The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Reverse Region-to-Entity Annotation for Pixel-Level Visual Entity Linking

Zhengfei Xu[1], Sijia Zhao[1], Yanchao Hao[2], Xiaolong Liu[2],
Lili Li[2], Yuyang Yin[2], Bo Li[2], Xi Chen[2], Xin Xin[1]*

[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]Platform and Content Group, Tencent, Beijing, China
{zhengfei, sijiazhao, xxin}@bit.edu.cn, {marshao, loongliu, irinali, yuyangyin, ryanbli, jasonxchen}@tencent.com

Abstract

Visual Entity Linking (VEL) is a crucial task for achieving fine-grained visual understanding, matching objects within images (visual mentions) to entities in a knowledge base. Previous VEL tasks rely on textual inputs, but writing queries for complex scenes can be challenging. Visual inputs like clicks or bounding boxes offer a more convenient alternative. Therefore, we propose a new task, Pixel-Level Visual Entity Linking (PL-VEL), which uses pixel masks from visual inputs to refer to objects, supplementing reference methods for VEL. To facilitate research on this task, we have constructed the MaskOVEN-Wiki dataset through an entirely automatic reverse region-entity annotation framework. This dataset contains over 5 million annotations aligning pixel-level regions with entity-level labels, which will advance visual understanding towards fine-grained. Moreover, as pixel masks correspond to semantic regions in an image, we enhance previous patch-interacted attention with region-interacted attention by a visual semantic tokenization approach. Manual evaluation results indicate that the reverse annotation framework achieved a 94.8% annotation success rate. Experimental results show that models trained on this dataset improved accuracy by 18 points compared to zero-shot models. Additionally, the semantic tokenization method achieved a 5-point accuracy improvement over the trained baseline.

Datasets — https://github.com/NP-NET-research/PL-VEL

Introduction

Visual Entity Linking (VEL) is an open-domain visual entity recognition task that expands the label space to web-scale knowledge bases. As a key task for achieving fine-grained visual understanding, VEL contributes to various tasks such as multimodal knowledge graphs completion (Wu et al. 2023), visual question answering (VQA) (Qiu et al. 2024), image caption (Zhang et al. 2024c), image retrieval (Sain et al. 2023; Saito et al. 2023) and so on.

Current VEL tasks (Hu et al. 2023; Caron et al. 2024a) relying on textual queries struggle with some complex scenes. For example, in fig. 1, a simple query like *what is on the plate?* cannot accurately refer to `Broccoli`, requiring more complex queries, such as *what is the small tree-like vegetable next to the fries on the plate?* Creating such queries demands extensive background knowledge and precise comprehension of visual relationships. This adds an additional burden on users, and we cannot assume that downstream models are equipped with such capabilities.

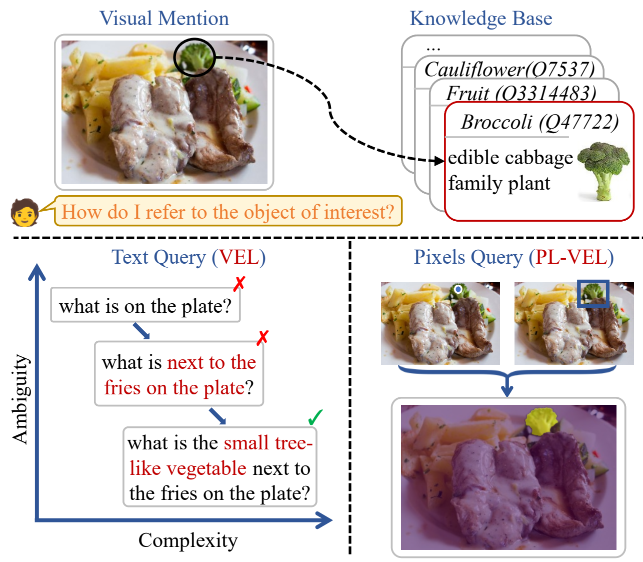


Figure 1: Overview of comparing text and pixel-based Visual Entity Linking (VEL) tasks

In such complex scenes, visual prompts such as clicks, boxes, and pixel masks can be supplementary methods for efficient and accurate reference. Therefore, this work introduces Pixel-Level Visual Entity Linking (PL-VEL), which uses pixel masks to refer to visual mentions and link them to knowledge-base entities, as shown in fig. 1. With promptable segmentation models like SAM (Kirillov et al. 2023) and SEEM (Zou et al. 2023), users or downstream models can create pixel masks through actions such as clicking, and drawing boxes. It makes PL-VEL more practical than traditional VEL tasks in real-world applications, such as VQA(Qiu et al. 2024) and visual reasoning (Chen and Wu 2024). To support the research on this task, an open-domain PL-VEL dataset that aligns pixel-level mask regions in images with entities in a knowledge base is required.

*Xin Xin (xxin@bit.edu.cn) is the corresponding author.

