Grounding Language Models for Visual Entity Recognition

Zilin Xiao[1], Ming Gong[2], Paola Cascante-Bonilla[1], Xingyao Zhang[2], Jie Wu[2], and Vicente Ordonez[1]

[1] Department of Computer Science, Rice University, USA
{zilin, pc51, vicenteor}@rice.edu
[2] Microsoft STCA, China
{migon, xingyaozhang, jiewu1}@microsoft.com

Abstract. We introduce AutoVER, an Autoregressive model for Visual Entity Recognition. Our model extends an autoregressive Multimodal Large Language Model by employing retrieval augmented constrained generation. It mitigates low performance on out-of-domain entities while excelling in queries that require visual reasoning. Our method learns to distinguish similar entities within a vast label space by contrastively training on hard negative pairs in parallel with a sequence-to-sequence objective without an external retriever. During inference, a list of retrieved candidate answers explicitly guides language generation by removing invalid decoding paths. The proposed method achieves significant improvements across different dataset splits in the recently proposed Oven-Wiki benchmark with accuracy on the Entity seen split rising from 32.7% to 61.5%. It demonstrates superior performance on the unseen and query splits by a substantial double-digit margin, while also preserving the ability to effectively transfer to other generic visual question answering benchmarks without further training.

Keywords: Language Model · Visual Entity Recognition

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated superior performance on a variety of vision-and-language tasks such as visual question answering, image captioning, zero-shot image classification, among others [40, 55]. Their abilities can be transferred with few-shot tuning [2, 69] or even learning in context without parameter updates [5, 73]. Given their remarkable generalization abilities and prior knowledge based on large-scale pre-training, we consider there is still great potential for using them on tasks that require knowledge grounding.

A recently proposed task requiring knowledge grounding is the Open-domain Visual Entity Recognition (Oven-Wiki) [28] task. In this task, given an input image and a question about the image, the goal is to answer with a very specific entity from Wikipedia. For instance, the answers could be a specific model of airplane, *e.g.* ATR 42, or British Aerospace 146. This is a very challenging

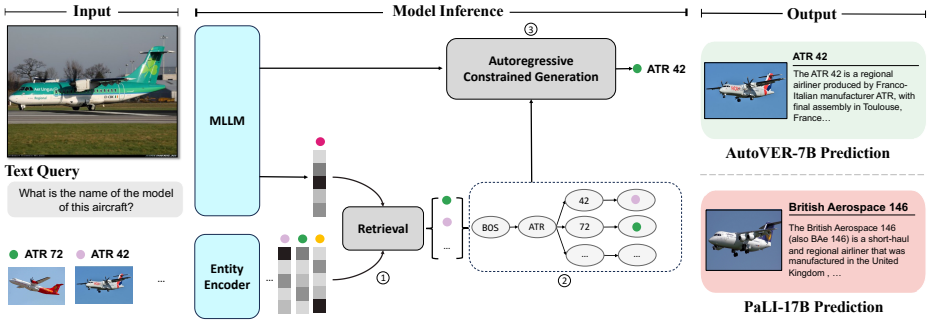

Fig. 1: A representative query-entity pair from OVEN-Wiki. We briefly illustrate the model inference process and compare predictions from PaLI-17B (in red) and AUTOVER-7B (in green) which obtains the correct answer: ATR 42. AUTOVER retrieves entity candidates without an external retriever (step 1), dynamically constructs a prefix tree (trie) (step 2), and performs decoding-time augmentation to guide autoregressive generation (step 3).

task where MLLM-based solutions are prone to *hallucinations*, or producing answers that are not at the right level of granularity. Importantly, visual entity recognition requires recognizing entities that never appear in the training data (UNSEEN split). Moreover, a portion of the OVEN-Wiki benchmark (QUERY split) extends the task beyond recognition, requiring non-trivial reasoning to resolve the query question.

Visual Entity Recognition (VER) presents unique challenges compared to general Visual Question Answering (VQA) [1, 45]. The first challenge lies in the answer label space comprising over 6 million Wikipedia entities, which is prohibitive for classifier-based VQA models [58] as many entities considered in this ontology are visually very similar. The second issue involves generation-based VQA solutions [18, 40], where *hallucinations* can lead to generated text that is not grounded to the entity space. Lastly, existing VQA approaches fail to consider the visual information of candidates [41, 65]. Entity images from the knowledge base also play a significant role when disambiguating between entities with similar identifiers but different visual appearance. These methods also struggle with out-of-domain generalization and multi-hop reasoning, challenges which are covered by the UNSEEN and QUERY split of the OVEN-Wiki dataset.

In this work, we introduce an Autoregressive Visual Entity Recognizer (AutoVER), the first approach that enables multimodal language models to perform accurate visual entity recognition over a massive knowledge base. AUTOVER addresses entity recognition by reformulating the problem as a sequence-to-sequence generation problem, as depicted in Fig. 1. The query image is translated into the token embedding space using a learnable projection layer, in a manner akin to treating images as a foreign language [43, 72]. This allows us to utilize pre-trained multi-modal language models and enormously improves performance on the QUERY split of OVEN-Wiki which requires reasoning over spatial relations, commonsense, and other visually-situated contexts.

To take advantage of visual clues on the entity side and enhance the generalization capability of the model on UNSEEN entities, we propose a unified and compact retrieval-augmented generation (RAG) framework upon AUTOVER. Specifically, a special token `<ret>` is added to the vocabulary, whose last-layer hidden states serve as the representation for the query side. A lightweight two-layer Transformer is responsible for fusing visual features from entity images and textual features from entity descriptions and producing a representation for the entities, allowing contrastive learning with query-entity positive pairs in parallel with language modeling. Unlike other RAG systems that directly use retrieved items as context [57, 75] or infuse them into the model's intermediate hidden states [17, 24, 39], AUTOVER dynamically constructs a prefix tree (trie) from the retrieved entity identifiers, and then generates entity identifiers by leveraging a methodology that constrains the next possible tokens and eliminates invalid options based on the trie, thus ensuring the generated text can always be grounded in retrieved candidates. To alleviate the entity granularity problem, two hard negative sampling strategies are proposed in our contrastive-generative framework to encourage the model to maximally distinguish between similar entities.

