WHAT DOES VISION SUPERVISION BRING TO LANGUAGE MODELS? A CASE STUDY OF CLIP

Anonymous authors
Paper under double-blind review

ABSTRACT

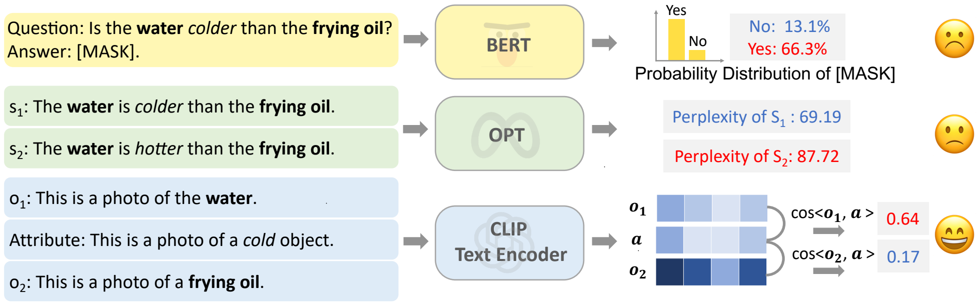
Vision-language (V+L) pre-training has shown promising performance in cross-modal tasks such as image-text retrieval and image captioning. On the other hand, these models surprisingly perform worse than text-only models (e.g., BERT) on widely-used text-only understanding tasks. The conflicting results naturally raise a question: What does vision supervision bring to language models? In this paper, we investigate this under-explored problem with one representative cross-modal model CLIP. We compare the text encoder of CLIP and widely-used text-only models on a wide range of tasks. We design a suite of evaluation tasks across three perception aspects, including the linguistic world featuring syntactic knowledge (e.g., dependency labeling), the visual world examining visual-related commonsense knowledge (e.g., color), and the embodied world featuring physical-related commonsense knowledge (e.g., mass). Experiments demonstrate that text-only models are not always better than CLIP on these perception tasks. Although the text encoder of CLIP falls far behind text-only models in linguistic-related tasks, CLIP achieves better zero-shot results in visual and embodied worlds with only 0.3% parameters compared to OPT-175B (one of the largest text-only models). This proves that CLIP can empower text encoders to learn rich visual and embodied knowledge through vision-text pre-training. Furthermore, qualitative studies show that CLIP pre-training yet restricts the text encoder from learning fine-grained semantics, like understanding ambiguous texts. These results shed light on future directions to improve V+L pre-training.

1 INTRODUCTION

Vision-language pre-training (V+L) has attracted increasing attention in recent years by jointly embedding images and text (Chen et al., 2019; Li et al., 2019; 2020; Jia et al., 2021). For example, CLIP (Radford et al., 2021), one of the representative V+L models, trains an image encoder and a text encoder separately via a contrastive objective. These pre-trained vision-language models excel at learning transferable visual and language representations and achieve promising results on downstream tasks such as image classification (Radford et al., 2021) and cross-modal retrieval (Li et al., 2021). It indicates that language supervision can indeed improve visual representations.

However, recent work shows that introducing visual supervision does not bring clear improvements on language tasks. Classical V+L models like VisualBERT (Li et al., 2019) and Oscar (Li et al., 2020) even under-perform vanilla text-only models like BERT on the Natural Language Understanding (NLU) benchmark GLUE (Tan & Bansal, 2020; Wang et al., 2019a). Theoretically, the vision-language models are expected to have greater potentials with the visual perception supervision (Bisk et al., 2020). In practice, they show worse performance than text-only models. These conflicting results naturally raise a question: What do visual signals bring to language models?

To figure it out, we first build a suite of tasks covering different perception levels to evaluate language models. Specifically, following the definition of Bisk et al. (2020), we define evaluation tasks in three world scopes: (1) *the linguistic world* ($\mathcal{W}_1$) probing syntactic and semantic knowledge, including tasks like dependency parsing and named entity recognition; (2) *the visual world* ($\mathcal{W}_2$), examining visual-related knowledge, including tasks like color-related commonsense understanding and material-related commonsense understanding; and (3) *the embodied world* ($\mathcal{W}_3$) evaluating physical-related knowledge that can only be learned via interaction with the world such as the mass



Question: Is the **water** *colder* than the **frying oil**? Answer: [MASK].

s_1: The **water** is *colder* than the **frying oil**.

s_2: The **water** is *hotter* than the **frying oil**.

o_1: This is a photo of the **water**.

Attribute: This is a photo of a *cold* object.

o_2: This is a photo of a **frying oil**.

Figure 1: An overview of the unified probing framework, where we design evaluation prompts for language model variants according to their pre-training characters. For BERT-like models with a masked language head, we convert the knowledge fact to a question and perform prediction with the head over yes or no. For the causal language model OPT, we evaluate the perplexity of different assertions and take the one with lower perplexity as a valid fact. For CLIP, we devise a matching-based probing framework, where a higher similarity of vectors with targeted attributes is adopted.

of objects. The tasks in the Linguistic world, the Visual world, and the Embodied world together compose our **LiVE** benchmark, **LiVE**-bench for short.

We then compare the text-encoder of CLIP and text-only models on **LiVE**-bench to examine the effects of visual signals on language models. We implement text-only model variants including masked language models BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b), and causal language models OPT-family (Zhang et al., 2022b) with parameters ranging from 125M to 175B. For $\mathcal{W}_1$, we adopt the widely-used edge probing framework (Tenney et al., 2019b) to investigate how much linguistic knowledge has been captured. For $\mathcal{W}_2$ and $\mathcal{W}_3$, we adopt prompt probing to evaluate how well language models understand visual and physical properties. In addition, we also design a unified probing framework to evaluate various language model variants according to their pre-training characters. More specifically, for masked language models, the visual and embodied knowledge facts are first converted into sentences, then the models predict whether the fact is valid by performing a cloze task (Schick & Schütze, 2021b). The causal language model OPT is evaluated by comparing the perplexities of different knowledge assertions. To evaluate the text encoder of CLIP which does not contain a language head, we design a matching-based framework, which compares the similarities between the objects and candidate attributes. The whole evaluation framework is illustrated in Figure 1.

Our key findings are: (1) In $\mathcal{W}_1$, CLIP and its enhanced variants with extra masked language modeling objectives including DeCLIP (Li et al., 2022) still fall far behind the masked language models BERT with the same configurations including model architecture and training corpus. (2) In $\mathcal{W}_2$, CLIP significantly outperforms text-only models with similar model sizes. Interestingly, we find that scaling OPT up to 175B can surpass CLIP on tasks like color and material recognition. (3) In $\mathcal{W}_3$, CLIP achieves significantly better results than text-only models. Notably, it even outperforms OPT-175B with few-shot demonstrations. Since previous work is limited to linguistic knowledge evaluation, we are the first to show visual supervision is beneficial for learning embodied knowledge. Finally, we provide some failure cases of CLIP under ambiguous text descriptions and discuss future directions to improve vision-text pretraining.

2 THE LIVE BENCHMARK

In this section, we introduce a comprehensive benchmark for evaluating language models from different perception aspects. Our LiVE benchmark is motivated by the well-known work of Bisk et al. (2020) that defines three-level perception abilities. We focus on the following three world scopes: $\mathcal{W}_1$: the linguistic world focusing on the syntactic and semantic knowledge (e.g., dependency parsing and named entity recognition), $\mathcal{W}_2$: the visual world covering the visual properties of objects (e.g., the color of an apple), and $\mathcal{W}_3$: the embodied world targeting at physical properties that can only be acquired via interaction with the world such as the mass of objects. The statistics and illustration of the benchmark datasets can be found in Table 1.