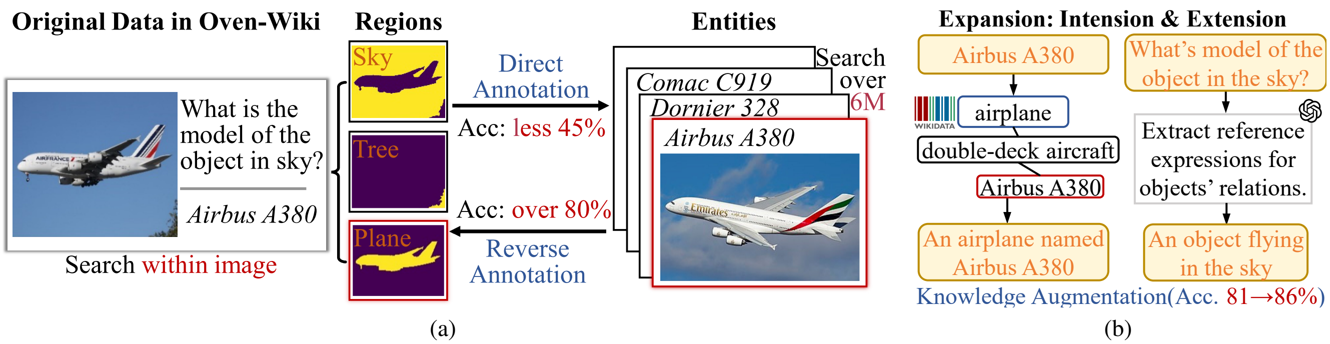


Figure 2: Overview of the annotation framework. (a) Comparison of direct and reverse annotation shows that direct annotation struggles to utilize existing entity labels effectively, whereas reverse annotation efficiently reduces the search space. (b) Knowledge-enhanced text prompt for segmentation models, built on intensional and extensional expansion.

The straightforward approach to constructing the dataset follows the VEL setup, **mapping visual objects to entities** by segmenting everything in the images and mapping each region to its corresponding entity. However, using GPT-4V to generate the entity names and searching within 6M entities achieves only about 25% accuracy (Xiao et al. 2024). Even powerful text-based VEL model AUTOVER-13B (Xiao et al. 2024) reaches only around 45% accuracy, making constructing a high-quality dataset challenging.

To address these challenges, we adopt a reverse approach by **mapping entities to visual objects**, as illustrated in fig. 2(a). We constructed the MaskOVEN-Wiki dataset based on the existing OVEN-Wiki dataset (Hu et al. 2023), which aligns entities with images. By segmenting pixel regions based on entity labels, we provide pixel references for visual mentions. This reverse annotation method leverages existing labels and reduces the search space from millions of entities to image regions. Pre-experiments with the segmentation pipeline model Grounded-SAM (Ren et al. 2024) show an annotation accuracy of approximately 80%.

Despite this, understanding long-tail entities remains challenging for segmentation models. We introduce a two-part knowledge augmentation method to improve annotation quality, as shown in fig. 2(b). For **intensional expansion**, we retrieve hypernyms from Wikidata[1] to provide broader semantics for entities in queries. For **extensional expansion**, we use GPT-3.5 to extract referring expressions from the original text questions that contain spatial or semantic relationships. This augmentation improves annotation accuracy from 81% to 86%. To address error propagation in the segmentation pipeline, we adopt the end-to-end model SEEM (Zou et al. 2023). We employ a model ensemble and heuristic rules to filter and correct low-quality annotations, thereby achieving an accuracy of approximately 95%. Finally, we developed a PL-VEL dataset with 5M visual mentions.

The PL-VEL task is more challenging than existing VEL tasks because it does not rely on textual queries with strong prior. To enhance visual feature utilization and region-interacted attention, we propose a visual semantic tokenization method based on Osprey (Yuan et al. 2023). Our approach produces more independent and complete image tokens than the fixed-size image patch sequence in ViT (Dosovitskiy et al. 2020). Experiments show our method improves model accuracy by about 5 points.

In summary, our main contributions are as follows:

- We introduce the PL-VEL task and construct MaskOVEN-Wiki, a large-scale dataset aligning pixel-level regions with entity-level labels.
- We design a reverse annotation framework that achieves 94.8% annotation accuracy through knowledge augmentation and model ensemble.
- We establish a PL-VEL baseline, achieving an accuracy improvement from 1.3% to 25.2% by fine-tuning on MaskOVEN-Wiki.

Related Work

Visual Entity Linking. Previous studies, such as Tag2Text (Huang et al. 2024) and RAM (Zhang et al. 2024b), generated common category tags for images but failed to recognize entity-level tags. To address this, OVEN-Wiki (Hu et al. 2023) was proposed as an open-domain visual entity linking benchmark, which links regions of interest to 6M Wikipedia[2] entities based on text queries. This benchmark also validated the effectiveness of the generative entity recognition framework (GER). Building on this, GER-ALD (Caron et al. 2024b) demonstrated that unAmbiguous Language-based Discriminative (ALD) entity codes offer a performance advantage within the GER framework. AUTOVER (Xiao et al. 2024) achieved an accuracy 11.9 points higher than GER-ALD on the OVEN-Wiki test set through retrieval-augmented constrained decoding.

In contrast to text-based references, Wikiperson (Sun et al. 2022), a VEL dataset using bounding box references, was introduced. However, Wikiperson is limited to "person" entities and is limited in scale. To address this, we propose an open-domain PL-VEL task, for advancing fine-grained visual understanding.