In summary, AUTOVER offers several advantages over baselines and prior work on VQA in the realm of visual entity recognition: (i) It refines the recognition process by integrating contrastive training into an MLLM. (ii) The proposed retrieval-augmented constrained decoding framework guarantees correct grounded entity prediction and enhances prediction over UNSEEN entities; (iii) AUTOVER leverages pre-trained visual models and knowledge graphs for hard negative mining which significantly strengthens our contrastive-generative framework in fine-grained entity recognition; (iv) Experimental results show that AUTOVER consistently outperforms fine-tuned CLIP and PaLI variants on all OVEN-Wiki splits. For instance, AUTOVER-7B achieves a 61.5% accuracy on the OVEN-Wiki ENTITY SEEN split over PaLI-17B's 30.6%, and 21.7% on ENTITY UNSEEN over PaLI-17B's 12.4%. Additionally, we evaluate AUTOVER alongside public multimodal LLMs on entity-related questions from the A-OKVQA dataset, demonstrating that our model can effectively zero-shot transfer to out-of-domain VQA datasets beyond OVEN-Wiki. Furthermore, ablation studies confirm the effectiveness of the introduced retrieve-generate framework. Code is released at https://github.com/MrZilinXiao/AutoVER.

2 Related Work

Visual Entity Recognition (VER) [28] is an emerging task for assessing the ability of a model to perform multi-modal knowledge grounding [29,53,79]. It can be regarded as a variant of visual question answering [1,45,59] with the key distinction that the choice set will comprise all entities in a specific knowledge base. Earlier solutions view this problem as an image-text-to-image-text retrieval task, where CLIP-based models [56] are fine-tuned and the top-scored answer in inference is treated as the model prediction. Alternatively, others rely on fine-tuning

generative language models such as PaLI [12] and GiT [8, 71] in an attempt to match the generated text with candidate entity identifiers using sparse matching approaches such as BM25 [61]. Entity Linking (EL) has been a long-standing text-only counterpart to VER, which entails locating mentions in the document and disambiguating them against a set of candidate entities. Our method bears the closest resemblance with the recent generative EL paradigm [7, 47, 78], which also reduces the knowledge-grounding problem into a sequence-to-sequence generation task. However, we stand out with several distinct advantages, including the seamless integration of retrieval capabilities into the language model and the use of explicit guidance for sequence generation.

Multimodal LLMs (MLLMs) are motivated by the remarkable reasoning abilities of Large Language Models (LLMs). LLaVA [43] builds upon LLaMA [68] and uses instruction-tuning to align visual features to the language space. The common practice of prompting MLLMs to produce proxy representations for downstream usage is to expand the vocabulary of language models. Such expansion augments the MLLM functionality beyond language generation. FROMAGe [34], GILL [33], GenLLaVa [26] and LISA [37] achieve satisfactory results over image-text-to-text retrieval, image generation and reasoning segmentation by inserting special tokens to augment the MLLMs original functionality. Our approach, built upon one of the latest MLLM model architecture, presents a compelling alternative to the bi-encoder shallow dot-product interaction or encoder-decoder sparse surface form matching employed in previous works on visual entity recognition. Our method also integratives contrastive learning on the outputs of entities and question pairs, analogous to the image-text contrastive learning used in CLIP-like models [54, 56, 66] and among visual features used for self-supervised contrastive learning [3, 11, 16, 25].

Retrieval-Augmented Language Models (RALMs) represent a class of NLP solutions focusing on knowledge-intensive tasks. RALMs typically condition a language model on relevant documents from a grounding corpus during generation, thereby enhancing the performance in knowledge-intensive language understanding tasks. Lewis *et al*. [39] jointly fine-tune a retriever with an encoder-decoder model, enabling the community to explore the RALM paradigm in language understanding. Guu *et al*. [23] train a bi-directional variant and also demonstrate superior performance. Apart from retrieved items in context, Févry *et al*. [17] are the first to integrate retrieved entity supervision by injecting intermediate representations. While augmented context helps in mitigating the well-known issue of *hallucination*, it does not ensure a faithful prediction with respect to an external knowledge base. Our retrieve-generate framework is distinct from existing retrieval-augmented methods. AutoVER reduces the candidate entity set from millions to hundreds for generative language models through retrieval [30] and dynamically constrains language model generation using a prefix tree. This framework coincides with the design philosophy of agents [22] in Interactive NLP [74], where an agent interacts with the dynamic environment by performing beam-search on all available options.

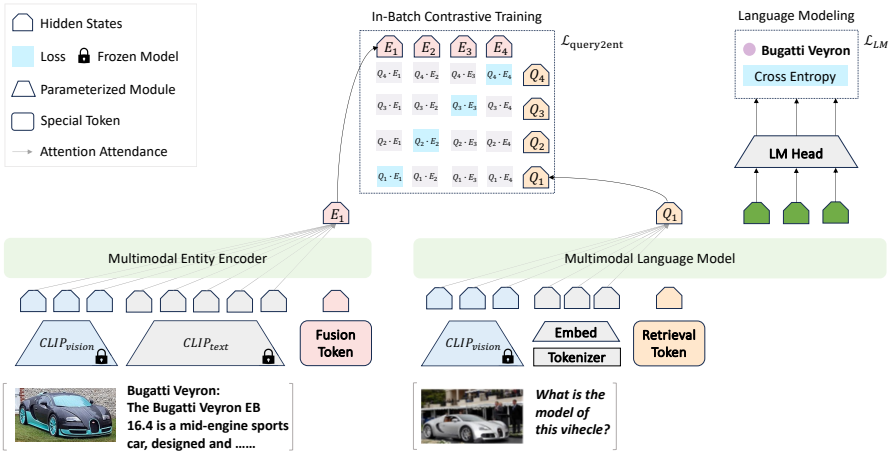

Fig. 2: Joint training of in-batch contrastive learning and language modeling in AUTOVER. For each training quadruple consisting of an entity image, an entity description, a query image and a query question, a lightweight Transformer encoder produces the fused entity representation E_i (left half). A special retrieval token prompts the multimodal language model to generate the query representation Q_i. The query-to-entity contrastive training ($\text{L}_{\text{query2ent}}$) encourages the correct retrieval of entities given the query pair, and the language modeling (L_{LM}) helps the successful entity grounding.

