

Visually-Augmented Language Modeling

Weizhi Wang¹ Li Dong² Hao Cheng² Haoyu Song² Xiaodong Liu²
 Xifeng Yan¹ Jianfeng Gao² Furu Wei²

¹University of California, Santa Barbara

²Microsoft Research

weizhiwang@ucsb.edu {lidong1, haocheng, xiaodl}@microsoft.com

Abstract

Human language is grounded on multimodal knowledge including visual knowledge like colors, sizes, and shapes. However, current large-scale pre-trained language models rely on the text-only self-supervised training with massive text data, which precludes them from utilizing relevant visual information when necessary. To address this, we propose a novel pre-training framework, named VALM, to Visually-augment text tokens with retrieved relevant images for Language Modeling. Specifically, VALM builds on a novel text-vision alignment method via an image retrieval module to fetch corresponding images given a textual context. With the visually-augmented context, VALM uses a visual knowledge fusion layer to enable multimodal grounded language modeling by attending on both text context and visual knowledge in images. We evaluate the proposed model on various multimodal commonsense reasoning tasks, which require visual information to excel. VALM outperforms the text-only baseline with substantial gains of +8.66% and +37.81% accuracy on object color and size reasoning, respectively.

1 Introduction

Large-scale pre-trained language models (PLMs) have achieved great success in promoting the state-of-the-art on various natural language understanding and generation tasks (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019; Yang et al., 2019; Brown et al., 2020a). PLM self-supervision training largely benefits from harvesting local context information in the pre-training corpus. To further strengthen such contextual self-supervision, recent seminal works, e.g. GPT-3 (Brown et al., 2020a) and Megatron-LM (Narayanan et al., 2021), focus on increasing the model size and the scale of pre-training corpus. With billions of parameters, these tremendous PLMs exhibit incredible ability as zero-shot or few-shot learners. More remarkably, PLMs can achieve human-parity performance on various downstream tasks, even without any task-specific supervision. Another major research line of PLMs is to enhance the language model with auxiliary knowledge (Wei et al., 2021), including entity knowledge (Yu et al., 2020), relational knowledge (Zhang et al., 2019; Qin et al., 2021), text chunk (Lewis et al., 2020b; Wu et al., 2022; Borgeaud et al., 2021), etc. The incorporation of various knowledge resources to PLMs mitigates the drawbacks of *local* contextual attention, bringing additional relevant *global* context that benefits both language understanding and generation tasks.

The research in psychology and cognitive science (Carey, 1978) has demonstrated the existence of *fast mapping* in children’s word learning process. Three- or four-year-old children can connect the object with a new color word via *fast mapping*. Since current unimodal PLMs lack visual knowledge grounding, they inevitably suffer from the *hallucination* problem, which refers to the inconsistent or false statements generated by PLMs with respect to the world knowledge (Logan et al., 2019). For instance, the PLMs may predict the color of the sky as red only due to the statistical contextual correlations between the token “color” and “red” in the pre-training corpus, neglecting the commonsense facts.

In this paper, we propose a novel framework to enable language model pre-training to take full advantage of both local text context and corresponding visual knowledge. Recent work on joint vision-language model pre-training (Su et al., 2020; Tan and Bansal, 2020) relies on *explicit* alignments between text and image, e.g. supervised image captioning data, which limits the cross modality fusion during fine-tuning/inference over text without accompanying images. Instead, we design a flexible text-image alignment mechanism via an image retrieval module that gathers related images for each token as visual-augmentation. To achieve better language-vision grounding, we propose a visual knowledge fusion layer to enable the joint attention across visually-augmented context including both textual tokens and retrieved images. Based on this, we build up a **V**isually-augmented **L**anguage **M**odel, VALM, with flexible on-the-fly visual knowledge enhancement.

We evaluate the effectiveness of the proposed VALM on various commonsense reasoning and language modeling benchmarks. Experimental results demonstrate that our model consistently outperforms the unimodal baseline in terms of both object color and size reasoning. Remarkably, our method significantly improves the baseline method by +8.66% and +37.81% accuracy on MEMORY-COLOR and RELATIVESIZE. Additional experiments on language modeling tasks also validate that the proposed method achieves comparable or slightly better performance on language modeling. We further conduct systematic ablation studies to analyze the effect of the visual augmentation, and show that the proposed VALM creates memory and sensitivity to the augmented images. Our code will be open-sourced at <https://github.com/Victorwz/VaLM>.

Our contributions are summarized as follows:

- We propose a novel visually-augmented casual language model, VALM, to enable language model to utilize visual knowledge flexibly and effectively. VALM achieves cross-modality semantic learning grounded on both text and corresponding retrieved images. Through the proposed visual knowledge fused language modeling, VALM is capable of accomplishing tasks with the high demand of cross-modality knowledge, such as reasoning object colors and sizes.
- We design a framework to construct flexible on-the-fly text-image alignments and fuse augmented images into the context of language modeling. We implement an image retrieval module to query token-level representation in a large-scale cached image database and retrieve its nearest neighbors as the augmentation. With the proposed visual knowledge fusion layer, VALM can effectively take full advantage of both language information from local text context and visual information from retrieved images.
- The experimental results demonstrate that VALM effectively alleviates the hallucination problem of PLMs via introducing visual knowledge in language model pre-training. VALM achieves significant performance improvements on inferring the colors and sizes of common objects under zero-shot evaluations.