[1]https://www.wikidata.org

[2]https://www.wikipedia.org/

Region-specific Visual Understanding. It focuses on semantic information in local image regions, including region-specific conversation (Rasheed et al. 2024), region captioning (Yuan et al. 2023), and referring expressions comprehension (Guo et al. 2024). Our PL-VEL is also a region-specific recognition task. Recent works on region-specific visual understanding focus on MLLMs. Although MLLMs like BLIP (Li et al. 2022), LLaVA (Liu et al. 2023a), and MiniGPT-4 (Zhu et al. 2023) extend LLMs' capabilities to vision. However, they struggle to comprehend effectively specific visual regions. Kosmos-2 (Peng et al. 2023) and Shikra (Chen et al. 2023) input bounding boxes as location-aware reference tokens into LLMs, while GPT4RoI (Zhang et al. 2024a) and GlaMM (Rasheed et al. 2024) use specialized visual modules for bounding box regions.

These models, however, cannot describe pixel-level features accurately. Osprey (Yuan et al. 2023) achieves pixel-level understanding with a mask-aware visual extractor. Expanding on this, we introduce cross-attention interactions of pixel-level features and train the model on MaskOVEN-Wiki to enhance pixel-level visual understanding and provide a baseline for PL-VEL.

Pixel-Level Visual Entity Linking Task

Task Definition

Original Task (PL-VEL) *The PL-VEL task takes an image I and a pixel mask m as input. The pixel mask m represents a visual object in I, referred to as a visual mention V^m. The goal of PL-VEL is to link this visual mention V^m to its corresponding entity e in the knowledge base $\mathcal{K}$.*

Reverse Annotation (Dataset Construction) *The dataset construction task is the reverse process of the PL-VEL task. Given an entity e, an image I containing e, and a text query q for e, it takes them as input, and its goal is to segment the pixel mask m of the visual object of the entity e in I.*

The PL-VEL task assumes that mask references for visual mentions are provided. Various visual and textual prompts can be processed into pixel masks using preprocessing models such as SAM (Kirillov et al. 2023) and SEEM (Zou et al. 2023). This integration enhances the PL-VEL system's adaptability and supports interactive and fine-grained visual entity comprehension.

The MaskOVEN-Wiki Dataset Construction

To define and address the PL-VEL task, we have developed the MaskOVEN-Wiki dataset, a benchmark with approximately 5 million annotations, covering various categories of entities. Each annotation includes an image, a visual mention represented by a pixel mask, a text query, and the corresponding entity label from Wikipedia.

For the source of data, we use an open-domain entity recognition dataset, OVEN-Wiki (Hu et al. 2023), where each sample includes an image, a text query for visual mention and its corresponding entity. This dataset uses a 6 million-entity set derived from Wikipedia. The dataset aggregates 14 existing datasets and is divided into two subsets based on the original tasks of the source datasets. **entity split** (ES) for image recognition/retrieval and **query split** (QS) for visual question answering. Additionally, OVEN-Wiki provides a high-quality evaluation dataset, the **human set**, which is manually annotated. Based on this data, we developed and employed an automated method to annotate pixel-mask visual references for visual mentions in those three subsets. Additionally, we enriched it by annotating visual mentions for entities with images on Wikipedia pages. This additional content serves as a supplement to the knowledge base, referred to as **wiki split** (WS).

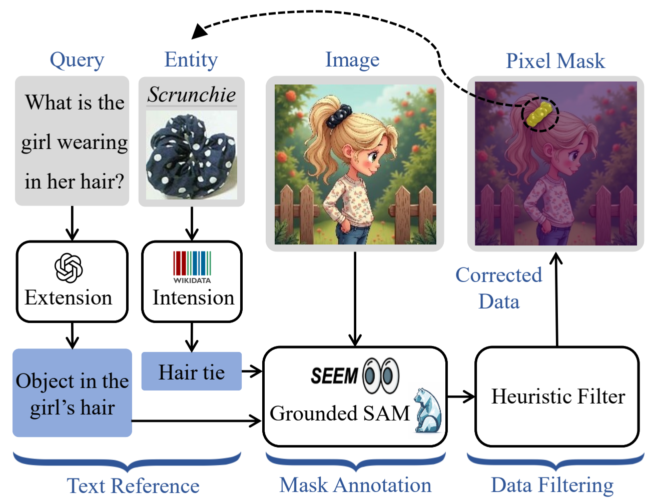


Figure 3: The procedure of building MaskOVEN-Wiki. The illustration image is generated by AI (Chang et al. 2024).

As illustrated in fig. 3, we have developed a knowledge-enhanced methodology for segmentation annotation. This workflow consists of three steps: text reference construction, mask annotation, and data filtering. For automated pixel-mask annotation, we utilize Grounded-SAM (Ren et al. 2024) and SEEM (Zou et al. 2023), which generate pixel masks from textual references. Finally, we apply a heuristic rule-based filter to remove or correct noisy data.

Text Reference Construction. We construct text references for each visual mention to guide the annotation model. A straightforward approach is to directly input the entity label and text query into the segmentation annotator, but this approach has limitations. Specifically, long-tail entities challenge the annotator's generalization performance. Therefore, we propose a two-part knowledge augmentation method to enhance the text reference.

For intensional description, we enrich the intensional description of entity labels by querying Wikidata. Specifically, we retrieve super-categories of the entity using two properties: `Instance of (P31)` and `Subclass of (P279)`. These super-categories are then combined with the original entity information through predefined templates to generate intension-enhanced textual references.

