Z-LaVI: Zero-Shot Language Solver Fueled by Visual Imagination

Yue Yang[1*], **Wenlin Yao**[2], **Hongming Zhang**[2], **Xiaoyang Wang**[2],
Dong Yu[2], **Jianshu Chen**[2]
[1]University of Pennsylvania, [2]Tencent AI Lab

yueyang1@seas.upenn.edu

{wenlinyao,hongmzhang,shawnxywang,dyu,jianshuchen}@global.tencent.com

Abstract

Large-scale pretrained language models have made significant advances in solving downstream language understanding tasks. However, they generally suffer from *reporting bias*, the phenomenon describing the lack of explicit commonsense knowledge in written text, e.g., "*an orange is orange*". To overcome this limitation, we develop a novel approach, Z-LaVI, to endow language models with visual imagination capabilities. Specifically, we leverage two complementary types of "imaginations": (i) recalling existing images through retrieval and (ii) synthesizing nonexistent images via text-to-image generation. Jointly exploiting the language inputs and the imagination, a pretrained vision-language model (e.g., CLIP) eventually composes a zero-shot solution to the original language tasks. Notably, fueling language models with imagination can effectively leverage visual knowledge to solve plain language tasks. In consequence, Z-LaVI consistently improves the zero-shot performance of existing language models across a diverse set of language tasks.[1]

1 Introduction

Large-scale Pretrained Language Models (PLMs) have achieved great success on various Natural Language Understanding (NLU) tasks and even exhibit impressive zero-shot capabilities without task-specific fine-tunings (Radford et al., 2019). And recent research suggests that such ability improves by further scaling up the model size (e.g., to hundreds of billions of parameters) and the amount of textual pretraining data (to TBs of raw texts) (Min et al., 2021; Brown et al., 2020; Chowdhery et al., 2022; Kaplan et al., 2020). However, zero-shot language learners solely trained on texts inevitably suffer from human reporting bias. For example, people tend not to write common or apparent things (Grice, 1975), and the frequency of a certain textual statement does not always correspond to their relative
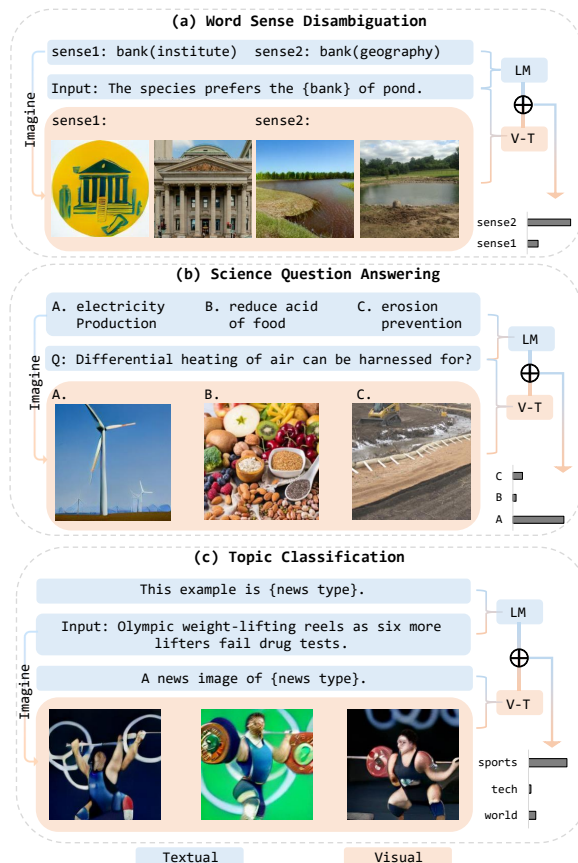

Figure 1: Our system endows language models with two complementary types of visual imagination capabilities: recalling existing images (through retrieval) and synthesizing nonexistent images (via image-to-text generation). They effectively alleviate the reporting bias issue and improves the zero-shot performance for solving *plain* language tasks. We experiment with three types of tasks: (a) Word Sense Disambiguation, (b) Science Question Answering, and (c) Topic Classification.

likelihood in the world (Gordon and Van Durme, 2013). Therefore, looking into other modalities to supplement the textual information is crucial.

In this paper, we focus on incorporating vision knowledge to facilitate the solution of plain lan-

---

*Work was done during the internship at Tencent AI Lab.

[1]Code and data are available at https://github.com/YueYANG1996/Z-LaVI

guage understanding tasks. Cognitive science has demonstrated that the human vision system is crucial to supplement, interact, and influence the language system (Dessalegn and Landau, 2013). For example, there exists a fast mapping between vision and language in the human language learning process (Altmann and Kamide, 2004). Inspired by this, we propose a visual imagination framework, Z-LaVI, to endow any PLMs (e.g., GPT, BERT, BART, etc.) with visual imagination capabilities.

Specifically, we apply two different types of "visual imaginations" to the input texts. Given input text, the first approach *recalls* existing images (e.g., through search engines), and the second one *synthesizes* nonexistent images via text-to-image generation models (e.g., DALL-E (Ramesh et al., 2021)). These two strategies mimic different types of human mental behaviors, i.e., recalling past memories and creative mental image construction. Interestingly, we find that these two mechanisms are highly complementary to each other. Our proposed visual imagination module tends to rely more on recalling when input texts are short because their corresponding objects or scenes generally exist and are easy to find. However, when input texts are long and complex, the module is more inclined to create new images. We develop a unified framework (Figure 1) that exploits both types of imaginations along with the original textual inputs to compose zero-shot solutions to a broad set of downstream language tasks. Note that our work differs from existing multi-modal tasks such as VQA (Antol et al., 2015; Wu et al., 2017) or Visual Dialog, (Das et al., 2017), which have both textual and visual inputs. Instead, we use visual imagination as machinery to facilitate the (zero-shot) solution of pure language tasks.

We show that on a diverse set of language understanding tasks, Z-LaVI consistently improves the performance of existing language models of different sizes and architectures. In particular, our Z-LaVI with SBERT can achieve a zero-shot F1 score of 87.5% on the WSD task without fine-tuning, even outperforming BERT-large, which is fine-tuned with three examples per sense, by 2.3%. Z-LaVI also beats all existing zero-shot models on four Science QA tasks and two Topic Classification tasks by a large margin. Our analysis demonstrates that Z-LaVI can complement language models and significantly alleviate PLMs zero-shot prediction errors by adaptively executing two visual imagination mechanisms - RECALL and SYNTHESIS.

(a) Imagine the stream of input labels.

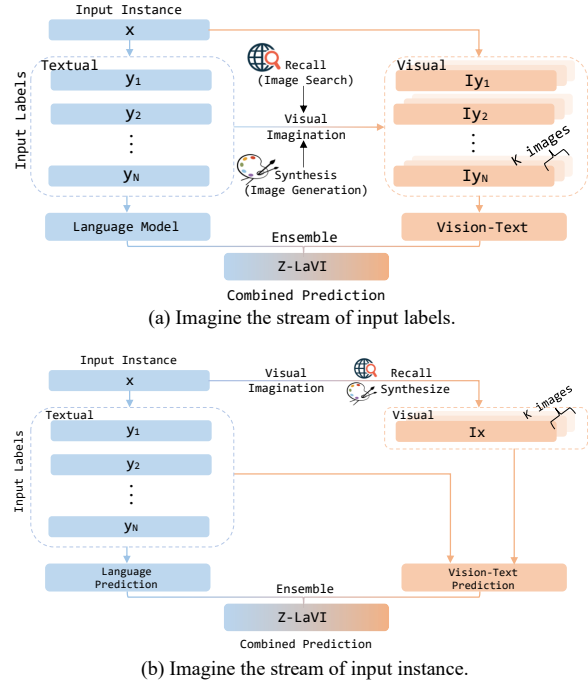(b) Imagine the stream of input instance.