2 Methods

We propose a novel multi-modal pre-trained language model, which is augmented with retrieved images, named VALM. The architecture of VALM is presented in Figure 1. VALM augments each token in pre-training text corpus with k retrieved related images. VALM uses an image retrieval module to retrieve corresponding images for each token. The image retrieval module deploys a pre-trained CLIP model, which is capable of unifying the textual query and image candidates into a joint embedding space. VALM constructs a cached large-scale image knowledge base using image encoder of CLIP, and uses the contextual representation of each token as textual query to search its nearest neighbors in image knowledge base. With the help of the unified text and image embedding space provided by CLIP, the image nearest neighbors are taken as augmented images of each token to construct text and image alignments. We then propose a visual-knowledge fusion layer to enable learned hidden state to attend to both texts and augmented images.

2.1 VALM: Visually-Augmented Language Modeling

Given an input text sequence $\{\mathbf{x}_i\}_{i=1}^N$, the embedding layer first encodes input vector $\{\mathbf{x}_i\}_{i=1}^N$ into embedding space and outputs the initial hidden state \mathbf{H}^0 to the successive Transformer decoder layers. Then the proposed VALM model encodes \mathbf{H}^0 into visual knowledge fused contextual representations

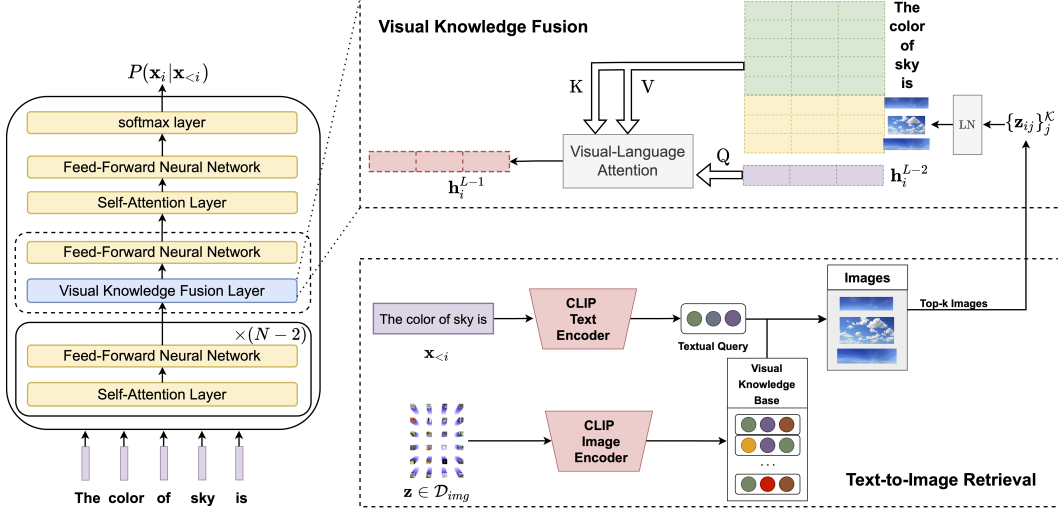


Figure 1: Overview of visually-augmented language modeling (VALM). We conduct dense retrieval to get top- k images for the input context at each time step. Then the visual knowledge fusion layer attends to both text tokens and retrieved images. The vision-language fused representation is fed back to Transformer for language modeling.

at difference levels $\mathbf{H} = \{\mathbf{H}^l\}_{l=1}^L$ via $L - 1$ Transformer decoder layers and one special visual knowledge fusion layer. Each Transformer decoder layer is identical to Vaswani et al. (2017), which outputs the contextual representations at different semantic levels given the representation from previous layer $\mathbf{H}^l = \text{Layer}_l(\mathbf{H}^{l-1})$, $l \in [1, L]$.

The visual knowledge fusion layer is proposed as a variant of Transformer decoder layer to incorporate visual knowledge in the contextual learning via joint-attention on both text contexts and augmented images. The visual knowledge fusion layer is injected in the second-to-last layer of VALM. The visual knowledge is stored in corresponding augmented image representations, obtained from image retrieval module $\{\{\mathbf{z}_{ij}\}_{j=1}^K\} = f_{rt}(\mathbf{x}_i)$. Then the visual knowledge fusion layer takes the input including both contextual representation of previous layer and augmented image sets and outputs a visual-knowledge fused contextual representation $\mathbf{H}^{L-1} = \text{VisualLayer}(\{\mathbf{H}_i^{L-2}, \{\mathbf{z}_{ij}\}_{j=1}^K\}_{i=1}^N)$. Finally, the output contextual representations are passed into output projection layer and a *softmax* function is used to compute the token probability $P(\mathbf{x}_i|\mathbf{x}_1, \dots, \mathbf{x}_{i-1}) = \text{softmax}(W\mathbf{H}^L + b)$.