For extensional relations, we leverage relationships between objects to resolve ambiguities. Such relationships are often encoded in the text queries in OVEN-Wiki (Hu et al. 2023), for example, `What is the brown item on` `the chair facing the camera?`. We use GPT-3.5 to analyze these queries and extract expressions describing the mention's spatial or relational context. This process generates extension-enhanced references, such as `the brown item on the chair facing the camera`.

Mask Annotation. We utilize two open vocabulary segmentation models, Grounded-SAM (Ren et al. 2024) and SEEM (Zou et al. 2023), for annotating masks with textual references. Grounded-SAM, as a pipeline tool, initially employs Grounding-DINO (Liu et al. 2023c) to identify bounding boxes based on text prompt, followed by the SAM (Kirillov et al. 2023) for segmentation. This pipeline achieves a labeling success rate of 81.4% in preliminary experiments, forming the foundation of our solution. On the other hand, SEEM, as an end-to-end model, is good at processing diverse inputs. We utilize it as a complementary strategy to mitigate potential error propagation in Grounded-SAM's annotation process.

Data Filtering. Upon analyzing the results, we have identified four primary issues: reference to non-visual entities, error propagation in the segmentation pipeline, incomplete entity depiction in images, and foreground-background confusion in dense object scenes. To improve the annotation quality, we have applied heuristic filtering rules, as follows:

- For non-visual entities, we deleted the annotations of specific entities, such as events, technology, games, chart reasoning, and so on.
- For error propagation in the pipeline, we identify and correct potential errors by analyzing the agreement between different types of reference and segmentation models using intersection over union (IOU) metrics. IOU values indicate potential errors, we correct these by sampling the most confident bounding box using the intersection with segmentation results.
- For incomplete entity depiction, we found that such errors mainly occur in location entities. To address this, we apply a confidence threshold constraint specifically for location entities and treat the entire image as the corrected mask.
- For foreground-background confusion, we found that such errors mainly occur in dense object scenes. To mitigate this, we employ a rule-based correction using morphological operations. When multiple bounding boxes of the same type cover a significant portion of the image, we apply erosion and dilation to the predicted mask. We then analyze the number of connected components to judge this error and invert the mask for correction.

The MaskOVEN-Wiki Dataset Analysis

Annotation Quality. To evaluate the efficacy of our annotation method, we randomly sampled 2,000 annotations for manual inspection. To ensure diversity, we limited each entity to a maximum of one sample and proportionally allocated samples from the entity, query, and wiki splits. As shown in table 1, our knowledge-enhanced text references and model-ensemble heuristic filtering rules improved the annotation accuracy from 81% to 95%.

Reference		Model	
		G-SAM	SEEM
original query	# entity label	81.4	69.3
	# text query	71.3	62.8
knowledge augmentation	# intension	86.1	65.5
	# extension	83.0	67.5
overall after filtering		94.8	

Table 1: Annotation accuracy under different settings.

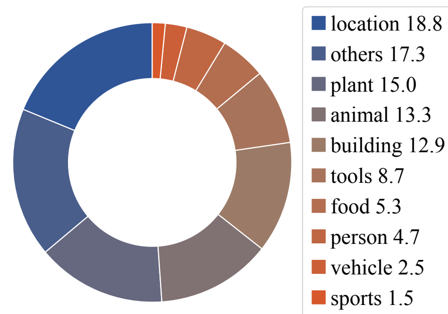


Figure 4: Entity category distribution in the evaluation set.

Figure 4 shows the distribution of entity types in this sample set. Compared to fig. 5(a), the entity category distribution of the sample set is similar but more balanced.

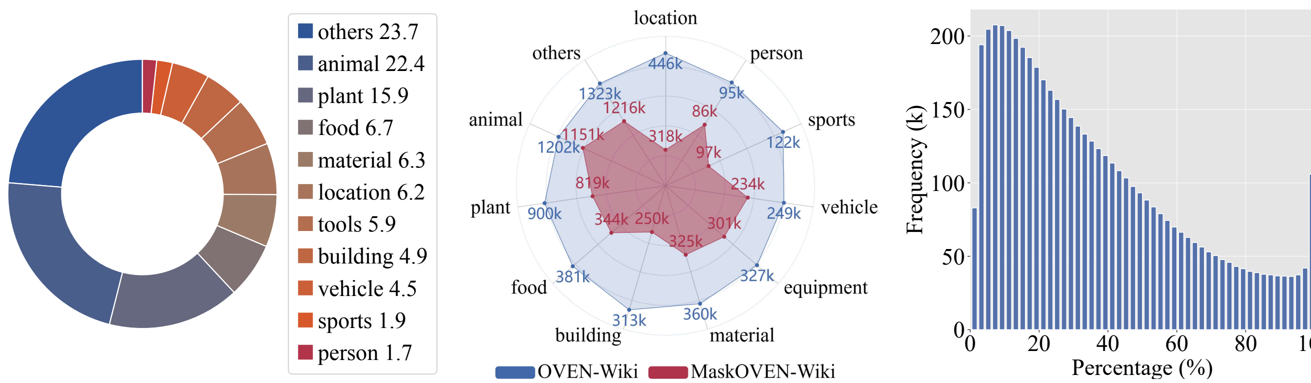
Table 2 summarizes the statistics of MaskOVEN-Wiki dataset. Our dataset contains 5,245,421 annotations for 5,214,965 images from OVEN-Wiki (Hu et al. 2023) dataset covering 20,077 entities. We reused the knowledge base of OVEN-Wiki, which contains 6,063,945 Wikipedia entities, of which 2,032,340 entities have a corresponding image.