Table 1: Statistics and example illustrations of the proposed LiVE benchmark. For $\mathcal{W}_1$, we focus on the syntactic or semantic labels of [**text spans**] and the relations between them. For knowledge facts in $\mathcal{W}_2$ and $\mathcal{W}_3$, we design two forms of probing tasks. The former (Color, Shape and Material) asks models to make a choice between two tail options for the given head object; and the latter (Size, Height, and all $\mathcal{W}_3$ tasks) is to predict whether the relation is valid given the head and the tail.

	Dataset	**Instance**	**Label**	**# Test Examples**
$\mathcal{W}_1$	Coref	That, [**he**]$_1$ says , is just fine with [**him**]$_2$.	True	27,800
	Deps.	[**Click**]$_2$ [**here**]$_1$ To view it .	advmod	25,049
	NER	Back to [**the Middle East**] tonight .	LOC	12,586
	SRL	[**Four Palestinians**]$_2$ were shot and [**killed**]$_1$.	ARG1	61,716
	RC	Seniors get much [**joy**]$_2$ from [**animals**]$_1$.	Cause-effect	2,717
$\mathcal{W}_2$	Color	Head: **melon**, Tail$_1$: **green**, Tail$_2$: **black**	green	574
	Shape	Head: **lemon**, Tail$_1$: **triangle**, Tail$_2$: **round**	round	140
	Material	Head: **guitar**, Tail$_1$: **wood**, Tail$_2$: **glass**	wood	922
	Size	Head: **ant**, Rel: **larger than**, Tail: **bird**	False	500
	Height	Head: **bottle**, Rel: **shorter than**, Tail: **truck**	True	500
$\mathcal{W}_3$	Mass	Head: **wooden spoon**, Rel: **heavier than**, Tail: **toaster**	False	654
	Temperature	Head: **water**, Rel: **colder than**, Tail: **frying oil**	True	422
	Hardness	Head: **pearl**, Rel: **softer than**, Tail: **glass**	True	1,016

2.1 Linguistic Knowledge

To avoid the confounding effect of fine-tuning on language models, we do not choose traditional supervised evaluation benchmarks (e.g., GLUE). Instead, we evaluate the linguistic knowledge of different language models by adopting the edge probing benchmark provided by Tenney et al., which covers linguistic knowledge ranging from syntax to semantics. Specifically, we probe the model on various fundamental NLP tasks, including dependency labeling (Deps.), named entity recognition (NER), semantic role labeling (SRL), coreference resolution (Coref), and relation classification (RC). Following Fayyaz et al. (2021), we adopt OntoNotes 5.0 (Weischedel et al., 2013) for NER, SRL, and coreference prediction, the English Web Treebank of the Universal Dependencies (Silveira et al., 2014) for dependency labeling, and SemEval 2010 Task 8 dataset (Hendrickx et al., 2009) for relation classification.

2.2 Visual Knowledge

Perception is necessary for language learning because it forms the basis for many of our semantic axioms (Bisk et al., 2020). Among the various types of perception, visual signals can provide abundant information for modeling a vastness of experiences in the world that cannot be stated by text alone (Harnad, 1990). In this work, we consider visual commonsense understanding to evaluate language models. Specifically, we combine the recently proposed visual knowledge probing datasets, including Spatial Commonsense (Liu et al., 2022) and ViComTe (Zhang et al., 2022a). The combined dataset requires understanding various aspects of the visual world: including color, shape, material, size, and height. According to the format of task definition, these visual-related tasks can be divided into two categories. The former includes color understanding, shape understanding, and material understanding. Given an object, it requires a model to answer which label the object is. For example, the input is `(The color of banana is [MASK])`. The model is supposed to make a correct choice between the ground-truth answer `yellow` and an alternative option such as `black`. The latter includes size understanding and height understanding. These understanding tasks require the model to perform the comparison between different objects. For example, the input is `(Ant is [mask] than table)`. The model is asked to compare the size of paired objects and make choices between the ground-truth `smaller` and the antonym `larger`.

2.3 Embodied Knowledge

Understanding the physical realities is also an important aspect of perception. The embodied world contains knowledge that revolves around physical realities (e.g., mass, temperature), which are held by humans intuitively (Bisk et al., 2020). This kind of knowledge is the basis of intelligence and enables agent models to explore challenging tasks in physical environments. Strictly speaking, visual commonsense knowledge is also an important part of the physical world, which is evaluated in $\mathcal{W}_2$. In addition to visual knowledge, the connection between language and the multi-modal world also relies on other interactions with the embodied environment (Thelen & Smith, 1994), like tactile sensation. In this part, we focus on this interaction-related knowledge that cannot directly be evaluated only according to visual signals. We are curious about whether current language models can capture embodied knowledge such as the physical properties of objects. To explore this, as a first step toward evaluating embodied knowledge, we construct evaluation datasets regarding basic physical properties including mass, temperature, and hardness. The details of the construction of embodied knowledge datasets are elaborated below.

Mass Dataset We use the Image2Mass dataset curated by Standley et al., which annotates common objects with corresponding weights. The most light-weight object is a red lego brick, weighing 0.026 lbs, and the heaviest object is a 2.664 lbs drill. As directly asking the language model for the absolute mass of objects can be challenging (Wallace et al., 2019), we define the task in a comparison format. Specifically, each comparison pair contains two objects with a weight gap greater than 1 lbs.[1] We build 654 pairs like `(hair dryer, heavier than, red lego brick)` for evaluation.

Temperature Dataset We design a temperature probing dataset by collecting the temperature of common objects from Wikipedia.[2] For example, the ice is $0°$C and the temperature of water vapor is $100°$C. We convert the object with temperature annotations into pairs, and each pair contains two objects and the corresponding temperature relation. For example, `(ice, colder than, water vapor)`. The temperature gap between two objects is required to be greater than a difference threshold, which is loosely set to $10°$C for assurance of thermal perception for human (Jones, 2009). The final temperature dataset consists of 422 pairs in total.

Hardness Dataset Hardness (antonym: softness) is a measure of the resistance to localized plastic deformation in material science. Different material differs in hardness, for example, hard metals such as titanium are harder than soft mineral such as talc. Humans can perceive the hardness of different materials in interaction with the environment by using tactile organs like fingers (Gueorguiev et al., 2016). To investigate whether language models can capture hardness knowledge, we build a hardness dataset by collecting the Mohs hardness scores of different objects from Wikipedia.[3]. We define the task in a comparison format. For example, `(talc, softer than, titanium).` Each pair contains two objects. The gap between two objects is greater than the threshold for human-level understanding. The final dataset contains $1,016$ pairs.

3 LIVE-BENCH: LANGUAGE MODEL EVALUATION

As different language models vary in pre-training paradigms, to faithfully examine the knowledge learned during pre-training, we devise evaluation methods specific to different language models and knowledge types. For $\mathcal{W}_1$, we adopt the edge probing method (Tenney et al., 2019b) by investigating the linguistic information encoded in the representations of language models (§3.1). For $\mathcal{W}_2$ and $\mathcal{W}_3$, we adopt the prompting (Schick & Schütze, 2021b) for models with language heads capable of predicting words over a vocabulary, such as BERT and OPT. For CLIP models without a language head, we design a matching-based prompting method to fit its pre-training objective (§3.2).

3.1 EDGE PROBING FOR LINGUISTIC KNOWLEDGE

Edge probing (Tenney et al., 2019b) is a commonly adopted technique for measuring linguistic knowledge in contextualized representations of language models (Tenney et al., 2019a; Fayyaz et al., 2021). It trains a probing classifier on top of the span representations specified by the dataset, and the task performance metric is taken as the quality of the encoded information about the linguistic tasks. As recent studies have shown that evaluation metrics like F1-score can be influenced by the complexity of probing classifier and hyper-parameter settings (Hewitt & Liang, 2019; Belinkov, 2022), we adopt a more robust information-theoretic probing method, minimum description length (MDL) Voita & Titov (2020). The idea behind MDL is to reformulate the probing task into a data transmission problem, where the MDL metric measures the cost of transmitting the data, i.e.,

[1] The threshold is set according to the Weber–Fechner laws (Fechner, 1948) to guarantee that the mass difference is perceivable for humans.