Figure 2: The overview of the proposed Z-LaVI system. Z-LaVI aims to solve the language tasks with two streams of inputs: one stream of label options and another stream of instance to be labeled. Z-LaVI converts one of the streams (either input labels or input instance) into images through visual imagination (RECALL and SYNTHESIS) to enable the vision-text model to solve language tasks. We ensemble the language and vision-text models to make the final prediction.

2 Method

2.1 Task Formulation

To provide a zero-shot solution for language tasks and solve them in a uniform way, we transform different tasks into multi-choice questions, where input stream x and candidate answers stream $y \in \mathcal{Y}$ are provided. The goal is to select the correct answer from $\mathcal{Y}$. In particular, for word sense disambiguation tasks, x is the instance sentence, and $\mathcal{Y}$ are all possible word senses of the target word; for science question answering tasks, x is the question, and $\mathcal{Y}$ are answer options; for text classification tasks, x is the input sentence, and $\mathcal{Y}$ is the pool of categories. To make a prediction, the model needs to estimate the plausibility of each tuple (x, y) for all $y \in \mathcal{Y}$ and select the best answer $\hat{y}$.

$$\hat{y} = \operatorname{argmax}_{y \in \mathcal{Y}} P(y|x). \qquad (1)$$

2.2 Language Models for Zero-shot Tasks

We consider three main approaches for employing language models to make zero-shot predictions on

language tasks:

Prompt-based Approach (Petroni et al., 2019; Schick and Schütze, 2021) treats Natural Language Understanding tasks as a cloze test using prompts. For example, we can format question-answering tasks into:

"*Question*: [x]*? The answer is* [y]."

We convert the input (x, y) into a sequence of tokens $\mathbf{W} = (w_1, ..., w_t, ...w_{t+k}, ..., w_{|\mathbf{W}|})$ via a prompt, in which $y = (w_t, ...w_{t+k})$.[2] We apply autoregressive language models such as GPT (Brown et al., 2020) to calculate the score:

$$\text{Score}_{\text{La}}(x, y) = \frac{1}{-\frac{1}{k}\sum_{i=0}^{k} \log P_{\text{La}}(w_{t+i}|\mathbf{W}_{<t+i})},$$

where $P_{\text{La}}(\cdot)$ denotes the probability given by the language model. Note that we adopt the standard token-length normalization to handle different lengths of answer choices. Finally, we apply softmax to $\text{Score}_{\text{La}}(x, y)$ to obtain the probability of each candidate:

$$p_{\text{La}}(y|x) = \frac{e^{\text{Score}_{\text{La}}(x,y)}}{\sum_{y \in \mathcal{Y}} e^{\text{Score}_{\text{La}}(x,y)}}. \quad (2)$$

For the prompt-based approach, we select GPT-Neo-1.3B/2.7B (Black et al., 2021), GPT-J-6B (Wang and Komatsuzaki, 2021) and OPT-30B (Zhang et al., 2022c) as our models. The GPT-Neo and GPT-J are trained on the Pile dataset (Gao et al., 2020), which contains 825 GB of English text data. Besides Pile, OPT concatenates the training data of RoBERTa (Liu et al., 2019) and PushShift Reddit (Baumgartner et al., 2020).

Natural Language Inference (NLI) Approach (Yin et al., 2019) propose a textual entailment framework for zero-shot text classification. The NLI approach considers the input pair (x, y) as a (*premise*, *hypothesis*) pair to predict the probability that the premise logically entails the hypothesis.

$$p_{\text{La}}(y|x) = p(\text{ENTAILMENT}|(x, y)). \quad (3)$$

Note that this approach requires language models to be fine-tuned on (*premise*, *hypothesis*) pairs. Here we select RoBERTa-large (Liu et al., 2019) and BART-large (Lewis et al., 2020) fine-tuned on Multi-genre NLI (MNLI) corpus, (Williams et al., 2018) consisting of 433k sentence pairs.

Latent Embedding Approach utilizes an off-the-shelf feature encoder f_θ to project the input tuple (x, y) into a shared latent space and determines their relevance based on a distance metric - cosine similarity scores:

$$\text{Score}_{\text{La}}(x, y) = \cos(f_\theta(x), f_\theta(y)). \quad (4)$$

Relevance scores are normalized with softmax (equation 2) to get the final probabilities.

We choose two state-of-the-art sentence encoders, i.e., Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) and SimCSE, (Gao et al., 2021) as our latent embedding models. For SBERT, we pick the `all-mpnet-base-v2` checkpoint,[3] which achieves the best performance on 14 sentence embedding datasets.[4] For SimCSE, we choose the best fully unsupervised model `unsup-simcse-roberta-large`.

2.3 Language with Visual Imagination

Visual Imagination aims to convert either x or y (depending on the task) in the textual input tuple (x, y) into an image. For WSD and QA tasks, we imagine the candidate options y. While for topic classification tasks, we imagine the instance sentence x. Here we illustrate our method through the example of imagining y. We propose two imagination mechanisms: 1) RECALL and 2) SYNTHESIS.

1) RECALL: We use the text input to query Bing Image Search[5] to recall the corresponding images. We set a maximum number of images for each query. When only limited images are available for some queries, we download all of them.

2) SYNTHESIS: We adopt DALL·E (Ramesh et al., 2021), a text-to-image generation model pretrained on image-caption pairs, to synthesize images. DALL·E constructs a codebook $\mathcal{V}$ using a discrete variational autoencoder (dVAE) (Rolfe, 2016) to map the image into tokens concatenated with the caption's text tokens. DALL·E models the joint distribution over the text and image tokens with an autoregressive transformer. During inference, DALL·E feeds the text tokens y into the transformer and generates a sequence of image tokens $(v_1, v_2, ..., v_m)$, where an image token v_i is

---

[2]We include the prompts of all tasks in Table 9.

[3]It is trained on 1B sentence pairs from diverse tasks, including NLI, QA, image captions, etc. The training data have no overlap with our evaluation data, so we still consider this approach as zero-shot.

[4]https://www.sbert.net/docs/pretrained_models.html

[5]We use the Azure Bing Image Search API.

predicted based on the previous ones:

$$v_i = \text{argmax}_{v \in \mathcal{V}} p(v|y, v_{<i}), \quad (5)$$

in which, $\mathcal{V}$ is the visual codebook. After we generate enough image tokens, we decode the tokens into images by looking up the vectors in the dVAE codebook to construct the pixels.

We iterate the SYNTHESIS process multiple times and combine with the images from RECALL to collect a set of K images $\{I_y^k|i = 1, ..., K\}$ for each textual input y.[6]

Vision-Text model for Zero-shot language tasks. After transferring an input stream into images, we modify a plain language task into a multimodal task. Thus we can apply vision-text models to solve the problems. We choose CLIP (Radford et al., 2021) as our vision-text model, which is pre-trained on 400M image-caption pairs with the contrastive learning strategy.

CLIP has a text encoder f_T and a visual encoder f_V, which can project text and image into the shared latent space. Similar to the latent embedding approach described in 2.2, we aggregate the K images collected previously and use CLIP to compute the relevance score of (x, y):

$$\text{Score}_{\text{VI}}(x, y) = \frac{1}{K} \sum_{k=1}^{K} \cos(f_T(x), f_V(I_y^k)), \quad (6)$$

and we obtain a probability distribution through softmax (over y):

$$p_{\text{VI}}(y|x) = \text{softmax}\left(\text{Score}_{\text{VI}}(x, y)\right). \quad (7)$$

Ensemble Language and Vision Prediction. Our system is designed for zero-shot tasks without labeled data to learn weights to ensemble the two models. Therefore, we adopt a weighted sum as the late fusion over the final output distributions of the language and multi-modal models:

$$p_{\text{LaVI}}(y|x) = (1-w) \cdot p_{\text{La}}(y|x) + w \cdot p_{\text{VI}}(y|x), \quad (8)$$

where we design a heuristic function to calibrate the weight w based on the relative size between the vision-text model and the language model:

$$w = \text{sigmoid}\left(\frac{\mathcal{P}_{\text{VI}}}{\mathcal{P}_{\text{La}}}\right), \quad (9)$$

---

[6]The two methods will produce more than K images, and we select the top-K based on their similarity with the text input calculated by CLIP.