Entity Distribution. Figure 5(a) shows the distribution of categories in the MaskOVEN-Wiki dataset. We have identified 10 primary categories and grouped less prevalent categories under the 'others' category. Figure 5(b) shows a more detailed distribution with numbers of each category both in the OVEN-Wiki and MaskOVEN-Wiki. As shown in fig. 5(b), we note that the highest proportion of unannotated entities is found in the location, building, and sports categories. Entities in these categories may hit the first and third data filtering rules and be dropped.

Visual Mention Distribution. Figure 5(c) shows a histogram of the area ratio of visual mentions in images, computed as a_m/a_i, where a_m and a_i represent the area of the mention and the image, respectively. The distribution exhibits a generally smooth profile, with an increase in frequency when the area ratio surpasses 95%, which is primarily caused by the third filtering rule.

	Train Set		Val Set		Test Set		Wiki Set	Human Set
	Entity	Query	Entity	Query	Entity	Query		
# SEEN entities	7,943	2,470	1,604	199	7,943	2,339	8,733	2,015
# SEEN examples	4,464,748	23,514	51,906	588	291,327	7,460	8,733	12,057
# UNSEEN entities	0	0	1,588	433	7,944	3,096	1,956,412	2,429
# UNSEEN examples	0	0	56,549	1,406	316,817	7,979	1,956,412	11,100
# Total examples	4,464,748	23,514	108,455	1,964	608,144	15,439	1,965,145	23,157

Table 2: Statistics of the MaskOVEN-Wiki.



(a) (b) (c)

Figure 5: Distribution of MaskOVEN-Wiki: (a) distribution of entity categories; (b) comparison of the entity category distribution between MaskOVEN-Wiki and OVEN-Wiki; (c) distribution of mask ratios for visual mentions in images.

Method

Model Architecture

Figure 6 illustrates our model overview. We employ visual instruction tuning to train the MLLM in autoregressively decoding the pre-constructed target entity ALD code. Following the generative entity recognition framework of GER-ALD (Caron et al. 2024b), we construct the ALD code for entity $e \in \mathcal{K}$ as

$$\mathrm{ALD}_e = \mathcal{S}^L\big(\mathcal{T}^T(e), \bigcup_{e_i \in \mathcal{K}} \mathcal{T}^T(e_i)\big) \tag{1}$$

Where $\mathcal{T}^T$ is the text tokenizer of LLM, and $\mathcal{S}^L$ denotes a function taking the first L tokens in ascending order of term frequency. L denotes the ALD code length. LLM autoregressively generates ALD_e with embedding matrix $\mathbf{Y}$, instruction $\mathbf{X}_{ins}$, image I's features $\mathbf{X}_I$ and mask query embedding $\mathbf{X}_m$ as follows

$$\mathrm{ALD}_i^{\hat{e}} = \mathrm{LLM}\big(\mathbf{X}_{ins}, \mathbf{X}_I, \mathbf{X}_m, \mathbf{Y}_{\mathrm{ALD}^{\hat{e}}_{0\leq j<i}}\big) \tag{2}$$

Our backbone is based on Osprey(Yuan et al. 2023), a pixel-level MLLM designed for general visual understanding. Following Osprey's settings, we employ the ConvNeXt CLIP (Liu et al. 2022) as the vision encoder, Vicuna (Chiang et al. 2023) as the foundational LLM, and a vision-language projector using a multilayer perceptron (MLP). Additionally, we reuse its mask-aware visual extractor for constructing regional-level features.

Our method utilizes visual semantic tokenization to extract the fine-grained semantic features from images. It achieves this by reusing feature maps from the vision encoder and parameters from the mask-aware visual extractor, enabling minimal computational and parameter overhead.

Visual Semantic Tokenization for Region-Interacted Attention

Current MLLMs (Liu et al. 2023b; Yuan et al. 2023) use vision encoders like ViT (Dosovitskiy et al. 2020) or ResNet (He et al. 2016). These encoders tokenize images based on spatial location rather than semantic content, so that the visual tokens contain incomplete and non-independent semantics, and require additional cross-modal projectors. While Osprey (Yuan et al. 2023) and GLaMM (Rasheed et al. 2023) use region encoders to represent user-specified regions, they do not enhance overall image understanding. PL-VEL focuses on pixel-level visual understanding, motivating us to tokenize images based on semantic content. This approach aligns the semantic granularity of image tokens with the instruction or entity text tokens by controlling each visual token to represent an object, enabling feature interaction within a unified semantic space.

