9
Marco-VL-Plus*	46.2	44.3	57.4	34.7	38.5	48.7	72.2	36.7
Gemini 1.0 Pro* [22]	47.9	-	-	-	-	-	-	-
Adept Fuyu-Heavy* [19]	48.3	-	-	-	-	-	-	-
Claude 3 Haiku* [72]	50.2	-	-	-	-	-	-	-
Reka Flash* [62]	53.3	-	-	-	-	-	-	-
Skywork-VL* [31]	51.4	46.2	61.4	39.6	36.6	50.8	71.6	40.2
Qwen-VL-MAX* [65]	51.4	46.8	64.2	39.8	36.3	52.5	70.4	40.7
HPT Pro* [28]	52.0	-	-	-	-	-	-	-
Claude 3 Sonnet* [72]	53.1	-	-	-	-	-	-	-
SenseChat-Vision-0423-Preview* [68]	54.6	50.3	62.7	44.1	42.3	55.7	74.7	**43.5**
Gemini 1.5 Flash* [23]	56.1	-	-	-	-	-	-	-
Reka Core* [62]	56.3	-	-	-	-	-	-	-
GPT-4V(ision) (Playground) [60]	56.8	**55.7**	**65.3**	**64.3**	**48.4**	**63.5**	**76.3**	41.7
Claude 3 Opus* [72]	59.4	-	-	-	-	-	-	-
Gemini 1.0 Ultra* [22]	59.4	-	-	-	-	-	-	-
Gemini 1.5 Pro* [23]	62.2	-	-	-	-	-	-	-
GPT-4o* [61]	**69.1**	-	-	-	-	-	-	-
Large Language Models (LLMs): Only Text as Input								
Llama2 7B [75]	30.1	28.7	30.7	27.2	26.7	27.7	32.6	29.8
FLAN-T5-XXL [14]	32.1	31.2	36.8	**28.9**	26.7	32.8	44.8	28.3
+ OCR	34.7	**31.9**	36.2	28.8	26.2	32.6	**50.5**	**29.7**
+ LLaVA Caption	**34.8**	**31.9**	**38.4**	27.8	**27.0**	**33.2**	49.9	28.7
Vicuna-13B [12]	33.3	31.0	35.1	**30.1**	24.7	31.4	44.8	30.1
+ OCR	**35.4**	31.9	37.1	28.6	**26.5**	32.0	49.3	30.0
+ LLaVA Caption	33.9	**32.7**	**42.0**	26.8	26.2	**33.4**	**49.4**	**31.4**
GPT-4 Text [59]	34.9	33.8	32.9	28.5	30.6	41.3	53.0	28.4

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Results

Apart from human-expert, GPT-4V is comparably performing best among the tested models

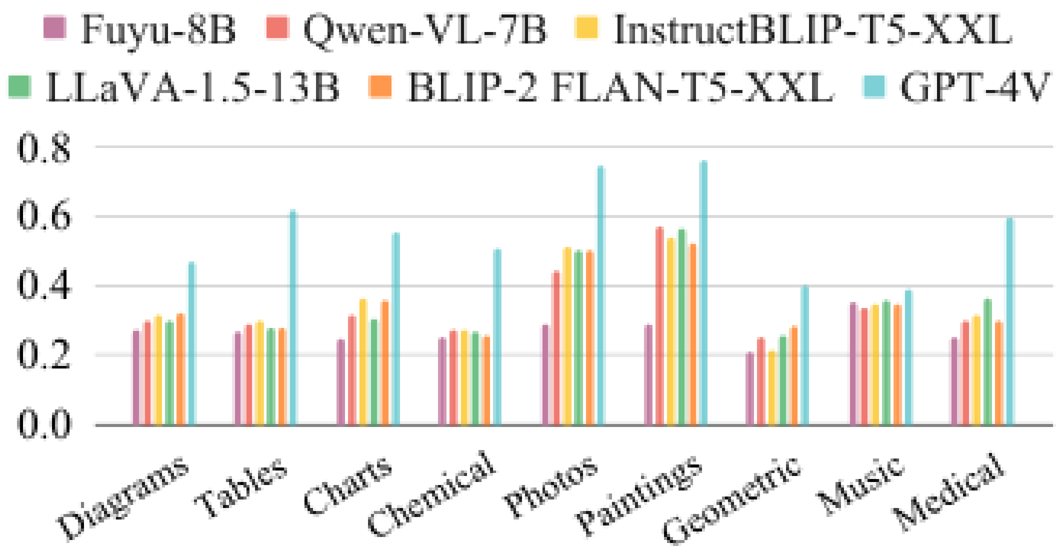


Figure 4. Performance of models on different types of images.

