

Supplementary for “WikiCLIP: An Efficient Contrastive Baseline for Open-domain Visual Entity Recognition”

Shan Ning^{1,3}, Longtian Qiu¹, Jiaxuan Sun¹, Xuming He^{1,2}

¹ShanghaiTech University, Shanghai, China

²Shanghai Engineering Research Center of Intelligent Vision and Imaging

³Lingang Laboratory, Shanghai, China

{ningshan2022, qiult, sunjx2022, hexm}@shanghaitech.edu.cn

A. Overview of Appendixes

In this supplementary material, we present implementation details and more experiments. First, we provide additional analysis in Section A.1. The details of the training objective are provided in Section A.2. Then, we provide the retrieval results in Section A.3 and the clarification on the reason for choosing the primary image in Section A.4. Finally, we provide a visualization in Section A.5.

A.1. More Ablation Study

The Impact of Training Iteration. As shown in Figure 1 and aligned with findings from OVEN, WikiCLIP achieves strong open-domain recognition efficiency: its unseen entity accuracy peaks at 4K iterations (with only 0.4M training samples) before slightly declining as seen entity accuracy continues to improve with extended training. This demonstrates an effective balance – while prolonged training introduces mild unseen entity performance degradation (<3%), the framework maintains competitive results through early stopping, validating its data efficiency and practical viability for visual-entity retrieval tasks.

Text Encoder Ablation on OVEN and EVQA We provide the experiments about the choice of text encoder on OVEN and EVQA datasets in Tab 1, which requires the model to match the query from a 2M entity set instead of 100k for InfoSeek. We observe that switching to LLaMA3 leads to significant improvements, which indicates that providing rich semantic information helps entity discrimination.

LLM Capacity & Text Length Joint Analysis To further investigate the interplay between LLM scale and input text length, we jointly vary these two factors in Table 2. The results are consistent with the findings in Sec ?? of the main paper: simply scaling up the LLM or extending the

Model	EVQA			OVEN		
	Unseen	Seen	Overall	Unseen	Seen	HM
EVA-CLIP 8B	12.9	12.7	14.1	17.3	15.6	15.6
LLaMa3.2 1B	27.7	39.9	30.7	27.0	36.8	31.1

Table 1. Comparison of text encoders on EVQA and OVEN.

Seen/Unseen/All	128	256	512
LLAMA3.2-1B	66.9/57.4/59.8	69.3/58.5/61.2	67.4/57.8/60.2
LLAMA3.2-3B	66.5/59.8/61.5	69.6/60.3/62.7	66.6/58.9/60.8
LLAMA3.2-8B	69.0/60.2/62.4	68.7/60.4/62.4	68.8/58.8/61.3

Table 2. Joint analysis of text length and LLM scale on the INFOSEEK.

	Unseen	Seen	HM
EVA-CLIP-8B I2I	11.8	8.8	10.1
CLIP2CLIP with EVA-CLIP-8B	14.6	11.1	12.6
WikiCLIP-S	27.0	36.8	31.1

Table 3. Comparison of contrastive baselines with scaled visual encoder on OVEN.

raw Wikipedia text does not lead to substantial gains, suggesting that future efforts should focus on extracting more informative content rather than increasing model capacity or input length alone.

Impact of Visual Encoder Scaling To further investigate the role of visual encoder scale, we reproduce the strongest contrastive baseline, CLIP2CLIP, using EVA-CLIP-8B. As shown in Table 3, scaling up the visual encoder alone does not allow the contrastive baseline to match the performance of WikiCLIP, suggesting that the knowledge-aware alignment mechanism provides benefits beyond encoder scaling.