We conduct *generative unsupervised pre-training* (Radford et al., 2019) for VALM on a large-scale text corpus. The training objective of VALM is the standard left-to-right language modeling objective, which maximizes the likelihood of the next word token based on the left context:

$$\max \sum_{\mathbf{x} \in \mathcal{D}} \sum_{i=1}^{|\mathbf{x}|} \log P(\mathbf{x}_i|\mathbf{x}_1, \dots, \mathbf{x}_{i-1}), \quad (1)$$

where \mathbf{x} represents a sentence randomly sampled from the large-scale pre-training text corpus \mathcal{D} .

2.2 Image Retrieval

The visual knowledge corresponding to a specific token is stored in its correlated images. Therefore, to prepare the fused visual knowledge, VALM deploys an image retrieval module to retrieve augmented images, denoted as $f_{rt}(\cdot)$. In order to achieve multi-modality text-image retrieval, it is of great importance to build up a discriminator to assess the correlation of every image in the extremely large-scale open image knowledge bases to the specific text representation. CLIP (Radford et al., 2021) proposed a simple-yet-effective method to connect images and texts into a unified multi-modal embedding space. We directly deploy the pre-trained CLIP model to encode the images and texts to enable a nearest neighbor text-image retrieval. Specifically, the pre-trained CLIP model we use in constructing the image retrieval module includes a ResNet-50x16 (He et al., 2016) model as an image encoder and a Transformer (Vaswani et al., 2017) model as a text encoder. Here, we only use the

CLIP model as the backbone of our image retrieval module, and the CLIP parameters are not updated during the pre-training process of VALM.

Image Knowledge Base Creation. The image knowledge base of the retrieval module is the cache of a set of image keys, which are the high-level visual representations of images. Given an image $\mathbf{z} \in \mathcal{D}_{img}$, such visual representation can be obtained via forwarding image \mathbf{z} to the pre-trained CLIP image encoder. Then the whole image knowledge base (\mathcal{Z}) is constructed by taking the output hidden state $f_{\theta_I}(\mathbf{x})$ as image keys: $\mathcal{Z} = \bigcup_{\mathbf{z} \in \mathcal{D}_{img}} \{f_{\theta_I}(\mathbf{z})\}$, where θ_I represents the image encoder parameters.

Textual Query. We take the contextual representation of each token as the query in the nearest neighbor search. For each sentence $\mathbf{x} \in \mathcal{D}$, the contextual representation of i -th token is computed via $f_{\theta_T}(\mathbf{x}_{<i})$, where θ_T represents the text encoder parameters. As the input sequence length of VALM generally exceeds the input length limitation of 75 tokens of CLIP text encoder, the long context $\mathbf{x}_{<i}$ is cut off into a context-chunk \mathbf{y}_i for fitting in CLIP text encoder:

$$\mathbf{y}_i = \begin{cases} \mathbf{x}_{[t, i-1]}, & i - t < 75, \\ \mathbf{x}_{[i-75, i-1]}, & i - t \geq 75, \end{cases} \quad (2)$$

where t is the index of the closest stop character. Then the textual query for i -th token is computed as its context-chunk representation as $f_{\theta_T}(\mathbf{y}_i)$.

k NN Text-Image Retrieval. The retrieval module uses the contextual representation to search the cached image knowledge base (\mathcal{Z}) and retrieves k nearest neighbor image keys w.r.t. dot product distance. As the pre-trained CLIP model has learned a joint embedding space for text and image domain, the retrieved images $\{\mathbf{z}_{ij}\}_{j=1}^{\mathcal{K}}$ are thus regarded as the top- k relevant images to the query.

2.3 Visual Knowledge Fusion

With the help of the image retrieval module, each token in the pre-training corpus is augmented with k corresponding images, and these augmented images are represented in the joint embedding space with texts. Then the augmented image representations are directly treated as auxiliary ‘‘context’’ in the learning process.

As the conventional Transformer decoder layer uses the multi-head self-attention (Vaswani et al., 2017) to learn the contextual representation, we extend it to a joint-attention mechanism and propose a novel visual knowledge fusion layer to enable each token to attend on both contexts and retrieval images jointly. In addition, due to the inconsistency in magnitude and distribution between contextual hidden states and retrieved image representations, we apply Layer Normalization (Ba et al., 2016) on retrieved \mathcal{K} image representations to alleviate such inconsistency, denoted as LN_{img} . Assume that the hidden state output for i -th token is \mathbf{h}_i and the corresponding retrieved images are $\{\mathbf{z}_{ij}\}_{j=1}^{\mathcal{K}}$, the hidden state \mathbf{H}_i^{L-1} is computed as:

$$\mathbf{Q} = \mathbf{H}^{L-2}W^Q + b^Q, \mathbf{K} = \mathbf{H}^{L-2}W^K + b^K, \mathbf{V} = \mathbf{H}^{L-2}W^V + b^V, \quad (3)$$

$$\dot{\mathbf{k}}_{ik} = \text{LN}_{img}(\mathbf{z}_{ik})W^K + b_{img}^K, \dot{\mathbf{v}}_{ik} = \text{LN}_{img}(\mathbf{z}_{ik})W^V + b_{img}^V, \quad (4)$$

