What does CLIP know about peeling a banana?

What does CLIP know about peeling a banana?

Claudia Cuttano[1], Gabriele Rosi[1,2], Gabriele Trivigno[1], Giuseppe Averta[1,2]
[1] Politecnico di Torino, [2] Focoos AI
[1] name.surname@polito.it, [2] name.surname@focoos.ai

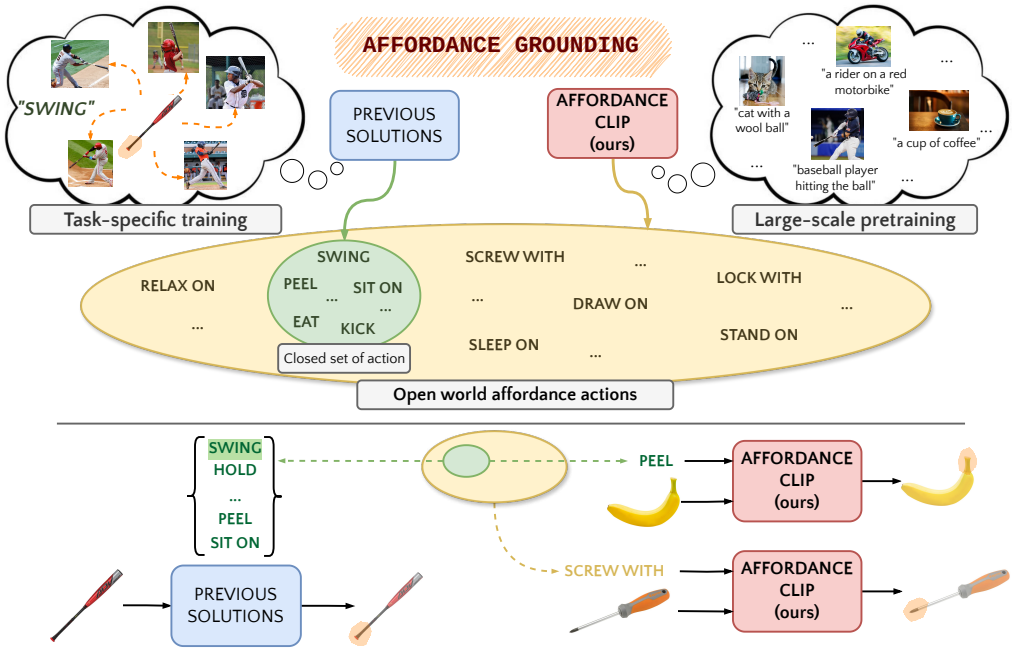


Figure 1. **Overview of AffordanceCLIP**. Our AffordanceCLIP unlocks the hidden affordance understanding capabilities within CLIP. Traditional techniques rely on task-specific supervised training, limiting them to a closed set of actions. Our key insight is that CLIP, instead, already embeds knowledge on how humans interact with objects, without the need for explicit finetuning. This enables open-vocabulary reasoning about a vast range of potential actions. Our open-vocabulary approach demonstrates promising performance in zero-shot, paving the way for broader and more flexible affordance understanding.

Abstract

Humans show an innate capability to identify tools to support specific actions. The association between objects parts and the actions they facilitate is usually named affordance*. Being able to segment objects parts depending on the tasks they afford is crucial to enable intelligent robots to use objects of daily living. Traditional supervised learning methods for affordance segmentation require costly pixel-level annotations, while weakly supervised approaches, though less demanding, still rely on object-interaction examples and support a closed set of actions. These limitations hinder scalability, may introduce biases, and usually restrict models to a limited set of predefined actions. This paper proposes AffordanceCLIP, to overcome these limitations by leveraging the implicit affordance knowledge embedded within large pre-trained Vision-Language models like CLIP. We experimentally demonstrate that CLIP, although not explicitly trained for affordances detection, retains valuable information for the task. Our AffordanceCLIP achieves competitive zero-shot performance compared to methods with specialized training, while offering several advantages: i) it works with any action prompt, not just a predefined set; ii) it requires training only a small number of additional parameters compared to existing solutions and iii) eliminates the need for direct supervision on action-object pairs, opening new perspectives for functionality-based reasoning of models.*

1. Introduction

Our daily lives are filled with objects and tools that we effortlessly manipulate to achieve goals. Our natural ability to link visual properties of an object (shape, material, and parts) with the actions it affords demonstrates a deep connection between perception and action. A concave shape, for instance, immediately suggests the ability to hold liquids, regardless of whether it is a cup or a coconut shell.