Task	Dataset	Split	# samples
WSD	CoarseWSD-20	test	10,196
	QASC	dev	926
Science	SciQ	dev	1,000
QA	ARC-E	dev	570
	ARC-C	dev	299
Text	AG-News	test	7,600
Classification	Situation	test	1,789

Table 1: Dataset statistics for the three tasks.

where $\mathcal{P}_{\text{VI}}$ and $\mathcal{P}_{\text{La}}$ are the number of parameters of the models. We hypothesize that when the language model's size increases, it will encode more knowledge and thus rely less on the vision model. The number of parameters of each model and their corresponding weight is listed in Table 8.

3 Experimental Setup

3.1 Datasets

We evaluate our methods on six datasets of three tasks. Table 1 shows dataset statistics.

CoarseWSD-20 (Loureiro et al., 2021) is a coarse-grained WSD dataset built from Wikipedia. The dataset consists of 20 nouns with 2-5 senses per noun (53 senses in total). Each sense is associated with a definition which is the first sentence on its Wikipedia page. CoarseWSD guarantees that every sense has test instances in the test set. On average, each sense has 192 test instances.

QASC (Khot et al., 2020) is a multi-hop, 8-way choice question answering dataset collected by decomposing sentences about scientific facts. We report the performance of the development set, which contains 926 questions.

SciQA (Welbl et al., 2017) is a dataset of 4-way multiple-choice science exam questions spanning from elementary to college-level covering chemistry, biology, physics, etc. We evaluate the development set with 1,000 questions.

ARC (Clark et al., 2018) consists of 7,787 natural, grade-school level science questions. The ARC dataset is split into easy (ARC-E) and challenge (ARC-C), where questions in the challenge set contain the ones that simple retrieval or word correlation methods cannot answer correctly. We evaluate the development sets of ARC-E and ARC-C, which contain 570 and 299 questions, respectively.

AG News (Zhang et al., 2015) is a news topic classification dataset, and each sentence is associated

with one of the four news types: *word*, *sports*, *business*, and *technology*. We run our models on the 7,600 examples in the test set.

Situation (Mayhew et al., 2018) is a event-type classification task. The dataset has 12 events: *need water*, *need infrastructure*, *crime violence*, etc. The original task on this dataset is multi-label classification and has an *out-of-domain* class. As the multi-label prediction requires a fine-tuned threshold to determine the predictions and is thus not suitable for zero-shot models, we remove those examples with more than one label and ones with the *out-of-domain* label, resulting in 1,789 instances.

3.2 Baselines

Aside from the zero-shot language models described in the section 2.2, we also evaluate on a random baseline and compare with previous work.

For CoarseWSD-20, we compare with the BERT-large few-shot (1-shot/3-shot per sense) results reported in Loureiro et al. (2021).

For QA tasks, we include the Information-Retrieval (IR) solver (Clark et al., 2016), which combines the question and option as a query and sends it to a search engine to check if they are explicitly written in some corpus. We also choose SMLM (Banerjee and Baral, 2020) as another baseline - a RoBERTa-large model fine-tuned on triplets extracted from knowledge graphs such as ATOMIC (Sap et al., 2019).

We compare topic classification with the TE-wiki (Ding et al., 2022), the state-of-the-art model on zero-shot topic classification trained on a dataset collected from Wikipedia.

3.3 Evaluation Metrics

We report the accuracy of all question-answering and topic-classification datasets. For CoarseWSD-20, we compute each word's accuracy and F1 score and take the mean score of all 20 words.

3.4 Implementation Details

Image Collection We adopt Bing Image Search to RECALL images. And for image SYNTHESIS, we utilize the newly released DALL·E-mini[7] which chooses VQGAN (Esser et al., 2021) as the image encoder/decoder and BART (Lewis et al., 2020) as the autoregressive transformer. For every textual input, we obtain 100 images from each of the two

methods. The 200 images are sorted using CLIP based on their similarity with the text input. We preserve each text input's top-10 images ($K = 10$) and feed them into the equation 6 to calculate the vision-text probabilities.

Model Implementation The GPT-style and NLI-based language models are built on top of the huggingface API.[8] For NLI models, we use the recently released zero-shot classification pipeline.[9] We use the official release of SBERT[10] and SimCSE[11] to implement the latent embedding approach. The CLIP model is adapted from the OpenAI's public repo,[12] and we select the ViT/B32 as the image encoder. The experiments were run on 3×8 NVIDIA V100 32GB, which can generate 24 images in 5 seconds. The majority of the running time of our model is image generation. In total, we employ DALL·E-mini to generate approximately 1.8M images which take around 104 hours.

4 Evaluation

4.1 Main Results

Z-LaVI boosts the performance of language models. Table 2, 3 and 4 show results on seven datasets of three tasks. Each dataset has two results columns: the original performance of the language models and the ensembled performance by adding our Z-LaVI model. We observe that in most cases, Z-LaVI consistently improves the performance of different language models. Especially in the WSD task, our Z-LaVI with SBERT can outperform the BERT-large fine-tuned with 3-shots of each sense. Z-LaVI also significantly enhances the language models on topic classification task where the best language model with Z-LaVI beats the SOTA zero-shot topic classification model TE-wiki by 2.8%. For science QA tasks, we can see Z-LaVI improves on QASC, SciQ, and ARC-E, but it struggles on the ARC-C, and adding Z-LaVI degrades the performance of a few language models. This is because the ARC-C questions are designed to be hard to answer using retrieval or correlation, and Z-LaVI uses CLIP, which is pre-trained on the image-text correlation only. Figure 7 (b) shows an example that needs multi-hop reasoning where Z-LaVI fails to answer correctly.

Model	# Param.	QASC		SciQ		ARC-E		ARC-C	
		Original	Z-LaVI	Original	Z-LaVI	Original	Z-LaVI	Original	Z-LaVI
Random	-	12.5	-	25.0	-	25.0	-	25.0	-
IR Solver†	-	18.6	-	-	-	30.4	-	20.3	-
SMLM†	355 M	26.6	-	-	-	33.4	-	28.4	-
RoBERTa-L-mnli*	355 M	19.3	_27.2_	44.7	_51.3_	48.4	_51.8_	34.4	33.4
BART-L-mnli*	400 M	21.7	_27.3_	48.8	_51.0_	54.7	_56.1_	**36.5**	**36.5**
GPT-Neo-1.3B	1.3B	29.3	_37.4_	57.5	_60.8_	46.3	_49.8_	27.4	26.1
GPT-Neo-2.7B	2.7B	29.6	_39.6_	64.0	_64.9_	49.6	_51.9_	31.8	30.4
GPT-J-6B	6B	36.3	_42.0_	73.2	_73.7_	55.1	_57.2_	34.8	34.1
OPT-30B	30B	39.7	**_42.1_**	72.7	**_74.0_**	58.2	**_59.5_**	34.8	34.1
SimCSE	355M	30.8	_33.2_	42.6	_48.6_	43.3	_49.3_	26.4	24.7
SBERT*	110M	36.7	_38.6_	57.7	_58.5_	54.4	_56.0_	30.1	27.1
Z-LaVI w/o LM	150M	-	32.7	-	49.5	-	50.2	-	26.7

Table 2: Zero-shot performance on Science QA tasks. Z-LaVI represents the performance with our Visual Imagination. Z-LaVI (w/o LM) is the model that only uses vision-text prediction. The best-performed number for each metric is **bolded**. The numbers are underlined if the original performance is improved with Z-LaVI. The models with * use labeled data for pre-training. The models with † indicate the results are from previous work.

Model	Accuracy		F1	
	Orig.	Z-LaVI	Orig.	Z-LaVI
Random	41.3	-	36.7	-
BERT-L-1shot†*	77.6	-	71.2	-
BERT-L-3shot†*	89.3	-	85.2	-
RoBERTa-L-mnli*	80.4	_83.0_	74.4	_78.1_
BART-L-mnli*	80.2	_82.4_	74.8	_77.9_
GPT-Neo-1.3B	84.7	_88.4_	78.3	_84.6_
GPT-Neo-2.7B	86.7	_88.9_	81.5	_85.3_
GPT-J-6B	84.1	_88.5_	79.3	_84.8_
OPT-30B	84.4	_88.8_	80.4	_85.1_
SimCSE	85.1	_89.7_	78.9	_86.0_
SBERT*	87.8	**_90.6_**	83.3	**_87.5_**
Z-LaVI w/o LM	-	87.7	-	83.8

Table 3: Zero-shot performance on CoarseWSD-20.

Model	AG-News		Situation	
	Orig.	Z-LaVI	Orig.	Z-LaVI
Random	25.0	-	9.1	-
TE-wiki†*	79.6	-	-	-
RoBERTa-L-mnli*	81.2	_81.7_	40.7	_41.8_
BART-L-mnli*	81.9	**_82.4_**	40.5	_41.1_
GPT-Neo-1.3B	59.1	_72.9_	17.8	_38.5_
GPT-Neo-2.7B	59.1	_74.5_	13.6	_35.2_
GPT-J-6B	61.0	_73.8_	21.9	_38.8_
SimCSE	58.1	_73.1_	42.1	_44.4_
SBERT*	77.8	_82.2_	42.6	**_46.6_**
Z-LaVI w/o LM	-	71.6	-	33.4

Table 4: Zero-shot Performance on Topic Classification.

Z-LaVI without language model is a strong baseline. Surprisingly, we also find that Z-LaVI w/o language model performs well on plain language tasks. In some datasets, such as QASC, Coarse-WSD, and topic classification tasks, Z-LaVI w/o LM outperforms the language models without fine-tuning on the downstream datasets (e.g., Sim-CSE, GPT-Neo-1.3B/2.7B). This indicates that the vision-text model pretraining on image-caption pairs learns the knowledge that can be leveraged to solve single modality tasks.

Ensembling two language models is not as good as Z-LaVI. To verify the effectiveness of using visual knowledge, we replace the visual imagination of Z-LaVI with another language model - SimCSE. We select SimCSE here because Sim-CSE is trained fully unsupervised and has the same contrastive learning objective as CLIP. We define the performance gain (PG) of model $\mathcal{M}_1$

(i.e., SimCSE) on top of model $\mathcal{M}_2$ by computing the relative improvement of the ensemble model $\text{Ens}(\mathcal{M}_1, \mathcal{M}_2)$ performance over the original model $\text{Orig}(\mathcal{M}_2)$.

$$\text{PG}(\mathcal{M}_1, \mathcal{M}_2) = \frac{\text{Ens}(\mathcal{M}_1, \mathcal{M}_2) - \text{Orig}(\mathcal{M}_2)}{\text{Orig}(\mathcal{M}_2)}$$
(10)

We include all the language models (exclude Sim-CSE) in the set $\mathbb{M}$[13] and calculate the average performance gain on a dataset by:

$$\text{avg-PG}(\mathcal{M}_1) = \frac{1}{|\mathbb{M}|} \sum_{\mathcal{M} \in \mathbb{M}} \text{PG}(\mathcal{M}_1, \mathcal{M}) \quad (11)$$

For fair comparison, we fix the ensemble weight $w = 0.5$ in equation (8) for both SimCSE and Z-LaVI. We also include the Z-LaVI with dynamic ensemble weight controlled by equation (8). The performance gain of SimCSE and Z-LaVI on all six datasets is shown in Figure 3. We observe

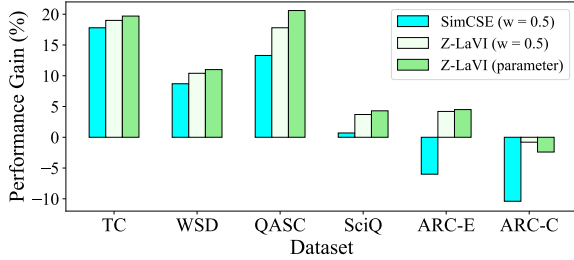[13]6 models in total, which are GPT-Neo-1.3B/2.7B, GPT-J-6B, RoBERTa-L-mnli, BART-L-mnli, and SBERT.

Figure 3: The average performance gain on each dataset. Z-LaVI (parameter) stands for accounting models' number of parameters (Equation 9) to adjust the weights.

Figure 4: The overlap of correctly predicted examples between each pair of models in the Situation dataset.

that Z-LaVI consistently has higher performance gain than SimCSE across all datasets, demonstrating that the visual information provided by Z-LaVI complements language models more hence boosts more on performance. Additionally, Z-LaVI with dynamic weights perform better than simply setting the weight to 0.5.

4.2 Analysis

Vision and Language models behave differently. We define the overlap of correctly predicted examples between two models as:

$$\text{overlap}(\mathcal{M}_1, \mathcal{M}_2) = \frac{|S_{\mathcal{M}_1^*} \cap S_{\mathcal{M}_2^*}|}{|S_{\mathcal{M}_1^*}|} \quad (12)$$

where $S_{\mathcal{M}^*}$ is the set of correctly predicted examples of model $\mathcal{M}$. Figure 4 shows the overlap of models' predictions in the Situation dataset. We observe that Z-LaVI (w/o LM) has an obviously smaller overlap with the other models, while different language models have a big mutual overlap. This difference explains the substantial performance gain after exploiting visual imagination.

	QASC	SciQ	ARC	WSD	AG	Situ
# TOK	2.4	2.6	4.9	28.4	54.7	56.8
✗	26.8	41.1	42.1	75.0	52.9	22.4
RECALL	<u>32.6</u>	**49.9**	<u>48.5</u>	80.3	52.5	21.1
SYNTHESIS	31.5	39.9	44.6	<u>81.5</u>	**71.6**	**34.2**
BOTH	**32.7**	<u>49.5</u>	**50.2**	**83.8**	**71.6**	<u>33.4</u>

Table 5: The performance of Z-LaVI (w/o LM) with different imagination methods. # TOK is the average number of tokens of text inputs in each dataset. ✗ means no image is provided to the model, and we only use the text encoder of CLIP. RECALL and SYNTHESIS represent using image search and image generation, respectively. BOTH means combining the two methods.
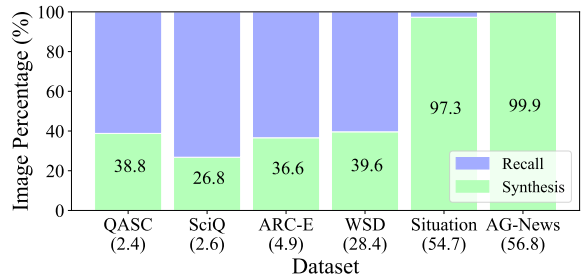
Figure 5: The percentage of recall and synthesis images within the top 10 images of each dataset. The average token length for each dataset is given on the x-axis.

RECALL vs. SYNTHESIS. We ablate on the imagination methods and compare the performance of only using one of the methods. Table 5 demonstrates the performance on each dataset with different imagination methods. We can see that for the dataset with short inputs for imagination (e.g., QA tasks), RECALL is better than SYNTHESIS. This is because short inputs of science QA datasets normally correspond to objects that exist in the real world and are easy to find on the web, such as *mollusca* and *porifera* shown in Figure 7 (a). However, for queries with long sentences (WSD and Topic Classification), the text inputs are too specific to match any real photo. Hence SYNTHESIS is preferable.[14] Figure 5 also indicates that the model prefers to choose RECALL images for short input and tends to use SYNTHESIS images when the input contains more tokens. We also find that without images, Z-LaVI has poor performance on all tasks, reflecting the necessity of imagination.

Performance vs. Image Quantities. We combine RECALL and SYNTHESIS to imagine 200 image candidates. We wonder whether the num-

---

[14]We notice only 328 examples (out of 1789) of the Situation dataset have more than 10 images by RECALL.

Model	COLOR				SHAPE				MATERIAL			
	Original		Z-LaVI		Original		Z-LaVI		Original		Z-LaVI	
	ρ	Acc@1	ρ	Acc@1	ρ	Acc@1	ρ	Acc@1	ρ	Acc@1	ρ	Acc@1
Random	-0.1	6.6	-	-	1.3	7.1	-	-	-1.5	6.0	-	-
BERT-L[†]	37.6	30.3	-	-	42.7	28.4	-	-	36.6	35.7	-	-
Oscar-L[†]	31.8	17.1	-	-	40.0	38.1	-	-	**39.2**	40.5	-	-
RoBERTa-L-mnli*	35.3	25.1	<u>41.4</u>	<u>38.2</u>	41.7	66.4	<u>43.5</u>	<u>67.9</u>	28.3	34.9	<u>31.6</u>	<u>37.7</u>
BART-L-mnli*	34.6	27.5	<u>38.2</u>	<u>32.6</u>	41.5	68.6	<u>42.0</u>	<u>69.3</u>	30.5	37.0	<u>32.6</u>	<u>40.5</u>
GPT-Neo-1.3B	40.1	31.7	<u>47.4</u>	<u>48.4</u>	44.2	52.9	<u>46.1</u>	<u>64.3</u>	35.1	34.5	<u>36.4</u>	<u>41.2</u>
GPT-Neo-2.7B	41.3	29.3	<u>47.0</u>	<u>43.6</u>	43.5	50.0	<u>45.2</u>	<u>62.1</u>	35.5	30.6	<u>36.1</u>	<u>37.7</u>
GPT-J-6B	44.3	38.0	**49.6**	<u>50.9</u>	46.9	65.7	<u>47.1</u>	<u>70.0</u>	38.5	42.3	37.9	<u>46.5</u>
OPT-30B	41.9	41.3	<u>48.1</u>	**52.1**	46.4	60.0	**48.4**	**72.1**	38.3	44.7	38.1	**47.5**
SimCSE	30.7	34.8	<u>36.3</u>	<u>40.4</u>	30.1	28.6	<u>34.7</u>	<u>40.7</u>	24.6	26.4	<u>29.2</u>	<u>33.5</u>
SBERT*	27.6	26.5	<u>38.2</u>	<u>40.6</u>	20.3	13.6	<u>33.6</u>	<u>35.0</u>	24.1	22.5	<u>30.4</u>	<u>34.9</u>
Z-LaVI w/o LM	-	-	37.2	39.4	-	-	33.8	32.1	-	-	24.9	32.7

Table 6: Zero-shot probing on the three relation types (COLOR, SHAPE and MATERIAL) in ViComTe (Zhang et al., 2022a) dataset. We report the average Spearman correlation (ρ) and top-1 accuracy (Acc@1).
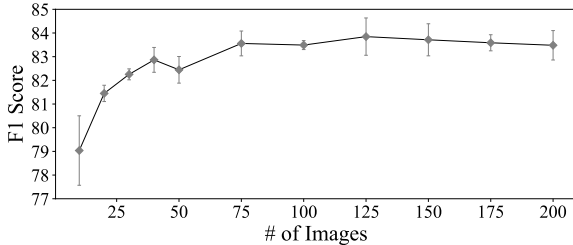
Figure 6: The performance on CoarseWSD-20 with the number of image candidates provided by imagination.

Relation	# Objs	# Test Subjs	Ex (subj, obj) pair
COLOR	12	574	(leave, green)
SHAPE	12	140	(egg, oval)
MATERIAL	18	284	(mug, glass)

Table 7: Statistics of ViComTe dataset.

ber of imaginations impacts the Z-LaVI's performance. Figure 6 reports Z-LaVI's performance on CoarseWSD-20 versus the number of images. We observe that Z-LaVI's F1 score increases with a higher number of images. While the improvement is marginal when the number is higher than 125.

Z-LaVI supplements visual commonsense knowledge. To further validate Z-LaVI helps to mitigate reporting bias problems of language models, we conduct experiments on ViComTe (Zhang et al., 2022a), which is a commonsense knowledge dataset containing different types of properties for over 5000 subjects, e.g., the subject "*egg*" has the property (object) "*oval*". We investigate three relation types (COLOR, SHAPE, and MATERIAL) and report the results on the test set (see Table 7 for details). We select the BERT-large, and Oscar-large (Li et al., 2020) as the baselines of which the results are directly obtained from Zhang et al. (2022a).[15] For a fair comparison, we adopt the same set of seven prompt templates provided by Zhang et al.

(2022a) and report the average performance over these prompts. Table 6 demonstrates the performance of Z-LaVI with language models. We can see Z-LaVI continue to consistently boost the performance of language models and outperform the baselines with significant margins. The results on ViComTe indicate Z-LaVI is a promising way to overcome the reporting bias of language models on visual commonsense.

Qualitative Examples. Figure 8 shows some qualitative examples from the two topic classification datasets. We observe Z-LaVI can effectively correct language models' prediction with more straightforward visual signals. However, we also notice that Z-LaVI fails on examples that cannot be solved by correlation, e.g., Z-LaVI wrongly relates *flooding* with the situation of *need water*. More examples are provided in the Appendix C.

5 Related Work

Visually Grounded Representation Learning Several studies have focused on learning visually grounded word or sentence representations. To learn better word embeddings, Kiros et al. (2018) introduce Picturebook that encodes images as vectors by querying all vocabulary words through Google image search. Lazaridou et al. (2015) optimize a multimodal skip-gram model, where visual

---

[15]We also include the random baseline by assigning a number between 0 and 1 for each class by chance. We iterate the random runs 7 times and report the average performance.

What phylum includes sponges, which are aquatic invertebrates?

A. hymenoptera   B. chordata   C. mollusca   D. porifera

LM: B. chordata ✘   Z-LaVI: D. porifera ✓

(a)

The Sun is about 1.5 x 10^8 km from Earth. The speed of light is 3 x 10^8 m/s. What's the distance from the Sun to Earth in light seconds?

A. 2.0 light-seconds   B. 0.5 light-seconds   C. 2 x 10^-3 light-seconds   D. 5 x 10^2 seconds

LM: D. 5x10^2 seconds ✓   Z-LaVI: C. 2 x 10^-3 ✘

(b)

Figure 7: Qualitative examples from (a) SciQ and (b) ARC-C. Z-LaVI can successfully answer the questions which can be solved by correlation. Z-LaVI fails to answer the question that requires multi-hop reasoning.

Text	Imagination
(a) Pharmacy chain CVS Corp. on Thursday said it would offer the world's first disposable digital camera with a bright color viewing screen that allows consumers to instantly preview pictures.	
LM: business news ✘	Z-LaVI: technology news ✓
(b) Boro left feeling Blue: Accepting mediocrity has been part and parcel of following Middlesbrough over the years, yet this campaign was supposed to bring something new.	
LM: world news ✘	Z-LaVI: sports news ✓
(c) He urged the people to start reconstructing, at least one room, with corrugated steel sheets as roofing material, till the massive rebuilding operations gets underway.	
LM: need Infrastructure ✘	Z-LaVI: need shelter ✓
(d) Earlier reports said nearly 50,000 people were driven from mudflat villages by the flooding as local rivers burst banks while the situation was exacerbated by thunderstorms.	
LM: danger & evacuation ✓	Z-LaVI: need water ✘

Figure 8: Qualitative examples from AG-News (a, b) and Situation (c, d).

information is presented together with the corpus contexts to produce word embeddings. Zablocki et al. (2018) leverage the visual context of objects to learn multimodal word embeddings. With respect to visually grounded sentence embeddings, previous work develops several strategies to enrich sentence representations with visual information, such as using the given sentence as captions to get image features (Kiela et al., 2018), capturing both cluster information and perceptual information in grounded space (Bordes et al., 2019), or exploiting multimodal contrastive learning objective (Zhang et al., 2022b). Zhang et al. (2021) retrieves images from COCO (Lin et al., 2014) as augmented visual features for language models. Lu et al. (2022b) augment sentence embeddings with VQGAN-(Esser et al., 2021) generated images and fine-tune them on GLUE Benchmark (Wang et al., 2018). Liu et al. (2022) probes the spatial commonsense knowledge (sizes, positions) of language models and vision-language models through image generation.