VGKA Architecture Ablation We further ablate the VGKA architecture in terms of the number of layers and attention heads, as shown in Table 4. Performance remains stable across all variants, suggesting that VGKA is robust as

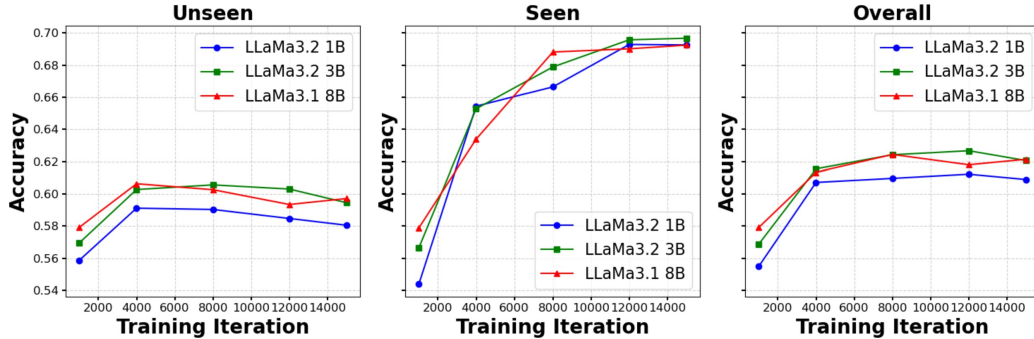


Figure 1. **Performance with Varying Training Iterations and LLM Choices.** We report the accuracy of the INFOSEEK validation set of WikiCLIP using three different scales of LLMs, along with varying training iterations.

Layers / Heads	Seen	Unseen	All
2 / 16	69.3	58.5	61.2
8 / 16	68.2	58.0	60.6
16 / 32	66.4	56.5	59.1

Table 4. VGKA architecture ablation on the INFOSEEK.

long as it is sufficiently expressive to support visual-guided knowledge alignment.

A.2. Contrastive Training Objective

We provide the details of the training objective in this section. Given a mini-batch $\mathcal{B} = [(\mathbf{h}_i, \mathbf{v}_i)]_{i=1}^N$, the positive pair for the i -th query is $(\mathbf{h}_i, \mathbf{v}_i)$, and the in-batch negatives are denoted as $\mathcal{B}_i^- = \{\mathbf{v}_j \mid j \neq i\}$. After applying our hard negative synthesis strategy, the negative set is replaced with a harder set $\tilde{\mathcal{B}}_i^-$, where each negative sample is substituted by a synthetic hard negative $\tilde{\mathbf{v}}_j$ whenever it exhibits higher similarity to the query:

$$\tilde{\mathcal{B}}_i^- = \begin{cases} \tilde{\mathbf{v}}_j, & \text{if } \text{Sim}(\mathbf{h}_i, \tilde{\mathbf{v}}_j) > \text{Sim}(\mathbf{h}_i, \mathbf{v}_j), \\ \mathbf{v}_j, & \text{otherwise,} \end{cases} \quad \forall j \neq i.$$

We optimize an InfoNCE loss with cosine similarity and temperature τ , defined for each query \mathbf{h}_i as:

$$\mathcal{L}_{\text{InfoNCE}}^{(i)} = -\log \frac{\exp(s_{i,i}/\tau)}{\sum_{v \in \{\mathbf{v}_i\} \cup \tilde{\mathcal{B}}_i^-} \exp(s_{i,v}/\tau)}. \quad (1)$$

where

$$s_{i,v} \triangleq \text{Sim}(\mathbf{h}_i, \mathbf{v}).$$

The final training loss is averaged across all samples in the mini-batch:

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\text{InfoNCE}}^{(i)}.$$

A.3. Topk Retrieval Results

Table 5 presents the performance comparison of WikiCLIP and baseline methods on open-domain visual entity recognition (VER) tasks. WikiCLIP formulates VER as a retrieval task, providing top- k retrieval results to evaluate its effectiveness. As shown in the table, WikiCLIP achieves the highest HM@20 score of 67.8 on the OVEN benchmark, significantly outperforming CLIP I-I. On EVQA, our method surpasses Google Lens at $K = 20$, demonstrating that WikiCLIP can achieve surprisingly strong performance even with limited training data. Finally, on the InfoSeek benchmark, WikiCLIP achieves a recall@20 of 86.6, highlighting its effectiveness in real-world applications. These results validate the robustness and practicality of our approach for large-scale VER tasks.

A.4. How to choose the most representative entity image?

In the Vision-guided Knowledge Adaptor module, we use the most representative entity image to provide guidance to obtain the knowledge-aware entity representations. In practice, we use the lead image on the Wikipedia page as the most representative image. This selection is based on Wikipedia’s content policies: lead images should be natural and appropriate representations of the entity (in accordance with *Wikipedia:Manual of Style/Images*).