3 Methodology

We first formulate the Visual Entity Recognition problem in Sec. 3.1. Then we present the overview design of AUTOVER in Sec. 3.2. We illustrate hard-negative mining with knowledge base source and pre-trained visual model in Sec. 3.3 and retrieval-augmented constrained decoding process in Sec. 3.4.

3.1 Problem Definition

Visual Entity Recognition can be viewed as a multimodal knowledge grounding task, in which the model is required to process an image-text pair input $x = (\mathbf{Q}_{\text{im}}, \mathbf{Q}_{\text{t}})$ and predict an entity e. $\mathbf{Q}_{\text{t}}$ describes the specific intent that prompts the model to ground some entity e in the image $\mathbf{Q}_{\text{im}}$ to a label space $\mathcal{E}$. Each entity $e \in \mathcal{E}$ is a member of the knowledge base $\mathcal{K} = \{(e, \mathbf{E}_{\text{im}}, \mathbf{E}_{\text{desc}}) \mid e \in \mathcal{E}\}$ where $\mathbf{E}_{\text{desc}}$ is a text description and $\mathbf{E}_{\text{im}}$ is a set of relevant images of the entity. To circumvent trivial solutions for some query questions, all image-text pairs are annotated so the question cannot be correctly answered without the image.

3.2 Model Overview

AUTOVER consists of two main modules: a multi-modal language model f_ϕ initialized with pre-trained weights ϕ, and a multi-modal entity encoder F_φ

that fuses the entity image and textual description. To integrate the retrieval functionality into f_ϕ, we insert a special token `<ret>` into the vocabulary of f_ϕ with its corresponding token embedding denoted as $\mathbf{H}_{\texttt{<ret>}}$.

We illustrate the training process in Fig. 2. For a training sample ($\mathbf{Q}_{\text{im}}, \mathbf{Q}_{\text{t}}$, $\mathbf{E}_{\text{im}}, \mathbf{E}_{\text{desc}}$), we first use a frozen pre-trained CLIP visual encoder g_{vision} and a learnable projection $\mathbf{W}_q$ to extract query image features $\mathbf{H}_{\text{im}}$ and the embedded query text representations $\mathbf{H}_{\text{t}}$ as follow:

$$\begin{aligned}\mathbf{H}_{\text{im}} &= \mathbf{W}_q \cdot g_{\text{vision}}\left(\mathbf{Q}_{\text{im}}\right), \\ \mathbf{H}_{\text{t}} &= \mathbf{W}_{\text{embed}} \cdot \mathbf{Q}_{\text{t}},\end{aligned}$$

where $\mathbf{W}_{\text{embed}}$ is the token embedding layer associated with f_ϕ, thus, treating the query image features as a foreign language. We concatenate the embedded query text representations and the query image features with the retrieval token embedding to organize the input instruction to model $\mathbf{H} = [\mathbf{H}_{\text{im}}, \mathbf{H}_{\text{t}}, \mathbf{H}_{\texttt{<ret>}}]$. The last-layer hidden states of `<ret>` token undergo dimension-matching projection and L2 normalization to a hyperspherical space, serving as the representation on the query side denoted as Q. The use of a causal attention mask in f_ϕ allows this representation to incorporate the multi-modal query without leaking any label information.

On the entity encoder side, a frozen CLIP visual encoder g_{vision} transforms entity image $\mathbf{E}_{\text{im}}$ into grid image features $\mathbf{Z}_{\text{im}}$, and the frozen CLIP text encoder g_{text} encodes the entity identifier and description into text features $\mathbf{Z}_{\text{text}}$. A two-layer Transformer encoder handles the fusion of two modalities by a fusion token as a soft prompt [38], whose design was also adopted in [32]. The final-layer hidden states of the fusion token are also subjected to dimension matching and L2 normalization, providing the entity representation E.

In-Batch Contrastive Training. Given normalized representations from both the query and entity side, Q and E, we formally introduce our in-batch contrastive training method specifically designed for query-to-entity retrieval. While image-text retrieval has been popular in learning joint vision and language representations [31, 56], few studies have delved into image-text-to-image-text retrieval, which happens to be the focus of query-to-entity retrieval. We adopt contrastive learning [13] and the InfoNCE loss [50] used in previous image-text retrieval work. Starting with computing the cosine similarity for a pair of representations,

$$\text{sim}(Q, E) = \frac{Q \cdot E}{\|Q\| \|E\|},$$

we minimize the InfoNCE loss for query-to-entity on a mini-batch consisting of N query-entity pairs (Q_i, E_i). Each corresponding pair is considered a positive pair and others are treated as negatives. τ is a learnable temperature parameter. Our formulation is as follows:

$$\mathcal{L}_{\text{query2ent}} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp\left(\text{sim}\left(Q_i, E_i\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(Q_i, E_j\right)/\tau\right)} \right).$$

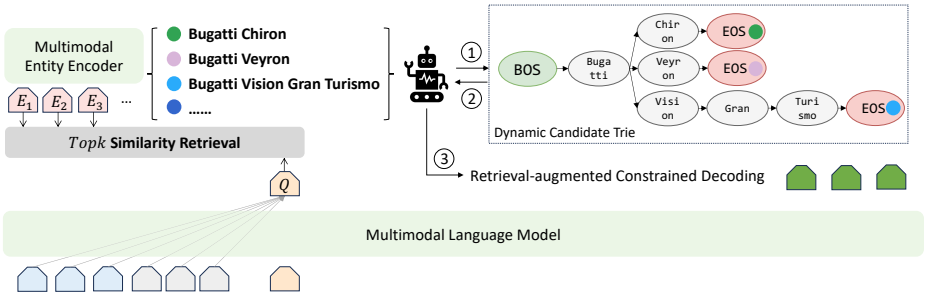

Fig. 3: Retrieval-augmented constrained decoding illustration of our proposed AutoVER inference process. The representation Q will query a pre-cached entity database constructed using the multimodal entity encoder, and get the top-k candidate entities. A prefix-constrained tree is dynamically built based on retrieved entity identifiers and guides the language model to autoregressively generate the next token, thereby ensuring the successful grounding of generated content.

Note that we do not train for the reversed objective, *i.e.* entity-to-query, as it does not align with our retrieval-augmented intuition.

In contrast to image-text retrieval which offers a unique correspondence between images and texts, query-to-entity mapping is subjective, *i.e.* multiple queries in the training data are associated with the same entity. This requires a particular sampling strategy during training to ensure that a batch does not include the same entity, referred to as a "conflicting batch". To preserve the integrity of the training data distribution, we implement a rejection sampling [20] method as an alternative to standard random sampling, resorting to resampling upon encountering conflicting batches.

Language Modeling. With organized input instruction $\mathbf{H}$, we optimize the following cross-entropy loss, which is known as the next token prediction loss in causal language modeling:

$$\mathcal{L}_{\text{LM}} = -\frac{1}{N-n}\sum_{i=n}^{N}\log P\left(y_i \mid \mathbf{H}, \ldots, y_{i-1}\right),$$

where n is the length of $\mathbf{H}$, N denotes the length of the concatenation of $\mathbf{H}$ and the expected output, *i.e.* entity text identifier (Bugatti Veyron in Figure 2). Note that we do not backward the next token prediction loss on the input sequence, but only on the target sequence.

The final training loss is a linear combination of the language modeling loss and in-batch contrastive loss, denoted as follows:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_r \cdot \mathcal{L}_{\text{query2ent}},$$

where λ_r is an empirically determined trade-off hyperparameter.

3.3 Hard-Negative Mining

To alleviate the entity granularity problem, two hard negative sampling approaches are proposed in our contrastive-generative framework, *i.e.* VISION-HARD and KB-HARD. Both approaches create entity groups used for constructing an in-batch sampler that prefers sampling similar entities in contrastive training, yet they differ in how they generate these similar groups. VISION-HARD relies on a pre-trained ViT [15] image classifier to identify visually similar entities based on shared prediction classes. In contrast, KB-HARD uses the category hierarchy of Wikidata as an external knowledge source, considering entities that share a parent node in the category hierarchy as knowledge-similar entities. We refer readers to the supplementary material for details about the construction of hard negative groups.

3.4 Retrieval-augmented Constrained Decoding

We illustrate the proposed model inference in Fig. 3. After training, all entity candidates are cached using the multi-modal entity encoder F_φ to construct an entity vector database $\mathcal{V} \in \mathbb{R}^{n \times d}$ for efficient retrieval. Given an evaluation sample $(\mathbf{Q}_{\text{im}}, \mathbf{Q}_{\text{t}})$, we first forward the MLLM for the normalized representation of the `<ret>` token, and query database $\mathcal{V}$ using top-k similarity search to get k entity candidates. The model dynamically generates a prefix tree (trie) that covers k entity candidates, and the trie will explicitly guide entity identifier generation by eliminating impossible decoding paths when autoregressively generating tokens. The retrieval process takes into consideration the image of entities and improves the model towards out-of-domain entities. Moreover, the retrieval-augmented constrained decoding guarantees the grounding of the generated content to the knowledge base, alleviating the issue of *hallucinations*.

4 Experiments

We describe the experimental setting in Sec. 4.1 and report the main results in Sec. 4.2. We present the zero-shot generalization results in Sec. 4.3 and ablation study in Sec. 4.4 with discussions. For an intuitive demonstration of our model, we present the case study in Sec. 4.5.

4.1 Settings

Metrics. We follow the standard setup in Hu *et al.* [28], which uses accuracy and harmonic mean to evaluate model performance on different data splits. The harmonic mean of each split will equally weigh the importance of the SEEN and UNSEEN subsets and penalize models that show weakness in either aspect. Finally, we report the overall harmonic mean on the ENTITY and QUERY splits as the final metric for the validation and test sets.

Table 1: Comparison among models on the OVEN-Wiki **validation** set. We report accuracies for the SEEN and UNSEEN subsets and the harmonic mean of each split (HM). Metrics of CLIP and PaLI variants are from [28]. For each subset, **bold** indicates the best metric.

		Entity Split			Query Split			Overall
Category	Method	SEEN	UNSEEN	HM	SEEN	UNSEEN	HM	HM
Discriminative	$\text{CLIP}_{\texttt{ViTL14}}$	5.4	5.3	5.4	0.8	1.4	1.0	1.7
	$\text{CLIP Fusion}_{\texttt{ViTL14}}$	32.7	4.3	7.7	33.4	2.2	4.2	5.4
	$\text{CLIP2CLIP}_{\texttt{ViTL14}}$	12.6	10.1	11.2	4.1	2.1	2.8	4.4
Generative	PaLI-3B	21.6	6.6	10.1	33.2	14.7	20.4	13.5
	PaLI-17B	30.6	12.4	17.6	44.2	22.4	29.8	22.1
Zero-shot	$\text{BLIP-2}_{\texttt{Flan-T5-XXL}}$	8.6	3.4	4.9	24.6	17.7	20.6	7.9
	GPT-4V	29.8	19.3	23.4	56.5	**52.7**	**54.5**	32.9
Ours	AUTOVER-7B	61.5	21.7	32.1	**69.0**	31.4	43.2	36.8
	AUTOVER-13B	**63.6**	**24.5**	**35.6**	68.6	32.3	43.9	**39.2**

Data Pre-processing and Models. We pre-process all query and entity images by resizing them into 336 × 336 pixels with padding to keep an identical aspect ratio. Entity descriptions are truncated to 77 tokens to fit the context window size of CLIP by cutting off sentences to the maximum available ones. Since query texts are typically short, no truncation is needed for query texts. Encoders on the entity side and visual encoder in MLLM are pre-trained `CLIP-ViTL/14-336px`. The MLLM is initialized with the `vicuna-7b-v1.5` or `vicuna-13b-v1.5` checkpoint and corresponding visual projectors pre-aligned on a 558k subset of LAION [63], CC [9] and SBU [52] curated by Liu *et al.* [42].

Training and Evaluation. We train on the entity train and query train splits of OVEN-Wiki, which consists of nearly 5 million query-entity pairs. We conduct all pieces of training in a batch size of 256 on 32 V100-SXM2-32GB GPUs. We refer to the supplementary material for hyper-parameter choices and training details. We set λ_r to 1 for all training settings. Due to a limited compute budget, we select 10% and 50% of the training data through weighted sampling according to entity occurrence frequency to facilitate ablation studies and reporting final results. For the same reason, we do not run a hyperparameter search. In evaluation, the number of retrieved entities k is set to 300 according to empirical trials on the validation split.

4.2 Main Results

Results on the validation set are presented in Tab. 1. Our experimental results demonstrate a consistent improvement in all data splits and subsets on OVEN-Wiki. Specifically, we observe a double accuracy improvement in the ENTITY SEEN subsets of both OVEN-Wiki validation and test splits. This indicates that

Table 2: Results of methods on the OVEN-Wiki **test** set and **human evaluation** set. Human evaluation results from [28] are highlighted in gray.

	Entity Split		Query Split		Overall	Human Eval		
Method	SEEN	UNSEEN	SEEN	UNSEEN	HM	SEEN	UNSEEN	HM
Human + Search	-	-	-	-	-	76.1	79.3	77.7
$\text{CLIP}_{\texttt{ViTL14}}$	5.6	4.9	1.3	2.0	2.4	4.6	6.0	5.2
$\text{CLIP Fusion}_{\texttt{ViTL14}}$	33.6	4.8	25.8	1.4	4.1	18.0	2.9	5.0
$\text{CLIP2CLIP}_{\texttt{ViTL14}}$	12.6	10.5	3.8	3.2	5.3	14.0	11.1	12.4
PaLI-3B	19.1	6.0	27.4	12.0	11.8	30.5	15.8	20.8
PaLI-17B	28.3	11.2	36.2	21.7	20.2	40.3	26.0	31.6
GER-400M [8]	31.5	17.7	-	-	-	-	-	-
`llava-v1.5-7b` [43]	7.5	2.1	41.9	**37.4**	6.2	-	-	-
AUTOVER-7B	62.8	16.0	63.7	31.9	31.8	64.7	39.9	49.4
AUTOVER-13B	**65.0**	**18.6**	**65.7**	32.0	**34.6**	**68.4**	**44.2**	**53.7**

AUTOVER can more effectively utilize its parameter capacity, achieving stronger in-domain visual entity recognition capabilities with fewer model parameters. The improvement on the QUERY SEEN subsets is slightly lower compared with the ENTITY SEEN improvement but still significant with +24.8% relative accuracy difference. We attribute this to the inherent reasoning and visual localization capabilities within the MLLM, a proficiency that has been extensively demonstrated across various benchmarks targeting MLLMs [10, 48].

On UNSEEN subsets, accuracy is severely impacted by queries whose answers involve out-of-domain entities, particularly in the Entity split. Nevertheless, AUTOVER still outperforms the largest model, PaLI-17B, which can be attributed to the retrieval-augmented framework eliminating many improbable options for the decision-making of the MLLM.

We report results on test and human evaluation set in Tab. 2. Similarly, AUTOVER-13B achieves the best performance on subsets and splits of the test set. However, it must be acknowledged that there remains a gap between our results and those of Human + Search in the human validation set, particularly in the human UNSEEN subset, where the difference reaches 35.1%.

In addition, we report the zero-shot visual entity recognition abilities of the largest BLIP-2 [40] checkpoint and GPT-4V[3] [51] on the validation set in Tab. 1, whereas no fine-tuning is performed on the training dataset. For BLIP-2, zero-shot generated outputs are grounded to Wikipedia using BM25, following generative baseline methods. As for GPT-4V, its elaborate generation, a result of instruction tuning, impedes effective evaluation with BM25. Hence, we adopt a partial matching evaluation strategy, treating the presence of the entity identi-

[3] Due to the limited budget, we only conduct experiments on 10% of $\text{Entity Split}_{\texttt{(Val)}}$ and 50% of $\text{Query Split}_{\texttt{(Val)}}$. Also, the detail of the image is set to low to minimize prompt token consumption.

fier in the generated text as a true positive. While BLIP-2 performance is anticipated, as its lack of fine-tuning hinders the acknowledgment of entity grounding intents, GPT-4V has surprisingly exceptional performance on the QUERY split. Nevertheless, GPT-4V poor performance on the ENTITY split still leaves it trailing behind AUTOVER. Furthermore, the opacity of its training data prevents us from determining if GPT-4V excellent zero-shot performance is attributable to possible data leakage.

4.3 Zero-shot Generalization Results

Although OVEN-Wiki has already incorporated 14 classic datasets[4] from image recognition and visual question answering, along with extensive annotation efforts that ground answers to the knowledge base, we still hope to test the generalization capability of our method using additional out-of-domain datasets. As such, we manually curated a subset from the A-OKVQA [64] validation dataset. The subset does not overlap with any dataset sources in OVEN-Wiki and covers all question-answer pairs in the A-OKVQA validation split whose answer can be grounded to entities in the Wikipedia knowledge base. We name the subset A-OKVQA-ENT to emphasize this distinct property. Following the design of OVEN-Wiki, we divide the subset into SEEN and UNSEEN splits depending on whether the answer entity is in the OVEN-Wiki train set. This subset includes 478 entries out of 1,145 in the A-OKVQA validation set, of which 322 are identified as SEEN, while the remaining 156 are classified as UNSEEN.