In artificial intelligence, the problem of associating functionality with objects is known as affordance grounding [2, 8]. It aims at locating the regions of an object that can be used to carry out a given action. To date, standard approaches [4, 6, 11, 24, 30] attempt to solve this problem with supervised learning, relying on manually annotated datasets to teach models about object functionalities. Each object in the picture should be provided with multiple segmentation masks, one for each "part" of the object associated with functional tasks [24, 28]. In the case of a glass, we may imagine to have the edge associated to the action *drink* and the handle with the action *hold*. This paradigm, while effective, presents practical limitations due to the resource-intensive nature of acquiring pixel-level annotations. Recognizing the need for more scalable and practical solutions, alternative techniques [13, 18, 19, 27, 34] formulated the affordance problem as a weakly supervised task, where the focus shifts towards learning object affordances through the observation of human-object interaction images [3]. For example, a model might learn how to *swing* a baseball bat after observing multiple images of humans grasping the bat (see Fig. 1, top-left).

Even though the annotations are simplified, we argue that weakly supervised approaches come with several limitations. First, these approaches mainly work with images representing a single object (*e.g*. a foreground baseball bat alone), limiting their use in real-world scenes with multiple objects. Additionally, they are trained on a closed-set of affordance actions (*i.e*. 36 on the popular AGD20K [18]), and cannot be used in an open vocabulary setting with arbitrary actions. Finally, in order to avoid introducing culture-dependent biases, they require a representative number of examples to learn from. For instance, the ways of carrying bags or chopping vegetables can be heavily influenced by the cultural habits [29].

In this paper we investigate whether it is possible to transfer affordance knowledge without direct supervision on a predefined set of classes. Our intuition relies on the observation that large pre-trained models may have already learnt how humans interact with functional objects by looking at millions of images. Even without datasets explicitly focused on affordances, these models potentially hold the key to identify affordances across a broader spectrum of actions than what can be achieved with annotated datasets. To assess the validity of this hypothesis, we experiment with CLIP [31], one of the most popular large Vision-Language model. However, unlocking affordance knowledge from CLIP is non-trivial, as it aligns the image representations with textual descriptions on a global level, discarding spatial information. This makes its embeddings unsuitable for the affordance grounding task, which requires to localize specific object details depending on the textual prompt.

Despite this, CLIP rich exposure to complex scenes and descriptive natural language suggests that it implicitly embeds local image semantics and concepts in its intermediate feature maps [42]. In this work, we address the challenge of extracting this latent affordance grounding knowledge in a zero-shot manner, *i.e*. without fine-tuning on datasets that explicitly focus on affordance localization task.

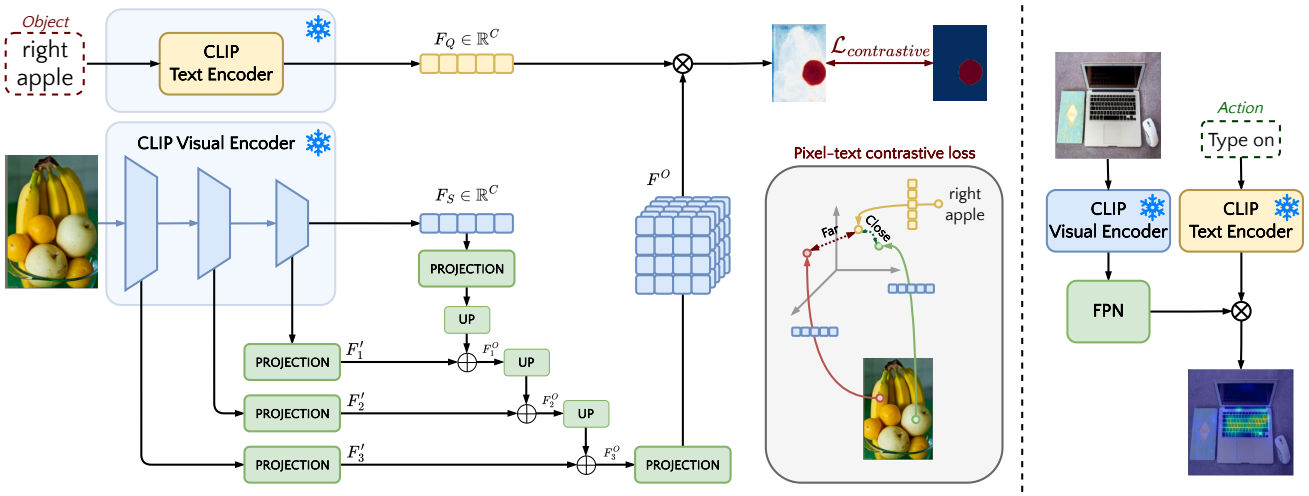
To this end, we start from a frozen CLIP model and we introduce a lightweight Feature Pyramid Network (FPN) [14], which gradually refines CLIP global descriptor with fine-grained spatial information from early layers of the visual encoder. To avoid introducing task-specific biases, we propose to train the FPN on the proxy task of referring image segmentation [36, 37, 39, 43], which provides binary masks of objects, referred by a textual prompt. Training on fine-grained segmentation masks exclusively for *objects*, our approach distills CLIP global understanding into pixel-level embeddings without direct action-affordance associations.