[2] `https://en.wikipedia.org/wiki/Orders_of_magnitude_(temperature)`

[3] `https://en.wikipedia.org/wiki/Mohs_scale_of_mineral_hardness`

the codelength for encoding the data. Due to the space limit, we refer readers to Voita & Titov (2020) for more details of MDL. In our study, following Voita & Titov (2020) and Fayyaz et al. (2021), we report a derived metric of MDL, i.e., the compression ratio c, which further eliminates the effect of the number of training examples. A higher compression c indicates the representations contain richer information regarding the linguistic property.

3.2 Zero-Shot Prompting for Visual and Embodied Knowledge

Different from linguistic knowledge in $\mathcal{W}_1$ which focuses on the syntactic and semantics within sentences, our evaluation tasks in $\mathcal{W}_2$ and $\mathcal{W}_3$ consist of knowledge facts about the visual and embodied environment. Inspired by recent studies showing that prompting methods are effective for probing the knowledge that language models acquire during pre-training (Petroni et al., 2019; Schick & Schütze, 2021b), we propose to evaluate the knowledge in $\mathcal{W}_2$ and $\mathcal{W}_3$ with prompts for models with language model heads. For the language model of CLIP without a language model head, we develop a matching-based prompt framework for probing its learned knowledge.

Prompting Masked Language Models Following PET (Schick & Schütze, 2021b;a), we probe the masked language models by converting the knowledge fact into a question-answering form. For example, a size knowledge fact `(coin, smaller than, table)` is converted into a sentence with a special mask token: `Question: is a coin smaller than a table? Answer: [MASK]`. We also implement other diverse prompts, like `Is a coin [MASK] than table`? Experiments show that such a question-answering form can better induce models to generate answers. Given masked inputs, the model is asked to predict the probabilities of the mask token over two choices, i.e., `yes` for confirming the knowledge fact is valid or `no` for an unreasonable assertion. Zhao et al. (2021) found that the prediction can be biased towards some answers. we calibrate the prediction by normalizing the probabilities of the two option tokens according to the estimated biased probability distribution given empty queries.

Prompting Causal Language Models Different from BERT, there is no special `[MASK]` token during the pre-training of causal language models like GPT (Radford et al., 2019). Therefore, introducing a special token would result in inconsistency between pre-training and evaluation. To remedy this, for each knowledge fact, we state it in natural sentences and evaluate the sentence perplexity as the proxy metric for its validity. Specifically, for the size-property evaluation, we convert it into a valid knowledge assertion $s1 =$ `A coin is smaller than a table`, and an invalid one by replacing the size relation with the antonym adjective $s2 =$ `A coin is larger than a table`. The sentence with lower perplexity is then chosen as the prediction. To better extract knowledge from language models, we design diverse prompt templates in this work. We evaluate the perplexity of each sentence $s = (w_0, w_1, \cdots, w_n)$ as:

$$\text{PPL}(s) = P_{\mathcal{M}}(s)^{-\frac{1}{n}} = \sqrt[n]{\prod_{k=1}^{n} \frac{1}{P_{\mathcal{M}}\left(w_k \mid w_0, w_1, \ldots, w_{k-1}\right)}}, \tag{1}$$

where $P_{\mathcal{M}}$ denotes the conditional word probability of the causal language model to be probed and n is the number of tokens in s. We compare the perplexity PPL(s_1) and PPL(s_2) and choose the sentence with lower PPL as a more valid assertion and calculate the prediction accuracy accordingly.

Prompting CLIP Models Unlike masked and causal language models with language head that supports diverse prompting templates, the text encoder in CLIP only has one sentence representation without any pre-trained language heads. To probe the learned knowledge of CLIP language models, we design a matching-based prompting method. In more detail, for the size fact stated before, we first obtain two object descriptions $o_1 =$ `a photo of a coin`, and $o_2 =$ `a photo of a table`. These two sentences are encoded to get the corresponding object vectors via the CLIP language encoder:

$$\mathbf{o}_1, \mathbf{o}_2 = \text{CLIP}(o_1), \text{CLIP}(o_2). \tag{2}$$

We then derive an attribute sentence $a =$ `a photo of a small object`, and encode it to an attribute adjective vector with the language encoder:

$$\mathbf{a} = \text{CLIP}(a). \tag{3}$$

The prediction is then performed by comparing the cosine similarity $\cos(\mathbf{o}_1, \mathbf{a})$ and $\cos(\mathbf{o}_2, \mathbf{a})$. The object with higher similarity with the attribute description is adopted as the answer, i.e., a coin is smaller than a table, if $\cos(\mathbf{o}_1, \mathbf{a}) > \cos(\mathbf{o}_2, \mathbf{a})$. Otherwise, we assume that the model thinks the reversed relation holds. We can also adopt the antonym adjective *large* for getting the attribute vectors. The results of the best performing adjective words for CLIP are reported and we discuss the influence of adjective options in (§4.3).

4 Experiments

4.1 Experimental Settings

Compared Models In $\mathcal{W}_1$, directly comparing the CLIP and BERT can be unfair due to the difference in the model architecture and the training corpus. To remedy this, we train CLIP and BERT models from scratch with an identical transformer architecture on the same text corpus, the captions in the YFCC-15M dataset, for the 32 epochs. We further add two variants of CLIP, i.e., DeCLIP (Li et al., 2022) and DeFILIP (Yao et al., 2021), to investigate the effect of extra language modeling supervision and fine-grained matching, respectively. The only difference between these models is the pre-training objective. In $\mathcal{W}_2$ and $\mathcal{W}_3$, we evaluate the original models for investigating the learned knowledge during large-scale pre-training. Text-only pre-trained language models includes BERT-base and BERT-large (Devlin et al., 2019), RoBERTa-base and RoBERTa-large (Liu et al., 2019b) for masked language models, and OPT models with parameters ranging from 125M to 175B for casual language models. For CLIP-like models, we adopt CLIP-ViT-B/32 and CLIP-ViT-L/14 (Radford et al., 2021) as a base and a large version, respectively. We also include an enhanced CLIP model with masked language modeling as self-supervision, DeCLIP-ViT-B/32 (Li et al., 2022) for a more comprehensive evaluation. Detailed parameters of these models are listed in Appendix A.

Prompts For each task in $\mathcal{W}_2$ and $\mathcal{W}_3$, we manually write 10 prompts to eliminate the side-effect of the variations of expression. For OPT models, we observe its high sensitivity to the prompts, where some prompts would results in significantly worse performance than random guessing. We discard those prompts for OPT and report the averaged performance over multiple prompts for all models. All prompts used can be found in Appendix B.

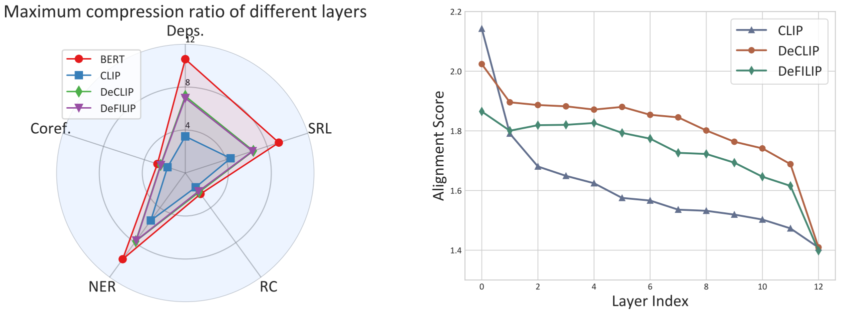


Figure 2: Linguistic probing results of BERT and CLIP-like models trained with YFCC15M data. (Left) Radar chart of the maximum compression ratio of different models in all layers. Higher scores represent richer linguistic knowledge. (Right) Layer-wise alignment score of text and visual features on CIFAR100. Lower alignment scores indicate better alignments between modalities.