$$e_i = \frac{\mathbf{Q}_i \mathbf{K}^T}{\sqrt{d}}, a_i = \frac{\exp(e_i)}{\sum_{j=1}^{\mathcal{L}} \exp(e_{ij}) + \sum_{k=1}^{\mathcal{K}} \exp(e_{ik})}, \quad (5)$$

$$e_{ik} = \frac{\mathbf{Q}_i \dot{\mathbf{k}}_{ik}^T}{\sqrt{d}}, a_{ik} = \frac{\exp(e_{ik})}{\sum_{j=1}^{\mathcal{L}} \exp(e_{ij}) + \sum_{k=1}^{\mathcal{K}} \exp(e_{ik})}, \quad (6)$$

$$\mathbf{H}_i^{L-1} = a_i \mathbf{V} + \sum_k a_{ik} \dot{\mathbf{v}}_{ik}, \quad (7)$$

where $\mathbf{Q}_i, \dot{\mathbf{k}}_{ik}, \dot{\mathbf{v}}_{ik} \in \mathcal{R}^E$, $\mathbf{K}, \mathbf{V} \in \mathcal{R}^{|\mathbf{x}| \times E}$, $e_i, a_i \in \mathcal{R}^{|\mathbf{x}|}$. The hidden state output from previous layer \mathbf{H}_i^{L-1} is linearly projected into contextual queries, keys, and values $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ separately. \mathcal{K} is the number of retrieved images for each token, and E is the embedding dimension for both context and image representations. In order to generate image-specific attention keys and values, we adopt image-specific bias b_{img}^K, b_{img}^V in linear projections and reuse the contextual projection weights

Task	Example Prompt	Object / Pair	Answer
Object Color Reasoning	<i>The color of [object] is [answer]</i>	<i>the sky</i>	<i>blue</i>
Object Size Reasoning	<i>Is [Item1] larger than [Item2]? [answer]</i>	<i>(sofa, cat)</i>	<i>Yes</i>

Table 1: Evaluation examples of object color reasoning and object size reasoning.

W^K, W^V to generate image-specific attention keys and values. Moreover, it is vital to mention that the image-specific attention keys and values are distinct for each query token, which is highly different from self-attention where the contextual keys and values are kept the same for each token. A secondary subscript k is used to denote different image representations for the i -th token.

3 Experiments

3.1 Setup

We evaluate our proposed VALM model on: a) *extrinsic* visual commonsense reasoning benchmarks of object color reasoning and object size reasoning; b) *intrinsic* language modeling benchmarks. We give two examples in Table 1 for visual commonsense reasoning tasks.

Text Corpus. We use the English corpus of CC-100 (Conneau et al., 2020) as the pre-training text corpus for both VALM and baseline models. CC-100 corpus is one of the largest high-quality web-crawl text data. The English monolingual dataset of CC-100 contains about 55 billion tokens, stored in 301 GiBs disk storage. Due to the limitation of computing resources, we only consume 15% of CC-100 English monolingual corpus for pre-training VALM and baselines.

Image Data. We use the LAION Open Image Dataset (Schuhmann et al., 2021) as the image knowledge base for dense retrieval. To the best of our knowledge, the LAION Open Dataset is the world’s largest openly available image-text-pair dataset with 400 million samples. In order to build up an efficient retrieval index and save storage space, we randomly select half of LAION images for the dense text-image retrieval, which is 200M images in total.

Pre-training Hyperparameters. The implementation of models and all experiments are based on the fairseq (Ott et al., 2019) toolkit. The proposed model VALM and baseline model GPT-BLIND share the same transformer decoder architecture with 123M trainable parameters, in which with $n_{\text{layer}} = 12, n_{\text{head}} = 12, d_{\text{embed}} = 768$. We deploy Adam (Kingma and Ba, 2015) ($\beta_1 = 0.9, \beta_2 = 0.98$) optimizer and train all models with $lr = 0.0005, t_{\text{warmup}} = 4000, \text{dropout} = 0.1, \text{bsz} = 128, \text{len} = 512$. The layer normalization over the retrieved image keys is initialized with ϵ of 0.00001. All the models are trained for 500k steps using 16 Nvidia Tesla V100-SXM2-32GB GPUs. VALM reuses the identical lower-cased byte pair encoding (BPE) (Sennrich et al., 2016) representation with a 49152 vocab size of CLIP text encoder.

Retrieval Module. For the implementation of the dense text-image retrieval module, we use the faiss (Johnson et al., 2021) toolkit to construct the efficient index. The faiss index contains the whole 200M image keys and provides the efficient nearest neighbor search. For efficiency purpose, we quantize all images keys to 32-bytes. faiss index stores image keys in clusters to speed up the search, which requires the additional training process to learn the cluster centroids. We use 10M keys for learning 131k cluster centroids and search 32 clusters to find the nearest neighbors during inference. We load the faiss index to GPU to achieve efficient dense text-image retrieval.