Our results demonstrate that our FPN enables the extraction of latent knowledge embedded in CLIP for zero-shot affordance grounding. We achieve competitive results w.r.t. existing supervised or weakly-supervised methods, with the additional benefits that: i) we don't need any sort of supervision on actions-objects pairs; ii) we are not bounded to a fixed set of actions and our method can work with open-vocabulary prompts; iii) our method introduces a very limited number of learnable parameters w.r.t. existing solutions. Summarizing, our contributions are the following:

- We demonstrate the feasibility to solve affordance segmentation without explicit (weakly-)supervised training;
- We showcase how large pre-trained Vision-Language models can naturally handle any action prompts;
- To adapt CLIP global descriptors to a dense task without finetuning, we propose a lightweight, low-overhead Feature Pyramid Network to extract multiscale, spatial features while retaining language-aligned embeddings.

2. Related works

Affordance Grounding The task of affordance grounding has gained increasing attention in computer vision, seeking to identify image regions that suggest potential interactions between humans and objects. Several works [4, 6, 11, 24, 30] have proposed to tackle the problem through supervised approaches, learning to identify relationship maps between local object regions and their as-



(a) Distill CLIP global understanding into pixel-level embeddings

(b) Inference

Figure 2. **Overview of the proposed AffordanceCLIP**. *Left*: We train a lightweight FPN to obtain dense feature maps from CLIP. Given an image, and a textual query referring an object, a frozen CLIP model extracts visual and linguistic features. Then, our FPN gradually refines the output visual vector with fine-grained spatial details, in order to retain both spatial information and local image semantics. Finally, a contrastive loss encourages pixel-level embeddings within the GT mask of the object to align with the corresponding linguistic features. *Right:* At inference, AffordanceCLIP can be directly queried with any textual prompt to obtain zero-shot affordance predictions.

sociated affordances. However, more recently there has been a consistent effort in searching alternative strategies to mitigate the challenges of collecting costly and extensive annotations. [34] introduced an innovative weakly supervised approach for affordance detection. By solving an Expectation-Maximization problem [5], their methodology relies on a sparse set of key points for weakly supervised affordance detection. [27], instead, proposes to use affordance labels only, to extract the interactions directly from videos. Notably, [18] annotates the first large-scale affordance dataset - AGD20K, with affordance/object categories and part-level annotations, serving as a benchmark for evaluating the efficacy of different methodologies. Existing weakly supervised object localization and affordance grounding methods [13, 18, 19] are mainly based on class activation mapping [41] (CAM). Unlike traditional methods, our solution avoids the requirement for task-specific, weakly-supervised data by utilizing the knowledge transferred to vision-language models during large-scale pre-training.

Dense prediction from Vision Language Models The shared visual-language embedding space learned from image-text pairs has enhanced open-world detection [16, 20, 23] and segmentation tasks [12, 17, 32, 36, 42]. LSeg [12] uses an image encoder trained on labeled segmentation data, which generates pixel-wise embeddings that align with the CLIP text embedding of the corresponding segmentation label. Fine-tuning methods like CRIS [36], CLIPSeg [17] and DenseCLIP [32] utilize an image decoder to create relevancy maps guided by CLIP text embeddings and the CLIP image encoder. However, the small datasets typically used for fine-tuning often limit the model's broader language understanding. MaskCLIP [42] extracts dense patch-level features from CLIP's image encoder without breaking the visual-language associations. Analogously, our approach directly leverages the multimodal representation learned by CLIP, without any finetuning of its original parameters.

3. Method

This research investigates the potential of CLIP, a powerful pre-trained multimodal model, to identify affordances of objects in an image (*i.e.* affordance grounding). Our framework leverages CLIP pre-trained image-language alignment, refining its output to obtain fine-grained spatial information for accurate localization.

We build on a frozen CLIP model to extract visual and language features, preserving its rich understanding of the relationship between images and text. However, as CLIP processes the image through a deep neural network, the output visual vector loses precise spatial information about where objects are located in relation to each other. Instead, intermediate feature maps extracted by the visual encoder retain both spatial information and local image semantics [42]. To recover this spatial information, we introduce a Feature Pyramid Network (FPN) [14] that operates on the