4.2 Findings

In $\mathcal{W}_1$, CLIP falls far behind the text-only counterparts in linguistic-related tasks, even with the help of text modeling objectives. The linguistic probing results are illustrated in the left of Figure 2. Detailed performance and compression results of all layers in each model can be found in Appendix C. We find that the vanilla CLIP falls far behind the BERT model, regarding all the evaluated linguistic tasks. Besides, even though DeCLIP enhances the original contrastive learning objective with extra mask language modeling, it still performs worse than the vanilla BERT, and DeFILIP trained with fine-trained matching objective yields no clear improvements. This phenomenon is consistent with previous studies which observe that V+L pre-training results in language models with inferior language understanding ability (Tan & Bansal, 2020; Yun et al., 2021). We speculate the behind reason is that the learned representations of CLIP language models become more specialized for image-text matching, hindering the learning of linguistic knowledge. To examine this, we

Table 2: Zero-shot probing accuracy of visual-related different tasks. Results are averaged over different prompts. The best results are shown in bold.

Model (# of Param.)	**Color**	**Shape**	**Size**	**Height**	**Material**	**Avg.**
BERT-base (110M)	49.29 ± 1.60	52.14 ± 4.22	49.94 ± 0.80	50.56 ± 0.59	48.08 ± 2.74	50.00
BERT-large (340M)	49.36 ± 1.88	51.21 ± 5.06	49.26 ± 1.60	49.08 ± 2.34	49.72 ± 0.58	49.73
RoBERTa-base (125M)	49.07 ± 1.62	49.36 ± 3.52	50.32 ± 0.57	49.58 ± 0.49	49.86 ± 1.44	49.64
RoBERTa-large (355M)	49.66 ± 0.54	50.68 ± 1.48	50.54 ± 1.46	50.14 ± 0.45	50.00 ± 0.14	50.20
OPT (125M)	70.02 ± 9.59	57.32 ± 6.46	45.98 ± 4.23	56.76 ± 1.36	82.43 ± 2.20	62.50
OPT (1.3B)	76.92 ± 5.97	65.00 ± 6.12	51.12 ± 2.66	57.82 ± 4.46	85.63 ± 3.49	67.30
OPT (13B)	79.62 ± 5.28	62.50 ± 6.44	57.56 ± 6.60	54.58 ± 4.53	**88.38 ± 3.14**	68.53
OPT (175B)	**83.10 ± 3.13**	65.71 ± 7.54	59.18 ± 9.05	55.84 ± 5.33	85.49 ± 2.01	69.87
CLIP-ViT/B-32 (63M)	80.07 ± 2.57	84.43 ± 2.57	61.40 ± 6.02	62.28 ± 6.40	80.07 ± 2.57	73.94
DeCLIP-ViT/B-32 (63M)	81.48 ± 2.63	84.07 ± 2.34	**76.92 ± 1.81**	**68.12 ± 2.15**	81.48 ± 2.63	**78.35**
CLIP-ViT/L-14 (123M)	80.33 ± 3.61	**85.00 ± 4.03**	63.96 ± 6.10	60.72 ± 5.56	80.33 ± 3.61	74.21

compute the layer-wise alignment score (Wang & Isola, 2020) of the learned textual representations with the visual representations on CIFAR100 (Krizhevsky et al., 2009) following Ren et al. (2022). As shown in the right of Figure 2, the alignment score decreases in deeper layers, indicating that learned language representations are fusing with the visual features more deeply in higher layers.

In $\mathcal{W}_2$, CLIP outperforms text-only models with similar model sizes on visual tasks, yet scaling the text-only models can overturn the game. The results of $\mathcal{W}_2$ are shown in Table 2. We observe that CLIP-like models achieve the overall best average performance on five tasks related to visual properties, which indicates that visual supervision introduced by CLIP helps the model learn visual knowledge better. Besides, adding extra self-supervision like DeCLIP and scaling up the model size are both beneficial. It is worth noting that the text-only casual language model OPT performs relatively well. Most surprisingly, when the model scales to 175B, the pure text-based model can even outperform CLIP models on visual properties like color and material. We speculate that the co-occurrence statistics of color and material as modifiers for objects are well captured by large language models. To figure it out, we need more rigorous investigation for future work.

Table 3: Zero-shot prediction accuracy of different tasks in the embodied world. We report the average performance associated with standard deviation over multiple prompts. The best results are shown in bold.

Model (# of Param.)	**Mass**	**Temperature**	**Hardness**	**Avg.**
BERT-base (110M)	50.35 ± 0.56	49.67 ± 0.56	50.20 ± 0.43	50.07
BERT-large (340M)	49.97 ± 1.31	49.83 ± 0.50	49.98 ± 0.06	49.93
RoBERTa-base (125M)	49.65 ± 0.51	50.00 ± 0.00	48.04 ± 2.04	49.23
RoBERTa-large (355M)	50.08 ± 0.23	50.07 ± 0.19	49.95 ± 0.15	50.03
OPT (125M)	50.00 ± 0.00	54.53 ± 4.33	46.16 ± 2.45	50.23
OPT (1.3B)	50.05 ± 0.10	50.90 ± 5.08	53.03 ± 2.69	51.33
OPT (13B)	50.14 ± 0.36	51.85 ± 6.34	52.38 ± 3.09	51.46
OPT (175B)	50.21 ± 0.24	59.83 ± 8.68	57.33 ± 3.41	55.79
CLIP-ViT/B-32 (63M)	65.20 ± 4.75	60.28 ± 6.83	59.43 ± 2.00	61.64
DeCLIP-ViT/B-32 (63M)	54.95 ± 2.00	**68.58 ± 3.08**	61.10 ± 4.14	61.54
CLIP-ViT/L-14 (123M)	**73.15 ± 6.34**	65.88 ± 2.31	**69.57 ± 2.26**	**69.53**

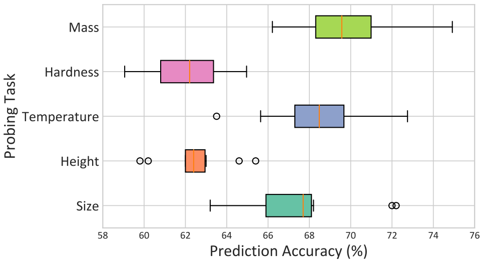


Figure 3: Few-shot prediction accuracy of OPT-175B with 16 instances as a demonstration on probing tasks.

In $\mathcal{W}_3$, CLIP performs the best on the embodied tasks, even outperforming the largest OPT-175B model with few-shot demonstrations. As shown in Table 3, CLIP-like models perform consistently better than pure text models, while text-based models like the OPT models struggle in $\mathcal{W}_3$. Besides, we notice that scaling up the model (CLIP-ViT/L-14) instead of adding extra language-size supervision (DeCLIP) is more effective for improving performance in the embodied world. We further conduct a few-shot prompt evaluation for OPT models by constructing the inputs with $k = 16$ randomly sampled instances, and the results are illustrated in Figure 3. We find that while the performance is boosted, the average results are still worse than the best-performing CLIP-ViT/L-14 model, which only utilizes 0.7% parameters of OPT-175B. These findings draw a conclusion that visual supervision is very beneficial for learning knowledge in the embodied world, and its role is irreplaceable by massive text data and billion-level parameters.