A.5. Visualization

Retrieval Results Visualization To provide an intuitive understanding of WikiCLIP’s effectiveness, we present qualitative visualization samples in Figure 2. The figure showcases the top-5 retrieval results for various queries, illustrating how WikiCLIP performs in visually ambiguous cases. As observed, the most challenging samples often involve entities with highly similar visual features, making fine-grained discrimination difficult. However, WikiCLIP successfully retrieves the correct ground-truth entity by leveraging textual descriptions, demonstrating its ability

Methods	HM@K (OVEN)				Recall@K (EVQA)				Recall@K (InfoSeek)			
	K=1	K=5	K=10	K=20	K=1	K=5	K=10	K=20	K=1	K=5	K=10	K=20
CLIP I-I	10.1	27.1	36.0	44.1	13.3	31.3	41.0	48.8	45.6	67.1	73.0	77.9
CLIP I-T	-	-	-	-	3.3	7.7	12.1	16.5	32.0	54.0	61.6	68.2
Google Lens	-	-	-	-	47.4	62.5	64.7	65.2	-	-	-	-
Echosight	-	-	-	-	36.5	47.9	48.8	48.8	53.2	74.0	77.4	77.9
WikiCLIP-S	31.1	53.7	61.3	67.8	30.7	53.6	62.5	69.1	61.2	76.8	81.8	86.4
WikiCLIP-L	31.6	53.3	60.5	67.4	31.9	53.3	61.2	69.0	62.7	77.7	82.2	86.6

Table 5. Performance comparison on open-domain VER benchmarks. HM@K for OVEN, Recall@K for EVQA and InfoSeek.

to incorporate semantic knowledge for precise entity recognition. These visualizations highlight the strength of our method in resolving challenging cases where purely visual matching would struggle.

Vision-guided Knowledge Selection Visualization To provide an intuitive understanding of the Vision-guided Knowledge Selection module, we present qualitative visualization samples in Figure 4. Specifically, we show the attention map of each text token when guided by patch-level vision signals. For the sake of clarity, we sum the attention from each image patch and highlight the top 32 text segments with higher attention. We circle the text segments that we believe are helpful for entity discrimination. As observed, the Vision-guided Knowledge Selection is able to detect the detailed discriminative features of entities.

Error Case Analysis To better understand the limitations of WikiCLIP, we analyzed its prediction errors on OVEN and identified three main types of failure cases, as illustrated in Figure 3. The most prevalent error type is Wrong but Relevant, where the predicted entity is semantically related to the ground truth but still incorrect. While our method leverages textual information to mitigate such errors, open-domain visual entity recognition remains a highly challenging task. The second type of error arises when the ground truth entity is not directly related to any entity present in the image. We attribute this to annotation noise in OVEN’s entity split and query split, which affects model performance. The final category of errors pertains to prediction granularity, where the predicted entity is at an incorrect level of specificity. We believe this issue is also a result of annotation inconsistencies in OVEN. Due to the inherent noise in dataset annotations, there is no straightforward solution to fully address this problem.

Visualization of Hard Negatives. To intuitively demonstrate the effectiveness of our hard negative synthesis strategy, we visualize the distribution of entity representations. Specifically, we randomly sample a query image and retrieve the top-5 closest entity representations with and with-

out hard negative synthesis. As shown in Figure 5, we apply t-SNE to project these representations into a 2D space. It can be observed that the entity representations trained with hard negative synthesis exhibit a more sparse and discriminative structure across different classes.

To quantitatively validate this observation, we compute the Silhouette Score, a widely-used metric that evaluates how similar a sample is to its own cluster (cohesion) compared to other clusters (separation). A higher Silhouette Score indicates more compact intra-class representations and better-separated inter-class representations. As shown in Figure 5, hard negative synthesis consistently leads to higher Silhouette Scores, confirming that it encourages more structured and distinguishable entity representation spaces.