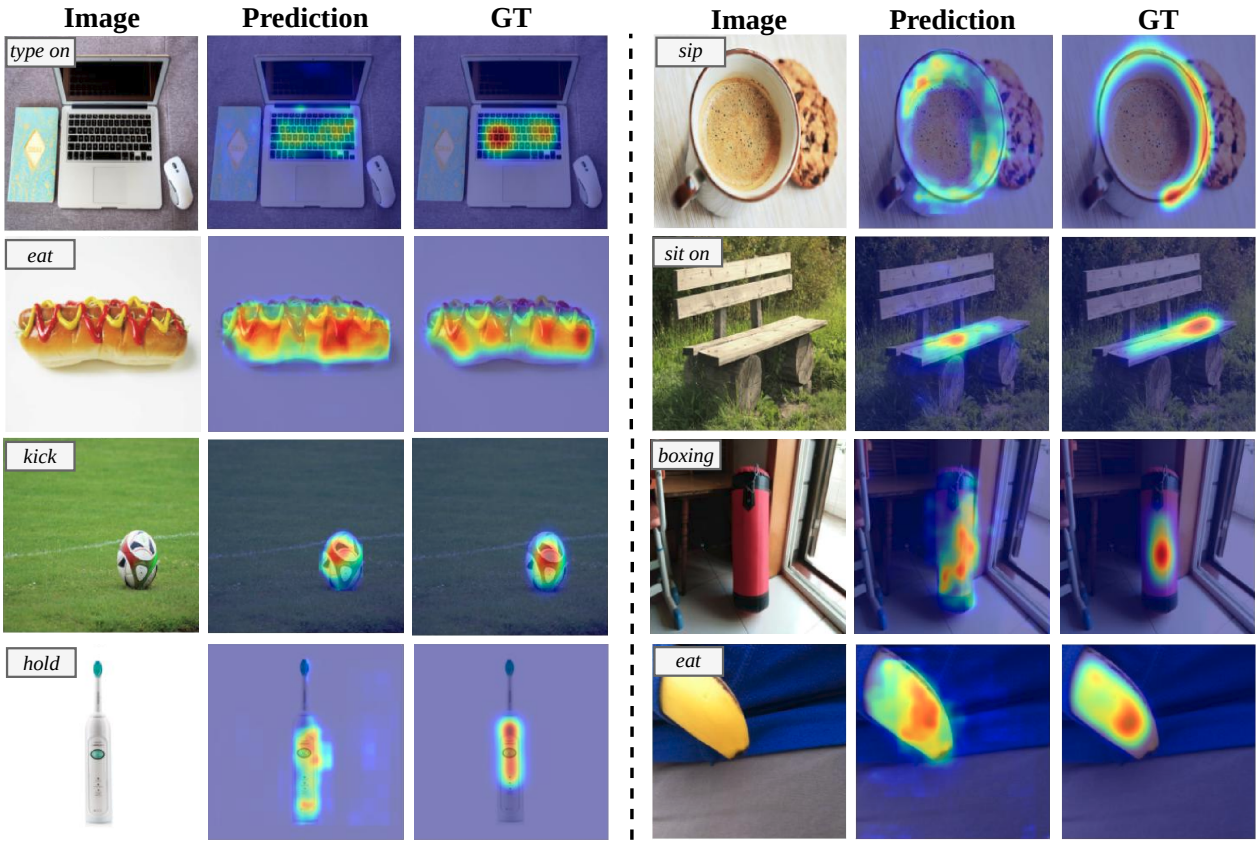


Figure 3. **Qualitative results**. Given an image and action, we show our model's prediction and the corresponding Ground Truth.

CLIP visual encoder at different depths. This FPN gradually integrates spatial details back into the global visual vector, allowing the model to recover crucial object localization information. Finally, inspired by CLIP original training, we introduce a contrastive learning objective to transfer CLIP image-level reasoning capabilities at the pixel level. The overall architecture is shown in Fig. 2.

3.1. Feature extraction

As feature extractor, we rely on the pre-trained backbone of CLIP to extract semantically aligned visual and textual representation for each input image and corresponding affordance expression.

Image encoder Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, we extract visual features from an image encoder. Specifically, we employ the frozen ResNet-101 [9] of CLIP [31] to obtain $F_S \in \mathbb{R}^C$, where C is the CLIP output dimension. This vector represents a compressed encoding of the image visual content. Additionally, we also consider the hierarchical feature volumes $F_1 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_1}$, $F_2 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_2}$ and $F_3 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C_3}$, where C_i represents the channel dimension at stage i and H and W are the height and the width, respectively. These features progressively encode higher-level abstractions of the image content.

Text encoder Given the textual query t, we extract the tokenized expression $T \in \mathbb{R}^L$, with L being the length of the expression. Note that the tokenization is obtained though lower-cased byte pair encoding (BPE) with 49152 vocabulary size and that the sequence is augmented by adding a global sentence representation token `[CLS]` and the end of sequence token `[EOS]`. A Transformer [35] modified by [31] processes T to extract the linguistic features $F_T^i \in \mathbb{R}^{L\times C}$, where C is the number of channels. The activation of the global contextual token `[CLS]` is further processed to generate the global textual representation $F_Q \in \mathbb{R}^C$.

3.2. Recovering spatial details

The output vector F_S of CLIP visual encoder captures the global context of the image but lacks the fine-grained details required for highlighting specific objects or regions. Pixel-level information is essential for the system to accurately identify contact points, to determine the relative positions of objects, and to analyze their orientations.