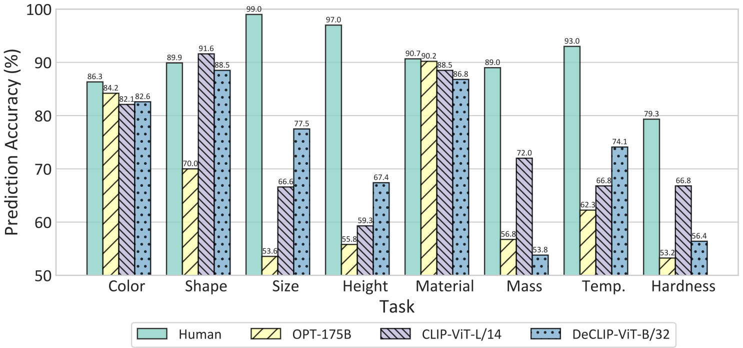


Figure 4: The comparison between the best-performing models and human annotators on sampled subsets of the visual and embodied tasks.

Compared with human annotators, CLIP and OPT achieve competitive performance in $\mathcal{W}_2$, while exhibiting great gaps in $\mathcal{W}_3$. We conduct a human evaluation to better understand the performance of different models. Specifically, we randomly sample 100 examples for each task and ask three annotators to label these examples. The comparison with best performing models, i.e., OPT-175B, CLIP-ViT/L-14 and DeCLIP is illustrated in Figure 4. We find that (1) In $\mathcal{W}_2$, both OPT and CLIP-like models perform closely to human annotators. CLIP and DeCLIP even outperform the human annotators on the shape task, which is potentially due to the noise introduced by the automatic construction of the dataset (Zhang et al., 2022a). Overall, the close-to-human results indicate that visual knowledge can be effectively acquired by large-scale cross-modal pre-training or even text-only pre-training with sufficient parameters. (2) In $\mathcal{W}_3$, the best-performing CLIP-ViT-L/14 model has an absolute 18.5% accuracy gap with the humans. The clear performance gaps reveal that there is still a long way to go in equipping language models with embodied knowledge.

4.3 Analysis

Embodied Knowledge in Image Representations We are interested in how the CLIP text encoder learns embodied knowledge. A potential answer is that the images contain some embodied knowledge like the heat of the object, and the knowledge can be propagated to the text encoder via the contrastive learning objective. To examine this, we perform a case study by calculating the attribute similarities over the images. We first take clips from a video of heating a pile of ice and then perform a binary classification by calculating the cosine similarities with text prompts *a photo of a hot object.* and *a photo of a cold object* for each frame. The left of Figure 5 shows that the probability of a hot object increases during the heating procedure. Similarly, we perform a binary classification over heavy and light-weight objects ranging from an elephant to a feather, and the illustration in the right of Figure 5 shows that the image representations are aware of the mass of different objects. This qualitative study shows clues that the learned embodied knowledge of CLIP's language model can be propagated from the image representations during the contrastive matching pre-training.

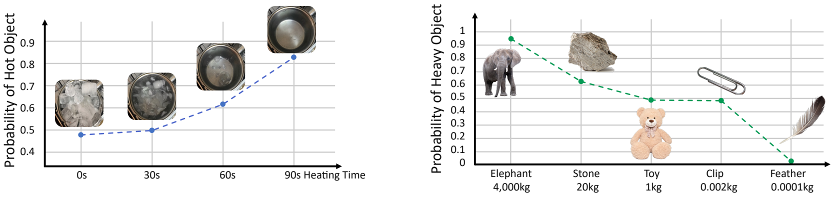


Figure 5: Case study showing that the image representations in CLIP exhibit embodied knowledge. (Left): The probability of being a hot object of clips taken from a video of heating ice in a boiler is increasing as the ice melts. (Right): The probability of being a heavy object of common objects with corresponding mass annotations.

Performance deterioration of CLIP with ambiguous attribute adjectives. During our preliminary study in $\mathcal{W}_3$, we observe that CLIP performs relatively worse for specific adjectives like *hard*. We further investigate this issue by checking the retrieved images with prompts with different at-

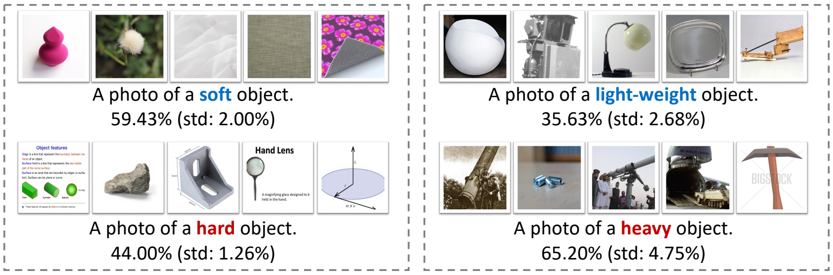

A photo of a soft object.
59.43% (std: 2.00%)
A photo of a light-weight object.
35.63% (std: 2.68%)
Object features
Hand Lens
A photo of a hard object.
44.00% (std: 1.26%)
A photo of a heavy object.
65.20% (std: 4.75%)

Figure 6: Top-5 retrieved images with different attribute prompts and the corresponding prediction accuracy. The accuracy of CLIP drops significantly when the text inputs contain ambiguous words and compound words, as the retrieved images are biased toward specific meanings.

tribute adjectives, on the CC12M dataset (Changpinyo et al., 2021). The results are illustrated in Figure 6. We find that for the text *a photo of a hard object*, the retrieved images are mostly about learning materials that are abstract and difficult, with only one rock image related to the hardness. Besides, for the text with the compound adjective word *light-weight*, the retrieved images are biased to the meanings related to lighting-bulb and light-toned color instead. Accordingly, the results with ambiguous texts are much lower. Possible remedies include devising better prompting methods like adaptively adjusting the prompts to eliminate ambiguity and developing better pre-training objectives taking the variations of text into consideration can also be promising.

5 Related Work

Vision-Language Pre-training The interest has grown recently in V+L pre-training for unifying cross-modal representations. Pilot studies adopt masked reconstruction to learn shared representations across modalities from a mixed sequence of visual region features and language token embeddings (Li et al., 2019; Tan & Bansal, 2019; Su et al., 2020; Chen et al., 2019; Li et al., 2020). CLIP (Radford et al., 2021) introduces a contrastive language-image pre-training framework, which utilizes language as supervision for learning transferable image representations with 400M image-text pairs. The ALIGN (Jia et al., 2021) utilizes more noisy image-text pairs up to 1.8B and achieves better performance. Further variants enhance the learned representations of CLIP by investigating self-supervision and data augmentation (Li et al., 2022), introducing fine-grained matching objectives (Yao et al., 2021), and modeling deep interactions between the modalities (Li et al., 2021).

Probing Language Models Understanding what language models know after large-scale pre-training is an active research area (Rogers et al., 2020). Various probing methods have been developed (Tenney et al., 2019b; Petroni et al., 2019), and investigations show that pre-trained language models like BERT capture linguistic (Tenney et al., 2019a; Liu et al., 2019a), factual (Petroni et al., 2019; Roberts et al., 2020; Dai et al., 2022), commonsense knowledge (Wang et al., 2019b; Forbes et al., 2019), and even learn complex reasoning ability (Srivastava et al., 2022). For V+L pre-trained models, studies demonstrate their potential in acquire spatial commonsense (Zhang et al., 2022a; Liu et al., 2022), yet performing worse on NLU tasks (Tan & Bansal, 2020) and achieving no significant on lexical grounding (Yun et al., 2021). We observe similar results in our LiVE benchmark when evaluating other V+L models such as Vokenization (Tan & Bansal, 2020) in Appendix D.