Vision-Language Pretraining Models To connect vision and language semantics, a line of work on multimodal masked language models (Li et al., 2019; Tan and Bansal, 2019; Lu et al., 2019; Su et al., 2020) explores vision-language pretraining and achieves SOTA fine-tuning performance on multimodal benchmarks. Tsimpoukelli et al. (2021) freeze a language model parameters to generate the appropriate caption by encoding each image into the embedding space to inject visual knowl-

edge into PLMs. To retrain knowledge in both vision and language pretrained models, Flamingo (Alayrac et al., 2022) freezes both pretrained models and brings in additional model components to do visually-conditioned autoregressive text generation. Tan and Bansal (2020) retrieve related images as vokens (visualized tokens) and then process large language corpora (e.g., Wikipedia) into voken-prediction tasks. FLAVA (Singh et al., 2022) is an alignment model that pretrains on both unimodal and multimodal data while optimizing cross-modal "alignment" objectives and multimodal fusion objectives. Unified-IO (Lu et al., 2022a) is a general-purpose model which can perform a wide range of vision, language, and multimodel tasks by unifying the inputs and outputs as sequences.

6 Conclusion

In this paper, we propose a novel approach, Z-LaVI, to alleviate the reporting bias problem of pretrained language models and enhance their zero-shot inference ability. We develop two complementary visual imagination mechanisms, i.e., RECALL that aims to retrieve existing objects or scenes and SYNTHESIS that generates nonexistent ones. Experiments on a wide range of language tasks show that our approach can significantly outperform existing zero-shot language models, pointing towards a promising direction to solve an unseen language task with visual imagination.

7 Limitations

Our experiments apply DALL·E-mini for synthesizing the images, but the quality and resolution of the generated images are still low, which can be the factor limiting Z-LaVI's performance. However, the recent breakthroughs of DALLE·E-2 (Ramesh et al., 2022), Imagen, (Saharia et al., 2022) and the open-sourced Stable Diffusion (Rombach et al., 2022) give us hope to obtain more realistic images and thus further unleash the potential of Z-LaVI. The negative results on ARC-C reveal the lack of complex reasoning ability in the current zero-shot vision-text model. At the same time, the success of Flamingo (Alayrac et al., 2022) on few-shot multi-modal tasks lets us sense the possibility of applying the framework of Z-LaVI with these powerful visual language models to solve broader language tasks. We can foresee the bright future of our method once these powerful resources are publicly available. In this paper, we focus on the zero-shot settings, and thus it is difficult to design a more effective approach to ensemble the language and vision without training data. However, when few-shot examples are available, it is possible to learn a mechanism to automatically calibrate the weights of imagination depending on the input examples.

In addition, the image generation model is trained on unfiltered data on the web, which may leak personal information such as the human face, etc. The generation model may also be biased toward stereotypes against minority groups. Furthermore, compared to language models, our approach requires extra resources such as an image search engine, pretrained text-to-image generation model, etc., which will increase the implementation cost. Finally, we evaluated our method in English datasets only, and we plan to incorporate other languages in the future with the help of multilingual multi-modal models (Huang et al., 2021).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Gerry TM Altmann and Yuki Kamide. 2004. Now you see it, now you don't: Mediating the mapping between language and the visual world. *The interface of language, vision, and action: Eye movements and the visual world*, pages 347–386.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Pratyay Banerjee and Chitta Baral. 2020. Self-supervised knowledge triplet learning for zero-shot question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 151–162, Online. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. 2019. Incorporating visual semantics into sentence representations within a grounded space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 696–707, Hong Kong, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and

Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Banchiamlack Dessalegn and Barbara Landau. 2013. Interaction between language and vision: It's momentary, abstract, and it develops. *Cognition*, 127(3):331–344.

Hantian Ding, Jinrui Yang, Yuqian Deng, Hongming Zhang, and Dan Roth. 2022. Towards open-domain topic classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 90–98, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2443–2459, Online. Association for Computational Linguistics.

Carlos E. Jimenez, Olga Russakovsky, and Karthik Narasimhan. 2022. CARETS: A consistency and robustness evaluative test suite for VQA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6392–6405, Dublin, Ireland. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning visually grounded sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 408–418, New Orleans, Louisiana. Association for Computational Linguistics.

Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933, Melbourne, Australia. Association for Computational Linguistics.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

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. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2365–2376, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022a. Unified-io: A unified model for vision, language, and multimodal tasks. *arXiv preprint arXiv:2206.08916*.

Yujie Lu, Wanrong Zhu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022b. Imagination-augmented natural language understanding. *arXiv preprint arXiv:2204.08535*.

Stephen Mayhew, Shyam Upadhyay, Wenpeng Yin, Lucia Huo, Devanshu Jain, Prasanna Poudyal, Tatiana Tsygankova, Yihao Chen, Xin Li, Nitish Gupta, et al. 2018. University of pennsylvania lorehlt 2018 submission. Technical report, Technical report, University of Pennsylvania, 2018. URL https://cogcomp . . . .

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jason Tyler Rolfe. 2016. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *CVPR*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. 2018. Learning multi-modal word representation grounded in visual context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022a. Visual commonsense in pretrained unimodal and multimodal models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5321–5335, Seattle, United States. Association for Computational Linguistics.

Lisai Zhang, Qingcai Chen, Joanna Siebert, and Buzhou Tang. 2021. Semi-supervised visual feature integration for language models through sentence visualization. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 682–686.

Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A Hedderich, and Dietrich Klakow. 2022b. Mcse: Multimodal contrastive learning of sentence embeddings. *arXiv preprint arXiv:2204.10931*.

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. 2022c. Opt: Open pre-trained transformer language models.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Implementation Details

A.1 Prompt Selection

We use the intuitive prompt templates of the three main tasks for each model shown in Table 9, and we do not tune the prompt for each dataset. For ViComTe dataset, we reuse the original 7 prompts provided by Zhang et al. (2022a).

A.2 Model Parameters

We include the number of parameters of all models we use in Table 8. We also list the ensemble weight w based on the relative sizes between the two models:

$$w = \text{sigmoid}\left(\frac{\mathcal{P}_{\text{VI}}}{\mathcal{P}_{\text{La}}}\right) = \frac{1}{1 + e^{-\mathcal{P}_{\text{VI}}/\mathcal{P}_{\text{La}}}} \quad (13)$$

Model	# of Parameters	Ensemble weight
CLIP	150 M	-
BART-L	400 M	0.59
RoBERTa-L	355 M	0.60
SBERT	110 M	0.80
SimCSE	355 M	0.60
GPT-Neo-1.3B	1.3 B	0.53
GPT-Neo-2.7B	2.7 B	0.51
GPT-J-6B	6 B	0.51
OPT-30B	30 B	0.50

Table 8: The number of parameters of each model and the ensemble weights calculated by equation 9.

A.3 Other Details

Predicted scores of prompt-based approach. The huggingface API can calculate the loss of the provided input, and we take the reciprocal of the loss as the prediction score in practice.

Speed of Bing Image Search. The efficiency of downloading images from Bing depends on your internet speed and the API plan you select. In our case, downloading 200 images of a query takes less than 10 seconds. The script of Bing Image Search is provided in our GitHub repo.