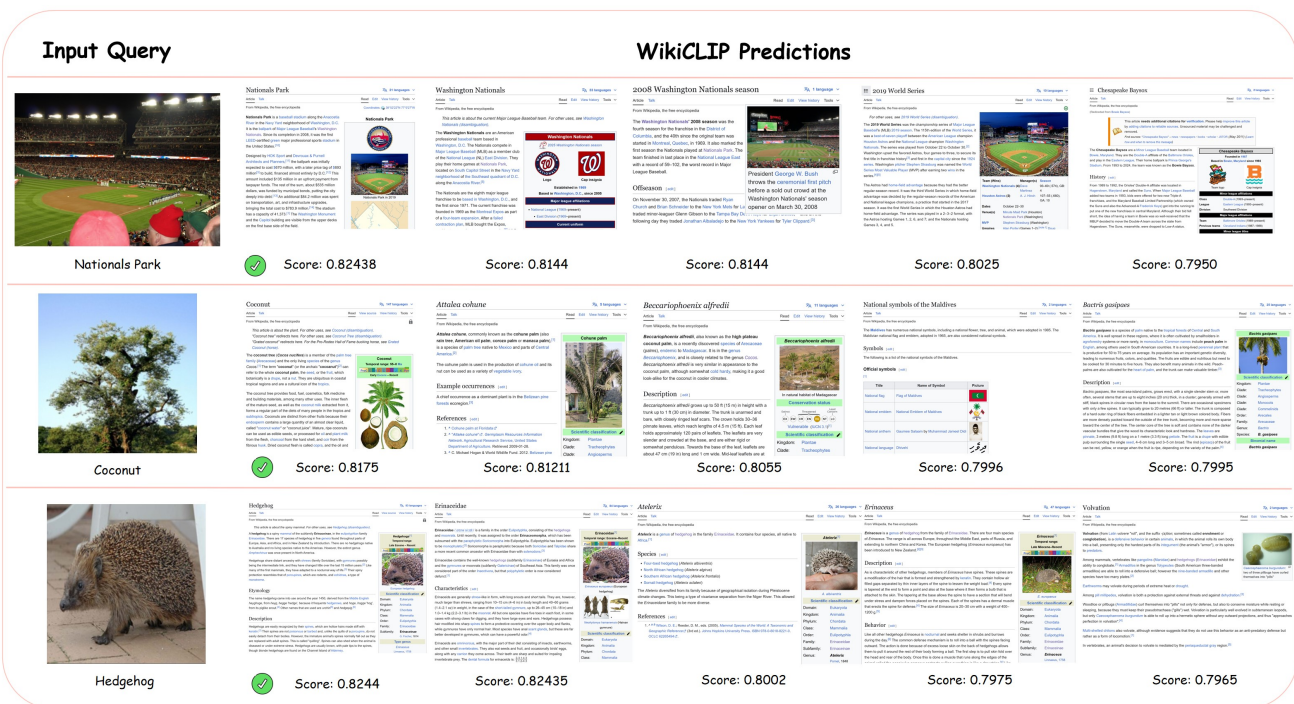


Figure 2. The Visualization of Topk prediction of WikiCLIP.

Error Type

(A) Wrong but Relevant

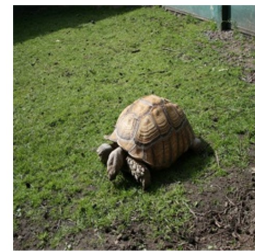


(B) Misunderstand Query



(C) Too Generic

Input query



Prediction

Brachycthon

"Botleee" redirects here. For other uses, see Botle tree (disambiguation).

Brachycthon (/brɑːkəˈtʃɒn/, *botle-tree*) is a genus of 31 species of trees and large shrubs, native to Australia (the centre of diversity, with 30 species) and New Guinea (one species). Fossils from New South Wales and New Zealand are estimated to be 50 million years old, corresponding to the Paleocene.

Description [edit]

They grow to 4 – 30m tall, and some are dry-season deciduous. Several species (though not all) are *parhyetal* plants with a very stout stem for their overall size, used to store water during periods of drought. The leaves show intraspecific variation and generally range from entire to deeply palmately lobed with long slender leaflet-like lobes joined only right at the base. Their sizes range from 4 – 20 cm long and wide.