Models	Easy (2946)	Medium (4917)	Hard (2637)	Overall (10500)
Fuyu-8B [6]	28.9	27.0	26.4	27.4
Qwen-VL-7B [5]	39.4	31.9	27.6	32.9
LLaVA-1.5-13B [44]	41.3	32.7	26.7	33.6
InstructBLIP-T5-XXL [16]	40.3	32.3	29.4	33.8
BLIP-2 FLAN-T5-XXL [35]	41.0	32.7	28.5	34.0
GPT-4V [60]	76.1	55.6	31.2	55.7

Table 3. Result decomposition across question difficulty levels.

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

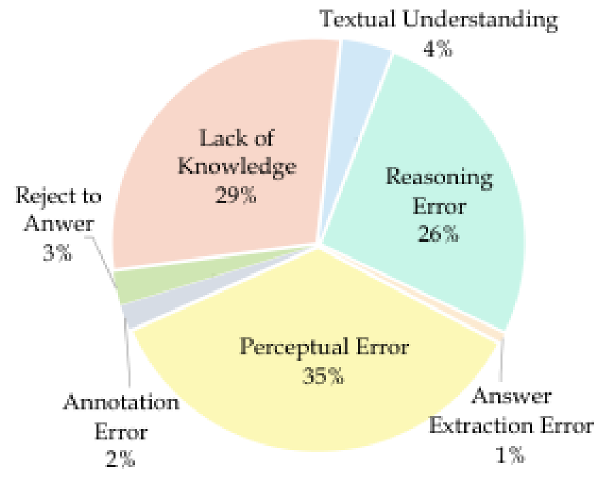
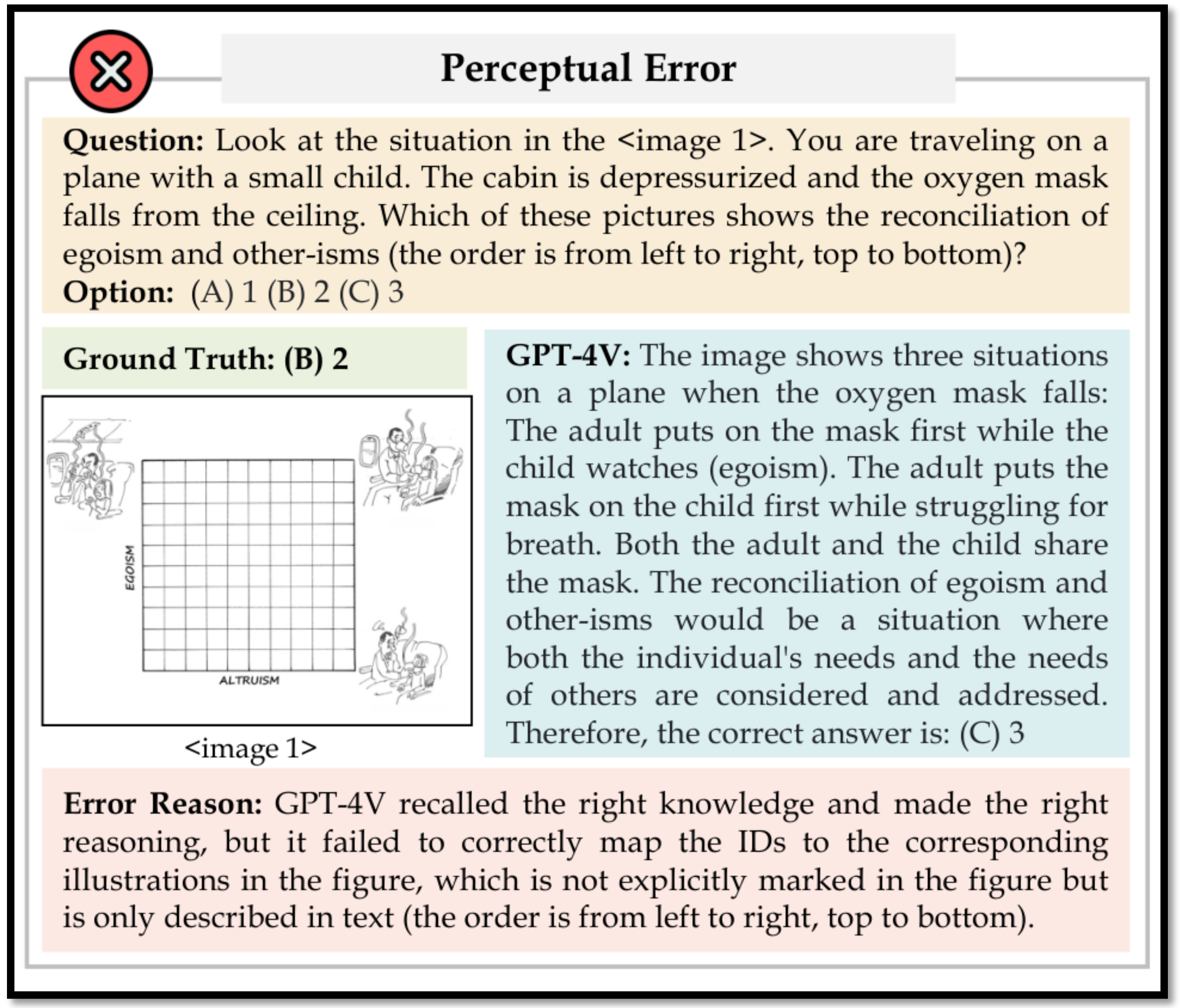