Baseline. We compare the proposed model with a unimodal baseline, GPT-BLIND, a text-only pre-trained auto-regressive language model (Brown et al., 2020b). We keep the training data, hyperparameter setting, and model size of baseline model the same as the proposed VALM. In addition, we consider two variants of VALM, VALM-distillation and VALM-random, as baselines. The two variants are only different from VALM in the inference stage. The former, **VALM-distillation**, blocks the proposed visual knowledge fusion process, making it degenerate into a text-only self-attention process during inference. VALM-distillation does not take augmented images as inputs and becomes a pure text language model during inference. The latter, **VALM-random**, replaces the

Model	\mathcal{K}	Object Color Reasoning (ACC \uparrow)		Object Size Reasoning (ACC \uparrow)
		MEMORYCOLOR	COLORTERMS	RELATIVESIZE
<i>Without Visual Augmentations</i>				
GPT-BLIND	N/A	44.14%	39.10%	47.22%
<i>With Visual Augmentations</i>				
VALM	4	52.70%	51.71%	85.03%
VALM	8	52.80%	49.36%	62.35%
<i>Ablation Variants of Our Method</i>				
VALM-distillation	4	41.39%	35.26%	46.19%
VALM-random	4	41.48%	38.46%	49.63%

Table 2: Accuracy on object color reasoning and object size reasoning benchmarks. \mathcal{K} represents for the number of images augmented into each token. Best performance is marked with bold.

corresponding images retrieved from image retrieval module with randomly selected images in the 200M image knowledge base during inference.

3.2 Object Color Reasoning

The visual information stored in retrieved images can play a significant role as useful knowledge to help language models perform high-quality commonsense reasoning. Such helpful visual information can be colors, positions, sizes, spatial relations, etc. The task of object color reasoning requires the pre-trained language models to predict the correct memory color of a given object. Due to reporting biases, such descriptive text about object colors rarely appears in text corpus, likely making this type of knowledge absent from text-only language models. We first evaluate the proposed model VALM on two public object color reasoning benchmarks, MEMORYCOLOR (Norlund et al., 2021) and COLORTERMS (Bruni et al., 2012) datasets. In order to demonstrate that VALM captures commonsense visual knowledge via pre-training, we evaluate the proposed model and the baseline in a zero-shot manner without any task-specific tuning. To achieve zero-shot object color reasoning, VALM takes the input consisting of textual prompts and objects during inference and predicts the color label as the last token. The prompts used in evaluating object color reasoning performance are listed in Appendix B. We use the top-1 accuracy as an evaluation metric and compute the average accuracy of all listed prompts to increase evaluation robustness.

MEMORYCOLOR Dataset. The *memory color* of a concrete object is the typical color an object appears in, e.g. the color of banana is mostly memorized as yellow. Norlund et al. (2021) proposed this dataset for evaluating visual knowledge transfer in multi-modal language models. The dataset contains 109 objects paired with their memory color labels, an illustrating picture, and a descriptor. The memory color dataset evaluates 11 color types.

COLORTERMS Dataset. The COLORTERMS dataset also contains a list of common items manually labeled with their commonsense color. COLORTERMS also holds a set of 11 color labels.

Results. The main results on MEMORYCOLOR and COLORTERMS datasets are presented in Table 2. The two variants of VALM ($\mathcal{K} = 4, 8$) significantly outperform baseline GPT-BLIND on both object color reasoning datasets. The proposed VALM model achieves the performance improvement of +8.66% and +12.61% on MEMORYCOLOR and COLORTERMS benchmarks respectively. The substantial improvements on both datasets demonstrate that VALM takes full advantage of visual knowledge (object color) to complete the corresponding visual commonsense reasoning. The results also present the effect of the number of augmented images, in which augmenting 8 images for each token does not lead to better improvement compared with fewer images of $\mathcal{K} = 4$. It is likely caused by more noise from a larger set of irrelevant images.

3.3 Object Size Reasoning

The task of object size reasoning requires the language model to predict the size relations between two given objects, e.g., an ant is smaller than an elephant. The size information is again rarely included

Model	\mathcal{K}	Wikitext-103 PPL↓	Lambada PPL↓	Lambada ACC↑
<i>Without Visual Augmentations</i>				
GPT-BLIND	N/A	36.44	42.46	42.17%
<i>With Visual Augmentations</i>				
VALM	4	35.78	42.51	42.65%
VALM	8	35.76	42.38	42.87%
<i>Ablation Variants of Our Method</i>				
VALM-distillation	4	39.22	44.59	41.20%
VALM-random	4	37.89	43.56	41.35%

Table 3: We report perplexity (PPL) on Wikitext-103 and Lambada, and also report final word prediction accuracy (ACC) on Lambada to evaluate the zero-shot language modeling capability on long texts. \mathcal{K} represents the number of images augmented into each token.

and described in text, while it is much easier to capture from the images. We convert the task to a binary classification form and construct 6 query prompts listed in Appendix B for the auto-regressive language model to perform the reasoning.

RELATIVESIZE Dataset. Bagherinezhad et al. (2016) proposed the relative size dataset, which includes a total of 486 object pairs between 41 physical objects. The RELATIVESIZE dataset requires the model to infer which object has a larger size for a given pair of objects.