Requirements for computing resources. Our approach is designed for zero-shot scenarios and does not involve training, thus most of the experiments can be conducted on a single GPU with more than 20GB of memory. The biggest challenge is to deploy the OPT-30B, which requires 70GB of memory. We successfully deploy OPT-30B using 4 P40 GPUs and released the code to implement OPT-30B across multiple GPUs in our GitHub repo.

B Additional Analysis

B.1 Relative Improvement of Each Word in Word Sense Disambiguation

To quantify the ability of Z-LaVI to disambiguate different words, we compute the relative improvement (RI) of each word by comparing the F1 score of SBERT with that of Z-LaVI:

$$\text{RI} = \frac{\text{F1}_{\text{Z-LaVI}} - \text{F1}_{\text{SBERT}}}{\text{F1}_{\text{SBERT}}} \quad (14)$$

As shown in Figure 10, Z-LaVI improves the F1 score of 16 words (out of 20), while for the rest of four words (*pound*, *chair*, *digit*, *club*), Z-LaVI hurts the performance. Furthermore, we notice those four words all contain abstract senses (see Figure 13) which are difficult to imagine, e.g., *pound* has two senses-*currency* and *mass*, which are both units and difficult to illustrate in images. Based on this finding, we think future work can design a dynamic ensemble method to calibrate the weights conditioned on the input.

B.2 Overlap of Incorrectly Predicted Examples

In Figure 4, we show that Z-LaVI w/o LM has a smaller overlap with other language models on **correctly** predicted examples. To tell a complete story of the different output distribution between Z-LaVI and language models, we compute the overlap of **incorrectly** predicted examples shown in Figure 9. We can see that Z-LaVI has a smaller overlap on incorrectly predicted ones.

Figure 9: The overlap of incorrectly predicted examples between each pair of models in the Situation dataset.

Model	Word Sense Disambiguation	Question Answering	Topic Classification
GPT NLI	[SENTENCE] The [TARGET WORD] mentioned before means [SENSE NAME].	Question: [QUESTION] The answer is [ANSWER].	[SENTENCE] This example is [CLASS NAME].
SimCSE SBERT	(SENTENCE, DEFINITION)	(QUESTION, ANSWER)	(SENTENCE, This example is [CLASS NAME].)
CLIP	(SENTENCE, DEFINITION)	(QUESTION, ANSWER)	(SENTENCE, A news image of [CLASS NAME].)

Table 9: The prompts we use for each model on WSD, QA, and Topic classification.

COLOR	MATERIAL	SHAPE
[SUBJ] can be of color [OBJ].	[SUBJ] is made of [OBJ].	[SUBJ] can be shape of [OBJ].
[SUBJ] has color [OBJ].	[SUBJ] can be made of [OBJ].	[SUBJ] has shape of [OBJ].
The color of [SUBJ] can be [OBJ].	[SUBJ] is made from [OBJ].	[SUBJ] is of shape [OBJ].
The color of the [SUBJ] is [OBJ].	[SUBJ] can be made from [OBJ].	The shape of [SUBJ] can be [OBJ].
[SUBJ] is [OBJ].	[SUBJ] is [OBJ].	The shape of the [SUBJ] is [OBJ].
This [SUBJ] is [OBJ].	This [SUBJ] is [OBJ].	[SUBJ] is [OBJ].
[SUBJ] is of color [OBJ].	[SUBJ] is made by using [OBJ].	This [SUBJ] is [OBJ].

Table 10: The prompts we use for ViComTe which are provided by Zhang et al. (2022a).
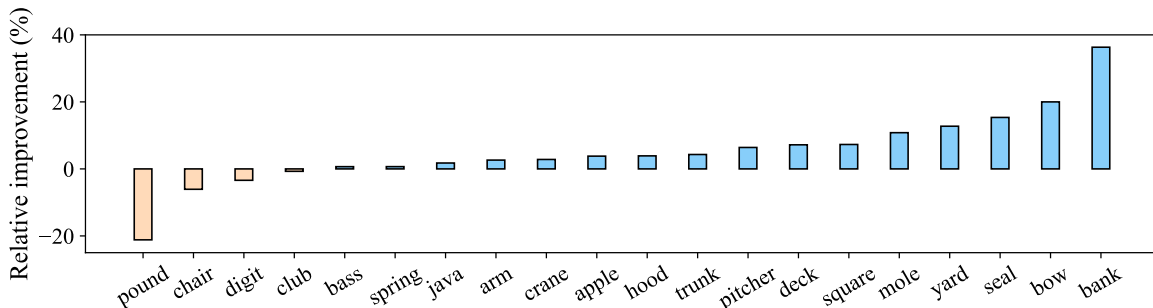
Figure 10: Relative improvement (F1) of the 20 words in CoarseWSD20.

B.3 RECALL vs. SYNTHESIS on ViComTe

We also ablate on the imagination methods on ViComTe shown in Table 11. As mentioned before, RECALL is preferred when the text input is short. This finding still holds for ViComTe where the text inputs are single words. We notice RECALL performs better than SYNTHESIS except for the COLOR relation. We find the images from SYNTHESIS are more prototypical in colors than the ones from RECALL. (See examples in Figure 15)

C Qualitative Examples

Figure 11 demonstrates the qualitative examples from the two topic classification dataset. Figure 12 shows the QA examples. Figure 13, 14 show the SYNTHESIS image of each word sense from CoarseWSD20. Figure 15 demonstrates the images (SYNTHESIS and RECALL) of different subjects from ViComTe. In addition, we release all the SYNTHESIS images generated by DALL·E-mini which can be downloaded from our GitHub repo.

ρ/Acc@1	RECALL	SYNTHESIS	BOTH
COLOR	35.5/35.7	<u>36.4</u>/<u>38.5</u>	**37.2/39.4**
SHAPE	**37.1/49.3**	27.9/31.3	<u>33.9</u>/<u>32.1</u>
MATERIAL	**29.9/34.9**	23.4/29.6	<u>24.9</u>/<u>32.7</u>

Table 11: The average Spearman correlation (left) and top-1 accuracy (right) of Z-LaVI w/o LM with different imagination methods. The highest number of each row is **bolded** and the second-best one is <u>underlined</u>.
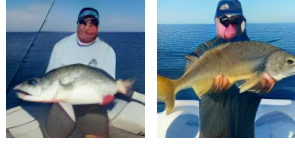
Text	Imagination	Text	Imagination
A second team of rocketeers competing for the 10 million Ansari X Prize, a contest for privately funded suborbital space flight, has officially announced the first launch date for its manned rocket.		JERUSALEM: The Israeli army is set to unveil a new weapon designed to get under the noses of Palestinians - a massive stink bomb. A report in the Maariv daily on Wednesday said that the military, which has	
LM: technology news ✔	Z-LaVI: technology news ✔	LM: technology news ✘	Z-LaVI: world news ✔
The Atlanta Braves have acquired standout righthander Tim Hudson from the Oakland Athletics in exchange for outfielder Charles Thomas, right-handed pitcher Juan Cruz and left-handed pitcher Dan Meyer.		Recreational anglers may be responsible for landing nearly 25 percent of over-fished saltwater species caught off US coasts, a study released on Thursday suggests.	
LM: business news ✘	Z-LaVI: sports news ✔	LM: technology news ✔	Z-LaVI: business news ✘
who is also providing water purification tablets to ensure a safe drinking water supply to the affected population.		residents have taken shelter in mosques, schools, local government offices and even in cemeteries after their homes were inundated by up to three meters (10 feet) of dirty brown water .	
LM: need water ✔	Z-LaVI: need water ✔	LM: danger evacuation ✘	Z-LaVI: need shelter ✔
international medical corps has already received vials of yellow fever vaccine for high risk workers and has ordered additional vaccines for community members.		Militants who tied to attack renaissance dam have been destroyed: 21 militants trained by Shaibia and Egypt entered to country by Benshangul gumez...	
LM: search people ✘	Z-LaVI: medical assistance ✔	LM: terrorist activity ✔	Z-LaVI: need infrastructure ✘

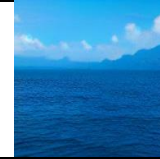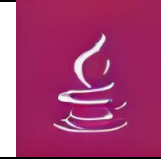Figure 11: Qualitative examples from AG-News (top-4) and Situation (bottom-4).