To this end, we introduce a lightweight Feature Pyramid

Network (FPN) that enriches CLIP output vector with detailed spatial information. Given the hierarchical nature of feature maps, where higher-context information is encoded progressively with lower spatial resolution, we propose to utilize the visual features extracted from the frozen image encoder (F_1, F_2, F_3) to augment the CLIP output vector F_S. Our approach involves an incremental fusion of lower-resolution features with higher-resolution ones, beginning with F_S.

To ensure compatibility across feature dimensions, both the global vector and the hierarchical visual features are first projected into a common representational space C'. Specifically, given F_S and the features F_i, we compute the projected features:

$$F'_S = \texttt{Conv}_{3\times3}(F_S) \tag{1}$$
$$F'_i = \texttt{Conv}_{3\times3}(F_i), \;\; i = 1, 2, 3 \tag{2}$$

where $\texttt{Conv}_{3\times3}$ denotes a convolution operation with kernel size 3, followed by a Batch Normalization and ReLU activation.

At this stage, we can directly execute the feature fusion operation. The global vector, F'_S, is initially fused with the lowest-resolution feature map, F'_3. The fusion process then continues by progressively integrating higher-resolution feature maps, F'_2 and F'_1. Formally, we compute the features F^O_i as:

$$F^O_1 = F'_1 + \texttt{Up}(F'_S) \tag{3}$$
$$F^O_i = F'_i + \texttt{Up}(F^O_{i-1}), \;\; i = 2, 3 \tag{4}$$

where $\texttt{Up}$ is a parameter-free upsampling operation that increases the resolution of F^O_{i-1} and F'_S.

To obtain the final visual feature, we project the C'-dimensional feature representation into the CLIP original feature space C. Formally, given the feature F^O_3, we compute the output feature $F^O \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C}$ as follows:

$$F^O = \texttt{Conv}_{1\times1}(F^O_3) \tag{5}$$

where $\texttt{Conv}_{1\times1}$ denotes a convolution operation with kernel size 1, followed by a Batch Normalization and ReLU activation.

3.3. Affordance Head

By keeping CLIP frozen, the visual-text alignment is preserved. While CLIP condenses the visual content into a single embedding aligned with a holistic description of the image, through our FPN we expand this representation to a higher resolution, in which individual pixels are semantically aligned with text. Hence, given the resulting visual features and a textual embedding, the corresponding activation map can be computed with a simple matrix multiplication.

Formally, the activation map $Y_{pred} \in \mathcal{R}^{H\times W}$ is obtained via matrix multiplication between the output of the text encoder (F_Q) and the output of the FPN (F^O):

$$Y_{pred} = F_Q \cdot (F^O)^T, \tag{6}$$

where T denotes the transpose operation.

3.4. Pixel-Text Contrastive Training

In its pre-training, CLIP employs a contrastive loss to learn a semantically rich joint representation space for images and their corresponding textual descriptions. The idea is to minimize the distance between the correct image-text associations (*i.e.* their embeddings are pushed closer in the shared space), while maximizing the distance of negative pairs in a batch of images. On the other hand, dense predictions tasks such as ours require pixel-level information to delineate the object referred by the query.

Our FPN is tasked with extracting this information from CLIP intermediate features. To do so, we apply the same concept of CLIP pre-training, and adopt a contrastive objective on our spatially augmented visual features, to distill CLIP global representation into pixel-level embeddings. Thus, we use a pixel-text contrastive loss [36, 37], to force the FPN to structure the resulting visual features in such a way that pixels referred by the query are precisely localizable (see Fig. 2).

Formally:

$$L^i_{con} = \begin{cases} -\log\sigma(Y^i_{pred}) & i \in \mathcal{P}, \\ -\log(1-\sigma(Y^i_{pred})), & i \in \mathcal{N}, \end{cases} \tag{7}$$

$$L_{con} = \frac{1}{|\mathcal{P}\cup\mathcal{N}|}\sum_{i\in\mathcal{P}\cup\mathcal{N}} L^i_{con}, \tag{8}$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the class of "1" and "0" in the ground truth, $|\mathcal{P}\cup\mathcal{N}|$ is the cardinality, σ is the sigmoid function.