Figure 5. Error distribution over 150 annotated GPT-4V errors.



Results

- Examine 150 randomly sampled error instances from GPT-4V's predictions.
- Perceptual Error:
  - **Basic:** fails in elementary visual interpretation
  - **Domain Specific:** error due to lack of knowledge

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Results

- Perceptual Error Example

History: Modern History

Question: In the political cartoon, the United States is seen as fulfilling which of the following roles?<image 1>
Option:
(A) Oppressor
(B) Imperialist
(C) Savior
(D) Isolationist

<image 1>

GPT-4V: Based on the political cartoon, the United States is seen as fulfilling the role of:
(B) Imperialist.

Error Category: Perceptual Error
Error Reason: GPT-4V wrongly interprets the image, because it doesn't correctly interpret the contradiction between the comics on the top half and the bottom half. The perspective of the cartoonist is that the United States has been a savior to the nations brought under its control, saving them from poverty or oppression and bringing them trade, prosperity, and democratic governments. This mistake also indicates that GPT-4V holds some wrong political beliefs.

Ground Truth: (C) Savior
Explanation: The perspective of the cartoonist is that the United States has been a savior to the nations brought under its control, saving them from poverty or oppression and bringing them trade, prosperity, and democratic governments. Although one might be tempted to cast the United States in the role of imperialist (B), the purpose of the cartoon is to highlight the positive transformation of the people due to their "rescue" by the United States, rather than the benefits to the United States. Because the cartoon claims the people are better off for having been "rescued" by the United States, the United States is not seen as an oppressor (A). Since isolationists do not support foreign intervention, (D) cannot be the correct answer.

Figure 67. A sample error case of History (subfield: Modern History). Error category: Perceptual Error

Back to List of Figures | Back to Table Index

<image 1>

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Take Aways

Comprehensive Multimodal Benchmark

Challenges for Current AI Models

highlights the significant gap between AI and human expert performance

Diverse Image and Question Types

highly heterogeneous image types, including diagrams, tables, medical images, and sheet music.

LIMITATIONS

- **Benchmark Does Not Fully Define Expert AGI** : Failed to show expert performance beyond academic assessments.
- **Biases in Data Curation:** 50 college students and co-authors may induce bias in question selection, difficulty categorization, and representation of domain knowledge

Source: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI