Tag-grounded Visual Instruction Tuning with Retrieval Augmentation

Daiqing Qi[1] **Handong Zhao**[2] **Zijun Wei**[2] **Sheng Li**[1]
[1]University of Virginia [2]Adobe
{daiqing.qi, shengli}@virginia.edu {hazhao, zwei}@adobe.com

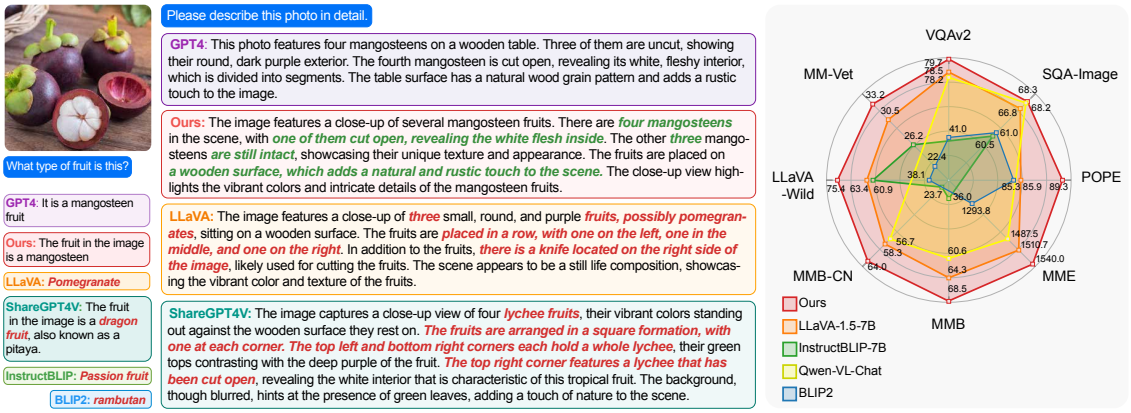


Figure 1: Examples on LLaVA-W (**left**), and quantitative comparison (**right**). Imprecise low-quality answers are marked in ***red*** and high-quality parts are marked in ***green***. Popular *open-source* MLLMs fail to identify the mangosteen (the first question), and list non-existent objects such as 'knife' and incorrect quantities and arrangements, while ours correctly identify 'mangosteens' with descriptions in detail.

Abstract

Despite recent advances in general visual instruction-following ability of Multimodal Large Language Models (MLLMs), when diving into low-level details, they still struggle with critical problems when required to provide a precise and detailed response to a visual instruction: (1) failure to identify novel *objects* or entities, (2) mention of *non-existent objects* and (3) neglect of *object*'s attributed details. Intuitive solutions include improving the size and quality of data or using larger foundation models. They show effectiveness in mitigating these issues, but at an expensive cost of collecting a vast amount of new data and introducing a significantly larger model. Standing in the intersection of them, we examine the three *object*-oriented problems from the perspective of the image-to-text mapping process by the multimodal connector. In this paper, we first identify the limitations of multimodal connectors stemming from insufficient training data. Driven by it, we propose to enhance the mapping with retrieval-augmented tag tokens, which contain rich *object*-aware information such as object names and attributes. With our **T**ag-grounded visual instruction t**UN**ing with retrieval **A**ugmentation, **TUNA** outperforms baselines that share same language model and training data on 12 benchmarks. Furthermore, we show the zero-shot capability of TUNA when provided with specific datastores.

1 Introduction

Multimodal Large Language Models (MLLM) have witnessed remarkable progress recently (Chen et al., 2023c; Liu et al., 2023a, 2024; Bai et al., 2023; Chen et al., 2023a; Dai et al., 2023; Ye et al., 2023; Zhu et al., 2023a; Zhang et al., 2023), exhibiting superior ability in following vision-and-language instructions. Despite their effectiveness in providing general responses, their performance often degrade when required to give a detailed and accurate answer to the question associated with an image with novel objects, named entities or complex scenes with rich and subtle details.

Specifically, they frequently encounter challenges (Fig. 1) in: 1. identifying novel objects and named entities, 2. preventing the generation of objects that do not align with the target images, and 3. delivering a comprehensive description that covers the details of the target images. We uncover the

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2008–2026
November 12-16, 2024

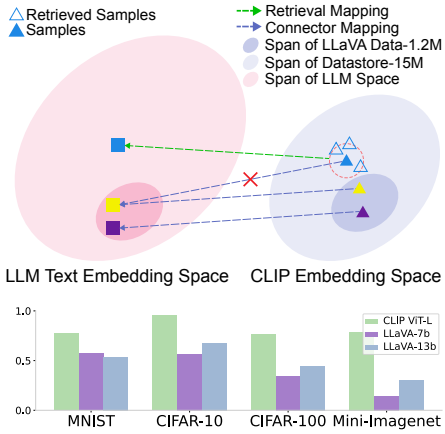


Figure 2: **Top**: the process of translating image embeddings to text embeddings (LLaVA (Liu et al., 2024)). **Bottom**: Image classification accuracy of CLIP (Radford et al., 2021) and MLLMs built on it.

some of the potential causes of above challenges starting from the commonly adopted two-branch structure and the two-stage training paradigm of MLLMs: the first-stage pre-training and second-stage supervised fine-tuning (SFT). Most existing MLLMs such as LLaVA (Liu et al., 2024) comprise two modules: (1) a vision branch consisting of a vision encoder and a multimodal connector, and (2) a Large Language Model (LLM). In the pre-training stage with large-scale image-text pairs, the multimodal connector often learns to translate the outputs of the vision encoder to text embeddings, followed by the SFT stage which enhances the multi-modal instruction-following capabilities with instruction-format data.

Despite the promising zero-shot capability of the vision encoder, such as CLIP (Radford et al., 2021), which is pre-trained with over 400M image-text pairs, its generalizability is bottlenecked by the learnt mapping of the multimodal connector when integrated into the MLLM framework. E.g., in the case of LLaVA (Liu et al., 2024), the two-stage training data is significantly smaller compared to the pre-training data of its vision encoder CLIP (1.2M vs. 400M), as a result, the connector often fails to effectively map the out-of-distribution (OOD) images to the corresponding LLM text embeddings. Therefore, LLM fails to successfully identify image contents. MLLMs' degradation on image classification performance (Zhai et al., 2023) is a simple illustration. In Fig. 2 (Bottom), an obvious classification performance gap between MLLMs and their frozen vision encoder (CLIP) is observed. The absence of similar classification objects in LLaVA's training data could be a critical factor, which makes it particular hard for the multimodal connector to translate OOD CLIP embeddings of test images to LLM text embeddings.

One intuitive solution is to enrich the training datasets with more image-text pairs, however, as high-quality instruction-format data is particularly critical for visual instruction tuning (Chen et al., 2023c), it is very expensive to build high-quality training data with hundreds of millions of image-text pairs of varying quality. Furthermore, the training could also become exceedingly burdensome.

Instead of directly improving the connector mapping with heavy training, could we build another lightweight new mapping as a complementary that effectively attends to objects, especially OOD ones? Motivated by retrieval augmented generation (RAG) (Ramos et al., 2023b,a; Yang et al., 2023; Hu et al., 2023; Lin et al., 2024; Li et al., 2023c; Yasunaga et al., 2022), we propose a retrieval mapping. As shown in Fig. 2 (Top), while the connector fails to correctly map the sample out of LLaVA training data span to its corresponding text embedding in LLM embedding space (i.e., the blue triangle sample is incorrectly mapped to the yellow square sample) , we introduce a large-scale external datastore with a better coverage of novel objects, named entities, and attributes, for the retrieval of useful knowledge towards the input image. In this way, a new retrieval mapping could be built from the input image to corresponding LLM text embeddings (green dashed line in Fig. 2).

While most existing works retrieve relevant captions as extra knowledge, it may not apply here because all three challenges mentioned above are oriented with *object*, where cleaner object-aware knowledge is urgent, instead of noisy captions. Therefore, we want to retrieve tags of the images that are similar to the input image as extra knowledge, where we can further enrich each tag representation with image region feature and adaptive weights to fulfill the potential of useful tags. To this end, we introduce a **T**ag-grounded visual instruction t**UN**ing with retrieval **A**ugmentation, termed **TUNA**, that performs a knowledge-aware and tag-grounded generation. With grounded tags, TUNA is effective in identifying novel objects, named entities, and generate tag-oriented response which pays more attention to image details.

We summarize our contributions as follows: (*i*) We identify potential factors hindering MLLMs and first propose a tag-grounded visual instruction tuning with retrieval-augmentation (TUNA) with enhanced knowledge on novel objects, more atten-

tion to details, and less mention of non-existent objects. (*ii*) To fulfill the potential of tags, We carefully designed the image-aware tag encoder, which produces tag embeddings enhanced by image features with an adaptive weight. (*iii*) We evaluate TUNA on extensive benchmarks along with a series of qualitative results, and show its zero-shot capability when provided with specific datastores.

2 Related Works

Multimodal Large Language Models. MLLMs evolve rapidly nowadays. With LLMs, while existing works (Li et al., 2022, 2023d) enable basic visual tasks like visual question answering, more recent works (Chen et al., 2023a; Liu et al., 2024) shows proficiency in image-text dialogues through alignment and fine-tuning. Subsequent research (Bai et al., 2023; Chen et al., 2023b; Dai et al., 2023; Li et al., 2023a; Peng et al., 2023; Ye et al., 2023; You et al., 2023) enhances LLMs by emphasizing data quality and diversity. With grounding data, a branch of works (Ye et al., 2023; You et al., 2023; Chen et al., 2023b; Peng et al., 2023) improves LLMs' grounding capability. Despite their evolution, as they share a similar multimodal connector module that performs image-to-text translation, a lingering fundamental problem persists: Out-of-distribution (OOD) images, such as novel objects, named entities, new scenes, etc., cannot be translated to text embeddings effectively, leading to misaligned answers, missing details or mention of non-existent objects from LLM.

Retrieval-Augmented Multimodal Learning. Retrieval-augmented language generation (RAG) consists of conditioning generation on additional information that is retrieved (e.g., with clustering (Zhao et al., 2017)) from an external datastore. Recently, A branch of works (Ramos et al., 2023b,a; Yang et al., 2023; Hu et al., 2023; Lin et al., 2024; Li et al., 2023c) integrate it into image captioning, where relevant captions are retrieved to guide the captioning. Distinct from them, in visual instruction tuning, where detailed and dense responses based on the multimodal instructions are often required, cleaner object-level information, such as names and attributes of novel objects, named entities, is urgent. We provide a more detailed discussion in Appendix A.

Multimodal Learning with Tags. Existing works (Huang et al., 2023; Zhou et al., 2020; Li et al., 2020; Hu et al., 2021; Qi et al., 2024a; Huang et al., 2022) show the effectiveness of introducing object tags as anchor points to help the learning of semantic alignments between images and texts in the training data. In the context of Fig. 2, they better align in-distribution data (yellow and purple samples) with tags. **Our goal is distinctive from them in that**, We do not aim to learn better representations of training data, instead, we want to (1) improve the tag-grounded generation capability of MLLMs and (2) acquire new knowledge with retrieved tags from *external datastore*. Besides, as they treat object tags as anchor points for feature learning, tags are commonly human-used ones (Huang et al., 2023) as guidance. For instance, Tag-to-Text (Huang et al., 2023) collects 3,429 well-used tags filtered by human annotation. While in our case, where the large coverage is the priority, less frequently used tags (e.g., named entities) are also desired, resulting in a total of 3M tags (details in Appendix A).

3 Tag-ground Visual Instruction Tuning

In this section, we first introduce how we extract tags from 15M captions from CC12M (Changpinyo et al., 2021) and CC3M (Sharma et al., 2018). Then we present how we build and use the datastore, followed by the illustration of TUNA.

3.1 Multimodal Retriever

From Captions to Tags. As introduced in Sec. 1, one of the fundamental challenges for MLLMs is to effectively translate image tokens to LLM text embeddings, especially for OOD images that contain novel objects. With better translation, LLMs would be less likely to confuse with them, which could improve the identification of objects. Thus in addition to the mapping learnt by the connector, we use a multimodal retriever to retrieve relevant information as an additional retrieval mapping (Fig. 2) to enhance the translation process. Therefore, the quality of the retrieval mapping is critical. As a result, object-oriented tags as retrieved information would be very helpful. Additionally, with tag-grounded generation, retrieved tags also serve as groundings or hints, which could prompt the LLM to generate tag-aware contents *if the tag is relevant to the input image*, which would also be helpful in alleviating missing objects or visual details.

Towards this end, we use CLIP image embeddings from image-text paired datasets as keys and corresponding *tags* as values. However, existing

Caption: Montepulciano, Italy - August 25, 2013: Old narrow street in the center of town with colorful facades. stock photography.
Results from Parser:
Entity (Attributes): facades (colorful) | lawn (narrow,old) | town
Relations: street (Subject) - in center of (Relation) - town (Object)
Results from NER: Montepulciano | Italy
Tags: ['facades', 'narrow', 'town', 'Italy', 'colorful', 'street', 'old', Montepulciano]

Figure 3: Examples of tags derived from parsing and NER results.

Number of Unique Tags	Characters per Tag	Tags per Image
3.2M	16.8	5.31

Table 1: Extracted tags from CC3M and CC12M

large-scale image-text datasets such as Conceptual Captions (Sharma et al., 2018; Changpinyo et al., 2021) only contain captions. To mine tags from texts, we parse each caption into a set of tags with a combination of FACTUAL scene graph parser (Li et al., 2023f) and Name Entity Recognition (NER) with spaCy, yielding 3M tags extracted from 15M captions in CC3M (Sharma et al., 2018) and CC12M (Changpinyo et al., 2021). We show several examples in Fig. 3. Details of the mining process are available in Appendix B. We also provide a statistics of the obtained tags in Tab 1.

Datastore and Cross-Modal Retrieval. With processed image-tags pairs, our datastore is indexed by FAISS library (Johnson et al., 2019) with image CLIP embeddings as keys and associated tags as values. Given a query image, a k-nearest neighbor retrieval with cosine similarity of embeddings between it and datastore images is performed. The *tags* of top-k retrieved images are input to TUNA as additional knowledge. In experiments, we use k=5. We consider CC12M (Changpinyo et al., 2021), CC3M (Sharma et al., 2018) and COCO (Lin et al., 2014) training set as our datastore, resulting in 15M image-text pairs. In experiments, we use a whole combination, as well as parts of them, as our datastore to study how different datastores affect results. For Fashion QA, we use a combination of fashion data as our retrieval datastore.

3.2 TUNA

Architecture. The framework of TUNA is illustrated in Fig. 4. Given a language instruction $\mathbf{X}_\text{q}$, and an input image $\mathbf{X}_\text{v}$, a set of images with associated tags are retrieved from the datastore. Assume there are M tags in total, they are mixed together and denoted as $\{\mathbf{X}_\text{t}^i\}_{i=1}^M$. For image, a frozen pre-trained CLIP vision encoder ViT-L/14 is employed to extract the visual feature $\mathbf{Z}_\text{v} = g(\mathbf{X}_\text{v}) \in \mathbb{R}^{[H\times W]\times D}$, followed by a MLP multimodal connector $h(\cdot)$ that translates the CLIP vision feature to text embeddings: $\mathbf{H}_\text{v} = h(\mathbf{Z}_\text{v})$. Similar to LLaVA (Liu et al., 2024), the grid visual features before the last Transformer layer are considered in our experiments. The language instruction $\mathbf{X}_\text{q}$ is tokenized and projected to text embeddings $\mathbf{H}_\text{q}$ by the pre-trained LLM's tokenizer and embedding layer. Specifically, tags $\{\mathbf{X}_\text{t}^i\}_{i=1}^M$ are encoded by our image-aware tag encoder.

Image-Aware Tag Encoder. Given a tag $\mathbf{X}_\text{t}^i$, its tag representation $\mathcal{H}_\text{i}$, which is encoded by our image-aware tag encoder, is a tuple of its text embedding $\mathbf{H}_\text{t}^i$ and the its tag-aware image token (embedding) $\mathbf{H}_\text{vt}^i$, which contains visual features of the *input query image* related to this tag. With this image token, LLM could better attend to details of the tag-related object in the input image. Same with $\mathbf{X}_\text{q}$, the tag $\mathbf{X}_\text{t}^i$ is tokenized and projected to $\mathbf{H}_\text{t}^i$ with the LLM's tokenizer and embedding layer. To obtain the tag-aware image token, the tag-aware image feature $\mathbf{Z}_\text{vt}^i \in \mathbb{R}^{1\times D}$ is first extracted from the grid visual features of the *input image* via the cross-attention module: $\mathbf{Z}_\text{vt}^i = \text{Cross-Att}(\mathbf{Q}_\text{t}^i, \mathbf{Z}_\text{v}, \mathbf{Z}_\text{v}) = \text{softmax}(\frac{\mathbf{Q}_\text{t}^i\mathbf{Z}_\text{v}^\intercal}{\sqrt{D}})\mathbf{Z}_\text{v}$, where $\mathbf{Q}_\text{t}^i \in \mathbb{R}^{1\times D}$ is the global CLIP text feature of tag $\mathbf{X}_\text{t}^i$, extracted by the frozen CLIP text encoder. Then we obtain the tag-aware image token $\mathbf{H}_\text{vt}^i = h(\mathbf{Z}_\text{vt}^i)$. Finally, the tag representation $\mathcal{H}_i$ consists of the tuple $(\mathbf{H}_\text{vt}^i, \mathbf{H}_\text{t}^i)$. Iterating over all tags, we have $\{\mathcal{H}_i\}_{i=1}^M$.

Adaptive Weight Tuner. As retrieved images may contain less relevant or irrelevant tags, e.g., the tag `durian` in Fig 4, we apply an adaptive weight tuner over them to give more attention to highly relevant tags while ignoring less related ones. Specifically, the score of $\mathcal{H}_i$ is the cosine similarity between $\mathbf{Q}_\text{t}^i$ and the global CLIP visual feature (i.e., the <CLS> token) of the input image. The scores are normalized to [0,1] as the final weights, which are applied to $\mathbf{H}_\text{vt}^i$ and $\mathbf{H}_\text{t}^i$ before input to the LLM.

Supervised Fine-Tuning. We consider Vicuna-7B (Chiang et al., 2023), a decoder-only LLM instruction-tuned on top of LLaMA (Touvron et al., 2023), as our language model. We use both image and text encoders from CLIP-ViT-L/14@336p. We initialize the pre-trained multimodal connector from LLaVA-1.5 (Liu et al., 2023a). During the instruction tuning, we always keep the weights of the vision encoder frozen, and update both the pre-trained weights of the connector and the LLM.

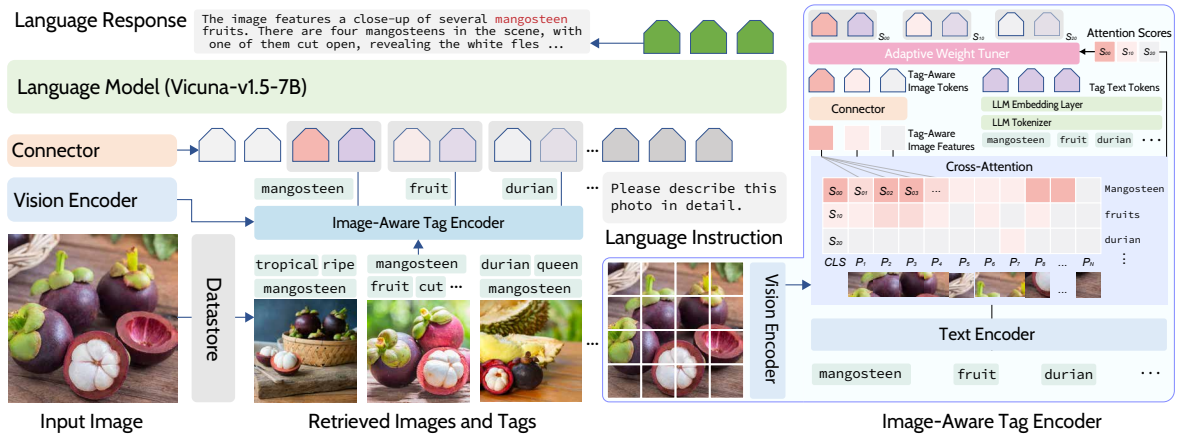


Figure 4: Framework of TUNA. **Left**: overall architecture. Given a language instruction, an image, and retrieved tags, they are transformed into tokens and input to the LLM. Only CLIP encoders are frozen. **Right:** architecture of the image-aware tag encoder, which produces tag representations with retrieved tags and the input image.

4 Experiment

In this section, we first present the training details of TUNA and benchmarks. Then we introduce quantitative and qualitative comparison with popular open-source models, followed by detailed analysis experiments and ablation studies.

Training Details. TUNA is finetuned on instruction data for one epoch, following existing works (Liu et al., 2023a; Chen et al., 2023c). We consider two different instruction-following datasets in our experiments: LLaVA-665K (Liu et al., 2023a) and ShareGPT4V-665K (Chen et al., 2023c) as our instruction-following data during fine-tuning separately, resulting in two versions of our model, TUNA and TUNA^{+}. ShareGPT4V-665K contains instruction-following data with higher quality. Details on datasets are available in Appendix C. We apply a learning rate of $2e$-5 and a batch size of 128. The training takes 12$\sim$14 hours with 8 A100 GPUs with ZeRO3. Details are available in Appendix C.

Benchmarks. We compare TUNA with baselines on 12 benchmarks, including VQA benchmarks and multimodal benchmarks designed for LLMs. Details are available in Appendix G.

4.1 Comparison with Baselines

Main Results. In Tab. 2, we provide a quantitative comparison of TUNA with popular open-source MLLMs. On 12 benchmarks, TUNA consistently outperforms previous LLMs that are finetuned from the same instruction-tuning datasets as ours with the same configuration on the vision encoder and language model (Vicuna-7B), especially on recent multimodal benchmarks with more notable improvements. As the size of LLM and different choices of instruction-following data can significantly improve the model performance, we mark the models gray that are equipped with a larger 13B language model or finetuned from currently unavailable datasets of higher quality and quantity. Specifically, LLaVA-1.6 (or LLaVA-NeXT)[1] is finetuned from larger instruction-following data of higher quality, with additional user instruct data. Besides, it equips the better vision encoder with dynamic high resolution, known as AnyRes (AR). Although it is not a fair comparison, we still outperform LLaVA-1.6 in MMB^{CN}, MMB and POPE, and the corresponding 13B models in MMB^{CN}, MMB, POPE and LLaVA-W.

How Can TUNA Improve the Recognition of Novel Objects and Entities? As visualized in Fig. 2 (Top), with our 15M large-scale datastore, the new retrieval mapping could greatly compensate for the original LLaVA multimodal connector that learns from around 1M data. With the additional mappings from retrieval data, TUNA is expected to show particularly improvements over questions towards novel objects or entities in the given input image. We show sub-tasks from MME (Fu et al., 2023) and MMB (Liu et al., 2023b) that consists of such questions in Tab. 3. We gain obvious improvements over the baseline in most sub-tasks. We also show several VQA examples from multimodal benchmarks in Fig 1 and Fig 5. In Fig 1, all of the baselines fail to correctly identify this fruit as a mangosteen, including LLaVA-1.5. It is reasonable as mangosteens do not appear in the its training data, which makes it particularly hard

[1] https://llava-vl.github.io

Method	LLM	V-Enc.	IT	VQA^{v2}	GQA	VizWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MMB^{CN}	SEED	LLaVA^{W}	MM-Vet
BLIP-2	Vicuna-13B	-	-	41.0	41.0	19.6	61.0	42.5	85.3	1293.8	-	-	46.4	38.1	22.4
InstructBLIP	Vicuna-7B	-	1.2M	-	49.2	34.5	60.5	50.1	-	-	36	23.7	53.4	60.9	26.2
InstructBLIP	Vicuna-13B	-	1.2M	-	49.5	33.4	63.1	50.7	78.9	1212.8	-	-	-	58.2	25.6
Shikra	Vicuna-13B	-	5.5M	77.4	-	-	-	-	-	-	58.8	-	-	-	-
IDEFICS-9B	LLaMA-7B	-	1M	50.9	38.4	35.5	-	25.9	-	-	48.2	25.2	-	-	-
IDEFICS-80B	LLaMA-65B	-	1M	60.0	45.2	36.0	-	30.9	-	-	54.5	38.1	-	-	-
Qwen-VL	Qwen-7B	-	50M	78.8	59.3	35.2	67.1	63.8	-	-	38.2	7.4	56.3	-	-
Qwen-VL-Chat	Qwen-7B	-	50M	78.2	57.5	38.9	68.2	61.5	-	1487.5	60.6	56.7	58.2	-	-
ShareGPT4V	Vicuna-13B	CLIP$^{V-L}_{336}$	665K(S)	81.0	63.4	55.6	71.2	62.2	85.9	1618.7	68.5	63.7	70.8	79.9	43.1
LLaVA-1.5	Vicuna-13B	CLIP$^{V-L}_{336}$	665K(L)	80.0	63.3	53.6	71.6	61.3	85.9	1531.3	67.7	63.6	61.6	70.7	35.4
LLaVA-1.6/NeXT	Vicuna-7B	CLIP$^{V-L}_{AR}$	760K(N)	81.8	64.2	57.6	70.1	64.9	86.5	1519.0	67.4	60.6	70.2	81.6	43.9
ShareGPT4V	Vicuna-7B	CLIP$^{V-L}_{336}$	665K(S)	80.6	63.3	57.2	68.4	60.4	85.3	1567.4	68.8	62.2	69.7	72.6	37.6
Ours^{+}	Vicuna-7B	CLIP$^{V-L}_{336}$	665K(S)	**81.1**	**63.4**	**57.4**	**70.8**	**60.4**	**89.6**	**1583.8**	**70.8**	**65.0**	**70.6**	**80.1**	**40.1**
LLaVA-1.5	Vicuna-7B	CLIP$^{V-L}_{336}$	665K(L)	78.5	62.0	50.0	66.8	58.2	85.9	1510.7	64.3	58.3	58.6	63.4	30.5
Ours	Vicuna-7B	CLIP$^{V-L}_{336}$	665K(L)	**79.7**	**62.6**	**50.0**	**68.3**	**58.4**	**89.5**	**1540.0**	**68.5**	**64.0**	**59.6**	**75.4**	**33.2**

Table 2: **Comparison with SoTA methods on 12 benchmarks.** Our model achieves the best performance on 12 benchmarks compared with LLMs that are finetuned from the same instruction tuning (IT) datasets with the same configuration on the vision encoder (V-Enc.) and language model (Vicuna-7B). Best results are in **bold**.

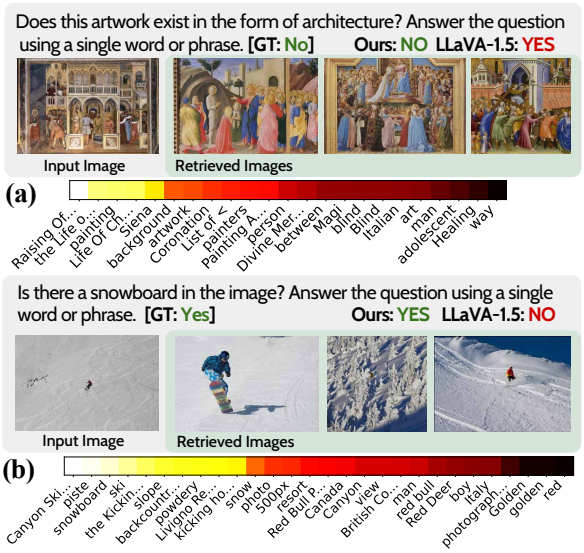


Figure 5: **VQA examples of TUNA.** For each example, we show top 3 retrieved images to save space. We show all tag set associated with all retrieved images as well as their tuned weights in heat map, where the brightest region for the highest weight 1 and darkest region for the lowest weight 0 (Zoom in for better view). Correct answers are marked green and wrong ones in red. More examples are available in Appendix E.

for the connector to map it to somewhere close to text embeddings of "mangosteen" in the LLM embedding space, as illustrated in Fig 2. When the question about the given image is a little tricky, e.g., in Fig 5 (a), the MLLM is asked if a painting of a building exists in the form of architecture, LLaVA-1.5 is confused on whether it is a real architecture or a painting. However, TUNA easily distinguished it from real architectures with additional knowledge from retrieved tags of similar images in datastore.

How Can TUNA Help to Identify the Existence of Objects? With an input image, the retrieved images are often similar to it or in the similar context.

Model	Posters	Celebrity	Artwork	landmark	Image Style	Celeb
LLaVA-1.5	146.6	137.1	119.5	163.8	69.1	83.8
Ours	**155.9**	**154.7**	**128.7**	**166.3**	**81.1**	**85.8**

Table 3: Results on sub-tasks of MME (Fu et al., 2023) and MMB (Liu et al., 2023b), where questions are towards novel objects, entities or scenes in the image. Otherwise mentioned, backbone LLM is Vicuna-7B.

Datasets	Metrics	Ours	Ferret	InstructBLIP	LLaVA	mPLUG-Owl
Random	Accuracy (↑)	**91.00**	90.24	88.57	88.00	53.97
	Precision (↑)	**98.05**	97.72	84.09	97.44	52.07
	Recall (↑)	**84.10**	83.00	95.13	78.80	99.60
	F1 Score (↑)	**90.93**	89.76	89.27	87.13	68.39
Popular	Accuracy (↑)	**90.16**	84.90	82.77	87.43	50.90
	Precision (↑)	**95.46**	88.24	76.27	95.24	50.46
	Recall (↑)	**84.20**	80.53	95.13	78.80	99.40
	F1 Score (↑)	**90.56**	84.21	84.66	86.24	66.94
Adversarial	Accuracy (↑)	**88.43**	82.36	72.10	85.50	50.67
	Precision (↑)	**91.99**	83.60	65.13	90.99	50.34
	Recall (↑)	**84.20**	80.53	95.13	78.80	99.33
	F1 Score (↑)	**87.63**	82.00	77.32	84.45	66.82
Average F1		**89.50**	85.32	83.75	85.94	67.38

Table 4: Results on POPE. We show most competing baselines. Full table is available in Appendix F. TUNA outperform Ferret (You et al., 2023), which is finetuned on grounding and referring data.

Intuitively, the retrieved images are very likely to contain similar elements or objects to the input image. Therefore, the tags could be helpful to provide additional hints to the LLM to pay special attention to them about their existence. We evaluate our model on POPE (Li et al., 2023e), a benchmark designed towards the existence of objects. Results are available in Tab. 4, we outperform competing baselines including referring and grounding MLLMs such as Ferret (You et al., 2023) and Shikra (Chen et al., 2023b). A simple example is shown in Fig 5 (b). When the object is visually imperceptible in the image, hints from tags are very helpful.

How Can TUNA Attend to Rich Details with Tags? While most of the traditional VQA bench-

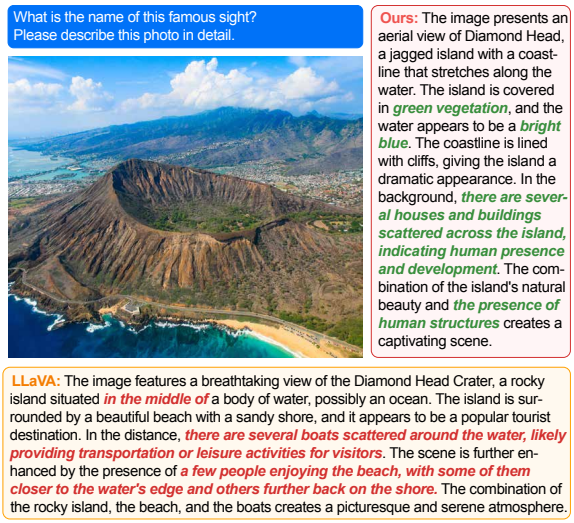


Figure 6: **TUNA on LLaVA-W examples**. Imprecise low-quality answers are marked in ***red*** and high-quality parts are marked in ***green***. TUNA does not mention non-existent objects and gives a more detailed description.

marks and multimodal benchmarks provide short questions answering pairs (Fig 5), LLaVA-W (Liu et al., 2023a) evaluates MLLM's capability of giving long detailed response. Quantitative results are available in Tab. 5. TUNA consistently outperforms baselines. We also provide one example in Fig. 6. While LLaVA mentions non-existent boats, people, TUNA accurately describes the water body, the existence of green vegetation, and interestingly, the presence of houses and buildings behind the mountain (zoom in for better view). More interestingly, there are no retrieved noun tags directly related "houses" or "buildings". By removing tags one by one, we finally identify that the tag "accessible" contributes to the the description of houses and buildings. It is an interesting phenomenon that somehow tells us that not only nouns can remind the LLM the existence of objects, relevant adjectives can also teach the LLM to pay attention to visual details. In this case, "accessible" means "human can access to this place", which might remind the LLM the existence of houses and buildings.

4.2 Ablation Study

Ablation of Adaptive Weight Tuner. Grounded on tags, intuitively, the quality of tags is critical to TUNA. However, retrieved tags could be noisy. E.g., the tag `durian` in Fig 4. To this end, we apply an adaptive weight tuner in our image-aware tag encoder to allocate more weight to more relevant tags and less weight to less relevant ones. We first ablate the tuner module to show its effectiveness of this simple but critical component in

Model	Average	Conversation	Reasoning	Detail
ShareGPT4V-13B	79.5	81.4	79.2	76.2
LLaVA-v1.5-13B	72.9	82.8	74.3	53.1
ShareGPT4V-7B	74.9	78.5	69.2	74.4
Ours$^+$	**80.1**	**87.0**	**80.2**	**77.2**
LLaVA-v1.5-7B	65.3	81.3	64.0	52.9
Ours	**75.4**	**82.0**	**77.2**	**62.5**

Table 5: Results on LLaVA-in-the-Wild (LLaVA-W) Bench. Our model consistently outperforms baselines that share the same LLM and instruction tuning data.

alleviating the noises of tags. Without the adaptive weight tuner, all retrieved tags would be equal important and their weights are set to the maximum value. The result is shown in Tab. 6 (w/o tuner). A clear performance drop is observed compared to the full method. It is reasonable because while related tags can provide useful information to the LLM, the irrelevant tags are misleading. Although it underperforms the full method, without the tuner, our model is still comparable or slightly better than LLaVA-1.5. This is favourable because it manifests that our model itself is somehow robust against less relevant tags without the tuner.

Effectiveness of Instruction Tuning. Since MLLMs are naturally in-context learners, we are interested in the effectiveness of our tag-grounded finetuning compared to the vanilla LLaVA-1.5, where tags are provided as in-context knowledge. For fair comparison, we apply the weight tuner to both models. Let's refer this model as TUNA$^-$. Results in Fig. 6 (w/o FT) indicates that, the LLM without tag-grounded instruction tuning cannot make effective use of informative tags.

Are Tags more Effective than Sentences? We compare TUNA with sentence-level retrieval in Tab. 6 (w/ captions). Instead of tags, we finetune TUNA with captions of retrieved images as additional knowledge. The image-aware tag encoder is also used, but the input tags are replaced by captions. Results show that sentence-level retrieval is not helpful. It is reasonable because tags provide cleaner and more object-related knowledge such as names, attributes, while captions are noisy.

Would Irrelevant Tags Hurt the Backbone during Inference? It is intuitive that a large-scale datastore often covers useful knowledge to the input image and question. Therefore, useful tags could be retrieved. However, there might be corner cases when retrieved tags are all irrelevant. To this end, we run experiments without tags and with random tags. Results are reported in Tab. 6. With irrelevant tags, TUNA is comparable to its backbone

Method	POPE	MMB^{CN}	MMB	MM-VET	LLaVA-W
Full method	**89.5**	**64.0**	**68.5**	**33.2**	**75.4**
w/o tuner	86.9	58.2	64.2	31.8	65.7
w/o tags	85.9	58.6	64.9	31.2	65.6
w/ random tags	85.3	58.0	64.7	31.1	65.0
w/o FT ($TUNA^{-}$)	85.9	58.1	64.2	30.8	66.5
w/ captions	85.5	59.3	65.4	30.6	65.7
LLaVA-1.5	85.3	58.2	64.3	30.5	65.3

Table 6: Ablation Studies on (1) the effectiveness of the adaptive weight tuner, (2) retrieved tags during inference and (3) tag-grounded finetuning.

Datastore	POPE	MMB	MMB^{CN}	MM-VET	LLaVA-W
All	**89.5**	**68.5**	**64.0**	**33.2**	**75.4**
CC12M	86.6	67.8	63.4	32.6	73.8
CC3M	86.2	67.5	62.9	32.1	69.2
COCO	87.9	65.9	60.2	31.4	65.2
w/o Datastore	85.3	64.3	58.2	30.5	65.3

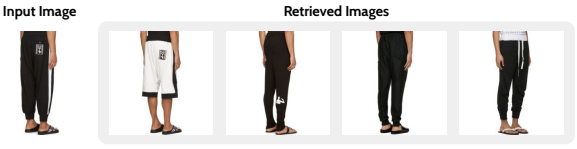
Table 7: Ablations on the choice of datastores.

LLaVA-1.5. It manifests that, our method notably improves the backbone performance with useful tags and will not hurt the backbone performance when only irreverent tags are available.

Different Choices of Datastore. We also study how different choices of datastores can affect the model performance. In the default setting, we use a combination of CC12M, CC3M and COCO training set. In addition, we perform the tag-grounded instruction tuning with different datastores, and use them for retrieval during inference, respectively. Results are available in Tab. 7. Default setting with largest datastore size outperforms other baselines. We have detailed analysis in Appendix H.

4.3 Zero-shot Inference on Fashion Domain

Fashion-Bench. To study TUNA on OOD data from another specific domain, we further collect data from FashionGen (Rostamzadeh et al., 2018) validation set and create a benchmark to measure the model's instruction-following capability in fashion domain, similar to LLaVA-Bench. Following LLaVA, we also collect a set of 24 images from FashionGen, with one question associated with each image. The questions are from one of the three types: (1) *Conversation*. We design a conversation between the assistant and a person asking questions about the product, including the product brands, categories, materials, etc. Only questions that have definite answers are considered. E.g., `What is the brand of this product?` (2) *Detailed Description*. We ask the assistant to give a comprehensive and detailed desperation of the given product. E.g., `Please describe the product in this image in detail.` (3) *Complex Reasoning*. The above two types focus on the visual content itself, based on which we further create reasoning questions. E.g., `What occasions is this clothing suitable for?`



Question: Please describe this product in detail, including its brand, category, etc.

LLaVA-1.5: The brand is called "H&M." The product is a pair of black and white pants with a logo on the side. The design is casual and comfortable, making them suitable for everyday wear. The pants are made of a blend of materials, providing both style and functionality. The logo on the side adds a touch of personality and brand recognition to the pants.

Ours: The product is a pair of black and white pants with a rubber and terry cloth applique. The pants are designed with a three-pocket styling and an elasticized waistband with a drawstring closure at the interior. The brand is KTZ, and the pants are made of 100% cotton. The design and composition of the pants suggest a casual and comfortable style, making them suitable for various occasions.

Figure 7: An example on Fashion-Bench. Precise answers in green and vague ones in red.

Evaluation. We follow LLaVA to perform GPT-assisted evaluation. After obtaining the responses from models, we feed the question, ground truth text information, and the generated responses, to a judge (GPT-4). The judge evaluates the quality of generated responses from models, and gives an overall score on a scale of 1 to 9. We report the relative scores w.r.t. the GPT-4 model that has the textural ground truth description as input. Details are available in Appendix D.

Model	Average	Reasoning	Conversation	Detail
LLaVA-v1.5-7B	57.9	73.2	62.8	55.4
LLaVA + sentence-level RAG	59.6	74.4	64.1	57.8
Ours	**68.0**	**78.9**	**74.4**	**65.9**

Table 8: Results on Fashion-Bench. Sentence-level RAG refers to using retrieved captions as in-context prompts for LLaVA-v1.5-7B.

Results. We use a combination of fashion data as our retrieval datastore, including: FashionGen (Rostamzadeh et al., 2018) training set, Fashion200k (Han et al., 2017) and PolyvoreOutfits (Vasileva et al., 2018), resulting in a total of 546.5K image-text pairs. We extract tags of a product from captions. Results in Tab. 8 demonstrates the effectiveness of TUNA.

5 Conclusion

In this paper, we discussed three challenges for MLLMs: (1) mention of non-existent objects, (2) neglect of visual details and (3) failure to identify novel objects and entities, and one of the potential causes: the bottleneck from the image-to-text translation. To alleviate these problems, we introduced TUNA, a tag-grounded visual instruction tuning framework with retrieval-augmentation, which achieves competing performance over 12 VQA and multimodal benchmarks, compared to baselines with the same LLM and finetuning data.

Limitations

Being lightweight and effective, our model could be easily further improved with simple modifications to overcome existing limitations. Our model is bottlenecked by the capability of CLIP (Radford et al., 2021), which can affect our model performance in two ways. First, the quality of retrieved images are highly related to it. As we use tags associated to the retrieved images as additional information, more relevant images we have, more relevant tags we obtain. Second, our adaptive weight tuner also relies on the knowledge of CLIP. For instance, even if we obtain a highly relevant tag, e.g., "Diamond Head" from the retrieved similar images, if image-text pairs containing "Diamond Head" do not exist in the 400M pre-training data of CLIP, CLIP cannot effectively align the text embeddings of "Diamond Head" to a photo of diamond head, subsequently, low weights would be assigned to the tag "Diamond Head" in our weight tuner, even though it is the ground truth. Fortunately in most cases, CLIP is capable of handling it. If not, we can easily replace CLIP with a more powerful vision-language models.

Our current design of the retriever is also simple, where we retrieve images regardless of the language instruction. A solution could be using Q-former (Li et al., 2023d), where instruction-aware visual features could be used for retrieval. We leave them for future work.

Acknowledgement

The work is in part supported by the National Science Foundation under Grants IIS-2316306 and CNS-2330215, and a gift from Adobe.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023c. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471.

Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. 2021. VIVO: visual vocabulary pre-training for novel object captioning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence*, pages 1575–1583.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A

Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23369–23379.

Yuncheng Hua, Yuan-Fang Li, Guilin Qi, Wei Wu, Jingyao Zhang, and Daiqing Qi. 2020. Less is more: Data-efficient complex question answering over knowledge bases. *Journal of Web Semantics*, 65:100612.

Xinyu Huang, Youcai Zhang, Ying Cheng, Weiwei Tian, Ruiwei Zhao, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Xiaobo Zhang. 2022. IDEA: increasing text diversity via online multi-label recognition for vision-language pre-training. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4573–4583. ACM.

Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. 2023. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. 2023c. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. *arXiv preprint arXiv:2311.15879*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023e. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zhuang Li, Yuyang Chai, Terry Zhuo Yue, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. 2023f. Factual: A benchmark for faithful and consistent textual scene graph parsing. *arXiv preprint arXiv:2305.17497*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2024. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Daiqing Qi, Handong Zhao, and Sheng Li. 2023. Better generative replay for continual federated learning. *Preprint*, arXiv:2302.13001.

Daiqing Qi, Handong Zhao, and Sheng Li. 2024a. Easy regional contrastive learning of expressive visual fashion representations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Daiqing Qi, Handong Zhao, Aidong Zhang, and Sheng Li. 2024b. Generalizing to unseen domains via text-guided augmentation. In *European Conference on Computer Vision (ECCV)*.

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

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023a. Retrieval-augmented image captioning. *arXiv preprint arXiv:2302.08268*.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023b. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Heng-Shiou Sheu, Zhixuan Chu, Daiqing Qi, and Sheng Li. 2022. Knowledge-guided article embedding refinement for session-based news recommendation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7921–7927.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. 2018. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European conference on computer vision (ECCV)*, pages 390–405.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. 2023. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Handong Zhao, Zhengming Ding, and Yun Fu. 2017. Multi-view clustering via deep matrix factorization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 13041–13049.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Ronghang Zhu, Dongliang Guo, Daiqing Qi, Zhixuan Chu, Xiang Yu, and Sheng Li. 2023b. Trustworthy representation learning across domains. *Preprint*, arXiv:2308.12315.

Ronghang Zhu, Dongliang Guo, Daiqing Qi, Zhixuan Chu, Xiang Yu, and Sheng Li. 2024. A survey of trustworthy representation learning across domains. *ACM Trans. Knowl. Discov. Data.* Just Accepted.

A Extended Related Works

A.1 Retrieval-Augmented Multimodal Learning

We are distinct from existing works on retrieval-augmented multimodal learning (Ramos et al., 2023b,a; Yang et al., 2023; Hu et al., 2023; Lin et al., 2024; Li et al., 2023c) in that we are motivated from the object-oriented challenges in visual instruction tuning, which leads to notable differences in (1) target task, (2) motivations, (3) retrieved knowledge and (4) usage of additional information.

Most existing works above focus on image captioning, where short captions (usually one or two sentences) are generated given an input image. While in our case, our model is asked to follow the given instruction, infer from the given image, and often provide a long and detailed response. The difference of tasks therefore lead to different challenges, thus the motivation of using retrieval-augmentation is also distinct. While existing models exploit retrieved captions for general purposes of providing related contents to help the captioning of the current image (e.g., help to better organize the language, or provide additional knowledge on image content or context), in our scenario, the retrieved *tags* aim to provide rich object-aware information to enhance the attention to object details, and help with the object or entity identification. Moreover, the capability of performing tag-grounded generation is enabled during our visual instruction tuning. In addition, we have meticulously crafted novel modules aimed at enriching the representation of retrieved tags and adaptively reallocating the attention to them based on their relevance.

A.2 Multimodal Learning with Tags

We are distinct from existing works (Huang et al., 2023; Zhou et al., 2020; Li et al., 2020; Hu et al., 2021; Huang et al., 2022) that introduce object tags as anchor points to help the learning of semantic alignments between images and texts in (1) substantially different objectives, (2) type of used tags and (3) the usage of them.

Existing works (Huang et al., 2023; Zhou et al., 2020; Li et al., 2020; Hu et al., 2021; Huang et al., 2022) use tags for the representation learning of semantic alignments between images and texts. For instance, OSCAR (Li et al., 2020) propose to use object tags to align the object-region features in the pre-trained linguistic semantic space. Wu et al. (Wu et al., 2016) utilize solely the predicted object tags as input to an LSTM for image captioning, whereas You et al. (You et al., 2016) incorporate both tags and region features. In contrast, Zhou et al. (Zhou et al., 2020) augment region features with the object prediction probability vector, leveraging salient regions identified by object detectors, to enrich the visual input for pre-training. In our case, object-oriented tags are used as groundings to provide additional information on the given input image, therefore alleviating neglect of object details and failure to identify novel objects or entities. Besides, the capability of tag-grounded instruction-following in our model is also unique. The large and abundant annotation-free tags we have (around 3.2M) also makes our work distinctive from the above. As we want to inform our model of more relevant object-oriented knowledge like object names, object attributes while ignoring less relevant ones, we also design new modules towards this end.

A.3 Continual Learning of Multimodal Large Language Models

Continual Learning aims to continuously learn a model from new data in different manners, such as class-incremental (Qi et al., 2023), data-incremental (Sheu et al., 2022; Hua et al., 2020) and domain-incremental (Qi et al., 2024b; Zhu et al., 2023b, 2024).

Zhai et al. (Zhai et al., 2023) studies the continual learning of multimodal large language models in the context of object classification. They demonstrate that the finetuned popular open-source MLLMs, such as LLaVA (Liu et al., 2024), exhibited degraded performance compared to their pre-trained frozen vision encoders, such as CLIP (Radford et al., 2021). It is an example of the problem caused by the misalignment between the CLIP embeddings of the input image and the LLM text embeddings, as we illustrated in the Introduction Section.

B Tag Mining

To mine tags from texts, we parse each caption into a set of tags with a combination of FACTUAL scene graph parser (Li et al., 2023f) and Named Entity Recognition (NER) with spaCy, yielding 3M tags extracted from 15M captions in CC3M (Sharma et al., 2018) and CC12M (Changpinyo et al., 2021). We show several examples in

Caption: Montepulciano, Italy - August 25, 2013: Old narrow street in the center of town with colorful facades. stock photography.
Results from Parser:
Entity (Attributes): facades (colorful) | lawn (narrow,old) | town
Relations: street (Subject) - in center of (Relation) - town (Object)
Results from NER: Montepulciano | Italy
Tags: ['facades', 'narrow', 'town', 'Italy', 'colorful', 'street', 'old', Montepulciano]

Figure 8: Examples of tags derived from parsing and NER results.

Hyperparameters	
Batch Size	128
Learning Rate	2×10^{-5}
Learning Rate Schedule	Cosine Decay
Learning Rate Warmup Ratio	0.03
Weight Decay	0
Epoch	1
Optimizer	AdamW
DeepSpeed Stage	3

Table 9: Hyperparameters for Instruction Finetuning.

Fig. 8.

Given that the FACTUAL scene graph parser (Li et al., 2023f) is built on a large language model, there is a slight probability that it may produce nonsensical lengthy sequences. We employ a filtering mechanism to exclude tags exceeding 30 characters in length.

C Training Details

C.1 Datasets

LLaVA-665K (Liu et al., 2023a) is collected and built with a variety of datasets, containing VQA, OCR, region-level VQA, visual conversation and language conversation data. In ShareGPT4V (Chen et al., 2023c), the supervised fine-tuning captions were collected from GPT4-Vision. Following Chen et al. (Chen et al., 2023c), a corresponding portion of detailed captions in the Supervised Fine-Tuning (SFT) datasets (i.e., LLaVA-665K) is replaced with a selection from the 100K GPT4-Vision-generated captions.

C.2 Hyperparameter

We follow the hyperparameter setting in LLaVA-1.5 (You et al., 2016). Details are summerized in Tab. 9.

D Zero-Shot Inference on Fashion Data

D.1 Fashion-Bench

To explore the effectiveness of TUNA on OOD data from another specific domain, we further collect data from FashionGen (Rostamzadeh et al., 2018) validation set and create a benchmark to measure the model's instruction-following capability in Fashion domain. Following LLaVA (Liu et al., 2024), we leverage GPT-4 to measure the quality of generated responses. Specifically, we create triplets consisting of image, ground-truth textual descriptions, and question. The candidate models (e.g., TUNA, LLaVA) predict the answers based on the question and the image. To provide an approximate upper bound, we build a reference prediction based on the question and the ground-truth textual descriptions, using the text-only GPT-4, following Liu et al. (Liu et al., 2024) . After obtaining the responses from both models, we feed the question, visual information (in the format of textual descriptions), and the generated responses from both assistants, to the judge (i.e., text-only GPT-4). The text-only GPT-4 evaluates the helpfulness, relevance, accuracy, and level of detail of the responses from the assistants, and gives an overall score on a scale of 1 to 9, where a higher score indicates better overall performance. We report relative scores w.r.t. the text-only GPT-4 model that uses the textural ground truth description as visual input.

Similar to LLaVA-Bench (In-the-Wild) (Liu et al., 2024), we also collect a set of 24 images from FashionGen (Rostamzadeh et al., 2018) validation set, with one question associated with each image. The questions are from one of the three types:

1. Conversation. We design a conversation between the assistant and a person asking questions about the product. A diverse set of questions are asked about the content of the image, including the product brands, categories, materials, etc. Only questions that have definite answers are considered. E.g., `What is the brand of this product?`

2. Detailed Description. We ask the assistant to give a comprehensive and detailed desperation of the given product. E.g., `Please describe the product in this image in detail.`

3. Complex Reasoning. The above two types focus on the visual content itself, based on

which we further create reasoning questions. E.g., `What occasions is this clothing suitable for?`

D.2 Experiments

Model	Average	Reasoning	Conversation	Detail
LLaVA-v1.5-7B	57.9	73.2	62.8	55.4
Ours	**68.0**	**78.9**	**74.4**	**65.9**

Table 10: Results on Fashion-Bench. Our model consistently outperforms the baseline.

We use a combination of fashion data as our retrieval datastore, including: Fashion-Gen (Rostamzadeh et al., 2018) training set, Fashion200k (Han et al., 2017) and PolyvoreOutfits (Vasileva et al., 2018), resulting in a total of 546.5K image-text pairs. To obtain the tags of a product, we extract them from the caption or associated product specifications (e.g., brand) of the product.

Results in Tab. 10 demonstrates the effectiveness of TUNA, especially on 'Conversation' and 'Detail', where retrieved tags on product specifications are very helpful to identify the related details of the input product. Examples are available in Fig. 9 and Fig. 10.

E More Examples

We present more examples with TUNA and LLaVA-1.5 in Fig. 11 and Fig. 10. In Fig. 11, we provide Out-of-Distribution (OOD) images of real-world products or television works, and ask TUNA and LLaVA-1.5 to provide answers to the question. In Fig. 10, we provide Out-of-Distribution (OOD) images in fashion domain, and ask the models to provide answers to the question.

When provided with OOD images, where novel objects or entities often appear, LLaVA-1.5 fails to correctly or precisely identify them due to a limited number of training samples. Although the CLIP vision encoder, which is pre-trained with over 400M samples, can effectively extract their visual features, the multimodal connector cannot effectively map them to text embeddings input to the LLM. In contrast, TUNA is effective in identifying unseen objects or entities, as the input OOD image is directly mapped to a set of retrieved tags from a large-scale external datastore, which has a better coverage of OOD data.

In examples in Fig. 10, where specific in-domain knowledge, i.e., fashion domain, is required for give a detailed and precise description of the given product, such as its brand, design, or composition (material), LLaVA fails to correctly identify them or response with detailed descriptions on them.

For instance, in the example in Fig 9, the only useful information about the given product itself is "a black jacket with white polka dots", where LLaVA-1.5 fails to precisely describe it as a "blazer". Moreover, LLaVA-1.5 does not mention its design and brand even if we explicitly ask it the brand of this product. In contrast, TUNA precisely describes its design details, style and the brand, benefiting from the retrieved products which are similar to the input product in design, brand, category or style. TUNA could effectively refer to the retrieved tags and learn from the useful ones with our tag encoder.

Cases are similar in examples from Fig 11, where TUNA correctly identifies the novel object in the input image with retrieved knowledge. Meanwhile, LLaVA-1.5 fails to identity the model of the Leica camera, Porsche car, and the name of the character and anime in the input images.

F Full Experiment Results

We show the full results on POPE in Tab. 11.

G Benchmarks

We compare TUNA with SoTA methods on 12 benchmarks, including five VQA benchmarks: VQA^{v2} (Goyal et al., 2017), GQA (Hudson and Manning, 2019), VizWiz (Gurari et al., 2018), ScienceQA-Image (SQA^{I}) (Lu et al., 2022), TextVQA (VQA^{T}) (Singh et al., 2019), and seven more recently multimodal benchmarks designed for LLMs: POPE (Li et al., 2023e), MME (Fu et al., 2023), MMBench (MMB) (Liu et al., 2023b), MMBench-Chinese (MMB^{CN}) (Liu et al., 2023b), SEED (Li et al., 2023b), LLaVA-in-the-Wild (LLaVA^{W}) (Liu et al., 2023a), and MM-Vet (Yu et al., 2023).

VQA^{v2} (Goyal et al., 2017) and VizWiz (Gurari et al., 2018) are benchmarks for traditional Visual Question Answering (VQA) tasks. MME (Fu et al., 2023) evaluates LLMs' assesses and cognition capabilities through a wide range of carefully crafted questions across 14 sub-tasks. MMBench (MMB) and MMBench-Chinese (MMB^{CN}) (Liu et al., 2023b) benchmarks manually design ques-

Question: Please describe this product in detail, including its brand, category, etc.

LLaVA-1.5: The image features a man wearing **a black jacket** with **white polka dots**. The jacket is likely a fashionable and stylish piece that stands out. The man is posing for the camera, showcasing the unique design of the jacket. The outfit suggests that the man is confident and comfortable in his clothing choice..

Ours: The image features a man **wearing a black blazer** with **a white star pattern**. The blazer is part of the **Neil Barrett brand**, and **it is a long-sleeve wool-blend garment**. The man is standing in front of a white background, which highlights the blazer's design. The blazer is a versatile piece that can be worn in various settings, making it a popular choice for both formal and casual occasions.

Figure 9: Examples of results on Fashion-Bench. Precise and accurate descriptions are marked green while vague or wrong descriptions are marked red.

Datasets	Metrics	Ours	Ferret	Shikra	InstructBLIP	MiniGPT4	LLaVA	MM-GPT	mPLUG-Owl
Random	Accuracy (↑)	**91.00**	90.24	86.90	88.57	79.67	88.00	50.10	53.97
	Precision (↑)	**98.05**	97.72	94.40	84.09	78.24	97.44	50.05	52.07
	Recall (↑)	84.10	83.00	79.26	95.13	82.20	78.80	**100.00**	99.60
	F1 Score (↑)	**90.93**	89.76	86.19	89.27	80.17	87.13	66.71	68.39
Popular	Accuracy (↑)	**90.16**	84.90	83.97	82.77	69.73	87.43	50.00	50.90
	Precision (↑)	**95.46**	88.24	87.55	76.27	65.86	95.24	50.00	50.46
	Recall (↑)	84.20	80.53	79.20	95.13	81.93	78.80	**100.00**	99.40
	F1 Score (↑)	**90.56**	84.21	83.16	84.66	73.02	86.24	66.67	66.94
Adversarial	Accuracy (↑)	**88.43**	82.36	83.10	72.10	65.17	85.50	50.00	50.67
	Precision (↑)	**91.99**	83.60	85.60	65.13	61.19	90.99	50.00	50.34
	Recall (↑)	84.20	80.53	79.60	95.13	82.93	78.80	**100.00**	99.33
	F1 Score (↑)	**87.63**	82.00	82.49	77.32	70.42	84.45	66.67	66.82
Average F1		**89.50**	85.32	83.94	83.75	74.53	85.94	66.68	67.38

Table 11: Results on POPE. We outperform competing baselines including Ferret (You et al., 2023), which is finetuned on grounding and referring data.

tions to evaluate the LLM's visual reasoning and perception abilities in English and Chinese, respectively. SEED (Li et al., 2023b) generated a dataset comprising around 19K questions with images and videos with the GPT4 assistance.

H Analysis on Choices of Datastores

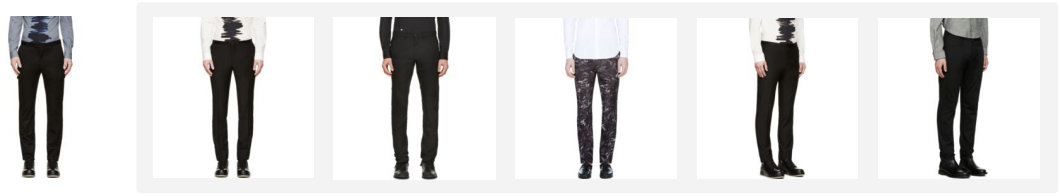
From Tab. 6 and previous analysis we know that the quality of retrieved tags is critical. Therefore, the datastore, where the images are (with corresponding tags) retrieved from is crucial. Here we study how different choices of datastores can affect the model performance. In the default setting, we use a combination of CC12M (Changpinyo et al., 2021), CC3M (Sharma et al., 2018) and COCO training set (Lin et al., 2014). Two of the three retrieval datasets, CC3M and the COCO training set, share overlaps with the LLaVA training data, which is a frequent scenario in retrieval-augmented generation, where a datastore with full or partial overlap with the training data is common (Ramos et al., 2023b,a; Yang et al., 2023; Hu et al., 2023; Lin et al., 2024; Li et al., 2023c). While CC12M and CC3M are different in size but similar in content style, COCO is different from them in both size and content. CC12M and CC3M consist of web image-text pairs, where the variance in caption quality and style is more significant. In COCO, captions are human-written, where the language style is more coherent, usually a short and plain description of the image. Consequently, tags extracted from COCO captions are often commonly used words and phrases and are very general, for instance, "boy", "girl", "plane" and "train", etc. It can provide the existence of objects in the image, which might help to alleviate the mention of non-existent objects. However, it is hard help to improve object or entity identification as these commonly seen phrases are very likely to be already included in LLaVA training data and new retrieval mappings cannot be established. On the contrary, CC12M and CC3M provide an ocean of novel ob-

jects and entities, which could greatly improve the image-to-text translation process with additional new retrieval mappings built from them.

We are curious to see how different datastore size and datastore style can influence our model performance. In additional to the default setting, we perform the tag-grounded instruction tuning with different datastores, and use them for retrieval during inference, respectively. Results are available in Tab. 7.

It is not surprising that the default setting with largest datastore size consistently outperforms other baselines. In most cases, the baseline with CC12M is the second best one while the one with COCO training set performs worst, except for on POPE. This is because POPE is built with COCO validation set, which shares the same style of the COCO training set. On other multimodal benchmarks, the improvements with COCO training set is less than CC12M and CC3M. Particularly, in LLaVA-in-the-Wild (LLaVA-W) benchmark, where all test images are not overlapped with COCO training and validation set, COCO training set as datastore does not help at all.

Input Image **Retrieved Images**

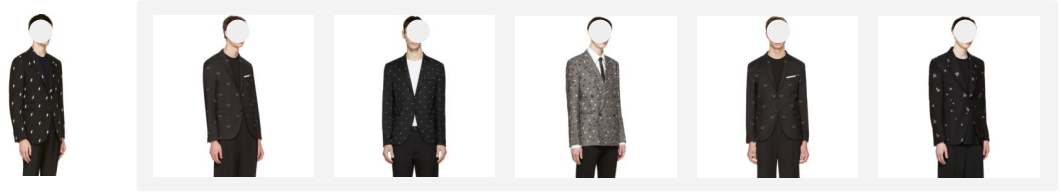


Question: Please describe this product in detail, including its brand, category, etc.

LLaVA-1.5: The image features a man wearing a black pair of pants with a blue design. He is standing in a white background, showcasing the **pants' design and fit**. The pants appear to be well-fitted and stylish, making them an attractive choice for the man.

Ours: The image features a man wearing a pair of black pants with a blue design. The pants are part of the **Calvin Klein Collection**, which is known for its stylish and fashionable clothing. **The pants are made of cotton and have a slim-fit design**, making them a popular choice for both casual and formal occasions. The man is standing in front of a white background, showcasing the pants' appearance and design.

Input Image **Retrieved Images**

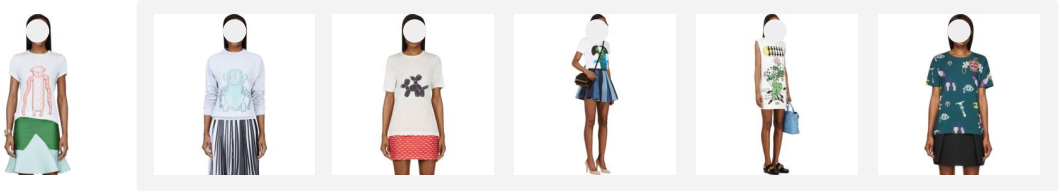


Question: Please describe this product in detail, including its brand, category, etc.

LLaVA-1.5: The image features a man wearing **a black jacket** with **white polka dots**. The jacket is likely a fashionable and stylish piece that stands out. The man is posing for the camera, showcasing the unique design of the jacket. The outfit suggests that the man is confident and comfortable in his clothing choice..

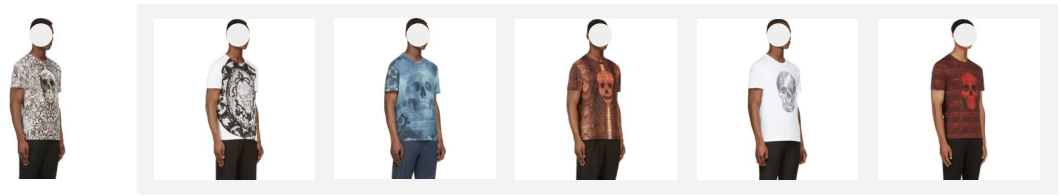
Ours: The image features a man **wearing a black blazer** with **a white star pattern**. The blazer is part of the **Neil Barrett brand**, and **it is a long-sleeve wool-blend garment**. The man is standing in front of a white background, which highlights the blazer's design. The blazer is a versatile piece that can be worn in various settings, making it a popular choice for both formal and casual occasions.

Input Image **Retrieved Images**



Question: What is the brand of this product?
LLaVA-1.5: The brand of this product is **Versace**.
Ours: The brand of this product is **Ostwald Helgason**.

Input Image **Retrieved Images**



Question: What is the brand of this t-shirt?
LLaVA-1.5: The brand of this product is **"Vans"**.
Ours: The brand of this product is **Alexander McQueen**.

Figure 10: Examples of results on Fashion-Bench. Precise and accurate descriptions are marked green while vague or wrong descriptions are marked red.

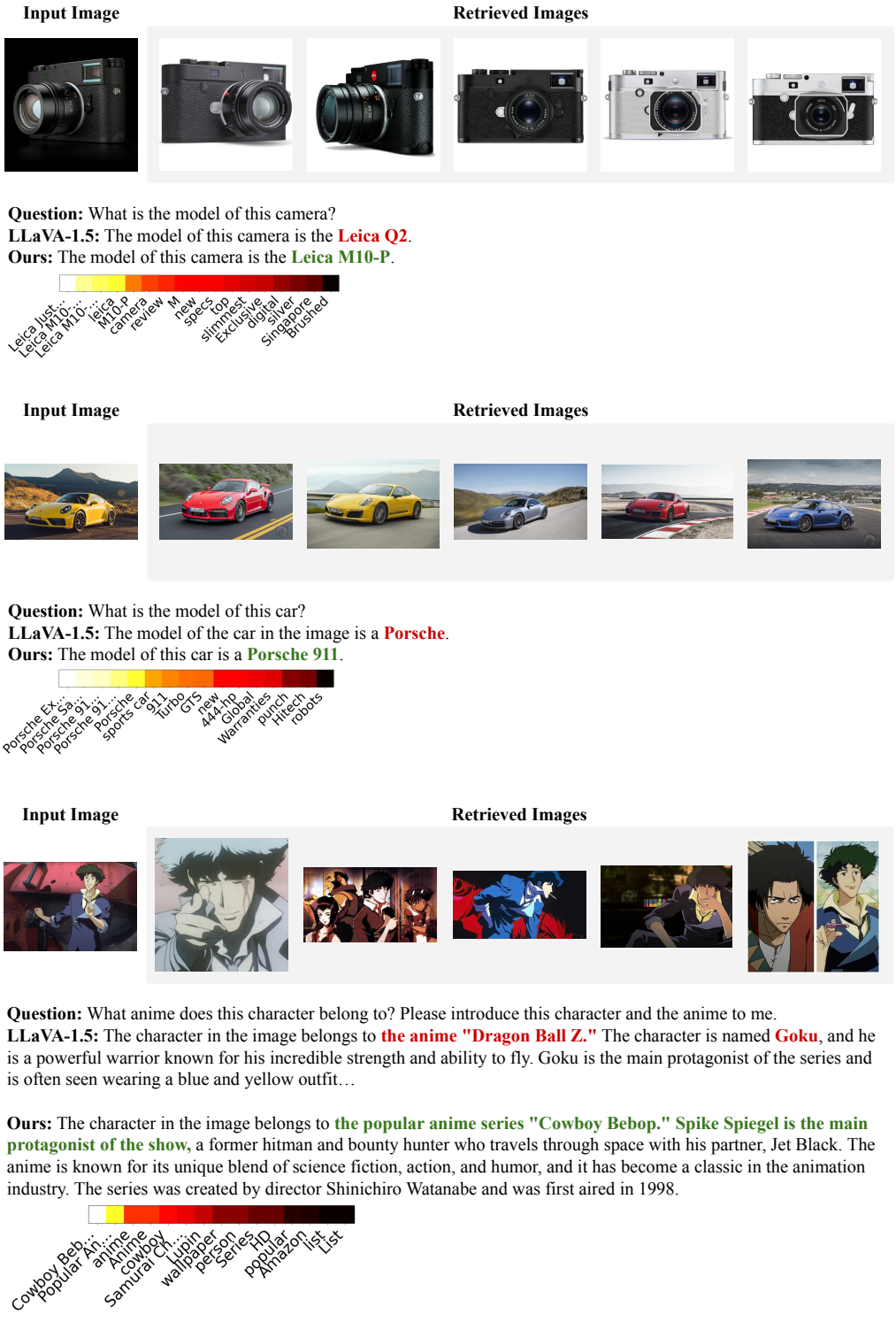


Figure 11: **TUNA examples.** For each example, we show top 5 retrieved images. We show the entire tag set associated with all retrieved images as well as their tuned weights in heatmap, where the brightest region for the highest weight 1 and darkest region for the lowest weight 0 (Zoom in for better view). Correct and precise answers are marked green while vague or wrong ones in red.