Results. The main results of proposed VALM model and baselines on RELATIVESIZE are shown in Table 2. The proposed VALM of $\mathcal{K} = 4$ achieves an encouraging result with +37.81% accuracy gain over the text-only baseline. In addition, both visually-augmented variants ($\mathcal{K} = 4$ and $\mathcal{K} = 8$) substantially outperform the singular-modality model. Similar to object color reasoning, retrieving more images for each token again results in a performance drop for object size reasoning. We attribute the degradation to the increased noise brought by augmenting each token with more images which causes model confusion when differentiates relevant visual information from irrelevant one.

3.4 Language Modeling

The ability of modeling long texts enables PLMs to utilize knowledge and semantics in the pre-training corpus to complete downstream tasks. Therefore, the zero-shot language modeling performance is universally adopted to evaluate the capability of PLMs (Radford et al., 2019). We follow Radford et al. (2019) to evaluate the language modeling performance on several benchmarks in a zero-shot manner, including Wikitext-103 (Merity et al., 2017) and Lambada corpus (Paperno et al., 2016). We report perplexity for two corpora and also report last word prediction accuracy for LAMBADA corpus.

The results of VALM and baseline models on language modeling tasks are summarized in Table 3. Compared with the text-only model (GPT-BLIND), our visually-augmented model, VALM ($\mathcal{K} = 8$), slightly improves the perplexity on both datasets, +0.68 on Wikitext-103 and +0.08 on Lambda. A similar trend is also observed for final word prediction accuracy on Lambada corpus. Different from previous extrinsic visual commonsense reasoning tasks (object color and size), we find that VALM models with varying size of retrieved image set ($\mathcal{K} = 8$ vs $\mathcal{K} = 4$) perform similarly on the intrinsic language modeling task. The difference is likely caused by the task properties. In general, visual commonsense reasoning tasks require more fine-grained fusion of text and image, i.e. locating the text object in the image set, extracting relevant vision information, and verbalizing reasoning output. In contrast, a certain portion of text from generic language modeling corpus is probably not visually related. Thus, only a coarse-grained fusion is sufficient here (e.g. deciding if the image set is useful), making the language modeling evaluation less affected by the retrieval noise from augmented images.

3.5 Ablation Studies

The experimental results demonstrate the effectiveness and superiority of proposed model in utilizing visual knowledge on object color and size reasoning. In order to figure out how the visual knowledge

\mathcal{K}	W_{img}	b_{img}	MemoryColor ACC \uparrow	ColorTerms ACC \uparrow	RelativeSize ACC \uparrow	Wikitext-103 PPL \downarrow	Lambada PPL \downarrow	Lambada ACC \uparrow
4	✗	✓	52.70%	51.71%	85.03%	35.78	42.51	42.65%
8	✗	✓	52.80%	49.36%	62.35%	35.76	42.38	42.87%
4	✓	✓	47.81%	45.09%	83.54%	35.95	42.28	42.67%
8	✓	✓	53.11%	52.78%	68.88%	36.12	42.46	41.14%
4	✗	✗	46.08%	46.79%	62.04%	35.95	42.33	41.74%
8	✗	✗	49.64%	50.00%	63.48%	35.97	42.60	41.32%

Table 4: Ablation studies on the effects of attention key and value projection layers for augmented images. Checking represents that the image-specific weight or bias is adopted for both attention keys and values in visual knowledge fusion layer.

and information take effects in the model, several key questions arouse our interests: 1) Whether the model has learned a pattern or intelligence to utilize the visual knowledge in the language modeling process or the model just memorized all the images in the pre-training, making the learning more like a knowledge distillation process? 2) Whether the model learns and forms the sensitivity to the retrieved images? As the random replacement, swapping, and random insertions can make a huge difference to the semantics (Lewis et al., 2020a), can the randomly augmented images also make a difference to the model? 3) We make several contributions and innovations to the transformer decoder architecture. What are the true effects of the architecture innovations? To figure out the above-mentioned interesting questions, we conduct very systematic ablation studies to evaluate the effects of different components.

The two baselines of VALM-distillation and VALM-random are proposed to answer the first two questions and the results are presented in the last two rows of Table 2. Concluded from the ablation study results, VALM-distillation gets a remarkable performance degradation on all evaluation benchmarks, even lagging behind baseline model GPT-BLIND. We find that removing the image augmentations during inference also makes a huge difference to the language modeling, which is believed to be more related to contextual learning on the corpus rather than augmented images. Therefore, we may partially draw a conclusion that the learned semantics in VALM might be from both the pre-training corpus and also augmented images. Meanwhile, the fundamental zero-shot task transferability of VALM also establishes partial dependence on augmented images rather than fully captured from contextual self-supervised learning.

We also observe a remarkable performance decrease of VALM-random on all evaluation benchmarks. VALM-random and VALM-distillation realize similar performance decreases on object size and color reasoning. These results further illustrate that achieving high-quality visual knowledge utilization cannot be realized by only augmenting unrelated images to language models, while only corresponding images for each token can take effects. Besides, VALM-random achieves better zero-shot language modeling perplexity than VALM-distillation, proving that the augmented images do make contributions to language understanding and semantics learning for language models.