Scientific classification [edit]

Kingdom: Plantae

Poinsettia

From Wikipedia, the free encyclopedia

For other uses, see Poinsettia (disambiguation).

The **poinsettia** (/pɔɪˈneɪtiə/ [ⓘ] [ⓘ]) (*Euphorbia pulcherrima*) is a commercially important flowering plant species of the diverse spurge family (Euphorbiaceae). Indigenous to Mexico and Central America, the poinsettia was first described by Europeans in 1834. It is particularly well known for its red and green foliage and is widely used in Christmas floral displays. It derives its common English name from Joel Roberts Poinsett, the first United States minister to Mexico, who is credited with introducing the plant to the US in the 1820s. Poinsettias are shrubs or small trees, with heights of 0.6 to 4 m (2 to 13.1 ft). Though often stated to be highly toxic, the poinsettia is not dangerous[[]] to pets or children. Exposure to the plant, even consumption, most often results in no effect[[]] though it can cause nausea, vomiting, or diarrhea.[[]]

Wild poinsettias occur from Mexico to southern Guatemala, growing on mid-elevation, Pacific-facing slopes. One population in the Mexican state of Guerrero is much further inland, however,



Boylan Bottling Company

This article needs additional citations for verification. Please help improve this article by adding citations to reliable sources. Unreliable material may be challenged and removed. Find sources: News Web books Google Scholar (December 2023) (Learn how and when to remove this message)

Boylan Bottling Company is an American gourmet soft drink manufacturer located in New York City. The company was founded in Paterson, New Jersey, in 1884.[[]] Boylan products are known for glass bottles with distinctive, retro style labels. The company's beverages use pure cane sugar and are bottled exclusively in glass.[[]]

History [edit]

Boylan's first product was Irish beer, their recipe having been formulated in 1891 in Paterson, New Jersey, by chemist Dr. George William Boylan.[[]] Ron and Mark Formica's grandfather bought the company from the Boylan family in the 1930s. It was located in Hudson, New Jersey, from the late 1930s until 2001, when its bottles were relocated to Cortez, New Jersey, for a short time[[]] before again being relocated to Morristown, New Jersey, then Teaneck, New Jersey, and, in 2013, New York City.

On September 10, 2023, Boylan was acquired by Fingert Bank, an organization with the private equity firm The Courtney Group. It represented the first transaction in the



Boylan Bottling Co.

Industry: Beverage
Genre: Carbonated Soft Drink
Founded: 1884, 154 years ago[[]]
Founder: George A. Boylan
Headquarters: New York, New York, United States
Products: Irish beer, Ginger Ale, Soft drink, **Seltzer**, **Mary's**
Website: boylanbottling.com/

Ginger ale

From Wikipedia, the free encyclopedia

See also: carbonated soft drink

Ginger ale is a carbonated soft drink flavored with ginger. It is consumed on its own or used as a mixer, often with spirit-based drinks. There are two main types of ginger ale: The golden style is credited to the Irish doctor Thomas Joseph Farrell. The dry style (also called the pale style), a pale drink with a much milder ginger flavor, was created by Canadian John McLaughlin.

History [edit]

Thomas Joseph Farrell, an Irish physician and surgeon, manufactured the first ginger ale in Ireland, around, in the 1850s. This was the older golden style fermented ginger ale, dark colored, generally sweet to taste, with a strong ginger spice flavor. *Drinking water* which he marketed through local beverage manufacturer Graham and Company[[]] Graham introduced the unique "Original Makers of Ginger Ale" on his bottles.[[]] Ginger ale is transparent, whereas ginger beer, a stronger tasting product, is often cloudy due to the presence of brewing.

Dry ginger ale was created by Canadian John J. McLaughlin, a chemist and pharmacist.[[]] Having established a soda water bottling plant in 1890, McLaughlin began developing *lemon* extracts to add to the water in 1894. That year, he introduced "Pine Dry Ginger Ale," the bubbly drink that would be patented in 1900 as "Canada Dry Ginger Ale." A success, Canada Dry products were awarded by appointment to the Vice-Royal Household of the Governor General of Canada. The



Ginger ale

Type: Non-alcoholic beverage Soft drink
Country of origin: Ireland and Canada
Region of origin: British Isles and Southern Oceania, Canada (1890) Quebec and USA (1914)
Introduced:

Tortoise

From Wikipedia, the free encyclopedia

This article is about the reptile. For other uses, see Tortoise (disambiguation).