To adhere to the evaluation setting of A-OKVQA, we adopted two evaluation approaches – **multi-choice** and **entity match**. In the **multi-choice** setting, AUTOVER constructs a prefix tree with four available options to guide generation, while for other models, options are numbered in the prompt and either the correct option number or an exact entity identifier match in the response is considered a true prediction. In the **entity match** setting, a partial match of the entity text identifier within the generated response qualifies a correct prediction. To ensure equitable comparison, baselines included are other generative models comparable in size with AUTOVER-7B, specifically LLaVA [43] and its improved `v1.5` version [42], OpenFlamingo [4], and InstructBLIP [14]. While we expect that all models have not seen the A-OKVQA dataset so that we can assess the zero-shot generalization ability, it is important to acknowledge that LLaVA-v1.5 has been fine-tuned with 50k A-OKVQA multi-choice instructions, potentially giving it an edge. For context, we still include LLaVA-v1.5 results although they are not directly comparable.

We present results on A-OKVQA-ENT in Tab. 3 and Tab. 4, and refer readers to the supplementary material for qualitative analysis on this dataset. Our method still outperforms baseline models across all evaluation settings, with the exception of multiple-choice LLaVA-v1.5 which has been tuned with in-domain

[4] ImageNet21k-P [60,62], iNaturalist2017 [70], Cars196 [35], SUN397 [77], Food101 [6], Sports100 [19], Aircraft [44], Oxford Flower [49], Google Landmarks v2 [76], VQA v2 [21], Visual7W [80], Visual Genome [36], OK-VQA [46], Text-VQA [67]

Table 3: Model accuracies on A-OKVQA-ENT under **multi-choice** evaluation. Here methods are provided with multiple choices for answers in the prompt. (*) Denotes unusual results due to failure to adhere to the provided multiple-choice prompts. **Bold** indicates the best metric under the same setting.

Supervision	Method	SEEN	UNSEEN	Overall
Zero-shot	OpenFlamingo-9B*	5.9	9.0	6.9
	InstructBLIP$_{\text{vicuna-7B}}$	53.7	49.4	52.4
	LLaVA-v1-7B*	13.0	10.3	12.1
	AUTOVER-7B (Ours)	**67.7**	**52.5**	**62.8**
Fine-tuned	LLaVA-v1.5-7B	72.4	73.7	72.8

Table 4: Model hit rates on A-OKVQA-ENT under **entity match** evaluation. Here methods are not provided with multiple choices in the prompt and are deemed successful if they produce the right entity anywhere in their output.

Supervision	Method	SEEN	UNSEEN	Overall
Zero-shot	OpenFlamingo-9B	39.8	30.8	36.8
	InstructBLIP$_{\text{vicuna-7B}}$	38.8	34.0	37.2
	LLaVA-v1-7B	45.7	36.5	42.7
	AUTOVER-7B (Ours)	**61.3**	**42.3**	**55.0**
Fine-tuned	LLaVA-v1.5-7B	47.8	42.9	46.2

instructions. Interestingly, we find that LLaVA-v1.5 in the **entity match** evaluation setting still falls short of AUTOVER-7B, although it has seen in-domain samples. We also observed that LLaVA-v1 and OpenFlamingo, with inadequate instruction tuning, find it challenging to comply with the multiple-choice setting, and instead generate content unrelated to the given options, leading to their significantly lower metrics. Those findings reveal that generative language models struggle to adapt well to out-of-domain instructions and that unrestricted generation is prone to hallucinations in tasks requiring precise recognition. This further emphasizes the advantages of retrieval-assisted decoding-time augmentation for such tasks.

4.4 Ablation Study and Discussion

We present ablation studies exploring the effect of retrieval augmented generation, constrained decoding, and hard negative mining in Tab. 5. Ablations are conducted with 10% of the training data and as such the metrics in ablations differ from our main results and we refer to our model as AUTOVER-7B-0.1.

Retrieval Augmentation. We first focus on the absence of retrieval augmentation in AUTOVER. Without retrieval, the model performs constrained decoding over a prefix tree composed of all Wikipedia entity identifiers instead of a set of retrieved candidates. We notice a slight performance gain on the SEEN subset,

Table 5: Ablation study of AutoVER-7B-0.1 on Oven-Wiki entity Split$_{\texttt{(Val)}}$.

Method	SEEN	UNSEEN	HM
AutoVER-7B-0.1	48.9	19.0	27.4
+ w/o retrieval	50.7	0.6	1.2
+ w/o constrained decoding	46.8	0.6	1.2
+ w/ LoRA	43.5	2.8	5.3
AutoVER-7B-0.1-[CLS]	12.8	0.1	0.2

likely due to the gold entity not being covered in retrieval. Conversely, the integrated retrieval design is found to significantly improve the performance of AutoVER on the UNSEEN subset from non-viable to viable. Retrieval-augmented constrained decoding narrows down the decision-making scope of the language model, allowing AutoVER to generate entity identifiers that are never seen during training.

Constrained Decoding. Upon the excluding of retrieval augmentation, we proceed to disable the entire constrained decoding mechanism, reverting AutoVER to greedily decode the next token, where grounding to an external knowledge base is no longer assured. The consequent decrease in SEEN accuracy demonstrates that constrained decoding effectively alleviates hallucinations of the model in predicting entities. Unfortunately, since neither checkpoints of encoder-decoder variants from [28] in Table 1 and 2 nor PaLI pre-trained weights are publicly available, we are unable to assess whether our proposed decoding-time augmenting methods can bring universal improvement on generative baselines.