Ablation study on the effect of deploying image-specific linear projections for attention keys and values. VALM proposes a series of innovations in model architecture to adapt singular-modality transformer decoder to a novel framework which is capable of learning both contextual and visual knowledge. The proposed joint self-attention mechanism is a key component of VALM. The separate linear projection layers are regarded as significant components to map contexts into different embedding spaces for attention keys and values. Therefore, the proposed joint self-attention mechanism would naturally hold three potential solutions to generate image keys and values: reuse contextual linear projections, establish image-specific linear projections, and only differentiate linear bias for augmented images. We conduct the ablation study to evaluate the effect of these three solutions on image linear projections and the experimental results are shown in Table 4. The results demonstrate that adopting image-specific linear projections outperforms the reuse of contextual projections. The two types of image-specific linear projections do not lead to remarkable performance differences. Thus, we take the solution of only adding additional linear bias for augmented images and

reuse contextual linear weights in generating visual attention keys and values with the consideration of convenient implementation and less parameter introduction.

4 Related Work

Large-Scale Pre-trained Language Models. Since the emergence of BERT (Devlin et al., 2019), large-scale language model pre-training became dominant approach in universal natural language understanding tasks. The majority of novel pre-trained language models are based attention module (Bahdanau et al., 2015) and Transformer architecture (Vaswani et al., 2017), with the representatives like BERT (Liu et al., 2019), GPT (Radford et al., 2018), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), ELECTRA (Clark et al., 2020), GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), and T5 (Raffel et al., 2020). Meanwhile, the paradigm of pre-training with language modeling objective on large-scale text corpus and fine-tuning on specific downstream tasks has been universally adopted in the NLP community. Moreover, with the exponential increase of model size, a surprising fact emerged that the pre-trained language models like GPT-3 (Brown et al., 2020b) could work as few-shot or zero-shot learners.

Vision-Language Models. Different from natural language understanding and text generation tasks, vision-language tasks are at the intersection area of both modalities of vision and natural language, including visual-question answering (Agrawal et al., 2015), visual commonsense reasoning, and image captioning (Chen et al., 2015). Resolving vision-language tasks requires multi-modal models to learn cross-modal or universal representations. VL-BERT (Su et al., 2020) firstly proposed to extract corresponding region-of-interest (RoI) as visual features and then concatenate text tokens and image RoIs as joint inputs. OSCAR (Li et al., 2020) proposed to introduce object tags detected in images as anchor points to solve the issue of high demand for image-text alignments. Another significant pathway for establishing vision-language models is to construct a unified embedding space for text and images with image captioning data in pre-training and then deploy textual prompts to extract useful labels in completing downstream tasks during inference. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) are the representative works in this research line and achieve remarkable performance on both vision tasks and vision-language tasks.

Visually-Grounded Language Learning. Visually-grounded language learning is an emerging research topic with the goal of building up PLMs with visual grounding, in which the proposed VALM could be categorized in this area. Visual information and knowledge could be memorized by the language model via fusion layer or concatenated inputs. However, extracting and deploying the visual information efficiently and effectively is still difficult for singular-modality language models. Vokenization (Tan and Bansal, 2020) and its successor iACE (Lu et al., 2022) are a parallel research line with our work. Vokenization concatenated tokens and token-related images as vokens, transferring sentence-level caption text to token-level voken with a Vokenizer model.

5 Conclusion

In this paper, we propose a multi-modal framework to enable auto-regressive language modeling to effectively utilize visual knowledge. We propose an effective text-to-image retrieval module to construct text-image alignments, which is possible to bring hints to future works on visually-grounded language models. Empowered by pre-training, VALM realizes zero-shot task transfer on downstream visually-grounded tasks and natural language tasks. Experimental results on various object color and size reasoning benchmarks demonstrate the effectiveness of our model. VALM also achieves comparable performance with text-only baseline models on language modeling tasks. We would like to adapt the model architecture to encoder-only and encoder-decoder Transformer backbones. Moreover, we can adapt VALM to more input modalities in future work.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. VQA: Visual question answering. *International Journal of Computer Vision*, 123:4–31, 2015.

- Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016.
- Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. Are elephants bigger than butterflies? reasoning about sizes of objects. *ArXiv*, abs/1602.00753, 2016.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggione, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. Improving language models by retrieving from trillions of tokens. *ArXiv*, abs/2112.04426, 2021.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020a.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020b.
- Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. Distributional semantics in technicolor. In *ACL*, 2012.
- Carey. Acquiring a single new word, 1978.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, 2015.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, abs/2003.10555, 2020.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2020.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.
- Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547, 2021.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2020.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020a.
- Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401, 2020b.
- Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.
- Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. Barack’s wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1598. URL <https://aclanthology.org/P19-1598>.
- Yujie Lu, Wanrong Zhu, Xin Wang, Miguel P. Eckstein, and William Yang Wang. Imagination-augmented natural language understanding. *ArXiv*, abs/2204.08535, 2022.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *ArXiv*, abs/1609.07843, 2017.
- Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Ali Patwary, Vijay Anand Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei A. Zaharia. Efficient large-scale language model training on gpu clusters using megatron-lm. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021.
- Tobias Norlund, Lovisa Hagström, and Richard Johansson. Transferring knowledge from vision to language: How to achieve it and how to measure it? *ArXiv*, abs/2109.11321, 2021.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*, 2019.
- Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and R. Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *ArXiv*, abs/1606.06031, 2016.
- Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. Erica: Improving entity and relation understanding for pre-trained language models via contrastive learning. In *ACL*, 2021.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683, 2020.

- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *ArXiv*, abs/2111.02114, 2021.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *ArXiv*, abs/1508.07909, 2016.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *ArXiv*, abs/1908.08530, 2020.
- Haochen Tan and Mohit Bansal. Vokenization: Improving language understanding via contextualized, visually-grounded supervision. In *EMNLP*, 2020.
- Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.
- Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew O. Arnold. Knowledge enhanced pretrained language models: A comprehensive survey. *ArXiv*, abs/2110.08455, 2021.
- Yuhuai Wu, Markus N. Rabe, DeLesley S. Hutchins, and Christian Szegedy. Memorizing transformers. *ArXiv*, abs/2203.08913, 2022.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019.
- Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. Jaket: Joint pre-training of knowledge graph and language understanding. *ArXiv*, abs/2010.00796, 2020.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. In *ACL*, 2019.

A Case Studies

We provide a case study in the object color reasoning task for VALM. In order to reason the correct commonsense color of objects sky and parsley, VALM takes the input combination of the prompt and the object as "The color of [object] is". Then we present the retrieval results of top-4 corresponding images to the textual query in Figure 2.

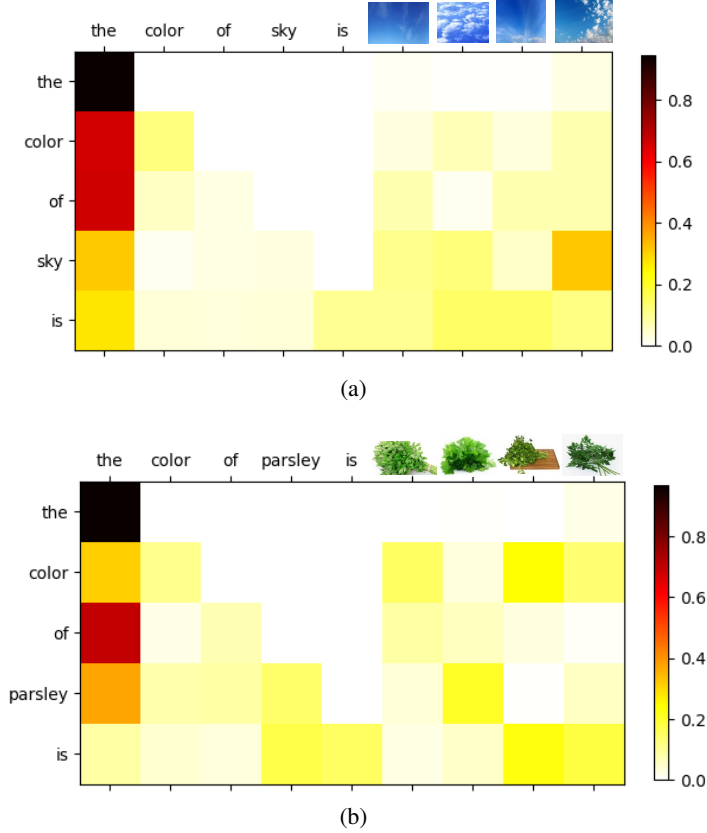


Figure 2: The attention matrix visualization given the query prompt “the color of [object] is” for VALM. VALM achieves accurate image retrieval of top-4 images corresponding to the objects of sky and parsley as augmented images, shown in the horizontal index of each subfigure.

B Probe Templates

Index	Prompt
1	Is [ITEMA] larger than [ITEMB]? [Label]
2	Is [ITEMA] taller than [ITEMB]? [Label]
3	Is [ITEMA] higher than [ITEMB]? [Label]
4	[ITEMA] is larger than [ITEMB], is it true? [Label]
5	[ITEMA] is taller than [ITEMB], is it true? [Label]
6	[ITEMA] is larger than [ITEMB], is it true? [Label]

Table 5: The prompts used to query the model predictions on the zero-shot evaluation of object size reasoning benchmark, RELATIVESIZE. As the task is transferred to a binary classification task, then the zero-shot evaluation task labels are Yes and No.

Index	Prompt
1	Q: What is the color of [DESCRIPTOR] [ITEM]? A: It is [Label]
2	What is the color of [DESCRIPTOR] [ITEM]? It is [Label]
3	What is the usual color of [DESCRIPTOR] [ITEM]? [Label]
4	What is the typical color of [DESCRIPTOR] [ITEM]?
5	The color of [DESCRIPTOR] [ITEM] is [Label]
6	The usual color of [DESCRIPTOR] [ITEM] is [Label]
7	The common color of [DESCRIPTOR] [ITEM] is [Label]
8	The typical color of [DESCRIPTOR] [ITEM] is [Label]
9	[DESCRIPTOR] [ITEM] usually has the color of [Label]

Table 6: The prompts used to query the model predictions on the zero-shot evaluation of object color reasoning benchmarks, MEMORYCOLOR and COLORTERMS datasets. The reasoning task prediction labels are a set of 11 color types.