6 Conclusion

In this paper, we investigate what vision supervision brings to language models with CLIP as a representative V+L model. We build a comprehensive evaluation benchmark named **LiVE**, covering linguistic, visual and embodied knowledge. Comparative experiments between CLIP and text-only models show that, while CLIP falls far behind regarding linguistic knowledge, CLIP performs significantly better in the visual and the embodied world, even outperforming OPT-175B with only 0.3% parameters. The evaluation proves that visual signals are beneficial for language models to acquire knowledge beyond the linguistic world. The further qualitative analysis demonstrates that the learned embodied knowledge is potentially from the image representations and CLIP struggles when dealing with ambiguous text inputs, shedding light on future directions for further improvements.

References

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 2022.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In *Proc. of EMNLP*, 2020.

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, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey 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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *ArXiv*, 2019.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proc. of ACL*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, 2019.

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. Not all models localize linguistic knowledge in the same place: A layer-wise probing on BERToids' representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2021.

Gustav Theodor Fechner. Elements of psychophysics, 1860. 1948.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. Do neural language representations learn physical commonsense? In *CogSci*, 2019.

David Gueorguiev, Séréna Bochereau, André Mouraux, Vincent Hayward, and Jean-Louis Thonnard. Touch uses frictional cues to discriminate flat materials. *Scientific reports*, 2016.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 1990.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, 2009.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proc. of EMNLP-IJCNLP*, 2019.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. Quora question pairs. *First Quora Dataset Release: Question Pairs*, 2017.

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 *Proc. of ICML*, 2021.

Lynette Jones. Thermal touch. *Scholarpedia*, 2009.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Proc. of Neurips*, 2021.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, 2019.

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 *Proc. of ECCV*, 2020.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *Proc. of ICLR*, 2022.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proc. of NAACL*, 2019a.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. Things not written in text: Exploring spatial commonsense from visual signals. In *Proc. of ACL*, 2022.

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 pre-training approach. *ArXiv preprint*, 2019b.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proc. of EMNLP-IJCNLP*, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 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 *Proc. of ICML*, 2021.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*, 2016.

Shuhuai Ren, Lei Li, Xuancheng Ren, Guangxiang Zhao, and Xu Sun. Rethinking the openness of clip. *arXiv preprint arXiv:2206.01986*, 2022.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proc. of EMNLP*, 2020.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 2020.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021a.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proc. of EACL*, 2021b.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP*, 2013.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, 2022.

Trevor Standley, Ozan Sener, Dawn Chen, and Silvio Savarese. image2mass: Estimating the mass of an object from its image. In *Proceedings of the 1st Annual Conference on Robot Learning*, 2017.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pretraining of generic visual-linguistic representations. In *Proc. of ICLR*, 2020.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proc. of EMNLP-IJCNLP*, 2019.

Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proc. of EMNLP*, 2020.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proc. of ACL*, 2019a.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *Proc. of ICLR*, 2019b.

Esther Thelen and Linda B Smith. *A dynamic systems approach to the development of cognition and action*. 1994.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proc. of EMNLP*, 2020.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In *Proc. of EMNLP-IJCNLP*, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. of ICLR*, 2019a.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. Does it make sense? and why? a pilot study for sense making and explanation. In *Proc. of ACL*, 2019b.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proc. of ICML*, 2020.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium*, 2013.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of ACL*, 2018.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *ArXiv preprint*, 2021.

Tian Yun, Chen Sun, and Ellie Pavlick. Does vision-and-language pretraining improve lexical grounding? In *Proc. of EMNLP Findings*, 2021.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. Visual commonsense in pretrained unimodal and multimodal models. In *Proc. of NAACL*, 2022a.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022b.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proc. of ICML*, 2021.

A Model Configurations

We provide the detailed configurations of models we evaluated in our main paper. Specifically, for the evaluation in $\mathcal{W}_1$, we train identical Transformer (Vaswani et al., 2017) language models configured with 12 Transformer layers and 512 hidden units with 8 attention heads, from scratch on the YFCC15M dataset. These models only differ in the pre-training objectives, as shown in Table 4.

Table 4: Pre-training objectives comparison of models evaluated in $\mathcal{W}_1$.

Model	Training Objective
BERT (Devlin et al., 2019)	Masked Language Modeling (MLM)
CLIP (Radford et al., 2021)	Contrastive Image-text Matching (CIM)
DeCLIP (Li et al., 2022)	MLM + CIM
DeFILIP (Yao et al., 2021)	MLM + Contrastive Token-Patch Matching

For tasks in $\mathcal{W}_2$ and $\mathcal{W}_3$, we evaluate the vanilla models for investigating the knowledge learned during the original pre-training. The detailed model architecture configurations of these models can be found in Table 5.

Table 5: Detailed configuration of models evaluated in $\mathcal{W}_2$ and $\mathcal{W}_3$.

Model	Hidden Layers	Hidden Size	Attention Heads	Total # of Parameters
BERT-base	12	768	12	110M
BERT-large	24	1,024	16	340M
RoBERTa-base	12	768	12	125M
RoBERTa-large	24	1,024	16	355M
OPT-125M	12	768	12	125M
OPT-1.3B	24	2,048	32	1.3B
OPT-13B	40	5,120	40	13B
OPT-175B	96	12,288	96	175B
CLIP-ViT/B-32	12	512	8	63M
DeCLIP-ViT/B-32	12	512	8	63M
CLIP-ViT/L-14	12	768	12	123M

B Details of Prompts

We provide the used prompts for evaluating different models based on their pre-training objectives. Examples of Head, Rel and Tail of each dataset is shown in Table 1. Due to the sensitivity of language models to prompts, we provide diverse prompts for each model on each task.

Prompts for Masked Language Models A `[MASK]` token is placed in the prompt and the models are asked to predict the probabilities of the `[MASK]` token. To avoid multiple mask tokens in prompts, we follow Schick & Schütze to convert knowledge fact into a cloze-question. For example, a temperature fact `(water, colder than, frying oil)` can be converted into `Q: is the water colder than frying oil? A: [MASK]!`. The models need to choose the token `yes` or `no` to fill the mask.

Prompts for Causal Language Models As there is no special `[MASK]` token during the pre-training of causal language models, we do not use `[MASK]` tokens in prompts for causal language models. For Color, Shape and Material datasets in $\mathcal{W}_2$, we construct two prompts for (Head, Tail$_1$) and (Head, Tail$_2$); while for other datasets in $\mathcal{W}_2$ and $\mathcal{W}_3$, we construct two prompts for (Head, Rel, Tail) and (Head, Rel$'$, Tail) where Rel$'$ is the antonym relation of Rel. The prediction is based on the prompt with lower perplexity.

Prompts for CLIP We follow Radford et al., to use prompts like `a photo of ...` here. As the language encoder of CLIP encodes sentences to a vector and can evaluate similarities between sentences. We use an attribute prompt like `a photo of a cold object` and construct same prompts for objects (water and frying oil) in the knowledge fact. We can determine the colder object if the prompt of this object has a higher similarity to the attribute prompt.