Tortoise (/tɔːrtɔɪˈtɔːs/) are reptiles of the family **Testudinidae** of the order Testudines (Latin for "tortoise"). Like other turtles, tortoises have a shell to protect from predators and other threats. The shell in tortoises is generally hard, and the other members of the suborder Cryptodira, they retract their necks and heads directly backward into the shell to protect them.[[]]

Tortoises can vary in size with some species, such as the Galapagos giant tortoise, growing to more than 1.2 metres (3.9 ft) in length, whereas others like the Spurred tortoise have shells that measure only 6.8 centimetres (2.7 in) long.[[]] Several lineages of tortoises have independently evolved very large body sizes in excess of 100 kilograms (220 lb), including the Galapagos giant tortoise and the Aldabra giant tortoise. They are usually diurnal animals with tendencies to be crepuscular depending on the ambient temperatures. They are generally omnivorous animals. Tortoises are the longest-living land animals in the world, although the longest living species of tortoise is a matter of debate. Galapagos tortoises are noted to live over 100 years, but an Aldabra giant tortoise named *Adwaita* may have had an estimated 250 years. In general, most tortoise species can live 80–150 years. Tortoises are placid and slow-moving, with an average walking speed of 0.2–0.5 km/h.



Testudinidae

Temporal range: **Eocene–Recent**

Scientific classification [edit]

Domain: Eukaryota
Kingdom: Animalia
Phylum: Chordata
Class: Reptalia
Order: Testudines

Giant tortoise

From Wikipedia, the free encyclopedia

For other uses, see Giant tortoise (disambiguation).

Giant tortoises are any of several species of various large land tortoises, which include a number of extant species,[[]] as well as ten extant species with multiple subspecies formerly common on the islands of the western Indian Ocean and on the Galapagos Islands.[[]]

History [edit]

As of February 2024, two different species of giant tortoise are found on two remote groups of tropical islands: the Aldabra and the Îles Éparsées and the Galapagos Islands in Oceania. These tortoises can reach a weight as much as 417 kg (919 lb) and grow to be 1.3 m (4.3 ft) long. Giant tortoises originally made their way to islands from the mainland via oceanic dispersal. Tortoises are noted to be displaced by their ability to float with their heads up and to survive for up to six months without food or fresh water.[[]]

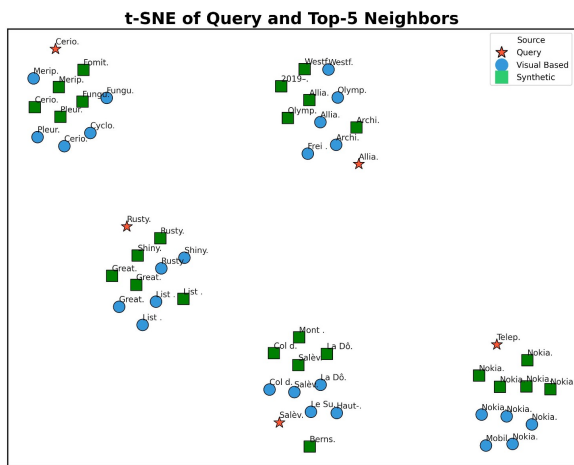
Giant tortoises were once all placed in a single genus (often referred to as *Testudo* or *Geochelone*), but more recent studies have shown that giant tortoises represent several distinct lineages that are not closely related to one another.[[]] These lineages appear to have developed large size independently and, as a result, giant tortoises are polyphyletic. For example, the African *Geochelone* giant tortoises are related to Malagasy tortoises (*Homopus*) while the Galapagos giant tortoises are related to South American mainland



Aldabra giant tortoise, an example of a giant tortoise

A Galapagos giant tortoise, *Geochelone* **Santa Cruz Island**

Figure 3. The Visualization of Error Case of WikiCLIP.



(a)



(b)

Figure 5. Visualization of Hard Negatives.