To achieve this, a SAM-like model, FastSAM (Zhao et al. 2023), executes "segment-everything" on the image I as a visual semantic tokenizer $\mathcal{T}^I$. Subsequently, the mask-aware visual extractor $\mathcal{M}$ takes the binary mask of the region r and the image I as input, encoding these into two embeddings,

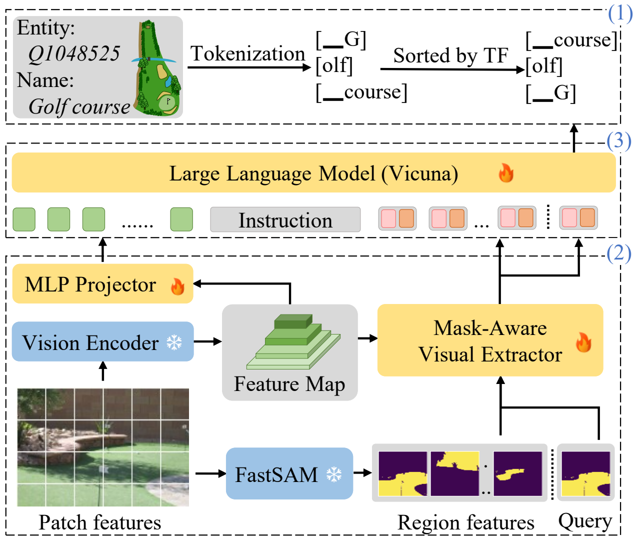


Figure 6: Model overview including 1) pre-built ALD codes for entities, 2) visual semantic tokenization, 3) autoregressive decoding target entity codes. Yellow denotes trainable parameters and blue denotes frozen parameters

$\mathbf{x}_r^{sem}$ and $\mathbf{x}_r^{pos}$, which correspond to semantic and positional feature, respectively. The region feature set $\mathbf{X}_I^{reg}$ is

$$\mathbf{X}_I^{reg} = \left\{\mathbf{x}_r^{sem}, \mathbf{x}_r^{pos} = \mathcal{M}(I, r) \mid r \in \mathcal{T}^I(I)\right\} \tag{3}$$

Compared to position-based tokenization, semantic tokenization loses the natural token order. Similar to human visual habits, which typically begin with an overview of larger image areas before concentrating on finer details, we arrange the $\mathbf{X}_I^{reg}$ in descending order based on their area a_r. This method emulates the human visual attention habit ensuring that larger areas receive broader attention within the autoregressive framework. Then we concatenate region features with the patch features $\mathbf{X}_I^{pat}$ to form the image feature $\mathbf{X}_I$.

$$\mathbf{X}_I = \left[\mathbf{X}_I^{pat}; \underbrace{(\mathbf{x}_{r_1}, \mathbf{x}_{r_2}, \cdots, \mathbf{x}_{r_{|\mathbf{X}_I^{reg}|}})}_{\mathbf{x}_r \in \mathbf{X}_I^{reg} \wedge a_{r_i} > a_{r_{i+1}}}\right] \tag{4}$$

Training

We have implemented a two-stage training strategy for our model. The vision encoder ConvNeXt CLIP (Liu et al. 2022) and the semantic tokenizer FastSAM (Zhao et al. 2023) remain frozen, while the mask-aware visual extractor $\mathcal{M}$ and the visual-language projector are fully fine-tuned. The base LLM is fine-tuned with the LoRA (Hu et al. 2022) approach. Both stages employ autoregressive language modeling loss to predict the next token (Liu et al. 2023a). In the first stage, we pre-train on the wiki split to embed entities from knowledge base $\mathcal{K}$ into the model parameters. In the second stage, we fine-tune the model on the entity and query splits to enhance its capability of fine-grained visual entity linking.

Experiments

Experimental Setting

Metrics. We evaluate model performance on the validation and test sets of MaskOVEN-Wiki using accuracy as the primary metric. Accuracy is computed for the entity and query splits, as well as the human set (test only). To address the challenges zero-shot models face in generating ALD codes and valid entity names, we use BM25 to search the 6 million Wikipedia entity names and take the top-1 result as the prediction.

Data Processing. The pre-train stage used about 2 million wiki split samples. Due to computational resource constraints and the large size of the dataset (approximately 4.5 million samples), we limited the number of annotated samples per entity to fewer than 50 during the fine-tuning stage. As a result, we used about 7% of the total samples (approximately 0.3 million) in the fine-tuning stage. In addition, all input images were uniformly preprocessed to 512 × 512. The length of the ALD code is limited to 4 tokens.

Main Results

In table 3, we compare the results of VEL models based on different types of prompts in the validation and test sets of OVEN-Wiki (Hu et al. 2023) (Text) and MaskOVEN-Wiki (Mask). Where the "None" prompt denotes that no prompt was utilized to reference the visual mention. Text-based results are from Hu et al. (2023) and Xiao et al. (2024).

Effectiveness of MaskOVEN-Wiki. In the box and mask prompts, $\mathcal{Z}$ denotes whether the result has been fine-tuned using our dataset. Osprey-7B (Yuan et al. 2023) achieves 1.3% in the zero-shot setting and 20.0% after fine-tuning, demonstrating the usefulness of our dataset. By introducing visual semantic tokenization, Osprey-7B-Seg improves the overall performance by 3.4% on the validation set and 5.2% on the test set.

Advances of Pixel Mask Reference. Results in table 3 verify the advantages compared with text and box. Compared with text-based results (6.4%-25.5%), our mask representation methods achieve similar performance (0.8%-25.2%), despite text prompts offering more detailed descriptions. Compared with box results (around 1.6%), mask prompts achieve better results. Additionally, we analyzed the limitations of mask methods when dealing with query split, where some questions include additional intents (e.g. "made of", "produced by") from original VQA datasets. These situations fall outside the scope of VEL.