Table 6: Prompts for Masked Language Models

Model	Task	Prompt
BERT & RoBERTa	Size, Height, Temperature, Weight, Hardness	is the [Head] [Rel] than the [Tail]? [MASK]! is the [Head] [Rel] than the [Tail]? [MASK]. is [Head] [Rel] than [Tail]? [MASK]! is [Head] [Rel] than [Tail]? [MASK]. is [Head] [Rel] compared with [Tail]? [MASK]. is [Head] [Rel] compared with [Tail]? [MASK]! compared with [Tail], is [Head] [Rel]? [MASK]. compared with [Tail], is [Head] [Rel]? [MASK]! is [Head] usually [Rel] than [Tail]? [MASK]. is [Head] usually [Rel] than [Tail]? [MASK]!
	Color	can [Head] be of color [Tail]? [MASK]! can [Head] be of color [Tail]? [MASK]. is the color of a [Head] [Tail]? [MASK]! is the color of a [Head] [Tail]? [MASK]. is [Head] [Tail]? [MASK]. is [Head] [Tail]? [MASK]! is [Head] typically in [Tail]? [MASK]. is [Head] typically in [Tail]? [MASK]! Q: is [Head] of color [Tail]? A: [MASK]. Question: is [Head] of color [Tail]? Answer: [MASK].
	Shape	can [Head] be the shape of [Tail]? [MASK]. can [Head] be the shape of [Tail]? [MASK]! does the [Head] have a shape of [Tail]? [MASK]. does the [Head] have a shape of [Tail]? [MASK]! is [Head] of [Tail]? [MASK]. is [Head] of [Tail]? [MASK]! Q: is [Head] of [Tail]? A: [MASK]. Question: is [Head] of [Tail]? Answer: [MASK]. [Tail] [Head]? [MASK]. is [Head] typically [Tail]? [MASK].
	Material	can [Head] be made of [Tail]? [MASK]! can [Head] be made of [Tail]? [MASK]. is [Head] made of [Tail]? [MASK]! is [Head] made of [Tail]? [MASK]. is [Tail] the necessary material for making [Head]? [MASK]. is [Tail] the necessary material for making [Head]? [MASK]! does [Head] consist of [Tail]? [MASK]. is [Head] made up of [Tail]? [MASK]. Q: is [Head] made of [Tail]? A: [MASK]. Question: is [Head] made of [Tail]? Answer: [MASK].

C Detailed Edge Probing Results

Here, we introduce the details of the compression metric c used in our main paper. Formally, given N instance representations with each label having K classes, we need to send their corresponding labels with a minimum description length (MDL). In *uniform encoding*, which assumes that each representation has a label with a probability of $1/K$ and does not require any learning, we can transmit the labels with codelength $N\log_2(K)$. However, when the representations exhibit some degree of regularity with respect to the labels, we can train a classifier to predict the labels given the representations and transmit the classifier's complexity, i.e., the classifier codelength, instead of sending the labels. As the classifier is usually not optimal, the final cross-entropy of the classifier over the data, i.e., the data codelength) will be added to the classifier's codelength, resulting in the final MDL metric. We further eliminate the effect of the number N on the total sum of data cross-entropy by reporting the compression metric, which is calculated as:

$$c = \frac{N\log_2(K)}{\text{MDL}}. \tag{4}$$

The MDL can be calculated via a Bayesian method or an online encoding framework. We adopt the latter due to its simplicity and report the compression ratio over a uniform coding in our main paper, where a higher compression ratio c indicates that the representations contain richer information regarding the target linguistic knowledge. The detailed probing results of different models in each layer, which is shown in Table 10, and layers are indexed from 0 to 12 indicating the embedding layer and the last Transformer layer. Further, following Tenney et al. (2019a), we derive the Lin-

Table 7: Prompts for Causal Language Models

Model	Task	Prompt
OPT	Size, Height, Temperature, Weight, Hardness	the [Head] is [Rel] than the [Tail]. [Head] is [Rel] than [Tail]. acutally, the [Head] is [Rel] than the [Tail]. acutally, [Head] is [Rel] than [Tail]. it is well-known that [Head] is [Rel] than [Tail]. [Head] is indeed [Rel] than [Tail]. the [Head] is indeed [Rel] than [Tail]. compared with the [Head], the [Tail] is [Rel]. a/(an) [Head] is [Rel] than a/(an) [Tail]. yes, [Head] is [Rel] than [Tail].
	Color	[Head] can be of the color [Tail]. the [Head] can be of color [Tail]. the color of a(an) [Head] is [Tail]. the color of [Head] is [Tail]. the [Head] is in [Tail]. [Head] is [Tail]. what color is the [Head]? [Tail]. [Head]'s color is [Tail]. usually, [Head] is in [Tail]. [Head] is typically [Tail].
	Shape	[Head] is usually [Tail]. what is the shape of [Head]? [Tail]. [Head] is typically [Tail]. [Head]'s shape is [Tail].
	Material	[Head] is made of [Tail]. the [Head] is made of [Tail]. [Head] consists of [Tail]. the main material of [Head] is [Tail]. [Tail] is necessary material for making [Head]. the [Head] consists of [Tail]. the [Head] can be made of [Tail]. the [Head] is built with [Tail]. the [Head] contains [Tail]. the [Head] is made up of [Tail].

Table 8: Prompts used for CLIP

Model	Task	Prompt
CLIP	All Tasks	a photo of a [Head]/[Attribute]. a photo of the [Head]/[Attribute]. a blurry photo of a [Head]/[Attribute]. a good photo of a [Head]/[Attribute]. a painting of a [Head]/[Attribute]. a bad photo of a [Head]/[Attribute]. a close-up photo of a [Head]/[Attribute]. a bright photo of the [Head]/[Attribute]. a photo of one [Head]/[Attribute]. a low resolution photo of a [Head]/[Attribute].

guistic Layer Center metric to check which layer contains the most information regarding the target linguistic property. Specifically, for layer ℓ of a language model and the corresponding compression ratio c_ℓ, the **linguistic layer center** which is similar to the physical concept **center of gravity**, is calculated as:

$$\mathbb{E}_c[\ell] = \frac{\sum_{\ell=0}^{L} c_\ell \cdot \ell}{\sum_{\ell=0}^{L} c_\ell}. \tag{5}$$

The linguistic layer center of different models in all linguistic probing tasks is shown in Figure 7. We observe that BERT has a higher linguistic layer center compared to CLIP models. This again verifies our assumption in the main paper that the learned language representations are fusing with the visual information more deeply in higher layers.

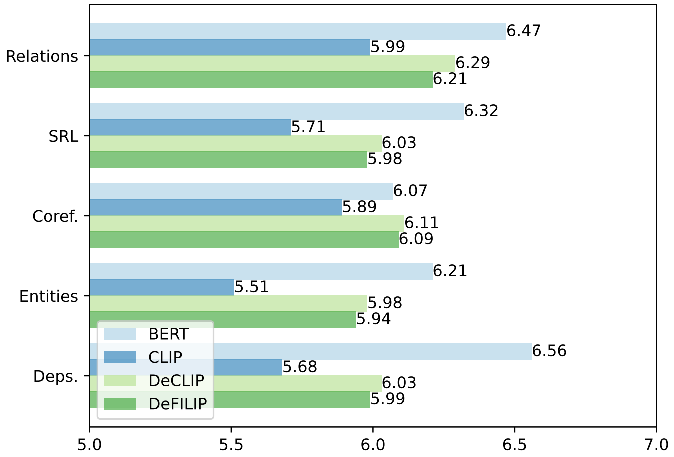


Figure 7: The center of gravity of different models in all linguistic probing tasks. The gravity is measured with the MDL compression. The linguistic information is centered in lower layers of CLIP models than BERT.

Table 9: Fine-tuned accuracy of other visual-informed pre-trained language models on NLU tasks and zero-shot results in $\mathcal{W}_2$ and $\mathcal{W}_3$.