Parameter-Efficient Tuning (LoRA). Our ablations extend to the impact of parameter-efficient tuning methods on AutoVER. Specifically, we configure a low-rank adapter (LoRA) [27] with a rank of 128 and alpha set at 256 and train this LoRA-variant of AutoVER. The results reveal that the efficacy of LoRA falls short of expectation and explicitly harms the recognition performance on the SEEN subset. We leave more specialized efficient tuning methods for the generative VER framework as future work.

Autoregressive Model or Classifier. One may attribute the success of AutoVER to the large model size of its underlying MLLM and accompanied expressive capacity instead of our proposed design. In response to that, we devise a [CLS] variant of AutoVER. It follows the classical design of treating a decoder-only LM as a classifier. Specifically, we introduce a new [CLS] token into the MLLM vocabulary and append this token at the end of each query image-question pair. The corresponding last-layer hidden states are then fine-tuned for classifying the query into the 20,549 existing entities in Oven-Wiki. We observed a significant performance degradation in this setting, particularly in the UNSEEN subset. This highlights the ineffectiveness of classifier-based VQA methods in the realm of visual entity recognition, indicating that our model performance stems from more than just its size.

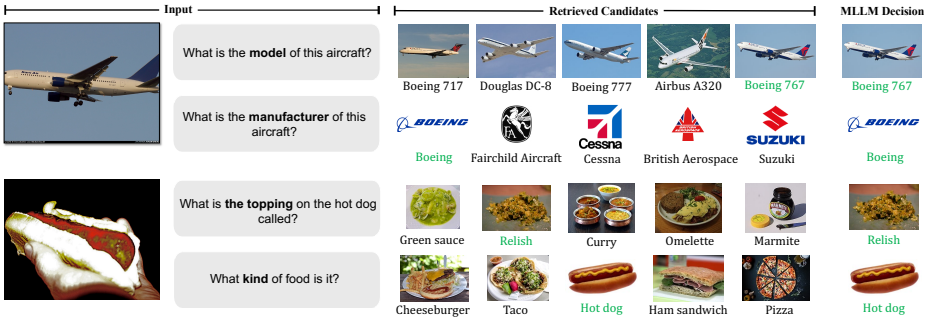

Fig. 4: Illustration of selected query image-question pairs, retrieved candidates and AutoVER-7B decisions. AutoVER adeptly captures slight variations in the query text and retrieves entirely different entity candidates, which forms the basis for the generative decisions from the language model.

4.5 Case Study

We illustrate the retrieved entity candidates and model decisions on four representative image-question pairs. We note that for the same query image, questions with different intents lead to retrieval representations with semantic relevance to the intent, laying a solid foundation for the subsequent constrained generation. For instance, in the upper part of Fig. 4, the question asking for the specific model of airplane prompts the model to retrieve various kinds of airplanes. With a slight modification from "model" to "manufacturer" in the query question, the model adapts to retrieve famous airplane manufacturer brands instead of plane models. Likewise, the model is also proficient in managing queries that demand visual localization and reasoning abilities, as depicted in the lower part of Fig. 4. We refer to the supplementary material for the error analysis and more case studies with comparisons against GPT-4V.

5 Conclusion

We present AutoVER, a compact retrieval-augmented generation framework specifically designed for visual entity recognition. Utilizing a novel constrained decoding technique, this approach effectively overcomes the challenges of low performance in recognizing out-of-domain entities while demonstrating remarkable proficiency in questions that require visually-situated reasoning. AutoVER marks a significant advancement in visual entity recognition by doubling accuracy on nearly all challenging subsets of benchmarks. We discuss limitations in the supplementary material.

Acknowledgements. This project was partially funded by an NSF CAREER Award #2201710, and support from the Ken Kennedy Institute at Rice University. We also thank reviewers for their feedback and encouragement.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: Vqa: Visual question answering. International Journal of Computer Vision (2015)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
3. Aubret, A., Teulière, C., Triesch, J.: Self-supervised visual learning from interactions with objects. arXiv preprint arXiv:2407.06704 (2024)
4. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv: 2308.01390 (2023)
5. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting. Advances in Neural Information Processing Systems **35**, 25005–25017 (2022)
6. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: ECCV (2014)
7. Cao, N.D., Izacard, G., Riedel, S., Petroni, F.: Autoregressive entity retrieval. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (2021)
8. Caron, M., Iscen, A., Fathi, A., Schmid, C.: A generative approach for wikipedia-scale visual entity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
9. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. Computer Vision and Pattern Recognition (2021)
10. Chen, L., Li, B., Shen, S., Yang, J., Li, C., Keutzer, K., Darrell, T., Liu, Z.: Large language models are visual reasoning coordinators. arXiv preprint arXiv: 2310.15166 (2023)
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
12. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D.M., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A.V., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B.K., Riquelme, C., Steiner, A., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: Pali: A jointly-scaled multilingual language-image model. International Conference on Learning Representations (2022)
13. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 539–546 vol. 1 (2005)
14. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2024)

15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (2021)
16. Fan, R., Poggi, M., Mattoccia, S.: Contrastive learning for depth prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3226–3237 (2023)
17. Févry, T., Baldini Soares, L., FitzGerald, N., Choi, E., Kwiatkowski, T.: Entities as Experts: Sparse Memory Access with Entity Supervision. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4937–4951. Association for Computational Linguistics, Online (Nov 2020)
18. Gao, F., Ping, Q., Thattai, G., Reganti, A., Wu, Y.N., Natarajan, P.: Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5067–5077 (2022)
19. Gerry: Sports100: 100 Sports Image Classification. https://www.kaggle.com/datasets/gpiosenka/sports-classification/metadata (2021)
20. Gilks, W.R., Wild, P.: Adaptive rejection sampling for gibbs sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics) **41**(2), 337–348 (1992)
21. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2017)
22. Gu, Y., Deng, X., Su, Y.: Don't Generate, Discriminate: A Proposal for Grounding Language Models to Real-World Environments. Annual Meeting of the Association for Computational Linguistics (2022)
23. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.W.: REALM: retrieval-augmented language model pre-training. International Conference on Machine Learning (ICML) (2020)
24. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval Augmented Language Model Pre-Training. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3929–3938. PMLR (13–18 Jul 2020)
25. Hernandez, J., Villegas, R., Ordonez, V.: Vic-mae: Self-supervised representation learning from images and video with contrastive masked autoencoders. arXiv preprint arXiv:2303.12001 (2023)
26. Hernandez, J., Villegas, R., Ordonez, V.: Generative Visual Instruction Tuning (2024), https://arxiv.org/abs/2406.11262
27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022 (2022)
28. Hu, H., Luan, Y., Chen, Y., Khandelwal, U., Joshi, M., Lee, K., Toutanova, K., Chang, M.W.: Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. ICCV (2023)
29. van Hulst, J.M., Hasibi, F., Dercksen, K., Balog, K., de Vries, A.P.: REL: An Entity Linker Standing on the Shoulders of Giants. In: Huang, J.X., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J., Liu, Y. (eds.) Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020. pp. 2197–2200. ACM (2020)