4. Experiments

4.1. Datasets

Following [13], we evaluate AffordanceCLIP on AGD20K. To assess CLIP's zero-shot performance in affordance grounding, we strictly avoid using any form of supervision (fully or weakly) derived from this dataset. While CLIP encoders are kept frozen, the Feature Pyramid Network (FPN) requires training to bridge CLIP's image-level reasoning to pixel-level predictions. For this reason, we train the FPN exclusively on the RefCOCO/+/g dataset [22, 25, 40], a

State-of-the-Art from Relevant Tasks	Test A (*Seen*)			Test B (*Unseen*)		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Fully Supervised Affordance Grounding						
AffordanceLLM [30]	-	-	-	1.463	0.377	1.070
Weakly Supervised Object Localization*						
EIL [21]	1.931	0.285	0.522	2.167	0.227	0.330
SPA [33]	5.528	0.221	0.357	7.425	0.169	0.262
TS-CAM [7]	1.842	0.260	0.336	2.104	0.201	0.151
Weakly Supervised Affordance Grounding						
Hotspots [26]	1.773	0.278	0.615	1.994	0.237	0.577
Cross-view-AG [18]	1.538	0.334	0.927	1.787	0.285	0.829
Cross-view-AG+ [19]	1.489	0.342	0.981	1.765	0.279	0.882
Locate [13]	1.226	0.401	1.177	1.405	0.372	1.157
Zero-shot Affordance Grounding						
AffordanceCLIP	1.628	0.335	0.791	1.812	0.301	0.760

Table 1. Comparison to state-of-the-arts methods on AGD20K dataset. Following LOCATE [13], we include state-of-the-art methods from a relevant task - weakly supervised object localization. Results of * are taken from [18]. (↑/↓ means higher/lower is better).

popular benchmark for referring image segmentation. RefCOCO/+/g focuses on segmenting objects, described in natural language, rather than affordances. Note that the provided GT are binary masks, whereas in the downstream task the objective is to obtain continuous activation maps highlighting affordance regions.

AGD20K dataset [18] provides a collection of 20,061 images captured from a third-person perspective (exocentric) and 3,755 images from a first-person perspective (egocentric). These images are annotated with labels for 36 different affordances, which represent the potential objects interactions. The AGD20K dataset is designed to evaluate model performance under two settings: Seen and Unseen. In the Seen setting, the categories of objects in the training and test sets are identical. Conversely, the Unseen setting contains novel object categories during testing. Notably, this distinction only applies to methods from the weakly or fully supervised category. In our work, we do not use the any supervision from the affordance dataset and therefore, both splits represent unseen object categories for our model. To reflect this, we use a revised nomenclature: **Test A** corresponds to the original *Seen* setting (1675 images), while **Test B** corresponds to the original *Unseen* setting (540 images).

RefCOCO, **RefCOCO+**, and **RefCOCOg** [22, 25, 40] datasets are widely used benchmarks for evaluating object reference understanding in images. RefCOCO comprises 142,209 short (3.6 words on average) textual descriptions for 50,000 objects in 19,994 images. RefCOCO+ introduces a greater challenge with 141,564 descriptions focused purely on appearance-based referencing, deliberately excluding location words. RefCOCOg expands the scope with 104,560 longer (8.4 words average) and more complex referring expressions, derived using crowdsourcing through Amazon Mechanical Turk. These expressions reference 54,822 objects across 26,711 images.

4.2. Evaluation metrics

Following [13, 18, 19, 30], we evaluate our model in terms of Kullback-Leibler Divergence, Similarity and Normalized Scanpath Saliency.

Kullback-Leibler Divergence (KLD) metric quantifies the discrepancy between the predicted affordance distribution (M) and the ground truth distribution (M').

$$\mathrm{KLD}\left(M, M'\right) = \sum_i M'_i \log\left(\epsilon + \frac{M'_i}{\epsilon + M_i}\right), \quad (9)$$

Similiary (SIM) measures the intersection between the predicted affordance map (M) and the ground truth (M').

$$\mathrm{SIM}\left(M, M'\right) = \sum_i \min\left(M_i, M'_i\right), \quad (10)$$

where $\sum_i M_i = \sum_i M'_i = 1$.

Normalized Scanpath Saliency (NSS) measures the correspondence between the prediction map (M) and the ground truth (M').

Methods	Params (M)
EIL [21]	42.41
SPA [33]	69.28
TS-CAM [7]	85.86
Hotspots [26]	132.64
Cross-view-AG [18]	120.03
Cross-view-AG+ [19]	82.27
Locate [13]	6.50
AffordanceCLIP	**2.71**

Table 2. Comparison of learnable parameters.

$$\mathrm{NSS}\left(M, M'\right) = \frac{1}{N} \sum_i \hat{M} \times M'_i, \tag{11}$$

where $N = \sum_i M'_i$, $\hat{M} = \frac{M - \mu(M)}{\sigma(M)}$. $\mu\left(M\right)$ and $\sigma\left(M\right)$ are the mean and standard deviation, respectively.

4.3. Implementation Details

We initialize the text and image encoder with CLIP, adopting ResNet-101 as visual encoder. The FPN is trained for 1 epoch with a batch size of 32 on a combination of RefCOCO, RefCOCO and RefCOCO+ images. Input images are resized to 416×416, following [36, 37]. We use Adam optimizer with a learning rate of $\lambda = 0.0001$.