Figure 12: Qualitative examples from SciQ (a, b), QASC (c), ARC-E (d), and ARC-C (e). (e) is a typical failure case of CLIP because multimodal models are not good at understanding negation (Jimenez et al., 2022).
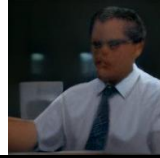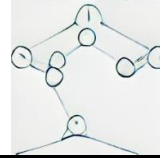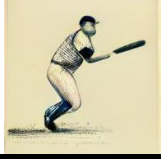
Figure 13: Qualitative examples from CoarseWSD20. We randomly sample 1 image (SYNTHESIS) for each sense.

seal			
emblem	*mechanical*	*musician*	*pinniped*

spring			trunk		
device	*hydrology*	*season*	*anatomy*	*automobile*	*botany*

square			
number	*company*	*shape*	*town*
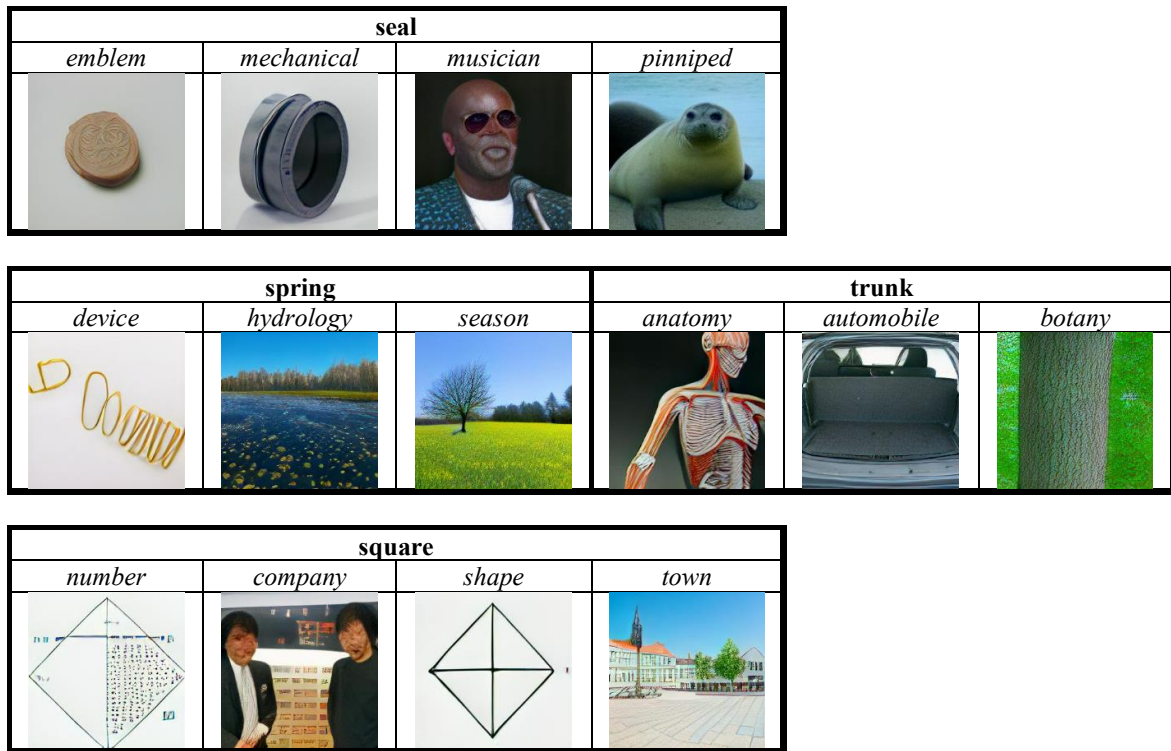
Figure 14: Qualitative examples from CoarseWSD20. (Continued)

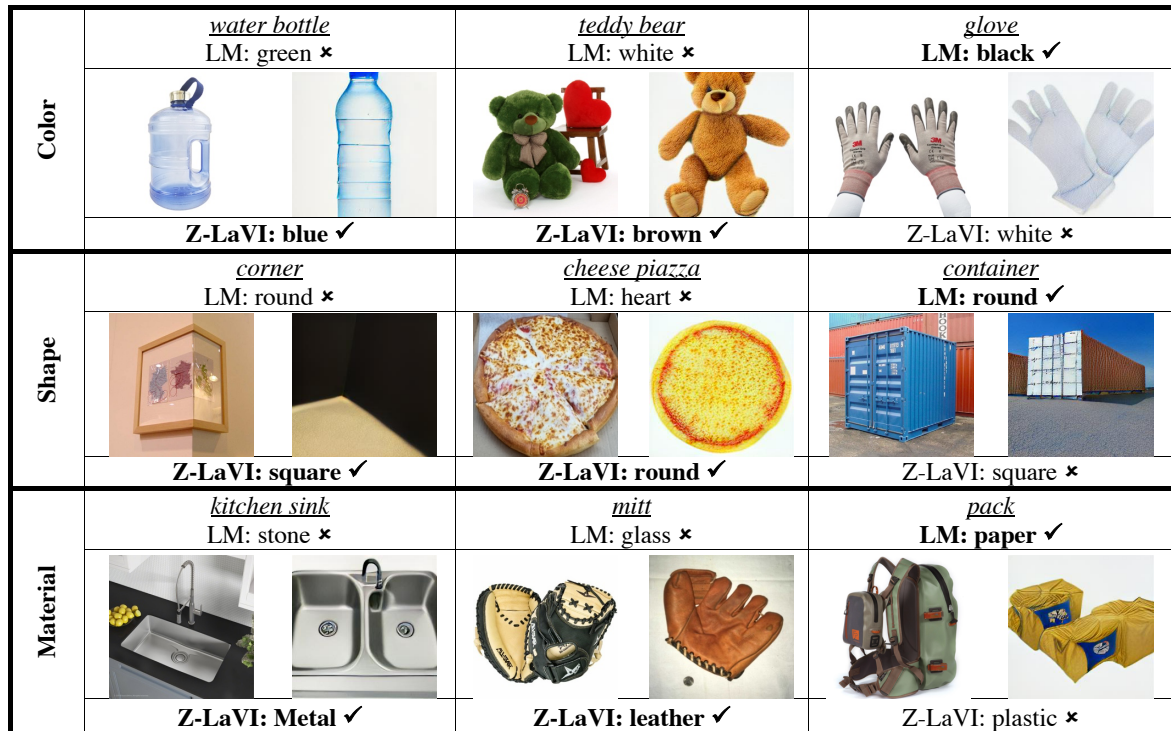|	water bottle LM: green ✘	teddy bear LM: white ✘	glove **LM: black ✓**
Color			
	Z-LaVI: blue ✓	**Z-LaVI: brown ✓**	Z-LaVI: white ✘
	corner LM: round ✘	cheese piazza LM: heart ✘	container **LM: round ✓**
Shape			
	Z-LaVI: square ✓	**Z-LaVI: round ✓**	Z-LaVI: square ✘
	kitchen sink LM: stone ✘	mitt LM: glass ✘	pack **LM: paper ✓**
Material			
	Z-LaVI: Metal ✓	**Z-LaVI: leather ✓**	Z-LaVI: plastic ✘

Figure 15: Qualitative examples from ViComTe. We randomly sample three subjects for each relation type (COLOR, SHAPE, and MATERIAL). For each subject, we present two images, one from RECALL (left) and another one from SYNTHESIS (right).