Analysis and Ablation Study

Direct versus Reverse Process. Comparing the experimental results in tables 1 and 3, we observe a performance gap between the direct PL-VEL methods and reverse annotation approaches. GPT-4V achieves an accuracy of 25.5% in the direct setting. The reverse annotation process, which is an open-vocabulary segmentation task, achieves an accuracy of 94.8%. These findings show the usefulness of our proposed reverse annotation approach for the PL-VEL task.

Prompt	Method	Category			Validation			Test			
		$\mathcal{R}$	$\mathcal{G}$	$\mathcal{Z}$	Entity	Query	Overall	Entity	Query	Human	Overall
None	CLIP (Hu et al. 2023)	✓	✗	✗	5.4	1.2	5.2	5.3	1.6	5.2	5.2
Text	CLIP Fusion (Hu et al. 2023)	✓	✗	✗	19.0	11.9	18.8	19.2	14.5	11.4	18.9
	CLIP2CLIP (Hu et al. 2023)	✓	✗	✗	11.4	2.8	11.2	11.6	3.5	12.7	11.4
	PaLI-3B (Hu et al. 2023)	✗	✓	✗	14.3	20.5	14.5	12.6	20.3	24.1	13.2
	PaLI-17B (Hu et al. 2023)	✗	✓	✗	21.8	29.2	22.0	19.8	29.5	34.1	20.5
	BLIP-2 (Xiao et al. 2024)	✗	✓	✓	6.1	19.8	6.4	-	-	-	-
	GPT-4V (Xiao et al. 2024)	✗	✓	✓	24.7	53.9	25.5	-	-	-	-
Box	GlaMM	✗	✓	✓	1.4	8.9	1.6	1.5	6.1	4.3	1.7
Mask	Osprey-7B	✗	✓	✓	0.6	9.8	0.8	1.0	8.2	5.6	1.3
	Osprey-7B-FT	✗	✓	✗	19.4	8.3	19.0	20.1	11.8	23.2	20.0
	Osprey-Seg-7B	✗	✓	✗	24.3	11.8	24.0	25.4	16.1	25.9	25.2

Table 3: Comparison of VEL models on OVEN-Wiki (Text, None) and MaskOVEN-Wiki (Mask, Box) validation and test sets. Method categories are denoted as follows: $\mathcal{R}$ for retrieval-based discriminative models, $\mathcal{G}$ for generative models, and $\mathcal{Z}$ for zero-shot models without fine-tuning. The gray line highlights our proposed method.

Semantic Tokenization and Training. The ablation experiments evaluate the effectiveness of visual semantic tokenization and training in table 4. The results indicate that the introduction of region features improves model accuracy in the entity split by 3.7% to 5.0% and in the query split by 3.5% to 5.5%. In addition, fine-tuning improves the accuracy of the model, whereas the impact of pre-training is relatively limited, with improvements ranging from 0.1% to 1.6%. This finding contrasts with those of GER-ALD (Caron et al. 2024b). We attribute the success of GER-ALD's pre-training to its larger pre-training dataset (Entity-WebLI, 55M) and the lighter model (GIT, 0.4B) (Wang et al. 2022).

Method	Entity	Query	Overall
Osprey-7B	0.6	7.7	0.8
+FT	19.0	6.2	18.7
+FT +Seg	<u>22.7</u> +3.7	<u>11.7</u> +5.5	<u>22.4</u> +3.7
+PT +FT	19.3	8.3	19.0
+PT +FT +Seg	**24.3** +5.0	**11.8** +3.5	**24.0** +5.0

Table 4: Ablation study on the validation dataset. PT refers to pre-training, FT refers to fine-tuning, and Seg represents visual semantic tokenization. Bold indicates the best results, and underline denotes the second-best results.

Retrieval versus Generation. Table 5 compares retrieval-based and generation-based methods. PL-VEL is a newly introduced task, so we primarily compare our results with text-based reference models. The absence of handcrafted text queries may create a disadvantage for our approach. AUTOVER (Xiao et al. 2024) is a recently proposed text-based VEL model and demonstrates approximately an 18% improvement in performance by combining retrieval augmentation ($\mathcal{R}$) and generative prediction ($\mathcal{G}$). Notably, AUTOVER-7B$_\text{p}$ is a peer version of AUTOVER model without retrieval augmentation, and its performance closely matches ours (-0.5%). This finding indicates that retrieval augmentation has the potential to benefit the PL-VEL task.

Method	Type		Dev		
	$\mathcal{R}$	$\mathcal{G}$	Entity	Query	Overall
CLIP Fusion	✓	✗	19.0	11.9	18.8
PaLI-17B	✗	✓	21.8	29.2	22.0
AUTOVER-7B	✓	✓	42.2	43.1	42.3
AUTOVER-13B	✓	✓	44.7	43.6	44.6
AUTOVER-7B$_\text{p}$	✗	✓	23.8	-	-
Osprey-Seg-7B	✗	✓	24.3	11.8	24.0

Table 5: Comparing the retrieval-based and generation-based methods on the validation dataset. Gray line represents peer results of AUTOVER.