5. Results

5.1. Comparison with State-of-the-art

In order to provide a comprehensive benchmark, we consider state-of-the-art methods on the affordance grounding task under varying supervision levels: fully supervised and weakly supervised. Additionally, following LOCATE [13], we include results from state-of-the-art methods from the related task of weakly supervised object localization. Results are presented in Tab. 1.

Our results showcase the strong generalization capabilities of AffordanceCLIP on the affordance grounding task. This suggests that even though CLIP itself was not explicitly trained for this task, it has implicitly captured relevant visual features and their relationships to concepts, including actions and interactions. Remarkably, AffordanceCLIP outperforms Weakly Supervised Object Localization approaches, on both the test splits, and is competitive with the Affordance Grounding methods that leverage weakly supervised data. Furthermore, Tab. 2 emphasizes the efficiency of our model. By training only a lightweight Feature Pyramid Network on top of CLIP, we significantly reduce the number of trainable parameters compared to competing approaches. Notably, the FPN, with only 2.71 M parameters, effectively bridges the gap between CLIP's global image understanding and the pixel-level precision required for affordance grounding.

Method	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
$\{F_1^O\}$	1.917	0.322	0.665	2.171	0.278	0.586
$\{F_1^O, F_2^O\}$	1.892	0.329	0.726	2.038	0.297	0.696
$\{F_1^O, F_2^O, F_3^O\}$	**1.628**	**0.335**	**0.791**	**1.812**	**0.301**	**0.760**

Table 3. Contribution of different levels of spatial detail from CLIP's intermediate ResNet-101 features.

5.2. Ablation study

To analyze the contribution of different levels of spatial detail from CLIP's intermediate ResNet-101 [10] features, we conducted an ablation study. Tab. 3 summarizes the results, where we progressively integrate higher resolution feature maps into the FPN. Results demonstrate a consistent performance improvement as we integrate additional, more spatially detailed features. This suggests that each feature map provides valuable complementary information, enhancing the model's ability to perform accurate localization of affordance regions. This experimental evidence confirms the value of latent knowledge encoded within CLIP's intermediate representations.

5.3. Qualitative results

In Fig. 3, we present qualitative results that highlight the remarkable capabilities of AffordanceCLIP. These results demonstrate the model's capabilities in two key areas. First, it accurately localizes the target object within the image, successfully differentiating it from other visually similar or contextually related objects. Second, AffordanceCLIP precisely identifies the specific region within the object where the queried affordance can be performed. Consider the query *type on*: AffordanceCLIP is first of all able to discern between multiple objects in the image (the mouse, the laptop, the notebook) to identify the object associated with the action; then, it disambiguates within the regions of the object (the display, the keyboard, the touchpad) to identify the part with which we perform the action.

5.4. Open-Vocabulary capabilities

In Fig. 4 we qualitatively demonstrate the open-vocabulary capabilities of AffordanceCLIP, by testing our model with new actions outside those present in the AGD20K dataset. Results show that our adaptation of CLIP to dense predictions has not compromised its knowledge of open-world concepts. For example, it can be queried with *lock* or *draw on* without requiring additional human-object interaction images or model finetuning for these actions. Due to its pre-training on complex scenes, CLIP demonstrates robust performance on in-the-wild images from the Internet

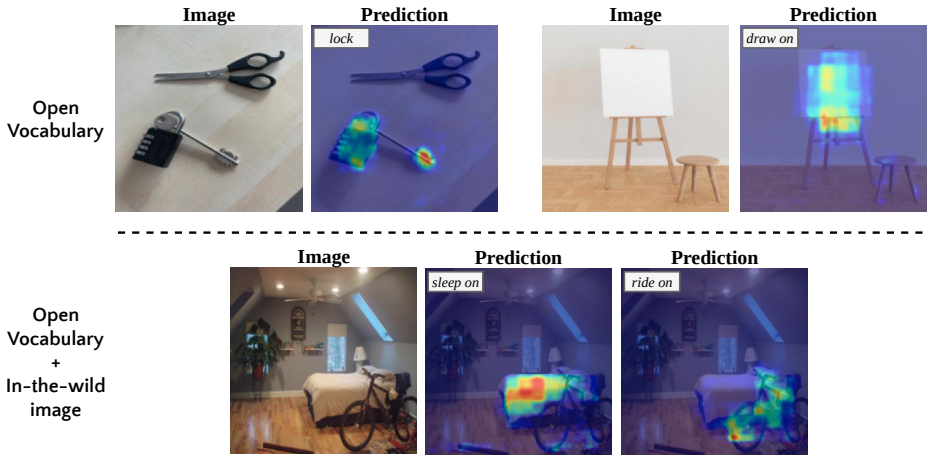


Figure 4. **Open Vocabulary capabilities.** *Top*: AffordanceCLIP is queried with actions outside the 36 of AGD20K dataset. *Bottom*: AffordanceCLIP is tested *in the wild*, on a challenging image from everyday settings.

featuring challenging, everyday settings. For example, in Fig. 4 (bottom), AffordanceCLIP successfully interprets a complex bedroom scene containing diverse objects and an unusual configuration (a bicycle near the bed). AffordanceCLIP is able to identify the bed when queried with *sleep on*, but also the bike if prompted with *ride on*.