30. Iscen, A., Caron, M., Fathi, A., Schmid, C.: Retrieval-enhanced contrastive vision-text models. In: The Twelfth International Conference on Learning Representations (2024)
31. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. International Conference on Machine Learning (2021)
32. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
33. Koh, J.Y., Fried, D., Salakhutdinov, R.: Generating images with multimodal language models. arXiv preprint arXiv: 2305.17216 (2023)
34. Koh, J.Y., Salakhutdinov, R., Fried, D.: Grounding language models to images for multimodal inputs and outputs. International Conference on Machine Learning (2023)
35. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 3dRR-13 (2013)
36. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
37. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv: 2308.00692 (2023)
38. Lester, B., Al-Rfou, R., Constant, N.: The Power of Scale for Parameter-Efficient Prompt Tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021)
39. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474. Curran Associates, Inc. (2020)
40. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742. PMLR (2023)
41. Lin, W., Chen, J., Mei, J., Coca, A., Byrne, B.: Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering. In: H.Larochelle, M.Ranzato, R.Hadsell, M.F.Balcan, H.Lin (eds.) Advances in Neural Information Processing Systems. Curran Associates, Inc. (2023)
42. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (June 2024)
43. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023)
44. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft (2013), https://arxiv.org/abs/1306.5151

45. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. Advances in neural information processing systems **27** (2014)
46. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 3195–3204 (2019)
47. Mrini, K., Nie, S., Gu, J., Wang, S., Sanjabi, M., Firooz, H.: Detection, Disambiguation, Re-ranking: Autoregressive Entity Linking as a Multi-Task Problem. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 1972–1983. Association for Computational Linguistics, Dublin, Ireland (May 2022)
48. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A.: A comprehensive overview of large language models. arXiv preprint arXiv: 2307.06435 (2023)
49. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. IEEE (2008)
50. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv: 1807.03748 (2018)
51. OpenAI: GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf (Sep 2023)
52. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: Neural Information Processing Systems (NIPS) (2011)
53. Ouyang, S., Huang, J., Pillai, P., Zhang, Y., Zhang, Y., Han, J.: Ontology enrichment for effective fine-grained entity typing. arXiv preprint arXiv:2310.07795 (2023)
54. Pan, X., Ye, T., Han, D., Song, S., Huang, G.: Contrastive language-image pretraining with knowledge graphs. Advances in Neural Information Processing Systems **35**, 22895–22910 (2022)
55. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv: 2306.14824 (2023)
56. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
57. Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., Shoham, Y.: In-Context Retrieval-Augmented Language Models. Transactions of the Association for Computational Linguistics (2023)
58. Ramakrishnan, S., Agrawal, A., Lee, S.: Overcoming language priors in visual question answering with adversarial regularization. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
59. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. Advances in neural information processing systems **28** (2015)
60. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. In: NeurIPS (2021)
61. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (2009)

62. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal on Comptuer Vision (IJCV) (2015)
63. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
64. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-OKVQA: A benchmark for visual question answering using world knowledge. In: ECCV (8). Lecture Notes in Computer Science, vol. 13668, pp. 146–162. Springer (2022)
65. Shao, Z., Yu, Z., Wang, M., Yu, J.: Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14974–14983 (June 2023)
66. Shrivastava, A., Selvaraju, R.R., Naik, N., Ordonez, V.: Clip-lite: Information efficient visual representation learning with language supervision. In: International Conference on Artificial Intelligence and Statistics. pp. 8433–8447. PMLR (2023)
67. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards VQA models that can read. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
68. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models. arxiv (2023)
69. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021)
70. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2018)
71. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: GIT: A generative image-to-text transformer for vision and language. Trans. Mach. Learn. Res. **2022** (2022)
72. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19175–19186 (2023)
73. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6830–6839 (June 2023)
74. Wang, Z., Zhang, G., Yang, K., Shi, N., Zhou, W., Hao, S., Xiong, G., Li, Y., Sim, M.Y., Chen, X., Zhu, Q., Yang, Z., Nik, A., Liu, Q., Lin, C., Wang, S., Liu, R., Chen, W., Xu, K., Liu, D., Guo, Y., Fu, J.: Interactive Natural Language Processing (2023), https://arxiv.org/abs/2305.13246
75. Wang, Z., Araki, J., Jiang, Z., Parvez, M.R., Neubig, G.: Learning to filter context for retrieval-augmented generation. arXiv preprint arXiv: 2311.08377 (2023)
76. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Conf. on Computer Vision and Pattern Recognition (CVPR) (2020)

77. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Conf. on Computer Vision and Pattern Recognition (CVPR) (2010)
78. Xiao, Z., Gong, M., Wu, J., Zhang, X., Shou, L., Pei, J., Jiang, D.: Instructed language models with retrievers are powerful entity linkers. arXiv preprint arXiv: 2311.03250 (2023)
79. Zhang, W., Hua, W., Stratos, K.: EntQA: Entity Linking as Question Answering. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022 (2022)
80. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)