Conclusion

In this paper, we introduce the Pixel-Level Visual Entity Linking (PL-VEL) task, which links visual mentions indicated by pixel masks to entities in a knowledge base. This task is a supplement to the text-based VEL, enhancing VEL's practicality for tasks like VQA, visual reasoning, and detailed image captioning. We developed the MaskOVEN-Wiki dataset, a multimodal dataset aligning pixel-level regions with entity-level labels, achieving 94.8% annotation accuracy. Models trained on this dataset achieved over an 18-point improvement in accuracy compared to zero-shot models, with our visual semantic tokenization method contributing an additional 5-point increase. Despite these gains, the final model's linking accuracy was about 25%, indicating both the effectiveness of reverse annotation and the potential of the MaskOVEN-Wiki dataset for enabling fine-grained visual understanding in MLLMs.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62172044). We thank the anonymous reviewers for their kind comments.

References

Caron, M.; Iscen, A.; Fathi, A.; and Schmid, C. 2024a. A Generative Approach for Wikipedia-Scale Visual Entity Recognition. arxiv:2403.02041.

Caron, M.; Iscen, A.; Fathi, A.; and Schmid, C. 2024b. A Generative Approach for Wikipedia-Scale Visual Entity Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17313–17322.

Chang, L.-W.; Bao, W.; Hou, Q.; Jiang, C.; Zheng, N.; Zhong, Y.; Zhang, X.; Song, Z.; Yao, C.; Jiang, Z.; Lin, H.; Jin, X.; and Liu, X. 2024. FLUX: Fast Software-based Communication Overlap On GPUs Through Kernel Fusion. arXiv:2406.06858.

Chen, K.; and Wu, X. 2024. VTQA: Visual Text Question Answering via Entity Alignment and Cross-Media Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27218–27227.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. arxiv:2306.15195.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Guo, Q.; De Mello, S.; Yin, H.; Byeon, W.; Cheung, K. C.; Yu, Y.; Luo, P.; and Liu, S. 2024. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13796–13806.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, H.; Luan, Y.; Chen, Y.; Khandelwal, U.; Joshi, M.; Lee, K.; Toutanova, K.; and Chang, M.-W. 2023. Open-Domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12065–12075.

Huang, X.; Zhang, Y.; Ma, J.; Tian, W.; Feng, R.; Zhang, Y.; Li, Y.; Guo, Y.; and Zhang, L. 2024. Tag2Text: Guiding Vision-Language Model via Image Tagging. In *The Twelfth International Conference on Learning Representations*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. arxiv:2304.02643.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, 12888–12900. PMLR.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 34892–34916. Curran Associates, Inc.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning. arxiv:2304.08485.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11976–11986.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. arxiv:2306.14824.

Qiu, J.; Madotto, A.; Lin, Z.; Crook, P. A.; Xu, Y. E.; Dong, X. L.; Faloutsos, C.; Li, L.; Damavandi, B.; and Moon, S. 2024. SnapNTell: Enhancing Entity-Centric Visual Question Answering with Retrieval Augmented Multimodal LLM. arxiv:2403.04735.

Rasheed, H.; Maaz, M.; Mullappilly, S. S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2023. GLaMM: Pixel Grounding Large Multimodal Model. arxiv:2311.03356.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. GLaMM: Pixel Grounding Large Multimodal Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13009–13018.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. ArXiv:2401.14159 [cs].

Sain, A.; Bhunia, A. K.; Chowdhury, P. N.; Koley, S.; Xiang, T.; and Song, Y.-Z. 2023. CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2765–2775.

Saito, K.; Sohn, K.; Zhang, X.; Li, C.-L.; Lee, C.-Y.; Saenko, K.; and Pfister, T. 2023. Pic2Word: Mapping Pictures to

Words for Zero-Shot Composed Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19305–19314.

Sun, W.; Fan, Y.; Guo, J.; Zhang, R.; and Cheng, X. 2022. Visual Named Entity Linking: A New Dataset and A Baseline. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2403–2415. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; and Wang, L. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *Transactions on Machine Learning Research*.

Wu, L.; Li, Z.; Zhao, H.; Wang, Z.; Liu, Q.; Huai, B.; Yuan, N. J.; and Chen, E. 2023. Recognizing Unseen Objects via Multimodal Intensive Knowledge Graph Propagation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, 2618–2628. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701030.

Xiao, Z.; Gong, M.; Cascante-Bonilla, P.; Zhang, X.; Wu, J.; and Ordonez, V. 2024. Grounding Language Models for Visual Entity Recognition. arxiv:2402.18695.

Yuan, Y.; Li, W.; Liu, J.; Tang, D.; Luo, X.; Qin, C.; Zhang, L.; and Zhu, J. 2023. Osprey: Pixel Understanding with Visual Instruction Tuning. https://arxiv.org/abs/2312.10032v2.

Zhang, S.; Sun, P.; Chen, S.; Xiao, M.; Shao, W.; Zhang, W.; Liu, Y.; Chen, K.; and Luo, P. 2024a. GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. arxiv:2307.03601.

Zhang, Y.; Huang, X.; Ma, J.; Li, Z.; Luo, Z.; Xie, Y.; Qin, Y.; Luo, T.; Li, Y.; Liu, S.; et al. 2024b. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1724–1732.

Zhang, Y.; Lin, C.; Cao, D.; and Lin, D. 2024c. End-To-End Spatially-Constrained Multi-Perspective Fine-Grained Image Captioning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3360–3364.

Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; and Wang, J. 2023. Fast Segment Anything. arXiv:2306.12156.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. https://arxiv.org/abs/2304.10592v2.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023. Segment Everything Everywhere All at Once. *Advances in Neural Information Processing Systems*, 36: 19769–19782.