6. Limitations

Despite AffordanceCLIP's impressive zero-shot affordance grounding abilities, it does have limitations. Fig. 5 highlights some interesting scenarios where it fails. In one case, when asked to identify a region to *write on*, the model focuses on the pencil tip rather than the part of the pencil we grasp with our hand. This suggests that CLIP has strongly associated the concept of writing with the tool used for the action, rather than the part of the object directly manipulated. Another interesting failure occurs when prompted with *ride*. AffordanceCLIP correctly locates the bike but excludes the bike seat. This weaker association between the seat and the concept of riding may be due to the fact that in many images used to train CLIP, the seat is often occluded by the rider.

7. Conclusions and future works

In this work, we explored an alternative approach to affordance grounding. We move away from traditional weakly-supervised learning methods and instead leverage the implicit knowledge within visual language models (like CLIP) to identify object activation regions based on action prompts. We believe that our work paves the way for future research towards open-vocabulary affordance grounding. These promising results highlight the potential of exploring more advanced Vision-Language models, like LLaVA [15], Flamingo [1] and GPT-4V [38]. In particular, these

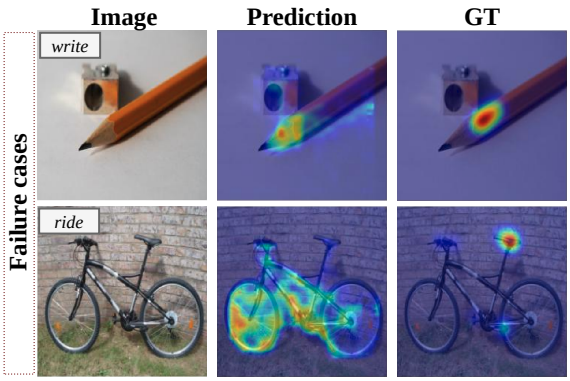


Figure 5. Examples of failure cases.

models are able to answer questions which require a deeper understanding of objects and their relationship to abstract concepts. This level of reasoning is essential in affordance grounding when dealing with complex images and queries, which demand taking into account object properties beyond mere geometric shape (*e.g.* material, inertial parameters) to associate them to functionalities.

Acknowledgements This study was carried out within the Sustainable Mobility Center (CNMS) and received funding from the European Union Next Generation EU (Piano Nazionale di Ripresa e Resilienza (PNRR), Missione 4 Componente 2 Investimento 1.4 "Potenziamento strutture di ricerca e creazione di "campioni nazionali di R&S" su alcune Key Enabling Technologies") with grant agreement no. CN_00000023. We also acknowledge FAIR - Future Artificial Intelligence Research which received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 8

[2] Paola Ardón, Èric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400*, 2020. 2

[3] Christopher J Burke, Philippe N Tobler, Michelle Baddeley, and Wolfram Schultz. Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, 107(32):14431–14436, 2010. 2

[4] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018. 2

[5] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977. 3

[6] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018. 2

[7] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *ICCV*, 2021. 6, 7

[8] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, 54(3):1–35, 2021. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[11] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013. 2

[12] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 3

[13] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *CVPR*, 2023. 2, 3, 5, 6, 7

[14] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 8

[16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[17] Timo Lüddecke and Alexander S Ecker. Image segmentation using text and image prompts. in 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7076–7086, 2021. 3

[18] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *CVPR*, 2022. 2, 3, 6, 7

[19] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *International Journal of Computer Vision*, pages 1–25, 2023. 2, 3, 6, 7

[20] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[21] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *CVPR*, 2020. 6, 7

[22] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 5, 6

[23] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 3

[24] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015. 2

[25] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 5, 6

[26] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *ICCV*, 2019. 6, 7

[27] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 2, 3

[28] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE, 2017. 2

[29] Chiara Plizzari, Toby Perrett, Barbara Caputo, and Dima Damen. What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13656–13666, 2023. 2

[30] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. *arXiv preprint arXiv:2401.06341*, 2024. 2, 6

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[32] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022. 3

[33] Johann Sawatzky and Jurgen Gall. Adaptive binarization for weakly supervised affordance segmentation. In *ICCV*, 2017. 6, 7

[34] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2795–2804, 2017. 2, 3

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4

[36] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022. 2, 3, 5, 7

[37] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *ICCV*, 2023. 2, 5, 7

[38] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 8

[39] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 2

[40] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 5, 6

[41] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

[42] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 2, 3

[43] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022. 2