Model	**SST-2**	**QQP**	**QNLI**	**MNLI (m / mm)**	**Avg.**
BERT (Wiki)	90.13	83.20	87.57	78.90 / 80.05	83.97
DistilledOscar	89.33	67.98	82.48	74.46 / 74.82	77.81
VLM-BERT-base	90.60	90.10	89.47	81.57 / 82.43	86.83
VLM-RoBERTa-base	90.13	80.37 / 80.43	87.91	88.44	85.46

Model	**Color**	**Shape**	**Size**	**Height**	**Material**	**Mass**	**Temperature**	**Hardness**	**Avg.**
BERT (Wiki)	49.41	48.07	51.70	49.46	52.39	48.85	51.07	52.34	50.41
DistilledOscar	49.97	53.61	49.07	49.80	51.46	51.22	47.94	51.23	50.54
VLM-BERT-base	50.69	50.07	51.00	50.92	53.89	44.83	50.64	49.22	50.16
VLM-RoBERTa-base	49.53	51.21	49.00	49.22	49.54	49.92	51.11	49.63	49.90

D Probing Other Visual-aided Language Modelings

We examine whether other V+L pre-training methods bring gains regarding visual and embodied knowledge. Specifically, following Zhang et al. (2022a), we distill the knowledge of Oscar (Li et al., 2020) into a BERT model by performing knowledge distillation (Hinton et al., 2015) on the image-caption pair dataset. Specifically, the paired text and image is fed into Oscar model for getting the vision-awared vocabulary distribution, and a student BERT model is performing masked language modeling on the text data only and learns from the soft labels provided by the Oscar teacher model. The distillation results in a DistilledOscar model supporting text-only inputs. We also evaluate VLM-BERT learned via Vokenziation (Tan & Bansal, 2020), which devises a fine-grained token-voken matching framework to utilize visual supervision. The models are evaluated on the four largest datasets in GLUE, including SST-2 (Socher et al., 2013), QQP (Iyer et al., 2017), QNLI (Rajpurkar et al., 2016) and MNLI (Williams et al., 2018) for stable results. As shown in Table 9, DistilledOscar performs worse than the vanilla BERT in both NLU tasks and probing tasks regarding visual and embodied knowledge. Besides, while VLM-BERT achieves improvements on NLU tasks, it performs at the random level on the probed tasks. We think the reason is not the differences in the training objectives and the model architecture, but the data scale used for pre-training, i.e., CLIP builds a 400M paired image-text dataset for training, yet the amount for Oscar and Vokenziation is less than 10M. As recent studies suggest that purely text language models acquire complicated reasoning abilities during scaling up in model parameters and training corpus (Brown et al., 2020; Wei et al., 2022), investigating the emergent abilities during the scaling up of multi-modal models can also be interesting.

Table 10: MDL Compression (Comp.) and Task F1 of different models trained with YFCC15M in different linguistic tasks. The maximum compression of a model is **bolded**.

Tasks	**BERT**		**CLIP**		**DeCLIP**		**DeFILIP**	
	Compression	Task F1	Compression	Task F1	Compression	Task F1	Compression	Task F1
Coref	2.41	89.90	**1.74**	**85.01**	1.78	85.52	1.81	86.33
	2.51	90.30	1.70	84.33	2.26	89.62	2.21	89.44
	2.53	90.60	1.71	84.43	2.37	89.85	2.31	89.84
	2.62	90.90	1.70	83.97	2.38	89.93	**2.39**	**90.08**
	2.66	91.04	1.70	84.37	**2.42**	**90.06**	2.38	89.78
	2.68	91.11	1.69	84.21	2.41	89.99	2.37	89.87
	2.70	91.02	1.69	84.01	2.41	89.90	2.37	89.82
	2.70	91.36	1.66	83.59	2.39	89.90	2.35	89.65
	2.72	91.50	1.65	83.28	2.40	89.95	2.35	89.78
	2.73	**91.51**	1.62	83.04	2.40	89.84	2.32	89.43
	2.68	91.32	1.61	82.49	2.38	89.73	2.30	89.28
	2.61	91.15	1.58	82.03	2.35	89.36	2.30	89.24
	2.55	90.66	1.58	80.84	2.25	89.08	2.23	88.91
Deps.	5.57	82.80	**3.41**	**70.10**	4.70	78.64	4.60	78.19
	6.86	86.80	2.76	63.60	6.50	86.03	6.12	84.86
	7.92	88.86	2.83	64.81	6.96	87.20	6.86	86.95
	8.39	89.53	2.86	65.24	6.98	87.21	**6.97**	**87.33**
	8.45	89.76	2.82	64.70	7.07	87.37	6.83	86.84
	8.58	89.88	2.78	64.03	**7.15**	**87.55**	6.67	86.56
	9.11	90.55	2.75	63.79	6.99	87.21	6.53	86.19
	9.51	90.92	2.70	63.15	6.85	86.96	6.43	85.86
	10.22	91.54	2.63	61.92	6.78	86.97	6.36	85.74
	10.42	91.70	2.55	60.86	6.67	86.71	6.27	85.49
	10.59	**91.71**	2.46	59.66	6.55	86.43	6.15	85.19
	10.43	91.45	2.37	58.38	6.42	86.03	6.00	84.82
	9.94	91.01	2.35	58.53	5.82	84.71	5.56	83.73
NER	7.29	90.16	**5.46**	**86.86**	6.49	88.81	6.50	88.76
	8.34	91.54	3.89	79.20	7.49	90.33	7.50	90.34
	8.96	92.16	3.88	79.10	7.74	90.51	7.74	90.66
	9.32	92.22	3.84	78.97	7.79	90.60	**7.76**	**90.63**
	9.42	92.31	3.79	78.59	7.84	90.65	7.71	90.54
	9.47	92.32	3.72	78.14	**7.81**	**90.70**	7.60	90.39
	9.55	92.26	3.66	77.73	7.75	90.67	7.53	90.21
	9.74	92.30	3.58	77.21	7.71	90.48	7.50	90.10
	9.91	**92.51**	3.49	76.62	7.63	90.46	7.39	89.98
	9.79	92.42	3.37	75.86	7.55	90.28	7.28	89.76
	9.79	92.65	3.23	74.65	7.39	89.95	7.18	89.66
	9.70	92.43	3.09	73.49	7.27	89.73	7.08	89.48
	9.31	92.05	3.02	73.01	6.81	89.27	6.65	89.03
SRL	6.27	81.51	**4.43**	**70.75**	4.93	74.87	4.88	74.49
	7.04	84.45	3.62	67.86	6.45	82.82	6.23	82.11
	7.76	86.31	3.67	68.55	6.62	83.24	**6.57**	**83.24**
	7.96	86.73	3.62	67.92	**6.66**	**83.23**	6.49	82.84
	7.89	86.39	3.62	67.97	6.61	83.00	6.44	82.71
	8.02	86.71	3.58	67.55	6.64	83.18	6.37	82.43
	8.24	87.08	3.48	65.96	6.52	82.68	6.24	81.83
	8.70	87.94	3.43	65.48	6.40	82.27	6.18	81.70
	8.86	88.14	3.42	65.70	6.43	82.37	6.10	81.34
	9.15	88.56	3.34	64.68	6.40	82.23	6.06	81.21
	9.07	**88.42**	3.22	62.74	6.30	81.98	5.99	80.98
	8.89	88.11	3.14	61.76	6.19	81.45	5.91	80.58
	8.52	87.49	3.12	61.27	5.94	80.51	5.69	79.58
RC	1.59	47.38	1.34	37.33	1.45	40.92	1.44	42.27
	1.68	50.57	1.49	42.44	1.66	51.51	1.60	49.75
	1.83	54.31	1.58	45.69	1.85	56.83	1.85	56.88
	1.91	55.92	1.60	46.30	1.89	58.08	2.01	60.19
	1.92	56.59	1.62	48.26	1.97	58.81	2.08	61.41
	1.99	57.97	**1.64**	**47.97**	2.11	61.36	2.09	60.95
	2.07	59.81	1.63	48.17	2.16	62.06	**2.10**	**60.57**
	2.13	61.33	1.60	47.85	**2.19**	**62.12**	2.06	60.95
	2.28	63.84	1.57	46.80	2.18	62.00	2.05	60.68
	2.31	64.54	1.56	46.04	2.16	61.15	2.03	59.95
	2.35	65.19	1.50	43.44	2.14	60.45	2.01	59.81
	2.39	65.74	1.44	42.65	2.09	58.84	2.01	59.69
	2.41	**65.71**	1.42	41.60	1.93	58.29	1.88	57.64